Programa de Pós-Graduação em Lingüística

PPGL

# *Investigação de métodos de sumarização automática multidocumento baseados em hierarquias conceituais*

Andressa Caroline Inácio Zacarias

SÃO CARLOS
2016

Universidade Federal de São Carlos

# INVESTIGAÇÃO DE MÉTODOS DE SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO BASEADOS EM HIERARQUIAS CONCEITUAIS

ANDRESSA CAROLINE INÁCIO ZACARIAS
Bolsista: FAPESP (2014/ 12817-4)

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística, área de concentração: Descrição, análise e processamento automático de línguas naturais.

Orientadora: Ariani Di Felippo

São Carlos - São Paulo - Brasil

2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

Z13i

Zacarias, Andressa Caroline Inácio
Investigação de métodos de sumarização automática multidocumento baseados em hierarquias conceituais / Andressa Caroline Inácio Zacarias. -- São Carlos : UFSCar, 2016.
128 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2016.

1. Sumarização Automática Multidocumento. 2. Métricas de grafo. 3. Hierarquia léxico-conceitual. I. Título.

## ERRATA

Em agradecimentos acrescentar:
E, finalmente, à Fapesp pela concessão da bolsa de mestrado (proc. 2014/12817-4).

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas

Programa de Pós-Graduação em Linguística

**Folha de Aprovação**

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de dissertação de mestrado da candidata Andressa Caroline Inácio Zacarias, realizada em 29/03/2016.

Profa. Dra. Gladis Maria de Barcellos Almeida
UFSCar

Prof. Dr. Oto Araújo Vale
UFSCar

Prof. Dr. Thiago Alexandre Salgueiro Pardo
USP

*“Aprender a lidar com o fracasso*
*é tão importante quanto necessário,*
*pois ele é apenas parte do processo de vitória.”*

## AGRADECIMENTOS

À **Deus** em primeiro lugar, por ter me dado forças e me iluminado em mais esta etapa da minha vida.

Aos meus pais **Sandra** e **Agnilson** pelo amor incodicional e amparo em todos os momentos difíceis.

Aos meus irmãos **Jéssica, Vanessa** e **Jean** por terem acreditado em mim mesmo nos momentos de fracasso, pela amizade, confiança e amor fraterno.

Aos meus tios **Marisol** e **Beto** por todo amor , incentivo e participação das minhas conquistas

Ao meu amor e companheiro **Caio** pelo amor imensurável, pelas palavras de carinho, pelo cuidado, paciência, e pela parceria de sempre.

Às minhas amigas **Renata, Carla, Letícia e Monique** que me acompanharam em toda a trajetória acadêmica e profissional.

À minha orientadora **Ariani Di Felippo**, meu coorientador **Thiago Pardo** e à professora **Gladis** pela paciência, sabedoria e por todo o conhecimento transmitido.

Aos colegas no NILC, em especial, ao **Fernando** e ao **Erick Fonseca** pelas inúmeras horas de resolução de problemas e prestatividade. À **Paula, Amanda, Lianet, Jackson, Fabrício** e **Renata** pelos momentos de descontração e palavras de consolo.

À família **Menzani** e à família **Domingues** por terem me acolhido em São Carlos.

À **Fapesp** pelo apoio financeiro.

## RESUMO

Na Sumarização Automática Multidocumento (SAM), busca-se gerar um único sumário, coerente e coeso, a partir de uma coleção de textos, de diferentes fontes, que tratam de um mesmo assunto. A geração de tais sumários, comumente extratos (informativos e genéricos), requer a seleção das sentenças mais importantes da coleção. Para tanto, pode-se empregar conhecimento linguístico superficial (ou estatística) ou conhecimento profundo. Quanto aos métodos profundos, destaca-se que estes, apesar de mais caros e menos robustos, produzem extratos mais informativos e com mais qualidade linguística. Para o português, os únicos métodos profundos que utilizam conhecimento léxico-conceitual baseiam na frequência de ocorrência dos conceitos na coleção para a seleção de conteúdo. Tendo em vista o potencial de aplicação do conhecimento semântico-conceitual, propôs-se investigar métodos de SAM que partem da representação dos conceitos lexicais dos textos-fonte em uma hierarquia para a posterior exploração de certas propriedades hierárquicas capazes de distinguir os conceitos mais relevantes (ou seja, os tópicos da coleção) dos demais. Especificamente, selecionaram-se 3 das 50 coleções do CSTNews, *corpus* multidocumento de referência do português, e os nomes que ocorrem nos textos-fonte de cada coleção foram manualmente indexados aos conceitos da WordNet de Princeton (WN.Pr), gerando, ao final, uma hierarquia com os conceitos constitutivos da coleção e demais conceitos herdados da WN.Pr para a construção da hierarquia. Os conceitos da hierarquia foram caracterizados em função de 5 métricas (de relevância) de grafo potencialmente pertinentes para a identificação dos conceitos a comporem um sumário: *Centrality*, *Simple Frequency, Cumulative Frequency, Closeness e Level.* Tal caracterização foi analisada de forma manual e por meio de algoritmos de Aprendizado de Máquina (AM) com o objetivo de verificar quais medidas seriam as mais adequadas para identificar os conceitos relevantes da coleção. Como resultado, a medida *Centrality* foi descartada e as demais utilizadas para propor métodos de seleção de conteúdo para a SAM. Especificamente, propuseram-se 2 métodos de seleção de sentenças, os quais compõem os métodos extrativos: (i) CFSumm, cuja seleção de conteúdo se baseia exclusivamente na métrica *Simple Frequency*, e (ii) LCHSumm, cuja seleção se baseia em regras aprendidas por algoritmos de AM a partir da utilização em conjunto das 4 medidas relevantes como atributos. Tais métodos foram avaliados intrinsecamente quanto à informatividade, por meio do pacote de medidas ROUGE, e qualidade linguística, com base nos critérios da conferência TAC. Para tanto, utilizaram-se os 6 *abstracts* humanos disponíveis em cada coleção do CSTNews. Ademais, os sumários gerados pelos métodos propostos foram comparados aos extratos gerados pelo sumarizador GistSumm, tido como *baseline*. Os dois métodos obtiveram resultados satisfatórios quando comparados ao *baseline* GistSumm e o método CFSumm se sobressai ao método LCHSumm.



**Palavras-chave:** Sumarização Automática Multidocumento, Métricas de grafo, Hierarquia léxico-conceitual.

## ABSTRACT

The Automatic Multi-Document Summarization (MDS) aims at creating a single summary, coherent and cohesive, from a collection of different sources texts, on the same topic. The creation of these summaries, in general extracts (informative and generic), requires the selection of the most important sentences from the collection. Therefore, one may use superficial linguistic knowledge (or statistic) or deep knowledge. It is important to note that deep methods, although more expensive and less robust, produce more informative extracts and with more linguistic quality. For the Portuguese language, the sole deep methods that use lexical-conceptual knowledge are based on the frequency of the occurrence of the concepts in the collection for the selection of a content. Considering the potential for application of semantic-conceptual knowledge, the proposition is to investigate MDS methods that start with representation of lexical concepts of source texts in a hierarchy for further exploration of certain hierarchical properties able to distinguish the most relevant concepts (in other words, the topics from a collection of texts) from the others. Specifically, 3 out of 50 CSTNews (multi-document corpus of Portuguese reference) collections were selected and the names that have occurred in the source texts of each collection were manually indexed to the concepts of the WordNet from Princenton (WN.Pr), engendering at the end, an hierarchy with the concepts derived from the collection and other concepts inherited from the WN.PR for the construction of the hierarchy. The hierarchy concepts were characterized in 5 graph metrics (of relevancy) potentially relevant to identify the concepts that compose a summary: Centrality, Simple Frequency, Cumulative Frequency, Closeness and Level. Said characterization was analyzed manually and by machine learning algorithms (ML) with the purpose of verifying the most suitable measures to identify the relevant concepts of the collection. As a result, the measure Centrality was disregarded and the other ones were used to propose content selection methods to MDS. Specifically, 2 sentences selection methods were selected which make up the extractive methods: (i) CFSumm whose content selection is exclusively based on the metric Simple Frequency, and (ii) LCHSumm whose selection is based on rules learned by machine learning algorithms from the use of all 4 relevant measures as attributes. These methods were intrinsically evaluated concerning the informativeness, by means of the package of measures called ROUGE, and the evaluation of linguistic quality was based on the criteria from the TAC conference. Therefore, the 6 human abstracts available in each CSTNews collection were used. Furthermore, the summaries generated by the proposed methods were compared to the extracts generated by the GistSumm summarizer, taken as baseline. The two methods got satisfactory results when compared to the GistSumm baseline and the CFSumm method stands out upon the LCHSumm method.



**Key Words:** Automatic Multi-Document Summarization, Graph Metrics, Lexical-Conceptual Hierarchy.

# LISTA DE FIGURAS

# LISTA DE QUADROS

## LISTA DE TABELAS

# ÍNDICE

# 1. Introdução

## 1.1. Contextualização

Atualmente, há uma imensa quantidade de informação disponível na *web* em formato textual. Muitos desses textos são bastante similares, já que veiculam um mesmo assunto ou tema. Esse é o caso, por exemplo, das notícias jornalísticas. Ao se fazer uma busca no *Google* sobre a "desclassificação do nadador Thiago Pereira no Pan de Toronto", por exemplo, ver-se-á que esse conteúdo é veiculado por inúmeras fontes jornalísticas distintas, sendo muitas das notícias similares entre si, como ilustrado no Quadro 1.

Quadro 1 – Notícias sobre um assunto veiculadas por fontes distintas.

**Texto 1**[1]
O nadador brasileiro Thiago Pereira foi desclassificado da final dos 400m medley dos Jogos Pan-Americanos, nesta quinta-feira em Toronto, e perdeu o título que faria dele o maior medalhista da história do evento, igualando a marca do esgrimista cubano Erick López (22). Com a punição de Thiago, que chegou em primeiro lugar da prova, a medalha de ouro foi para brasileiro, Brandonn Almeida.

**Texto 2**[2]
O brasileiro Thiago Pereira completou nesta quinta-feira em primeiro os 400m medley e se tornaria tricampeão pan-americano da prova. No entanto, os juízes encontraram irregularidades e ele foi desclassificado. A organização alegou que o nadador não tocou a parede da piscina com as duas mãos na virada do nado peito para o livre, o que é ilegal. O ouro , então, foi parar com o também brasileiro Brandonn Almeida, de apenas 18 anos. "Não sei o que aconteceu. Uma pena. Queria fazer uma dobradinha com o Thiago. Eu o vi no alto do pódio em 2011 (Guadalajara). Mas estou muito feliz, pois não esperava. Não era a prova que eu treinei e não consigo descrever o que estou sentindo", disse Brandonn. Com a decisão, Thiago Pereira segue com 21 medalhas na história do Pan, a uma do maior medalhista da história da competição, o ex-ginasta cubano Erick Lopez. O brasileiro ainda terá mais duas oportunidades para igualar e/ou bater a marca: nos 200m medley e 4x100m medley

Diante da falta de condições (tempo e capacidade de processamento) que o usuário tem para lidar com tal cenário, o interesse pela Sumarização Automática Multidocumento (SAM), uma das subáreas do Processamento Automático das Línguas Naturais (PLN), intensificou-se nos últimos anos, pois é importante ter acesso a um texto que resuma as informações centrais veiculadas pelas notícias. Para ilustrar, tem-se, no Quadro 2, um resumo produzido manualmente a partir das notícias do Quadro 1.

[1]http://esporte.uol.com.br/ultimas-noticias/afp/2015/07/16/thiago-pereira-e-desclassificado-e-perde-medalha-de-ouro-no-pan.htm
[2] http://esportes.terra.com.br/jogos-pan-americanos/thiago-pereira-e-desclassificado-e-recorde-de-mr-pan-e-adiado,20fe6e5b03b96fba7a1130cedf8951abefinRCRD.html

Quadro 2 – Sumário humano construído a partir das notícias do Quadro 1.

| |
|---|
| O nadador brasileiro Thiago Pereira foi desclassificado da final dos 400m medley dos Jogos Pan-Americanos, nesta quinta-feira em Toronto. A organização alegou que o nadador não tocou a parede da piscina com as duas mãos na virada do nado peito para o livre, o que é ilegal. Com a punição, a medalha de ouro foi para o brasileiro Brandonn Almeida. Além disso, Thiago perdeu a chance de se tornar tricampeão pan-americano da prova e o maior medalhista da história do evento, igualando a marca do esgrimista cubano Erick López (22 medalhas). O brasileiro ainda terá mais duas oportunidades para igualar e/ou bater a marca. |

As aplicações de SAM buscam gerar, a partir de uma coleção de dois ou mais textos (cada um advindo de um jornal distinto) sobre um mesmo tópico, um sumário coeso e coerente (MANI, 2001).

O foco da SAM tem sido a produção de extratos informativos e genéricos (isto é, voltados para uma audiência genérica). Os sumários extrativos ou extratos do tipo informativo são compostos por sentenças extraídas integralmente dos textos-fonte por veicularem a ideia central da coleção e justapostas na ordem determinada de acordo com o método utilizado.

A questão central na SAM extrativa tem sido selecionar as sentenças relevantes para compor o sumário. No geral, a metodologia de seleção segue 2 etapas. Primeiro, as sentenças são pontuadas e ranqueadas por um critério de relevância que busca capturar a redundância da informação na coleção, pois esse é comprovadamente o principal critério utilizado pelos humanos (MANI, 2001). Em seguida, as sentenças no topo do ranque são selecionadas para o sumário, buscando-se eliminar a redundância entre elas, até que se atinja a taxa de compressão (isto é, tamanho desejado do sumário).

Para capturar a redundância e ranquear as sentenças, há várias estratégias ou métodos. Para o português, há métodos de SAM desenvolvidos segundo os 3 paradigmas de sumarização automática (SA), os quais atingem e, por vezes, superam o estado-da-arte. Especificamente, há (i) métodos superficiais, que usam pouco conhecimento linguístico ou estatística para selecionar as sentenças; (ii) métodos profundos, que fazem uso massivo de conhecimento linguístico e (iii) métodos híbridos, que unem conhecimento linguístico e estatístico.

Os métodos profundos, em especial, são os mais caros e têm aplicação mais restrita que os superficiais, pois dependem de recursos (p.ex.: gramáticas, léxicos e modelos de discurso) e ferramentas linguístico-computacionais auxiliares (p.ex.: *parser* discursivo), porém geram sumários mais coerentes, coesos e informativos.

Os métodos profundos para o português pautam-se majoritariamente em conhecimento discursivo, principalmente porque os pesquisadores dispõem da CST (*Cross-document Structure Theory*) (RADEV, 2000), que é uma teoria (e modelo) multidocumento robusta e computacionalmente tratável para representar os textos de uma coleção em nível discursivo. Aliás, o melhor método profundo para o português, o RC-4 (CARDOSO, 2014), recebe esse nome porque seleciona as sentenças com base em informações advindas da anotação dos textos-fonte de acordo com a CST e também de acordo com a RST (*Rhetorical Structure Theory*) (MANN, THOMPSON, 1987).

Além desses, destacam-se os 2 métodos profundos de SAM mutilíngue (português-inglês) de Tosta (2014), os quais, para gerar extratos em português, baseiam-se em conhecimento léxico-conceitual. Especificamente, tais métodos partem de coleções compostas por 1 texto em português e 1 em inglês, cujos nomes são indexados à WordNet de Princeton (WN.Pr) (FELLBAUM, 1998), uma base léxico-conceitual em inglês. Na sequência, as sentenças dos textos-fonte são pontuadas e ranqueadas com base na frequência de ocorrência de seus conceitos constitutivos na coleção. A partir do ranque, um dos métodos seleciona apenas as sentenças em português com pontuação mais alta para compor o sumário, até que a taxa de compressão desejada seja atingida. O outro método seleciona as sentenças mais bem pontuadas independentemente de sua língua-fonte e, caso sentenças em inglês sejam selecionadas, faz-se a tradução destas para o português. Segundo Tosta (2014), tais métodos se mostraram muito promissores, gerando extratos com boa qualidade linguística e também informatividade.

Além de terem sido testados somente no cenário multilíngue, os métodos profundos de Tosta (2014) utilizam apenas a frequência de ocorrência de conceitos na coleção como critério para capturar a redundância e, por conseguinte, selecionar as sentenças para o sumário. Na literatura, no entanto, há métodos como o de Hennig *et al.* (2008) para o inglês que, a partir da indexação das palavras de conteúdo das sentenças de uma coleção de textos-fonte a uma hierarquia conceitual (representada em grafo), utilizam informações estruturais da hierarquia para delimitar os conceitos mais relevantes e, por conseguinte, as sentenças que os veiculam.

Assim, buscando contribuir para o avanço das pesquisas sobre SAM do português, sobretudo com as pesquisas baseadas em conhecimento semântico-conceitual, os objetivos descritos a seguir foram traçados, assim como as hipóteses sobre eles.

## 1.2 . Objetivos e Hipóteses

Neste trabalho, visa-se, de forma geral, investigar métodos profundos de SAM cuja seleção de conteúdo pauta-se na utilização de conhecimento léxico-conceitual codificado em uma hierarquia léxico-conceitual. Assim, os objetivos específicos são:

a) Investigar a pertinência de algumas propriedades estruturais (codificadas em métricas) das representações conceituais para a distinção dos conceitos ou tópicos mais relevantes dos demais conceitos de uma coleção de textos-fonte;
b) Propor métodos baseados nas métricas mais relevantes.

Tais objetivos, aliás, relacionam-se diretamente à meta de pesquisa de um projeto maior denominado SUSTENTO[3] (CNPq 483231/2012-6), que é a de produzir e/ou sistematizar conhecimento linguístico para subsidiar a SAM do português.

Os objetivos traçados nesse trabalho pautam-se em 3 hipóteses sobre a utilização de hierarquias conceituais na SAM:

a) A organização léxico-conceitual de uma coleção de textos-fonte sobre um mesmo assunto reflete os fenômenos multidocumento, como a redundância;
b) As métricas relativas às estruturas léxico-conceituais, representadas em grafos, permitem identificar o conteúdo principal da coleção dos textos-fonte, e
c) As propriedades das estruturas léxico-conceituais possibilitam a seleção do conteúdo mais importante da coleção, produzindo, assim, sumários informativos e com qualidade linguística.

Visando alcançar os objetivos, estabeleceram-se as etapas metodológicas a seguir.

## 1.3. Metodologia

O equacionamento metodológico engloba 6 tarefas, as quais são apresentadas a seguir:

- Tarefa 1 – Revisão da literatura: consistiu da leitura da bibliografia fundamental e demais referências pertinentes ao projeto que surgiram no decorrer da pesquisa. A revisão é composta basicamente por (i) métodos de SAM, sobretudo os baseados em conhecimento léxico-conceitual, e (ii) critérios de relevância aplicáveis a representações léxico-conceituais (dos textos-fonte).

[3] http://www.nilc.icmc.usp.br/arianidf/sustento/

- Tarefa 2 – Seleção do *corpus:* consistiu na seleção de um *corpus* adequado à realização do projeto. Tendo em vista os objetivos traçados, esse *corpus* tem as seguintes características: (i) multidocumento (isto é, composto por coleções que possuem diversos textos-fonte e seus respectivos sumários humanos multidocumento) e (ii) escrito em português. Dado o tamanho do *corpus* selecionado e, sobretudo, a complexidade das demais tarefas previstas, a Tarefa 2 englobou uma fase de recorte, em que uma parcela do *corpus* foi selecionada para a realização do trabalho.
- Tarefa 3 – Representação conceitual do *corpus*: essa tarefa consistiu em representar o conteúdo de coleções textuais multidocumento em uma estrutura conceitual. Para tanto, os textos-fonte foram indexados a um recurso léxico-conceitual já existente e a parcela pertinente foi recortada. Sendo assim, foi preciso selecionar o tipo de palavra a ser indexado, o método (manual ou semiautomático) de indexação e, o recurso léxico-conceitual.
- Tarefa 4- Investigação de critérios de relevância a partir de estruturas conceituais: consistiu em investigar alguns critérios (p.ex.: localização do conceito, número de relações do conceito, etc.) que são potencialmente relevantes para a distinção dos conceitos mais relevantes e menos relevantes a partir de estruturas conceituais. Para tanto, os conceitos que constituem a estrutura relativa às coleções multidocumento foram caracterizados em função dos critérios selecionados. Na sequência, a pertinência dos critérios foi analisada inicialmente de forma manual e, posteriormente, de forma automática.
- Tarefa 5 – Proposição de métodos de SAM: consistiu em propor métodos extrativos de SAM cujo processo de seleção de conteúdo é baseado nas propriedades mais relevantes identificadas na Tarefa 4. Tais métodos se caracterizam pela representação dos textos-fonte em uma estrutura léxico-conceitual e pela pontuação e ranqueamento das sentenças em função das propriedades dos seus respectivos conceitos na estrutura.
- Tarefa 6 - Avaliação dos métodos de SAM: consistiu em avaliar intrinsecamente os métodos propostos na Tarefa 5. Assim, os sumários produzidos pelos métodos aqui propostos foram avaliados automaticamente quanto à informatividade e manualmente quanto à qualidade linguística.

## 1.4. Estrutura da dissertação

Esta dissertação está organizada em 8 Seções. Na Seção 2, apresentam-se a revisão da literatura sobre (i) a sumarização automática multidocumento, (ii) a SA profunda baseada em conhecimento léxico-conceitual, (iii) a sumarização automática de ontologias e (iv) considerações sobre a revisão da literatura. Na Seção 3, apresenta-se o *corpus* CSTNews, na Seção 4 a parcela do mesmo utilizada neste trabalho, e a representação léxico-conceitual do corpus, isto é, a construção da hierarquia conceitual por meio da indexação dos nomes que ocorrem na parcela selecionada do corpus aos conceitos de uma ontologia pré-existente. Na Seção 5, descreve-se a investigação manual e automática da pertinência de medidas ou métricas de grafos para a identificação dos conceitos mais relevantes de uma coleção multidocumento, os quais devem compor um sumário correspondente à coleção. Na Seção 6, descrevem-se os critérios relevantes selecionados com base nas hierarquias conceituais apreendidos pelo AM. Na Seção 7, propõem-se métodos de SAM com base nas métricas mais pertinentes e apresentam-se os métodos de geração de extrato com base em duas abordagens, e descreve-se a avaliação intrínseca automática e manual dos mesmos. Por fim, na Seção 8, tecem-se algumas considerações finais sobre a pesquisa, delineam-se trabalhos futuros e destacam-se contribuições do trabalho.

# 2. Revisão da literatura

Na subseção 2.1, apresenta-se o cenário da SAM, focalizando (i) os conceitos básicos da área, (ii) os diferentes paradigmas de SA e os métodos desenvolvidos segundos esses diferentes paradigmas, principalmente para o português, e (iii) as estratégias de avaliação em SA. Na subseção 2.2, em especial, destacam-se os métodos do paradigma profundo baseados em conhecimento conceitual. Na subseção 2.3, descrevem-se trabalhos de Sumarização Automática de Ontologias (SAO), que lidam com representações ontológicas e critérios de relevância.

## 2.1. A Sumarização Automática Multidocumento

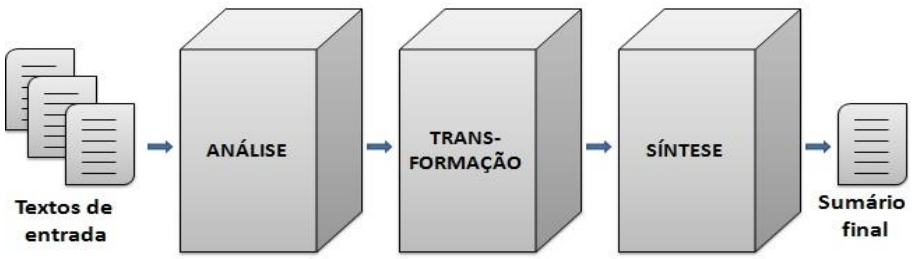
### *2.1.1 Conceitos básicos*

A SAM pode ser vista como uma especificação da SA monodocumento na medida que se parte não de um único texto, mas de uma coleção de textos (de fontes distintas) que abordam mesmo assunto para produzir um sumário (MANI, 2001, ORĂSAN, 2009).

Consequentemente, a SAM precisa lidar com questões clássicas da SA monodocumento, como a coesão e a coerência, e também com os fenômenos gerados pela multiplicidade de textos-fonte, como a ocorrência de informações redundantes, complementares e contraditórias, o emprego de estilos de escrita variados, a ordenação temporal dos eventos/fatos e as diferenças de perspectivas e focos.

Buscando emular na máquina as etapas de sumarização humana identificadas por Cremmins (1996) e Endres-Nieggemeyer (1998), a SAM envolve idealmente os 3 processos previstos por Mani e Maybury (1999) para a SA monodocumento, a saber: (i) análise dos textos-fonte, (ii) transformação e (iii) síntese (Figura 1).

Figura 1 - Arquitetura genérica de um sistema de SAM.




Fonte: Sparck Jones (1993).

A análise visa interpretar os textos-fonte e extrair uma representação formal dos mesmos. A transformação é a etapa principal, pois, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário. Essa etapa engloba a seleção do conteúdo central da coleção que irá compor o sumário até que a taxa de compressão seja atingida. A síntese visa à construção do sumário em língua natural a partir da representação interna gerada na transformação. Para tanto, métodos de justaposição, ordenação, fusão e correferenciação dos segmentos textuais selecionados podem ser utilizados.

As pesquisas em SAM intensificaram-se no início dos anos 2000 para a língua inglesa, devido à relevância da SA para a recuperação e extração de informação, como em buscadores de notícias (p.ex.: *Google News*) e em sintetizadores de informação (p. ex.: *Wolfram Alpha*). Atualmente, há várias abordagens e estratégias de sumarização, eventos científicos dedicados ao tema, competições e avaliações conjuntas e sistemas de amplo uso disponíveis na *web*.

Para o português brasileiro, as pesquisas são mais incipientes, iniciando-se em 2007. Apesar disso, elas já produziram: (i) recursos linguístico-computacionais, como o *corpus* de referência CSTNews (CARDOSO *et al.*, 2011a); (ii) ferramentas, como o *parser* discursivo CSTParser (MAZIERO, PARDO, 2011), e (iii) sistemas/métodos baseados nos diferentes paradigmas de SA, ou seja, no superficial, que usa pouco conhecimento linguístico (PARDO, 2005) ou puramente estatística (AKABANE *et al.*, 2011; RIBALDO *et al.*, 2011); no profundo, que faz uso massivo de conhecimento linguístico (CASTRO JORGE, 2010; CASTRO JORGE, PARDO, 2010; CARDOSO *et al.*, 2011b; CARDOSO, 2014) e, no híbrido, que une conhecimento linguístico e estatístico (CASTRO JORGE, PARDO, 2011; CASTRO JORGE *et al.*, 2011; RIBALDO *et al.*, 2012; RIBALDO, 2013; CAMARGO, 2013; SILVEIRA E BRANCO, 2012). Aliás, tais métodos/sistemas atingem e, por vezes, superam o estado-da-arte.

Assim, a SAM do português está em plena ebulição e as conquistas recentes resultam principalmente de 2 projetos relacionados e de longo prazo do NILC[4]: o **Sucinto**[5], que objetiva produzir recursos, ferramentas e aplicações, e o **Sustento**[6] (CNPq 483231/2012-6), que busca gerar conhecimento linguístico para a SAM.

[4] Núcleo Interinstitucional de Linguística Computacional (http://www.nilc.icmc.usp.br).
[5] http://www.icmc.usp.br/pessoas/taspardo/sucinto/
[6] http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23

As aplicações de SAM para o português, assim como as desenvolvidas para outras línguas, caracterizam-se por produzirem sumários informativos e genéricos a partir de coleções de notícias jornalísticas. Para tanto, elas adotam a abordagem ou paradigma extrativo, selecionando integralmente as sentenças mais importantes das coleções para compor os correspondentes sumários, que recebem o nome de "extratos". A essa abordagem, opõe-se à abstrativa, que visa à geração de sumários compostos pela reescrita da ideia central dos textos-fonte.

A ênfase dada à sumarização extrativa decorre do fato de que a geração textual é custosa para as máquinas, uma vez que conteúdo não-linguístico precisa ser realizado textualmente e, para isso, faz-se necessário o uso de recursos linguísticos como (i) gramaticais, que permitem a estruturação de sentenças, e (ii) estruturas conceituais, que permitem certo grau de generalização.

Para a produção dos extratos, a seleção do conteúdo central de uma coleção parte da pontuação e do ranqueamento das sentenças dos textos-fonte em função de um critério de relevância. A partir do ranque, as sentenças mais relevantes são selecionadas para o sumário, desde que haja pouca similaridade entre elas e até que a taxa de compressão seja atingida (MANI, 2001). Tal ranque pode ser construído com base em conhecimento linguístico simples ou estatística ou com base em conhecimento linguístico profundo. Aliás, a quantidade e o nível de conhecimento linguístico envolvidos na sumarização definem os diferentes paradigmas de SA.

Quanto à síntese, ressalta-se que os métodos superficiais, ao produzirem extratos, justapõem as sentenças extraídas dos textos-fonte na ordem em que estas ocorrem nos mesmos. Para a geração de *abstracts*, utilizam-se operações complexas de *cut-and-paste*, p.ex.: redução de sentença, combinação de sentença, transformação sintática, paráfrase lexical, generalização/especificação e reordenação (OTTERBACHER *et al.*, 2002).

Os métodos de SAM podem seguir um dos 3 diferentes paradigmas de SA, os quais são descritos a seguir. Além de apresentar os paradigmas, destacam-se os principais métodos extrativos desenvolvidos para o português segundo os paradigmas.

### *2.1.2. As abordagens ou paradigmas de SA no cenário multidocumento*

Segundo Mani (2001), a tarefa de sumarização automática pode ser abordada com base em um dos três paradigmas: superficial, profundo e híbrido.

Quanto paradigma superficial, Gupta e Lehal (2010) e Kumar e Salim (2012) estabelecem que os métodos se organizam em 3 grupos.

O primeiro grupo de métodos superficiais engloba os que se baseiam em atributos linguísticos (do inglês, *feature-based methods*), que podem variar em número e combinação (p.x.: LIN, HOVY, 2002) e apresentar pesos diferentes em função do tipo/gênero dos textos-fonte (cf. SUANMALI *et al.*, 2011). Um desses atributos é a frequência de ocorrência das palavras de classe aberta na coleção, sendo assim, a etapa de análise para estes métodos consiste em segmentar as sentenças e calcular a frequência de ocorrência de cada palavra nos textos-fonte da coleção. A etapa de transformação consiste em pontuar e ranquear as sentenças em função da soma da frequência de suas palavras constitutivas. Por conseguinte, os sumários são compostos pelas sentenças constituídas pelas palavras de classe aberta mais frequentes da coleção.

O segundo grupo engloba os trabalhos baseados em *cluster* e centroide (do inglês, *cluster-based methods*) (p.ex.: RADEV *et al.*, 2004). Neles, a análise consiste em agrupar as sentenças de dada coleção em *clusters* (conjuntos) com base na similaridade lexical. Assim, os *clusters* são formados por sentenças semelhantes entre si, que veiculam os tópicos da coleção. Cada *cluster* é representado por um centroide, ou seja, um conjunto de palavras estatisticamente importantes. De cada *cluster*, seleciona-se a sentença que contém o maior número de palavras em comum com o centroide.

O terceiro grupo engloba os métodos cuja análise consiste em modelar os textos-fonte de uma coleção em um grafo[7] (do inglês, *graph-based methods*) (p.ex.: SALTON *et al.*, 1997; MIHALCEA, TARAU, 2005). Nele, as sentenças são modeladas como nós e a similaridade lexical entre elas é modelada como arestas que conectam os nós. Assim, as sentenças mais conectadas a outras são extraídas para a construção do sumário.

Segundo a abordagem superficial destacam-se, para o português, os trabalhos de Pardo (2005), Akabane *et al.* (2011) e Ribaldo *et al.* (2012).

O método subjacente ao sumarizador GistSumm (PARDO, 2005) pode ser definido como "baseado em atributo linguístico", pois pontua e ranqueia as sentenças dos textos de uma coleção pela frequência de ocorrência de suas palavras na coleção. A

[7] A Teoria dos Grafos (DIESTEL, 2005) é um ramo da matemática que estuda as relações entre os objetos de determinado conjunto, constituindo um modelo matemático (grafo) para estudar as relações entre os objetos. A definição de um grafo pode ser representada por um par de conjuntos G = (V, E), em que V é o conjunto de vértices e E é o conjunto de arestas, formada por pares de vértices (BONDY, MURTY 1976, RUOHONEN, 2013).

sentença com a maior pontuação é considerada a *gist sentence* (isto é, sentença que expressa o conteúdo principal da coleção) e, consequentemente, selecionada para iniciar o sumário. As demais sentenças são selecionadas para o sumário caso satisfaçam a dois critérios: (i) conter pelo menos um radical em comum com a *gist sentence* e (ii) ter pontuação maior que a média das pontuações de todas as sentenças.

Os métodos desenvolvidos por Akabane *et al.* (2011) e Ribaldo *et al.* (2012) enquadram-se no grupo dos "baseados em grafos".

Akabane *et al.* (2011) desenvolveram um método que engloba grafo e redes complexas[8] (RC), o qual fundamenta o sistema RCSumm. No caso, os autores partem da representação dos textos-fonte em um grafo, nos moldes de Ribaldo *et al.* (2012), e aplicam 3 medidas de RC para ranquear e selecionar os sentenças, a saber: (i) grau de um nó *x* (isto é, quantidade de arestas diferentes conectadas a *x*), que indica quão um nó é conectado a seus vizinhos, sendo que, quanto maior o valor do grau, maior será a sua informatividade; (ii) coeficiente de aglomeração, que verifica se há uma conexão entre os vértices *x* e *k*, quando um vértice *x* está conectado a um vértice *y* e o vértice *y* a um vértice *k*, e (iii) caminho mínimo, isto é, a sequência mínima de arestas que leva um nó *x* ao outro, sendo que, quanto menor a distância, mais próximo, em média, o vértice *x* está dos demais (ou seja, mais informações relevantes *x* possui).

Ribaldo *et al.* (2012) desenvolveram o sumarizador RSumm, cujo método modela os textos-fonte em um grafo, em que as sentenças são representadas como nós e a proximidade lexical entre as sentenças por arestas. A seleção das sentenças segue o caminho denominado "denso" (do inglês, *bushy path*), segundo o qual os nós (sentenças) mais altamente conectados são selecionados para o sumário, posto que estes representam as informações mais relevantes da coleção.

Os métodos profundos de SAM extrativa podem ser organizados em 2 grupos de acordo com o tipo de conhecimento linguístico predominante.

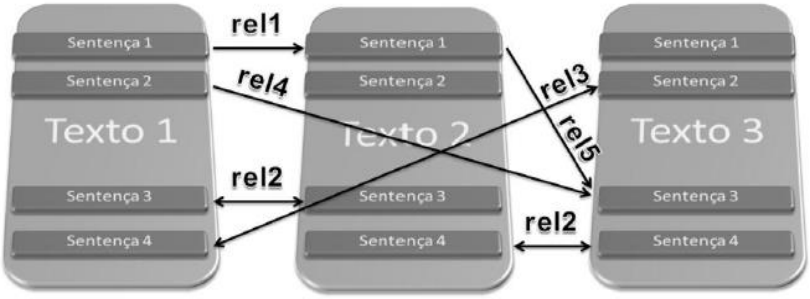
Em um grupo, há os métodos baseados em conhecimento discursivo. Para a sumarização de notícias em português, há inúmeros métodos baseados nesse tipo de conhecimento (p.ex.: CASTRO JORGE (2010); CASTRO JORGE E PARDO (2010); CARDOSO *et al.* (2011), etc.). Neles, a análise consiste em modelar os textos-fonte de

[8] Uma RC difere de um grafo tradicional pela grande quantidade de vértices, princípios complexos de organização (não aleatórios) e características topográficas particulares, como (i) coeficiente de aglomeração (ou fenômeno de transitividade), que indica a presença de um número elevado de triângulos (ou ciclos) na rede; (ii) distribuição de graus, função de distribuição probabilística que indica a probabilidade de determinado vértice ter grau (número de arestas) fixo, etc. (METZ *et al.* 2007).

uma coleção em um grafo, em que as sentenças são representadas por nós e as relações discursivas entre as sentenças são codificadas por arestas.

A Figura 2 ilustra tal modelagem.

Figura 2 - Esquema genérico de análise multidocumento.




Fonte: Maziero (2012).

As conexões entre as sentenças são rotuladas pelas relações da teoria/modelo *Cross-Document Structure Theory* (CST) (RADEV, 2000). Para o português, utiliza-se o conjunto de 14 relações proposto por Aleixo e Pardo (2008) quando da anotação manual do *corpus* CSTNews (CARSODO *et al.*, 2011) (Quadro 3).

Quadro 3 – Conjunto de relações CST de Aleixo e Pardo (2008).

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

Uma vez que as relações CST tenham sido identificadas na análise, a seleção das sentenças para compor o sumário extrativo da coleção, realizada durante a transformação, é feita com base no princípio da redundância. Segundo esse princípio, as sentenças cujo conteúdo mais se repete na coleção veiculam a informação principal da mesma e, por isso, devem compor o sumário, evitando-se que haja redundância entre elas. Nos métodos/sistemas baseados na CST, a redundância é capturada pelo número de relações CST que as sentenças possuem com as demais da coleção. Assim, as sentenças são pontuadas e ranqueadas pelo número de conexões no grafo, sendo que as sentenças com mais conexões ocupam o topo do ranque. As mais bem pontuadas são

selecionadas para compor o sumário desde que não sejam redundantes entre si e até que se atinja a taxa de compressão.

Apesar de seguir a mesma modelagem em grafo como o trabalho de Ribaldo *et al.* (2012), na teoria CST, o grafo representa apenas a modelagem do texto e, a partir dessa modelagem são aplicadas as informações discursivas provenientes da CST. Em contrapartida, o trabalho de Ribaldo *et al.*(2012), por exemplo, não utiliza nenhum conhecimento profundo, muito pelo contrário, é com base na estrutura do grafo e nas medidas calculadas de acordo com este que as informações para o sumário são extraídas.

No sistema CSTSumm de Castro Jorge e Pardo (2010), aplicam-se, ao ranque original de sentenças, operadores de seleção de conteúdo que codificam preferências do usuário, como "apresentação de informação contextual". Uma vez ativado, um operador reordena as sentenças, privilegiando a informação relevante para o usuário. As sentenças são selecionadas a partir do novo ranque.

Há também os métodos propostos por Cardoso (2014), que combinam informações codificadas pelas relações CST e pelas relações da *Rhetorical Structure Theory* (RST) (MANN, THOMPSON, 1987). A RST é uma teoria/modelo linguístico descritivo que classifica os segmentos discursivos de um texto individual em núcleo (informação principal) ou satélite (informação adicional), relacionando-os por meio de relações retóricas, como *Elaboration, List, Causa, Result, Justify*, etc. Uma vez que um texto tenha sido representado com base na RST, obtém-se uma árvore retórica do mesmo. Em Cardoso (2014), utilizou-se o CSTNews, cujos textos-fonte foram anotados com base nas 14 relações CST do Quadro 3 e nas 32 relações RST de Pardo e Nunes (2008). O melhor método de SAM de Cardoso (2014), o RC-4, pontua e ranqueia as sentenças de cada coleção multidocumento combinando dois critérios de relevância: (i) a saliência[9] da sentença em seu respectivo texto-fonte, codificada via RST, e (ii) a

[9] Cardoso (2014) utiliza o modelo de saliência de Marcu (1997), que se baseia no uso de um conjunto promocional formado pelas unidades mais salientes de cada nó interno de uma árvore RST. Dada uma árvore RST, os conjuntos promocionais são construídos no sentido *bottom-up*. Assim, o conjunto promocional de um nó-folha é composto por ele mesmo. O conjunto promocional dos nós internos da árvore é composto pela união dos conjuntos promocionais de seus filhos nucleares. Para pontuar cada segmento, atribui-se à raiz uma pontuação correspondente ao número de níveis da árvore. Na sequência, percorre-se (sentido *top-down*) a árvore em direção ao segmento sob avaliação e, cada vez que o segmento não está no conjunto promocional de um nó durante o percurso, o segmento tem a pontuação decrementada de 1. A ideia é que as unidades textuais que estão no conjunto promocional do topo de uma árvore são mais importantes do que as unidades encontradas mais abaixo.

correlação com os fenômenos multidocumento, indicada pela CST. Atualmente, o RC-4 é o método profundo de melhor desempenho para o português.

Em outro grupo, estão os métodos baseados em conhecimento léxico-conceitual, que são descritos com destaque em 2.2 (pág. 19) por serem o foco deste trabalho.

Quanto aos métodos de SAM híbridos, Schiffman *et al.* (2002) caracteriza-se por unificar informações superficiais (localização da sentença nos textos-fonte e tamanho da sentença) e conhecimento léxico-conceitual. Além dos atributos superficiais, os autores relacionam as palavras dos textos-fonte pela sinonímia e hiponímia (cf. 4.2, pág. 50) para delimitar os conceitos mais representativos da coleção.

Para o português, destacam-se os trabalhos de Castro Jorge e Pardo (2011), Castro Jorge *et al.* (2011), Ribaldo *et al.* (2012), Silveira e Branco (2012), Ribaldo (2013), Camargo (2013) e Castro Jorge (2015).

Castro Jorge e Pardo (2011), Castro Jorge *et al.* (2011) e Castro Jorge (2015) investigam, de modo geral, métodos híbridos em que a SAM é modelada em uma história gerativa via o modelo estatístico *Noisy-Channel* (Canal Ruidoso) (SHANNON, 1948). Segundo os autores, a modelagem gerativa da SAM é entendida da seguinte forma: um sumário é emitido por uma fonte e em seguida sofre um processo de transformação, o que geraria um conjunto de textos-fonte a partir dos quais esse sumário foi gerado. Presume-se que a fonte irá produzir um sumário multidocumento informativo, ou seja, um texto condensado que contém as informações mais relevantes de um conjunto de textos sobre um mesmo assunto. Fatores como gramaticalidade, coerência, coesão, e relevância da informação são expressos por meio de um modelo de linguagem que avalia a informatividade do sumário, como a medida ROUGE (cf. pág. 17). A ideia principal dos trabalhos baseados na modelagem estatística gerativa via *Noisy-Channel* é recuperar esse sumário original a partir da modelagem de transformação, o que representaria a história gerativa da SAM. Na referida modelagem, o canal ruidoso representa a etapa de transformação, em que a seleção de conteúdo ocorre. Sendo assim, o "ruído" é constituído por elementos provenientes dos fenômenos multidocumento. No caso, os autores testaram vários atributos superficiais (p.ex.: localização da sentença no texto-fonte) e profundos (p.ex.: relações CST) de seleção de conteúdo nessa modelagem. As relações CST, por exemplo, são representadas por fatores multidocumento como a (i) redundância, que pode ser modelada por meio das relações *Identity*, *Equivalence*, *Subsumption*, *Overlap* e *Summary*, (ii) a contradição que pode ser modelada pela relação *Contradiction*, e (iii) a complementaridade, que pode

ser modelada pelas relações *Elaboration*, *Historical background* e *Follow-up*. Em Castro Jorge (2015), em especial, o canal ruidoso engloba ainda um modelo que captura padrões de boa construção de um sumário multidocumento em termos de coerência. As vantagens desses modelos são a possibilidade de se explorar a geração de sumários analisando-se diferentes fatores que podem ser representados pela CST ou qualquer outro modelo semântico-discursivo, e a busca pelo sumário mais provável explorando-se fatores que influenciam na informatividade desses sumários.

No trabalho de Ribaldo *et al.* (2012), desenvolvido para o português, utiliza-se grafo (e suas medidas) em combinação com as relações CST para representar e sumarizar textos. Nos métodos propostos, os textos são modelados em grafo, em que as sentenças são representadas como nós e as relações CST entre as sentenças são codificadas em arestas. Em um dos métodos, as arestas codificam o número de relações CST de uma sentença, independente de seu tipo. Em outro, as arestas possuem diferentes pesos em função do tipo das relações CST; relações que codificam redundância (p.ex.: *Identity*, *Equivalence*, *etc.*) possuem pesos mais altos que as demais. Em Ribaldo (2013), utiliza-se grafo em combinação com subtópico. Especificamente, após a modelagem dos textos-fonte em grafo, realiza-se a segmentação topical com base em uma das estratégias disponíveis na ferramenta *TextTiling* (Hearst, 1997) para o português. Feito isso, realiza-se o agrupamento de subtópicos semelhantes expressos por diferentes trechos textuais com base na ocorrência de palavras-chave entre eles. Na sequência, aplica-se o método de seleção de conteúdo denominado *caminho denso segmentado* (*segmented bushy path*), baseado em Salton *et al.* (1997), que consiste em selecionar as sentenças mais importantes (mais conectadas) de cada subtópico. Feita a primeira escolha da sentença do primeiro conjunto (subtópico), é necessário selecionar uma sentença de transição antes mesmo da escolha da segunda sentença mais relevante de outro subtópico. A sentença de transição é escolhida de forma que a mesma seja cronologicamente anterior à sentença principal subsequente para que a passagem de um subtópico para outro se dê de forma coerente. Para eliminar a redundância, utiliza-se uma medida de similaridade. Por fim, o sumário é gerado até que se atinja a taxa de compressão.

O trabalho de Silveira e Branco (2012) apresenta um sistema para a sumarização automática multidocumento extrativa, o SIMBA, que combina uma abordagem de duplo *cluster* com simplificação de sentença, com o intuito de produzir sumários mais informativos. Nesse método, as sentenças são agrupadas em *clusters* por similaridade e

apenas uma sentença do *cluster* é extraída para reduzir a redundância, que é calculada com base em uma medida de similaridade. Em seguida, as sentenças são novamente agrupadas mas, dessa vez, em tópicos e, assim, formam outro *cluster* considerando a coleção de palavras-chave que assegura a preservação da ideia principal do conteúdo do texto e, a organização final do sumário. Para a construção do sumário, as sentenças são simplificadas com o intuito de reduzir o conteúdo original e trazer sentenças mais informativas. A simplificação diminui as sentenças retirando partes consideradas dispensáveis, que contêm informações pouco relevantes tratando-se da mensagem geral transmitida pela sentença. Por fim, o sumário é gerado até que se atinja a taxa de compressão que é determinada pelo usuário.

Os resultados das avaliações do SIMBA atestaram que os sumários preservam a ideia da coleção original de textos e contêm, ao mesmo tempo, sentenças incisivas e simples que transmitem a informação mais importante presente nos textos de entrada.

Camargo (2013), por sua vez, não propôs exatamente um método híbrido de SAM. A autora realizou um estudo de *corpus*, cujos textos-fonte e sumários manuais (humanos) foram alinhados manualmente em função do compartilhamento de conteúdo, o que evidencia a origem das sentenças que compõem os sumários. As sentenças dos textos-fonte e sumários alinhadas (e não-alinhadas) foram caracterizadas em função de uma série de atributos superficiais e profundos. Tais dados foram submetidos a algoritmos de Aprendizado de Máquina (AM) para que as características mais salientes das sentenças alinhadas fossem aprendidas por meio de regras. O conjunto de regras com melhor acurácia evidencia que os sumários mutidocumento são compostos por sentenças que possuem características linguísticas híbridas, as quais revelam estratégias humanas de sumarização. Segundo Camargo (2013), os humanos selecionam as sentenças que satisfazem as seguintes características: (i) localização no início dos textos-fonte, (ii) redundância (alta) de seu conteúdo, (iii) ocorrência das palavras mais frequentes da coleção e (iv) tamanho médio ou pequeno (em número de palavras). As regras compostas por esses atributos híbridos produzem resultados iniciais que são competitivos com um método do estado da arte para o português. Em outras palavras, as estratégias de SHM de Camargo (2013), codificadas em regras que salientam certas propriedades linguísticas, podem subsidiar um método de SAM cujo desempenho será compatível com o estado da arte para o português. Esse método, no entanto, será o primeiro baseado em estratégias de SHM.

A seguir, apresenta-se a revisão literária sobre como ocorre a avaliação de sistemas/métodos em SA.

### *2.1.3. As estratégias de avaliação em SAM*

As estratégias de avaliação já foram bastante exploradas no cenário do PLN, pois permitem verificar o avanço do estado da arte dos sistemas/métodos. Quanto à SA, as conferências internacionais, como a SUMMAC[10] (*Text Summarization Evaluation Conference*) e a DUC (*Document Understanding* Conference) (DANG, 2005), iniciada em 2001 e que passou a se chamar TAC[11] (*Text Analysis Conference*) em 2008, desempenharam papel central no estabelecimento dos parâmetros de avaliação.

De um modo geral, a avaliação de sistemas de SA pode ser classificada em intrínseca ou extrínseca. A primeira foca a avaliação do desempenho dos sistemas por meio da análise de seus resultados (sumários). A segunda foca a avaliação da utilidade dos sumários em alguma tarefa específica, por exemplo, na recuperação de informação (SPARCK JONES; GALLIERS, 1996).

Reconhece-se, na literatura, que a avaliação extrínseca é uma tarefa demorada, cara e que requer um planejamento cuidadoso (HALTEREN; TEUFEL, 2003) e que a intrínseca deve focar a qualidade e a informatividade dos sumários (MANI, 2001). A avaliação intrínseca, aliás, é a mais frequentemente realizada nos trabalhos de SA.

Há dois aspectos principais que são o alvo da avaliação intrínseca dos sumários produzidos automaticamente: a informatividade e a qualidade linguística (MANI, 2001). A informatividade diz respeito à quantidade de informação relevante que está contida nos sumários e esse tipo de avaliação é geralmente realizada de forma automática. A qualidade diz respeito a fatores relacionados à gramaticalidade, coesão, coerência, etc. Fatores esses que são avaliados de forma manual.

Para a avaliação de informatividade, uma das medidas automáticas amplamente usadas é a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation* (LIN, 2004), pois, além de ser de domínio público, é a medida mais adotada em conferências internacionais. O princípio dessa medida é basicamente comparar a quantidade de n-gramas (palavras) em comum entre o sumário produzido automaticamente e um ou mais sumários humanos, também chamados sumários de referência. Essa medida fornece

[10] http://www-nlpir.nist.gov/related_projects/tipster_summac/
[11] http://www.nist.gov/tac/about/index.html

resultados em termos de precisão, cobertura e medida-f, cujas métricas estão descritas em (1).

A precisão diz respeito ao número de n-gramas em comum entre o(s) sumário(s) de referência e o sumário automático em relação ao total de n-gramas do sumário automático. A cobertura diz respeito ao número de n-gramas em comum entre o(s) sumário(s) de referência e o sumário automático em relação ao total de n-gramas do(s) sumário(s) de referência. Em outras palavras, a precisão captura a quantidade de informação do sumário de referência que está no sumário automático e a cobertura, por sua vez, captura o quanto da informação do sumário de referência foi coberto pelo sumário automático. Essas duas medidas são complementares e são ponderadas pela medida-f, que calcula a média harmônica da precisão e cobertura. Como precisão e cobertura são inversamente relacionadas, uma tende a diminuir quando a outra sofre aumento.

(1)

$$\text{Precisão} = \frac{\text{n-gramas em comum entre sumário automático e humano}}{\text{n-gramas do sumário automático}}$$

$$\text{Cobertura} = \frac{\text{n-gramas em comum entre sumário automático e humano}}{\text{n-gramas do sumário humano}}$$

$$\text{Medida-f} = \frac{2 \times \text{Precisão} \times \text{Cobertura}}{\text{Precisão} \times \text{Cobertura}}$$

A ROUGE é muito popular no PLN porque é barata e facilmente aplicável a qualquer tipo de sumário. Uma desvantagem dessa medida é que, por apenas avaliar correspondência de n-gramas, ela não considera todo aspecto relacionado à qualidade dos sumários.

Para a avaliação da qualidade linguística dos sumários automáticos, a TAC sugeriu 5 aspectos: (i) gramaticalidade, que se refere aos padrões de boa ortografia, pontuação e sintaxe, (ii) coerência, que se refere à manutenção da organização textual de forma que preserve o sentido do texto, (iii) não redundância, que se refere ao fato de que não existam informações repetitivas no sumário, (iv) foco, que se refere ao fato de que as partes do texto devem estar relacionadas com o todo e (v) clareza referencial, que se refere a presença adequada de componentes linguísticos que liguem apropriadamente

os elementos do sumário. Para avaliar os sumários de acordo com esses critérios, a TAC sugere que cada um dos aspectos seja pontuado com valores entre 1-5, sendo que 1 significa “muito ruim” e 5 significa “muito bom”.

Além dessas possibilidades, vários autores têm investigado outras, já que não há consenso sobre a melhor forma de se avaliar um sistema dessa natureza. Dentre eles, citam-se, por exemplo, Nenkova e Passonneau (2004) e Louis e Nenkova (2013).

O método da pirâmide de Nenkova e Passonneau (2004) considera um conjunto de sumários de referência a partir dos quais são extraídas “unidades de conteúdo do sumário” (*summarization content units* - SCU). As SCU são organizadas em uma pirâmide cujo topo representa as que aparecem na maioria dos sumários de referência. As SCU são pontuadas de acordo com a posição na pirâmide, sendo mais bem pontuadas as SCU localizadas mais no topo do que as demais. Os sumários automáticos mais informativos são os que têm maior número de SCU próximas do topo da pirâmide.

Louis e Nenkova (2013) propuseram 3 métodos de avaliação, visando reduzir a influência da subjetividade humana na tarefa e a dificuldade de se obter dados provenientes de humanos. No primeiro, mede-se a similaridade entre os textos-fonte e os sumários automáticos, assumindo que um bom sumário é similar aos textos dos quais foram gerados. No segundo, adicionam-se, a um pequeno conjunto de sumários de referência, sumários automáticos escolhidos por humanos (pseudomodelos). No terceiro método, faz-se uso somente de sumários automáticos para construir o conjunto de sumários de referência, seguindo um critério similar ao método da pirâmide. Assim, as informações relevantes são aquelas que aparecem na maioria dos sumários automáticos, e os sumários que mais possuam essas informações relevantes são os novos sumários de referência. Com a proposição desses métodos, Louis e Nenkova salientam que as avaliações humanas podem ser reproduzidas por essas métricas totalmente automáticas com alta precisão.

A seguir, apresentam-se os principais métodos profundos baseados em conhecimento léxico-conceitual.

## 2.2. A SA profunda baseada em conhecimento conceitual

Como mencionado, os métodos profundos utilizam principalmente conhecimento discursivo ou léxico-conceitual. Este último foi explorado na SA monodocumento, por exemplo, por Reimer e Hahn (1988, apud MANI, 2001) e Wu e Liu (2003).

Reimer e Hahn (1988, apud MANI, 2001) desenvolveram o TOPIC, um sistema para o alemão capaz de identificar os principais conceitos de um texto-fonte sobre "computadores", os quais podem ser aplicados para a produção de sumários.

É importante ressaltar que apesar de MANI (2001) agregar o TOPIC à sumarização, tal sistema não sumariza textos. Ele apenas identifica os conceitos mais importantes dentro de um único texto através de uma hierarquia conceitual.

O TOPIC inicialmente identifica o núcleo dos sintagmas nominais das sentenças do texto-fonte. Na sequência, indexa as palavras que funcionam como núcleo sintagmático aos conceitos de uma ontologia[12] em alemão do referido domínio que fora previamente construída por especialistas. O sistema, então, aumenta o peso do conceito na medida que ele ocorre no texto e, por conseguinte, é indexado à ontologia. Ao final, para determinar os conceitos mais salientes ou importantes do texto, o sistema calcula: (i) a frequência de ativação de um conceito *x* em relação à frequência de ativação dos demais conceitos indexados e (ii) o número de instâncias (ou conceitos subordinados) de *x* ativadas em relação ao número total de instâncias.

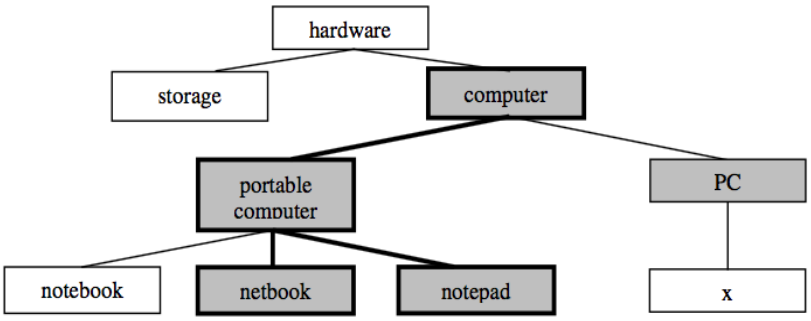
Na Figura 3, ilustra-se parte da referida ontologia do domínio "computador". Para ilustrar o cálculo da relevância realizada pelo TOPIC, considera-se que os núcleos sintagmáticos "*computer*", "*netbook*", "*notepad*" e "*PC*" da Figura 3 ocorreram uma única vez em seu respectivo texto-fonte e que "*portable computer*" ocorreu duas vezes. Assim, os conceitos correspondentes a essas palavras possuem pesos que refletem a sua frequência de ocorrência. Após calcular a saliência dos conceitos ativados com base nos critérios (i) e (ii), o TOPIC, baseando-se na representação arbórea[13] da ontologia, identifica o ramo da árvore composto pelos principais conceitos ocorridos no texto. No caso, o ramo da árvore conceitual mais saliente é o que está negritado na Figura 3.

É possível ainda generalizar o conceito principal, "*portable computer*", para "*hardware*", mesmo que este não tenha ocorrido no texto.

[12] O termo "ontologia" é controverso, denominando objetos bem distintos entre si e de ampla utilidade em várias áreas do conhecimento. Em seu sentido filosófico, ó termo "ontologia" é assim definido no Dicionário Oxford de Filosofia: '[...] o termo derivado da palavra grega que significa "ser", mas usado desde o século XVII para denominar o ramo da metafísica que diz respeito àquilo que existe' (BLACKBURN, 1997). Na Ciência da Computação, a definição mais difundida é "uma especificação explícita de uma conceitualização" (GRUBER, 1995) (cf. 2.3, pág. 27).

[13] Na teoria dos grafos, uma árvore é um grafo simples, no qual não existem ciclos.

Figura 3 - Saliência de um ramo da estrutura conceitual.

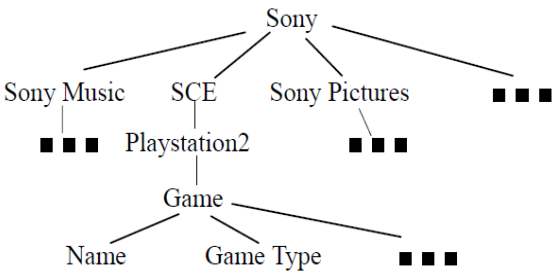



Fonte: Reimer e Hahn (1988, apud MANI, 2001)

O método profundo de Wu e Liu (2003) pauta-se na indexação das palavras de conteúdo dos parágrafos de um texto aos conceitos de ontologia, sendo que os conceitos mais frequentes do 2º nível ontológico (sentido *top-down*) correspondem aos tópicos do texto e, por isso, os parágrafos que os contêm são selecionados para o sumário. Especificamente, os autores utilizam uma ontologia do domínio *Sony Corporation*, composta por 142 conceitos, a qual está ilustrada na Figura 4. Do ponto de vista formal, a ontologia da Figura 4 é uma árvore, cujos conceitos, como *Sony*, *Sony Music* e *Sony Pictures*, são codificados em nós e as relações entre eles em galhos.

Quanto à indexação, ressalta-se que, a cada palavra indexada a um conceito, este e seus conceitos superordenados são pontuados. Por exemplo, se *Game* da Figura 4 for pontuado, pontuam-se também *Playstation2*, *SCE* ("*Sony Computer Entertainment*") e *Sony*. Ao final, os conceitos de maior pontuação do 2º nível correspondem aos tópicos do texto e seus respectivos parágrafos são selecionados para o sumário.

Figura 4 - Hierarquia do domínio *Sony Corporation* de Wu e Liu (2003)

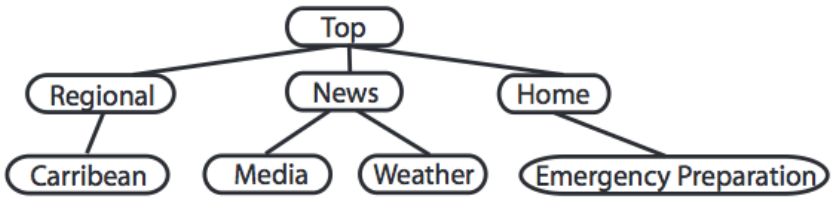



Fonte: Wu e Liu (2003).

O método de Wu e Liu (2003) é chamado “método baseado em ontologia” e foi comparado a outro dos mesmos autores, chamado “método de seleção aleatória”, que se baseia em frequência de palavras e peso de sentenças, sem utilizar informação ontológica. A comparação dos dois métodos comprovou que a performance do método baseado em ontologia é muito melhor em relação ao aleatório. Sendo assim, o baseado em ontologia pode extrair informação útil para a tarefa de sumarização.

Destaca-se também o trabalho de Hennig *et al.* (2008). O método de SA proposto por esses autores parte da indexação dos textos-fonte a uma ontologia para, na sequência, identificar os conceitos mais representativos dos textos-fonte por meio de características ou atributos relativos às propriedades ontológicas, como as relações semânticas entre os conceitos do texto e da ontologia. Ao final, as sentenças que veiculam tais conceitos são selecionadas para o sumário. Nesse trabalho, Hennig *et al.* (2008) construíram uma ontologia de 1036 conceitos. Cada conceito dessa ontologia é representado por um rótulo (do inglês, *tag*), como *Weather* (“Clima”), e por um “saco de palavras” (do inglês, *bag-of-words*), ou seja, um conjunto de palavras de conteúdo semanticamente relacionadas ao rótulo. Na Figura 5, ilustram-se simplificadamente os conceitos superiores da ontologia representados por rótulos simples.

Figura 5 - Os conceitos superiores da taxonomia de Hennig *et al.* (2008).




Fonte: Hennig *et al.* (2008).

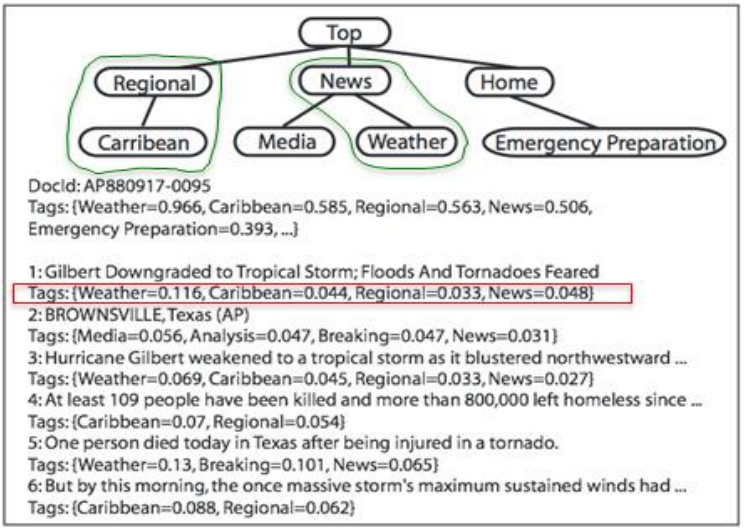
Para mapear as sentenças dos textos-fonte à ontologia, calcula-se inicialmente o quão cada conceito/rótulo da ontologia representa o conteúdo dos textos-fonte. Esse cálculo é feito por uma medida similar à *tf-idf*[14] (do inglês, *term frequency-inverse document*

[14] A *tf-idf* combina as medidas *tf* (*term frequency*) e *idf* (*inverse document frequency*). A primeira é a frequência de ocorrência simples de uma palavra em um documento. A segunda é a frequência inversa do documento, definida pela fórmula $idf=\log(N/n_i)$, em que $N$ é o número de documentos do *corpus* e $n_i$ é o número de documentos em que a palavra $i$ ocorre (assim, quanto menor o número de documentos em que $i$ ocorre, maior será o valor da *idf* de $i$). A média final *tf-idf* é data pela multiplicação do *tf* pelo *idf* (*idf=tf*

*frequency*) e, como resultado, valores de 0 a 1 são associados aos conceitos, sendo que os valores próximos a 1 indicam alta relevância. Na Figura 6, em que se ilustra o mapeamento das sentenças de uma notícia sobre o *Hurricane Gilbert* ("furacão Gilbert"), observa-se que, dentre os conceitos/rótulos que representam o conteúdo do texto (localizados imediatamente abaixo da taxonomia), *Weather* é o de pontuação mais alta, isto é, 0.966, sendo, portanto, o mais representativo do texto.

Na sequência, mapeia-se cada sentença dos textos aos conceitos da ontologia. Esse mapeamento é feito pela sobreposição entre as palavras de conteúdo da sentença e os "sacos de palavras" que representam cada conceito. Como resultado, a sentença é indexada aos conceitos correlatos, os quais recebem valores de 0 a 1, sendo que os valores próximos a 1 indicam alta correlação entre S e o "saco de palavras" do conceito. Por exemplo, a sentença 1 da Figura 6, "*Gilbert downgraded to tropical storm; Floods and tornadoes feared*"[15], foi mapeada a 4 conceitos, com os seguintes valores de similaridade: *Weather*=0.116, *Caribbean*=0.044, *Regional*=0.033 e *News*=0.048 (destaque em vermelho), o que configura o mapeamento a 2 subárvores ou ramos (*Weather*→*News* e *Caribbean*→*Regional*) (destaque em verde).

Figura 6 - Indexação das sentenças-fonte à hierarquia conceitual.




Fonte: Hennig *et al.* (2008).

Para o cálculo da relevância das sentenças, Hennig *et al.* (2008) utilizam 2 atributos pautados na representação das mesmas na hierarquia. Um deles é a localização das

x *idf*). Dessa forma, a *tf-idf* pauta-se na ideia de que palavras muito frequentes têm *idf* baixo e, apesar de um *tf* alto, são menos importantes que palavra raras. Em outras palavras, essa medida prefere palavras que são frequentes em um documento corrente *j*, mas são raras no *corpus* (JURAFSKY, MARTIN, 2008).

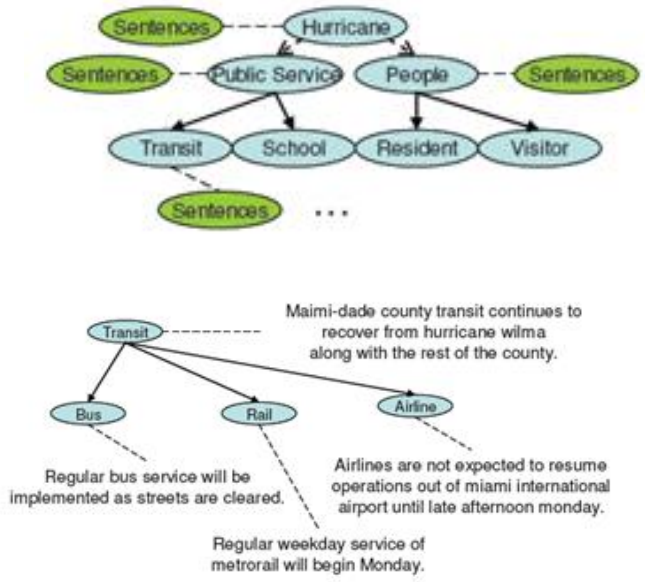
[15] "Gilbert rebaixado para tempestade tropical; Inundações e tornados temidos" (tradução nossa).

subárvores (em inglês, *subtree depth*) a que a sentença foi mapeada, que busca capturar a "especificidade do conteúdo" da sentença. O outro atributo é o número de ramos distintos a que a sentença foi mapeada (em inglês, *subtree count*), que busca capturar "a quantidade de informação distinta expressa". Os autores não fornecem informações detalhadas sobre como esses atributos são pesados para determinar a relevância. No entanto, um sistema de base (os autores não nomeiam) e os resultados do melhor sistema da DUC 2002 foram utilizados como comparativos ao método em questão. Apesar de os resultados não serem tão bons quanto os da DUC 2002 pelo fato de que o classificador hierárquico não foi treinado com os dados da DUC , o mapeamento de sentenças aos nós da ontologia, com treinamentos *offline* é muito eficiente e, sendo assim, importante para um sistema *online* de sumarização.

Em Li *et al.* (2010), um método *query-based* de SAM é proposto. Para tanto, os autores mapeiam as sentenças de uma coleção de textos-fonte aos conceitos de ontologia. Dada uma *query* ("consulta") do usuário, que também é mapeada à ontologia, o método seleciona apenas as sentenças dos textos indexadas aos conceitos a que as palavras da *query* foram mapeadas e/ou a conceitos subordinados. O sumário resultante não é genérico, mas sim "focado no interesse do usuário".

A ontologia utilizada por Li *et al.* (2010) é do domínio "desastre" e foi construída manualmente por especialistas; os conceitos que a constituem são expressos por rótulos únicos (Figura 7). Não há muitos detalhes em Li *et al.* (2010) sobre como efetivamente a indexação é feita.

Figura 7 - Indexação e seleção de conteúdo em Li *et al.* (2010).




Fonte: Li *et al.* (2010).

Na parte superior da Figura 7, vê-se um exemplo em que as sentenças (elipses verdes) dos textos-fonte sobre a "passagem do furacão Wilma por Atlanta em 2005" foram indexadas aos conceitos (elipses azuis) *Transit*, *Bus*, *Rail* e *Airline*. Na parte inferior da Figura, observa-se que, diante de uma *query* como "*get all the information related to transit in Miami-Dade County after Hurricane Wilma passed*"[16], indexada ao conceito *Transit*, apenas as sentenças indexadas a esse conceito e a seus subordinados, *Bus*, *Rail* e *Airline*, seriam selecionadas para compor o sumário.

Basicamente, o trabalho de Li *et al.* (2010) foi avaliado através da medida ROUGE comparando os sumários gerados pelo método com sumários humanos que foram construídos com base no tópico em questão. Os resultados indicaram que a eficiência da sumarização é significativamente melhorada através da adoção de ontologia.

Para o português, Silva (2006) propôs o sumarizador monodocumento ExtraWeb, que gera extratos de documentos provenientes da *web*. O método do ExtraWeb possui semelhança com o de Wu e Liu (2003) quanto ao critério de relevância para selecionar as sentenças. Quanto ao recurso conceitual, Silva (2006) utiliza uma versão enriquecida da ontologia em português do *Yahoo*, cuja estrutura e conteúdo se assemelham à de Hennig *et al.* (2008). Nela, um conceito é representado por um único identificador (ou rótulo) e um conjunto de palavras semanticamente relacionadas, denominadas descritores (ou "saco de palavras"). No caso, o identificador é uma palavra de língua natural (nome, verbo, etc.) que representa o conceito (p.ex.: "futebol") e os descritores são palavras que sinalizam a presença do conceito, sendo que o conjunto de descritores engloba o identificador (p.ex.: futebol, pênalti, gol, etc.).

Para selecionar as sentenças a compor o sumário, as palavras do documento-fonte são indexadas aos identificadores da ontologia pela similaridade ou sobreposição com os descritores, sendo que a propagação da pontuação de um conceito subordinado indexado aos seus conceitos superordenados também foi adotada por Silva (2006). Ao final, os rótulos/conceitos mais pontuados (ou seja, mais frequentes) são considerados os tópicos do documento. Por conseguinte, as sentenças do documento constituídas pelos conceitos mais frequentes são selecionadas para compor o sumário.

Os resultados do ExtraWeb evidenciaram que, quando apenas o peso da sentença é utilizado como critério para produzir um sumário, este tem problemas de textualidade

[16] "Obter todas as informações relacionadas ao trânsito em Miami-Dade County após o furacão Wilma". (trad. nossa)

enquanto que, o processamento ontológico parece filtrar melhor as sentenças relevantes contribuindo para a informatividade dos extratos.

Além de Silva (2006), destaca-se Tosta (2014), que propôs 2 métodos de SAM multilingue (SAMM). Objetivando gerar sumário em português a partir de coleções bilíngues, o autor construiu o *corpus* CM2News (*Corpus* Multidocumento Bilíngue de Textos Jornalísticos)[17] com 20 coleções, cada uma delas composta por 1 notícia em português, 1 notícia em inglês e 1 sumário multidocumento de referência (humano). Os métodos caracterizam-se por, dada uma coleção, indexar os nomes que ocorrem nos 2 textos à WN.Pr[18]. Na sequência, as sentenças dos textos são pontuadas e ranqueadas com base na frequência de ocorrência de seus conceitos constitutivos na coleção.

A partir do ranque, o Método 1 seleciona apenas as sentenças em português com pontuação mais alta para compor o sumário, até que a taxa de compressão desejada seja atingida. O Método 2 seleciona as sentenças mais bem pontuadas independentemente de sua língua-fonte. Caso sentenças em inglês sejam selecionadas, faz-se a tradução automática para o português. Os Métodos 1 e 2 foram avaliados de forma intrínseca, considerando-se a qualidade linguística e a informatividade dos sumários. Para avaliar a qualidade linguística, 15 linguistas computacionais analisaram manualmente a gramaticalidade, a não-redundância, a clareza referencial, o foco e a estrutura/coerência dos sumários e, para avaliar a informatividade, os sumários foram automaticamente comparados a sumários de referência pelo pacote de medidas ROUGE. Em ambas as avaliações, os resultados evidenciam o melhor desempenho do Método 1, o que pode ser justificado principalmente pelo fato de que as sentenças selecionadas são provenientes de um mesmo texto-fonte. Além disso, ressalta-se o melhor desempenho dos dois métodos baseados em conhecimento léxico-conceitual frente aos métodos mais simples de SAMM, os quais realizam a tradução automática integral dos textos-fonte.

Na sequência, descrevem-se trabalhos desenvolvidos na área de pesquisa denominada Sumarização Automática de Ontologias (SAO). Tais trabalhos são revisados porque utilizam diversos critérios para determinar a relevância dos conceitos em uma estrutura ontológica que podem ser aplicados neste trabalho.

[17] http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23
[18] A WN.Pr é descrita em detalhes em 4.2 (pág. 50).

## 2.3. A Sumarização Automática de Ontologias

### *2.3.1. Ontologia: definição*

Na Ciência da Computação (e, por conseguinte, na SAO), a definição mais utilizada de "ontologia" é "uma especificação explícita de uma conceituação" (GRUBER, 1995). "Conceituação" significa um modelo abstrato para representar uma visão simplificada de um domínio para algum propósito, o qual descreve conceitos e instâncias particulares do domínio, bem como as relações que se estabelecem entre esses elementos. A característica "explícita" ("formal") significa que os elementos do modelo são expressos em uma linguagem formal (isto é, linguagem explícita e computacionalmente tratável de representação do conhecimento) (STUDER *et al.*, 1998). Borst (1997) refina a definição de Gruber (1995) dizendo que essa ontologia é "uma especificação formal de uma conceituação compartilhada" e, nesse caso, a descrição "compartilhada" significa que é resultado de uma visão consensual sobre o conhecimento em questão.

As ontologias têm sido utilizadas para uniformizar o compartilhamento de informações em várias aplicações, pois promovem a compreensão de um conhecimento comum de um domínio. Para tanto, o conhecimento representado em uma ontologia é especificado formalmente para a interpretação e compreensão por máquinas (e pessoas) (STAAB, STUDER 2003). Várias linguagens para a descrição de ontologias foram desenvolvidas, como: lógica descritiva, RDF (*Resource Description Framework*) e OWL[19] (*Ontology Web Language*). A OWL, em especial, pode ser vista como uma estrutura abstrata ou como um grafo de RDF (W3C, 2012). Baseada na Lógica Descritiva, a OWL é composta por três tipos centrais de elementos: as classes (categorias ou conceitos), as propriedades (relações) e os indivíduos (do domínio). Por exemplo, (i) uma classe *Person* pode ser usada para representar o conjunto de todas as pessoas, (ii) a uma propriedade *parentOf* pode ser usada para representar a relação pai-filho, e (iii) o indivíduo *Peter* pode ser utilizado para representar determinada pessoa chamada "Peter". Na Figura 8, ilustra-se uma ontologia OWL em grafo.

[19] A OWL é a uma linguagem desenvolvida pelo *World Wide Web Consortium* (W3C) (http://www.w3.org/) para promover a Web Semântica, proposta feita por Berners-Lee (2001) para a estruturação dos documentos da *Web*.

Figura 8 - Ontologia OWL em grafo.


Fonte: http://www.kramirez.net/SMA_Maestria/Material/Presentaciones/Exposicion%20Ontologias/pizza/

Nela, somente os elementos conceituais essenciais para a organização do conhecimento (isto é, classes e propriedades) são ilustrados. A ontologia pode ser assim interpretada: (i) a hierarquia *is-a* (relação inversa *subClassesOf*) (ou taxonomia) contém a classe *Thing*, que é a superclasse de todas as outras do domínio (*DomainEntity*), responsável por organizar elementos independentes (*SelfStandingEntity*); (ii) classe *Thing* tem as subclasses disjuntas *Pizza*, *PizzaTopping* and *PizzaBase*; (iii) classe *PizzaBase* tem as subclasses *ThinBase* e *HeavyBase*; (iv) classe *Pizza* tem as subclasses *VegetarianPizza*, *CheesePizza*, etc.; (v) classe *PizzaTopping* tem as subclasses *FishTopping*, *MeatTopping*, etc.; (vi) classe *Pizza* se relaciona a *PizzaTopping* e *PizzaBase* por meio das relações ou propriedades *hasTopping* e *hasBase.*

Sobre o grafo da Figura 8, ressalta-se que se trata apenas de representação visual de um arquivo em formato OWL. Em outras palavras, isso significa que o código OWL está subjacente ao grafo. Normalmente, as ontologias em OWL são construídas por meio de ferramentas computacionais, denominadas editores, os quais geram o código

OWL e representações gráficas, como a da Figura 8. Um dos editores mais difundidos é o Protégé-OWL[20] (HORRIDGE *et al.*, 2004, KNUBLAUCH *et al.*, 2004). Na Figura 9, ilustra-se a linguagem OWL, destacando algumas informações utilizadas para descrever a classe *Pizza*. Especificamente, o código em OWL explicita as informações: (i) a classe *Pizza* tem as propriedades *hasTopping* e *hasBase*, que a relaciona às subclasses *PizzaTopping* e *PizzaBase*, (ii) as classes *Pizza*, *PizzaTopping* e *PizzaBase* são disjuntas[21] (*disjointWith*), e (iv) a classe *Pizza* é uma subclasse de *SelfStandingEntity*, e (iii) o rótulo (*Label*) da classe *Pizza* na ontologia é "Pizza".

Ao ser baseada na lógica descritiva, uma ontologia OWL codifica formulações lógicas. Por exemplo, com base na ontologia do domínio "pizza", que apenas está expressa de forma diferente nas Figuras 8 e 9, pode-se formular as seguintes sentenças em língua natural (i): "*Pizza* tem *PizzaBase* como sua base" e (ii) "*Pizza* e *PizzaBase* são classes disjuntas". Segundo a lógica subjacente à OWL, tem-se as seguintes representações lógicas para as sentenças: (i) $\texttt{Pizza} \subset \ni \mathsf{hasBase}\ \texttt{PizzaBase}$ e (i) $\texttt{Pizza} \cap \texttt{PizzaBase} \equiv \bot$. Daí dizer que tais ontologias são objetos formais.

Figura 9 - Ontologia na linguagem OWL.

```
</owl:Ontology>
</owl:Class>
<owl:Class rdf:ID="Pizza">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="#hasTopping"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#PizzaTopping"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:someValuesFrom>
        <owl:Class rdf:about="#PizzaBase"/>
      </owl:someValuesFrom>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="hasBase"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#PizzaBase"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="SelfStandingEntity"/>
  </rdfs:subClassOf>
  <rdfs:label>Pizza</rdfs:label>
</owl:Class>
```

Fonte: http://www.kramirez.net/SMA_Maestria/Material/Presentaciones/Exposicion%20Ontologias/pizza/

[20] O Protégé foi desenvolvido pelo Centro de Informática Biomédica (do inglês, *Stanford Center for Biomedical Informatics Research*) da Escola de Medicina da Universidade Stanford (do inglês, *Stanford University School of Medicine*) no final da década de 1990. Do ponto de vista computacional, o Protégé é um software livre e de código aberto, que auxilia o desenvolvimento de ontologias e sistemas baseados em conhecimento. Além disso, esse editor é uma aplicação *stand-alone* (ou *desktop*).

[21] São ditos **disjuntos** os conjuntos que se não possuem nenhum elemento em comum.

### *2.3.2. Representação de ontologias em grafos*

Uma ontologia é comumente representada em um grafo. Seguindo-se Pires *et al.* (2010), uma ontologia *O* em OWL2, por exemplo, é modelada em um grafo direcionado com conexões rotuladas $O = (C, R)$, em que $C = (c1,..., cn)$ é um conjunto finito de vértices (conceitos) e $R = (r1,..., rn)$ é um conjunto finito de arestas (relações entre conceitos).

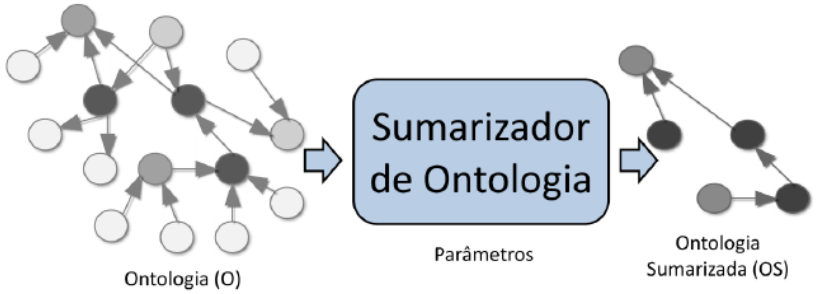
Um relacionamento $r_k \in R$, representa uma relação direcionada entre dois conceitos adjacentes $c_i$ e $c_j \in C$, ou seja, $Rk = (c_i \times c_j)$. Dois conceitos $c_i$ e $c_j \in C$ são adjacentes em *O* se existir $r_k \in R \mid r_k = (c_i \times c_j)$ ou $r_k = (c_j \times c_i)$. Uma aresta direcionada com rótulo é definida de $c_i$ para $c_j$ se $c_i$ é um subconceito direto de $c_j$. De forma similar, se $c_i$ é um conceito de domínio e $c_j$ é um conceito alcançável, então, uma aresta direcionada com rótulo é adicionada de $c_i$ para $c_j$.

Da mesma forma, o resumo de uma ontologia *OS* é uma representação de um subgrafo ou caminho de *O,* tal que $OS \subset O$. Um caminho (*Path*) de um grafo *O* é uma sequência de arestas que ligam vértices. Formalmente, $OS = (CS, RS)$, em que $CS \subset C$ e $RS \subset R$ e OS é um caminho em O.

### *2.3.3. Sumarização Automática de Ontologias: definição*

Segundo Zhang *et al.* (2007), o objetivo das pesquisas sobre SAO é desenvolver métodos/ferramentas computacionais capazes de, dada uma ontologia *O*, gerar uma versão resumida da mesma, denominada *OS* (ontologia sumarizada), composta pelo subconjunto dos conceitos mais representativos de O. A Figura 10 ilustra a SAO.

Figura 10 - Processo geral de Sumarização Automática de Ontologia.




Fonte: Sousa (2011).

As pesquisas sobre SAO são motivadas pelo fato de uma ontologia na íntegra ser de difícil visualização, compreensão, reuso, carregamento por certos *softwares* e

comparação. Ademais, segundo Sousa (2014), a SAO utiliza fundamentos advindos das seguintes áreas: (i) Ciência Cognitiva, que utiliza a noção de categoria natural para identificar conceitos, (ii) Topologia de rede, para identificar conceitos que têm sido ricamente caracterizados com propriedades e relacionamentos taxonômicos, (iii) Estatísticas Lexicais, para identificar conceitos que são susceptíveis de serem mais familiares aos usuários e (iv) Teoria dos grafos, para utilizar medidas de centralidade na identificação de conceitos centrais das ontologias representadas em grafos.

Dada a relevância da SAO, várias ferramentas têm sido propostas. Dentre elas, citam-se (i) a solução de SAO implementada no motor de busca *Falcons Ontology Search* (CHENG *et al.*, 2011), (ii) o aplicativo *Web* chamado *Key Concepts Extraction* (KCE) (PERONI *et al.*, 2008), para identificar os conceitos mais relevantes de uma ontologia, mostrando os nomes dos conceitos em formato de texto na página *Web*, etc.

Partindo-se da concepção de que a produção de um resumo, no geral, consiste em (i) identificar as partes importantes do conteúdo, podendo avaliar a importância da informação em relação a todo o conteúdo apresentado ou em relação à opinião de um usuário, e (ii) utilizar as partes identificadas como importantes para produzir uma versão reduzida do conteúdo, mantendo a coerência e o sentido das informações originais, pode-se dizer que a SAO é uma tarefa similar à SA, tendo como principal diferença o tipo de informação (ou objeto) a ser resumida. Especificamente, os textos-fonte a serem resumidos pelas aplicações de SA são objetos desestruturados, enquanto que as ontologias são altamente estruturadas. A falta de estrutura dos textos é um dos fatores que faz do processamento automático das línguas naturais tarefa tão complexa.

De forma mais específica, a SAO tem início com o cálculo da relevância dos conceitos, com base em certos parâmetros e medidas fornecidas pela ferramenta e selecionadas pelo usuário, resultando em uma hierarquia de importância dos conceitos de *O*. Na Figura 10, a hierarquia de relevância é representada pelas tonalidades em cinza. Em seguida, realiza-se a geração de *OS*, que corresponde a uma subontologia de *O*, concentrando o número máximo de conceitos de maior relevância de acordo com o tamanho especificado. Como os conceitos de maior relevância podem não ser adjacentes em *O*, é possível que conceitos menos importantes (tons mais claros de cinza) sejam introduzidos em *OS*. Tais conceitos são necessários para manter a integridade e preservar os relacionamentos entre os conceitos de grande relevância da ontologia original. Por isso, *OS* corresponde a uma subontologia de *O*, contendo os conceitos de maior relevância devidamente interconectados, como ilustrado na Figura 10.

Segundo Alencar (2008), o processo de SAO deve aceitar diferentes tipos de parâmetros. Dependendo dos valores providos, diferentes sumários podem ser gerados para uma mesma ontologia O. Os parâmetros podem ser: (i) medidas de relevância, ou seja, a relevância é calculada com base nas medidas *Centrality*, *Frequency*, etc.; (ii) tamanho de referência do sumário, segundo o qual se verifica o número de conceitos relevantes adjacentes, agrupando-os em um superconceito (conceito que engloba outros), (iii) variação do tamanho do sumário, segundo o qual se verifica o número de superconceitos, e havendo algum conceito relevante não adjacente, verifica-se todos os caminhos possíveis entre os conceitos ou superconceitos relevantes, e (iv) limite de conceitos relevantes, segundo o qual se estabelece o melhor caminho entre os conceitos relevantes, e consequentemente, determina-se o limite de conceitos relevantes dentro de determinada ontologia.

Guardadas as devidas ressavas, destaca-se que o processo de SAO, assim como fora equacionado metodologicamente, assemelha-se ao processo de SAM em vários aspectos. O principal deles, tendo em vista o objetivo deste trabalho, é a etapa de especificação da relevância das unidades que compõem o objeto de origem para a sumarização. No caso da SAO, tais unidades são os conceitos, cuja relevância precisa ser definida para decidir quais unidades integrarão a OS. Assim como a SAM, a SAO busca produzir uma versão condensada de seu objeto de origem que seja representativa do mesmo, quanto à qualidade e abrangência dos conceitos.

### *2.3.4. Medidas de relevância em SAO*

Para o cálculo da relevância, há uma série de parâmetros ou critérios distintos na literatura sobre SAO, os quais são pautados na representação das ontologias em grafos.

Dentre eles, dois critérios baseiam-se na noção de "categoria básica" da Ciência Cognitiva. De acordo com os fundamentos cognitivistas, a categorização é o processo cognitivo de identificação, classificação e nomeação de entidades como membros de uma categoria. Baseando-se em um "conceito" (isto é, representação prototípica da categoria) de uma categoria (p.ex.: FRUTO), o ser humano identifica os elementos dessa categoria, desde os mais prototípicos (p.ex.: *laranja, maçã, pera* e *banana*) até os menos prototípicos (p.ex.: *tomate* e *azeitona*). Dessa capacidade, surgem as hierarquias de categorias linguísticas. Estudos psicolinguísticos sobre as hierarquias lexicais (p.ex.: ROSCH *et al.* 1976) verificaram que há um nível de hierarquização cognitivamente mais saliente, o "nível básico". Do ponto de vista da aquisição da linguagem, as

categorias básicas (p.ex.: FRUTO e CARRO) são as mais rapidamente aprendidas, sobretudo porque seus conceitos se associam a imagens e seus referentes são concretos. Linguisticamente, as categorias básicas são expressas por termos morfologicamente curtos e simples (p.ex.: *carro* em relação *carro de passeio* ou *carro esporte*).

Assim, os dois critérios utilizados na SAO codificados nas medidas *name simplicity*[22] ("simplicidade de denominação") e *basic Level* ("nível básico)" capturam, de um modo geral, conceitos que são informativamente ricos (do inglês, *information-rich*) do ponto de vista psicolinguístico (PERONI *et al.*, 2008, LI *et al.*, 2010a; LI *et al.*, 2010b). A seguir descrevem-se as medidas.

1. *Name simplicity*

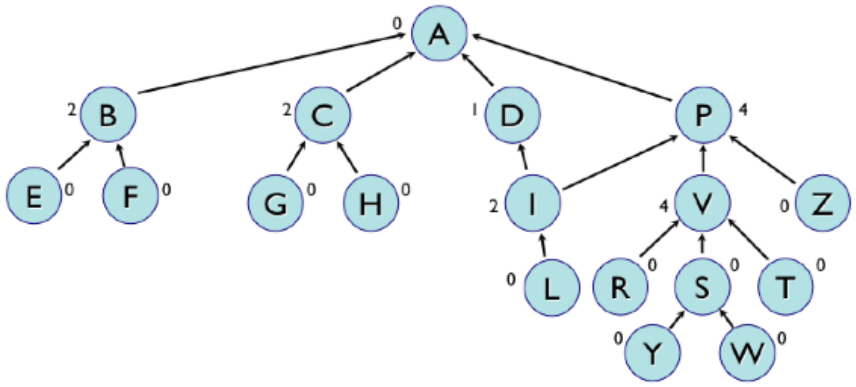
A medida *Name Simplicity (NS)* favorece conceitos de rótulos simples e penaliza conceitos rotulados por expressões multipalavra. A *NS* de um conceito *C* é expressa por valores entre 0 e 1 ($NS(C) \in [0..1]$), sendo que um rótulo simples possui valor 1; os rótulos multipalavra possuem valores menores que 1, resultantes da aplicação da fórmula $NS(C) = 1-c(nc-1)$, em que *nc* é o número de elementos do rótulo e *c* é uma constante empírica. Em PERONI *et al.*, 2008, por exemplo, usam $c = 0.3$, assim, a *NS* de "*artist*" é 1 e de "*musical artist*" é 0.7, pois NS(*musical artist*)=1-0.3(2-1).

2. *Basic Level*

A medida *Basic Level* (*BL)* de *(C)* indica o quão *C* é "central" na taxonomia de uma *O*, sendo expressa por valores entre 0 e 1. A *BL(C)* é calculada contando-se, dado um ramo da taxonomia que contém *C*, quantas vezes C é encontrado no meio do caminho entre um conceito "raiz" e um "folha". No ramo da Figura 11 composto pelos conceitos A, B, E e F, por exemplo, o conceito A é a raiz, possuindo *BL=0*, E e F são nós terminais, possuindo *BL=0*, e o conceito B está no meio dos caminhos entre E e A e F e A, possuindo, portanto, *BL=2*.

[22] Optou-se por utilizar os termos originais em inglês porque estes são assim utilizados na literatura geral. Sempre que possível, no entanto, apresenta-se uma tradução livre desses termos.

Figura 11 - Cálculo da BL(C) em uma grafo direcionado de uma ontologia (O).




Fonte: Peroni *et al.* (2008).

Diante de tais medidas, há dois passos necessários para identificar o conjunto de conceitos que correspondem às categorias naturais em uma *O*. Primeiro, os valores das medidas *basic level* e *name simplicity* são usados para gerar um conjunto de conceitos candidatos, escolhendo aqueles para os quais $W_{BL}*BL(C) + W_{NS}*NS(C)$ for maior que um limiar (*threshold*) $T_{nc}$[23]. Na sequência, esse conjunto de candidatos é filtrado, dando prioridade aos conceitos que são raízes e folhas em um ramo da árvore conceitual, e assumindo que há somente uma categoria natural para cada ramo. Se um ramo contém mais de um conceito candidato, o que maximiza $W_{BL}*BL(C) + W_{NS}*NS(C)$ é escolhido.

Outros 2 critérios da literatura foram delimitados a partir da topologia das ontologias: *density* ("densidade") e *coverage* ("cobertura") (PERONI *et al.*, 2008). A *density D(C)* mede o quão ricamente *C* é descrito em *O*. Essa medida é expressa por valores entre 0 e 1 (isto é, $D(C) \in [0..1]$), sendo calculada com base no número de conceitos subordinados, propriedades e instâncias de *C*. Quando se calcula a *density* total de um conceito, utilizam-se as submedidas *global density* e *local density*.

3. *Global density*

Medida que considera o quão rico um conceito *C* é descrito no cenário global de O. A *global density(C)* $\in [0..1]$ é calculada pela fórmula em (2), ou seja, pela agregação simples e ponderada (os valores $w_s$, $w_p$ e $w_i$ na fórmula abaixo) sobre o número (*n*) de subconceitos (*SubClasses*), propriedades (*Properties*) e instâncias (*Instances*) de *C*.

[23] Peroni et el. (2008) utilizaram $T_{nc} = 0.5$, $W_{BL} = 0.8$ e $W_{NS} = 0.2$; em que W significa agregação ponderada (do inglês, *weighted aggregation*).

(2) $globalDensity(C,0) = \frac{aGlobalDensity(C)}{max\ (\{\forall\ N_i \in O \rightarrow aGlobalDensity\ (N_i)\})}$

$$aGlobalDensity\ (C) = n.SubClasses(C) * w_s + n.Properties\ (C) * w_p + n.Instances(C) * w_1$$

4. <u>*Local density*</u>

A *local density(C)* $\in$ *[0..1]* de um conceito *C* em uma *O* é a densidade em relação aos conceitos vizinhos de *C*. A justificativa dessa medida é a de que, em uma mesma *O*, a riqueza de descrição de um conceito pode variar drasticamente, isto é, algumas áreas de *O* podem conter muitos conceitos densos, o que capturado pelo *global density*, enquanto outras áreas contêm somente conceitos superficiais (menos densos). Assim, a *local density* considera os conceitos mais densos de uma região *x* de O como os mais potencialmente relevantes da região *x* de O. Essa medida é calculada por meio da fórmula em (3), em que os "conceitos mais próximos" ("*nearest concepts*") a *C* são referidos como o conjunto que inclui os sub- e superconceitos acessíveis por um caminho de tamanho máximo 2 na hierarquia de *C*. Finalmente, a densidade total calculada pela combinação das medidas *local density* e *global density*, cada uma delas associada a certo peso. Em (4), tem-se a fórmula para o cálculo da densidade total. Nela, $W_G$ e $W_L$ são os pesos das respectivas medidas *global* e *local density.*

(3) $localDensity(C) = \frac{globalDensity\ (C)}{maxGlobalDensityNearestClasses\ (C)}$

(4) $density(C) = globalDensity(C) * w_G + localDensity(C) * w_L$

5. <u>*Coverage*</u>

Essa medida considera o quanto conceitos bem avaliados participam dos relacionamentos *is-a* em *O*, ou seja, o quanto os conceitos têm cobertura, por meio de relacionamentos, com todos os conceitos de *O*. A justificativa para esse critério/medida é que não se quer identificar somente os conceitos pertinentes, mas sim os conceitos pertinentes em extensão, propiciando a melhor ilustração possível da ontologia. Em (5), tem-se a fórmula para se determinar a cobertura de um conjunto de conceitos, *Coverage(S),* dada uma ontologia *O* . Nela, *Covered(C)* é o conjunto de conceitos cobertos por um conceito *C*, isto é, *Covered(C)= C ∪ allSubClasses(C)∪ allSuperClasses(C),* e |O| é o tamanho da ontologia *O* (isto é, o número de conceitos contidos em *O*).

(5) $Coverage(\{C_1, \ldots . C_n\}) = \frac{|\{Covered(C_1) \cup \ldots \cup Covered(C_n)\}|}{|O|}$

Outro critério da literatura, cuja medida é *popularity* ("popularidade"), foi delimitado com base em noções de estatística lexical (p.ex.: Peroni *et al.*, 2008).

6. *Popularity*

Em Peroni *et al.*, (2008), por exemplo, a *popularity* de um conceito *C* que pertence a uma ontologia *O* é determinada pelo número de resultados retornados por consultas feito ao buscador *Yahoo* com o rótulo de *C* como palavra-chave. Tendo em vista que as palavras-chave usadas como indexadores na web tendem a constituir as chamadas "categorias naturais", essa medida permite identificar os conceitos mais populares ou comuns da ontologia em questão. Mesmo que *popularity* não seja de fato de uma medida de grafo, esta se mostra relevante porque contribui para a identificação dos conceitos mais relevantes de uma estrutura ontógica.

Outro critério bastante difundido na SAO é *Centrality* ("centralidade" ou "conectividade"). Existem 2 centralidades amplamente utilizadas: *degree Centrality* ("grau de centralidade") e *betweenness Centrality* ("centralidade de intermediação") (FREEMAN, 1978; BORGATTI, EVERET, 2006; OPSAHL *et al.* 2010; NEWMAN, 2010).

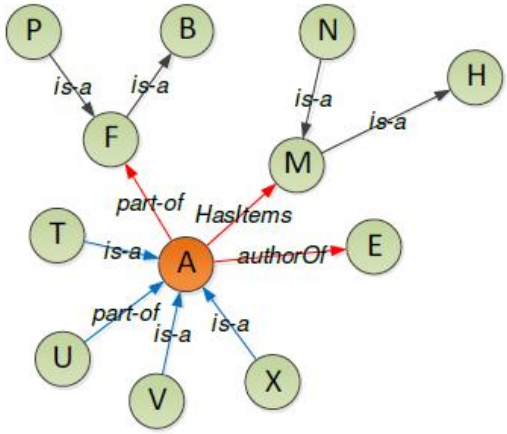
7. *Degree Centrality*

A *degree centrality* ("centralidade de grau") é definida como o número de ligações incidentes sobre um nó. Trata-se de uma medida de saliência de vértices que se pauta na ideia de que o número de relacionamentos proporciona uma ampla cobertura de acesso entre os conceitos da ontologia e que tipos de relacionamento podem ter pesos diferentes. Quando se usa pesos para os diferentes tipos de relações, essa medida passa a ser denominada *weighted-degree centrality* ("centralidade de grau ponderada"). Em (6), tem-se a fórmula da *degree centrality*, desconsiderando-se a distinção entre arestas de chegada e saída. Na fórmula, a *Centrality* é calculada como a razão entre a quantidade de arestas de *n* e o total de nós do grafo menos 1. Seja *G* um grafo e *n* um nó de G.

(6) $$Centrality(C,n) = \frac{\sum <arestas\ de\ n>}{<total\ de\ nós\ de\ G-1>}$$

Vale ressaltar que, no caso de uma rede direcionada como a ilustrada na Figura 12, é possível definir duas medidas separadas para representar a *degree centrality*, a saber: *in-degree*, que conta o número de ligações direcionadas ao nó, e *out-degree*, que conta o número de relações direcionadas de um nó aos outros. Ademais, os diferentes tipos de relações que rotulam as arestas, como *is-a*, *part-of*, *has-items* e *author-of*, na Figura 12, podem ser receber pesos distintos, de acordo com o interesse do usuário.

Figura 12 - Ilustração de grafo com diferentes tipos de relacionamentos.




Fonte: Sousa (2014).

8. *Betweenness Centrality*

Medida de centralidade definida pela ocorrência de um vértice nos menores caminhos formados dentro do grafo pelos outros vértices. Essa medida valoriza os vértices que possibilitam os menores caminhos entre os vértices. Para um grafo com *n* vértices, o cálculo de *betweenness Centrality* ($C_B$) requer: (i) determinar, para cada par de vértices ($\alpha,\mu$), os menores caminhos entre eles, (ii) determinar, para cada par de vértices ($\alpha,\mu$), a fração de menores caminhos que passam pelo vértice em questão, e (iii) somar todas as frações dos pares de vértices ($\alpha,\mu$). A medida é representada pela fórmula em (7), onde $\sigma_{\alpha\mu}$ é a quantidade de menores caminhos de $\alpha$ para $\mu$ e $\sigma_{\alpha\mu}(v)$ é o número de menores caminhos de $\alpha$ para $\mu$ que passam pelo vértice $v$.

(7) $$C_B(v) = \sum_{\alpha \neq v \neq \mu \in V} \frac{\sigma_{\alpha\mu}(v)}{\sigma_{\alpha\mu}}$$

Em Zhang *et al.* (2009), tem-se a medida *reference* (“referência”), descrita a seguir.

9. *Reference*

A *Reference* de um conceito *C* provê um valor normalizado do número de entidades dinamicamente compiladas da *Web* Semântica pela máquina de busca Watson[24], cuja referência depende do conceito *C*. Ela conta o axioma que o conceito tem sobre o lado direito, ou seja, o número de afirmação <s, p, o> de tal modo que *o* representa o conceito *C*, e *s* e *p* representam os axiomas presentes ao ao lado direito de *o*. Esses axiomas potencialmente envolvem propriedade de domínio, e variam bem como as relações instanciadas além do relacionamento “é-um”, porque as ontologias coletadas da Web Semântica podem conter essas relações. Sendo assim, *reference* deve fornecer uma indicação mais precisa de quão denso um conceito é descrito no escopo da *Web* Semântica.

Em Sousa (2011, 2014), outras duas medidas distintas são definidas: *frequency* (“frequência”) e *Closeness* (“proximidade”).

10. *Frequency*

Essa medida é aplicada especificamente a uma ontologia integrada $O_i$, obtida pela fusão de várias ontologias $O_1$,... *On*, (do inglês, *ontology merging*). Uma ontologia unificada pode ser descrita por arquivos de mapeamentos ontológicos. A Figura 13 representa um exemplo de arquivo de mapeamento (PIRES, 2007), em que o conceito *faculty*, pertencente à *O* de origem denominada CLO1, foi mapeado para os conceitos *phd* e *professor* das ontologias de origem LO1 e LO2, respectivamente (PIRES *et al.*, 2010).

[24] Watson é um sistema computacional de Inteligência Artificial capaz de responder a perguntas em linguagem natural. Ele foi desenvolvido no âmbito do projeto *DeepQA* da IBM e recebeu esse nome em homenagem ao primeiro CEO da IBM, Thomas J. Watson. O sistema foi desenvolvido especificamente para responder a perguntas no programa de TV Jeopardy!. Em 2011, Watson competiu no Jeopardy! contra antigos vencedores, recebendo o prêmio de $ 1 milhão pelo primeiro lugar (Wikipedia).

Figura 13 - Mapeamento de classes para o cálculo da *frequency.*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<CLUSTERONTOLOGY clo="CLO1">
<CLOCLASS>
<LABEL>faculty</LABEL>
<LOCLASS>
<LABEL>phd</LABEL>
<LO>L01</LO>
</LOCLASS>
<LOCLASS>
<LABEL>professor</LABEL>
<LO>L02</LO>
</LOCLASS>
</CLOCLASS>
...
</CLUSTERONTOLOGY>
```

Fonte: Alencar (2008).

Para o para o cálculo da *frequency* de um conceito *Cn*, aplica-se a fórmula em (8), na qual essa medida é a razão entre o número de conceito correspondentes que envolvem $C_n$ (|*correspondences(Cn)*|) e o número de ontologias-fontes de O (|$O_1$,...,$O_n$|), sendo o resultado expresso em um valor entre 0 e 1.

(8) $$frequency(c_n) = \frac{|correspondences(c_n)|}{|O_1,\ldots,O_n|}$$

11. <u>*Closeness*</u>

O valor da *Closeness* (“proximidade”) de um conceito *Cn* é proporcional à quantidade de conceitos com grande valor de relevância que estão próximos de *Cn*. Essa medida necessita que os conceitos tenham um valor de relevância previamente determinado (p.ex.: via *degree Centrality*). O objetivo da medida é capturar os conceitos relevantes, considerando a distância destes para os demais conceitos relevantes. A fórmula que provê um valor ponderado entre a distância e a relevância de um conceito *Cn*, com todos os conceitos (*C*) da ontologia *O*, é definida em (9).

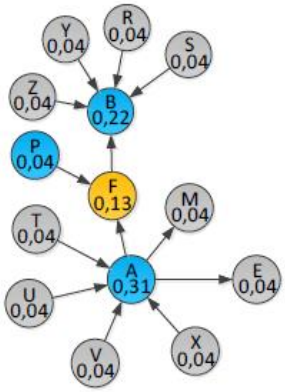
(9) $$Closeness\ (Cn) = \frac{\sum Cx \in (C-Cn) \frac{revelance\ (Cx)}{distance\ (Cn,Cx)}}{\sum Cx \in (C-Cn)\ \ 1/distance\ (Cn,Cx)}$$

Na fórmula, *Closeness(Cn)* $\in$ [0,1] é uma média ponderada formada pelo valor de relevância dos conceitos – *relevance(Cx)* – com o peso representado pelo inverso da

distância – *distance(Cn,Cx)* – de um conceito *Cn* para o conceito *Cx*. No caso, *Cx* é uma variável que é ocupada por todos os conceitos em *C*, menos o conceito *Cn*, indicado por *Cx* ∈ (*C-Cn)* . Por fim, *distance(Cn,Cx)* é a menor distância, em número de relações (ou arestas), que interliga os conceitos *Cn* e *Cx*. Ilustra-se o cálculo de *Closeness* na Figura 14.

Na Figura 14, ilustra-se especificamente o cálculo de *Closeness* do conceito F (isto é, *Closeness*(F)) da ontologia em questão. Na fórmula, vê-se que a variável *Cx* representa todos os conceitos de O, exceto F (isto é, A, B, P, T, Z, etc.). Assim, *relevance(Cx)* é preenchida pelos valores de relevância (previamente identificados) de cada um desses conceitos. Por exemplo, *revelance(A)=0,31*, *revelance(B)=0,22*, etc. A *distance(Cn,Cx),* por sua vez, é preenchida pela distância entre F e os demais conceitos. Por exemplo, *distance(F,B)=1*, *distance(F,T)=2*, etc.

Figura 14 - Cálculo da medida *Closeness*




$$Closeness(\mathrm{F}) = \frac{\overset{\mathrm{A}}{\frac{0,31}{1}} + \overset{\mathrm{B}}{\frac{0,22}{1}} + \overset{\mathrm{P}}{\frac{0,04}{1}} + \overset{\mathrm{T}}{\frac{0,04}{2}} + \overset{\ldots}{\frac{0,04}{2}} \ldots}{1/1 + 1/1 + 1/1 + 1/2 + 1/2 \ldots}$$

Fonte: Silva (2014).

Cada um dos critérios/medidas aqui revisados produz uma pontuação para cada conceito *C* de uma ontologia *O* e a pontuação final atribuída a um *C* é um somatório ponderado das pontuações resultantes dos critérios individuais. As medidas aqui revisadas são aplicadas em quantidade e combinações variadas nos diferentes trabalhos da área. Sousa (2011), por exemplo, propôs uma extensão da ferramenta OWLSum, denominada OWLSumBPR, pauta-se na combinação de 4 dos vários critérios aqui revisados, considerados os mais eficazes na literatura, a saber: (i) *Centrality*; (ii) *frequency*; (iii) *name simplicity*, e (iv) *Closeness*.

É importante ressaltar que a maioria das medidas citadas anteriormente são mais mais relevantes quando se trata de ontologias mais sofisticadas, como, por exemplo, diferentes tipos de relações conceituais, que configuram estruturas cíclicas.. Esse não é o caso da ontologia utilizada neste trabalho, que possui apenas a relação hierárquica ou hiponímica (*is-a*).

Sendo assim, selecionou-se as mais utilizadas na literatura e que possibilitaram a aplicação dentro da modelagem aceitável para o formato de árvore conceitual.

As principais medidas de relevância revisadas estão sistematizadas no Quadro 4.

Quadro 4 – Principais medidas de relevância em SAO

| **Medida** | **Qt.** | **Critério** |
|---|---|---|
| *Cognitiva* | 1 | *Name simplicity* |
| | 2 | *Basic Level* |
| *Topológica* | 3 | *Global density* |
| | 4 | *Local density* |
| | 5 | *Coverage* |
| | 6 | *Closeness* |
| | 7 | *Degree Centrality* |
| | 8 | *Betweenness Centrality* |
| | 9 | *Reference* |
| | 10 | *Frequency* |
| *Estatística Lexical* | 11 | *Popularity* |

A seguir, tecem-se algumas considerações sobre a revisão da literatura antes de apresentar a proposta deste mestrado.

## 2.4. Considerações sobre a revisão da literatura

O único trabalho de SAM profunda para o português baseado em conhecimento léxico-conceitual não utiliza uma representação estruturada dos conceitos das coleções. Tosta (2014) utiliza textos anotados em nível conceitual, ou seja, cujos conceitos subjacentes às palavras estão explícitos por meio dos *synsets* da WN.Pr.

No entanto, alguns métodos mono e multidocumento da literatura geral modelam o conteúdo dos textos-fonte em uma estrutura léxico-conceitual, denominada "ontologia". Sob esse rótulo, estão abrigados objetos relativamente distintos entre si quanto à organização, já que se tem (i) organizações exclusivamente hierárquicas

(taxonomias) (REIMER, HAHN, 1988, *apud* MANI, 2001), (ii) organizações predominantemente hierárquicas (WU, LI *et al.*, 2003) ou (iii) organizações de tópicos relacionados (SILVA, 2006; HENNIG *et al.*, 2008; LI *et al.*, 2010).

As relações taxonômicas que caracterizam o recurso de Reimer e Hahn (1988, *apud* MANI, 2001), Wu e *Et al.*(2003) e Silva (2006) são aquelas que ocorrem entre conceitos genéricos (superordenados) (p.ex.: "*portable computer*") ("computador portátil") e específicos (subordinados) (p.ex.: "*notebook*"), resultando em uma hierarquia. Essa relação também é conhecida como *is-a* ou *é-um*, já que se pode afirmar que "*notebook*" é um "*portable computer*".

Em Wu e Li *et al.* (2003), o recurso conceitual armazena outras informações além da relação hierárquica. Por exemplo, "*Name*" e "*Game Type*" codificam características de "*Game*", e "*Sony Music*", "*SCE*" e "*Sony Music*" são segmentos de mercado da empresa *Sony Corporation*. Assim, essa ontologia é de fato uma visão do domínio *Sony Corporation* compartilhada pelos especialistas que a construíram.

Em Silva (2006), Hennig *et al.* (2008) e Li *et al.* (2010), as ontologias são organizações de tópicos relacionados. A de Hennig *et al.* (2008), por exemplo, foi construída a partir do DMOZ (do inglês, *Directory Mozilla*), um diretório multilíngue de *links* da *Web* organizados em categorias ou tópicos. Silva (2006) construiu seu recurso a partir da ontologia do Yahoo, cujas categorias[25] e subcategorias[26] também são utilizadas para organizar *sites*. Na ontologia de Silva (2006), os rótulos originais dos conceitos foram reduzidos às formas canônicas e os descritores acrescidos de palavras (nomes, verbos e adjetivos) relacionadas ao conceito via sinonímia e hiponímia.

Ainda no que diz respeito ao conteúdo, tais objetos diferem quanto à estratégia de nomeação dos conceitos. Enquanto algumas ontologias denominam os conceitos somente por meio de um rótulo simples, como *"notebook"* e "*game*" (REIMER, HAHN, 1988, *apud* MANI, 2001), outras utilizam rótulos simples associados a um conjunto de descritores (ou palavras relacionadas) (HENNIG *et al.*, 2008; SILVA, 2006). No primeiro caso, tais ontologias são manuais e dedicadas ao *corpus* sob análise e, no segundo, elas são recursos genéricos já existentes, a partir dos quais é preciso identificar a parcela conceitual que representa o *corpus* em questão. E esse recorte é comumente feito pela indexação automática dos textos ao recurso. Nesse caso, a

[25] Artes e Cultura, Esportes, Educação, Ciência, Regional, Business to Business, Fontes de referência, Saúde, Compras e Serviços, Lazer, Informática, Internet, Notícias, Finanças, Governo e Sociedade.
[26] A categoria Artes e Cultura, por exemplo, possui as subcategorias Artes Cênicas > Artistas > Atores e Atrizes > Bibi Ferreira.

descrição de um conceito por meio de rótulos simples associados a descritores aumenta a possibilidade de a ferramenta automática responsável pela indexação identificar os conceitos correspondentes às palavras do *corpus.*

Todos esses objetos, no entanto, compartilham três características. Uma delas é a baixa formalização, já que a descrição dos conceitos e relações não segue uma linguagem formal. Em outras palavras, pode-se dizer que as ontologias utilizadas na SA são objetos semiformais, já que os conceitos e relações estão expressos em língua natural, mas de forma estruturada. A outra é relativa ao escopo. A maioria das ontologias aqui investigadas podem ser consideradas de domínio, contendo conceitos e vocabulários relacionados a domínios particulares (p.ex.: computadores, *Sony Corporation* e desastres). E a terceira característica é o fato de a representação dos domínios ser baseada em conceitos expressos por nomes ou expressões nominais. As exceções são os recursos utilizados por Hennig *et al.* (2008) e Silva (2006)

Sobre os critérios de relevância para a seleção do conteúdo, destaca-se que a maioria dos métodos de SA utiliza a frequência de ativação de um conceito *x* na estrutura conceitual que representa os textos-fonte para determinar a sua relevância. Por vezes, a frequência de ativação reflete a frequência de ocorrência das diferentes palavras que expressam o conceito *x* nos textos-fonte, como no sistema TOPIC de Reimer e Hahn (1988, *apud* MANI, 2001). A frequência também pode ser cumulativa para os conceitos superordenados, cuja frequência final é a soma da frequência de seus subordinados, com em Wu e Li *et al.* (2003) e Silva (2006).

Além da frequência, os outros critérios mais utilizados são (i) o número de conceitos subordinados a *x* que foram ativados e (ii) a localização (nível ou altura) do conceito na estrutura hierárquica.

Na SAO, as "ontologias" são de fato objetos formais, pois seus elementos constitutivos são representados por linguagens formais, como OWL. Apesar da diferença no grau de formalidade, a relação hierárquica (*is-a*) é central à organização do conhecimento e ao processo de sumarização na SA e SAO. Em LI *et al.*(2010b), por exemplos, vários experimentos promissores foram feitos levando-se em conta apenas a hierarquia conceitual, excluindo propriedades e instâncias. Além disso, as ontologias da SAO também são de domínio, como o da Biosfera e Finanças.

Sobre a identificação dos conceitos relevantes à construção das versões resumidas das ontologias, vê-se que os pesquisadores possuem um leque bastante grande de critérios codificados em medidas ou métricas, uma vez que as ontologias são

representadas formalmente como grafos aos quais essas medidas são aplicadas. Alguns desses critérios/métricas podem ser agrupados, segundo sua fundamentação teórica, em: (i) medidas cognitivas (*basic Level* e *name simplicity*), (ii) medidas topológicas (*densities*, *coverage*, *centralities* e *proximity*) e (iii) estatísticas lexicais (*popularity*).

Assim, nas seções seguintes, apresenta-se a proposta deste mestrado, que é, de um modo geral, usar o conhecimento léxico-conceitual estruturado, subjacente aos textos-fonte, para identificar o conteúdo relevante de uma coleção multidocumento com base em critérios que exploram essa estrutura, as quais serão formulados com base nos trabalhos de SAO. Acredita-se que certas medidas utilizadas na SAO sejam pertinentes para capturar o conteúdo central para a geração de sumários extrativos multidocumento.

Na sequência, inicia-se a apresentação da proposta pela seleção do *corpus*.

## 3. Seleção do *corpus* multidocumento

Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de maneira que possa ser processado por computador com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SINCLAIR, 2005. Por essa definição, um *corpus* é um artefato produzido para a pesquisa e que, por isso, a maioria de suas características é dependente dos objetivos da pesquisa.

Tendo em vista o trabalho ora descrito, o *corpus* tinha de apresentar as seguintes características: (i) monolíngue: especificamente do português, já que este trabalho focaliza a investigação de métodos de SAM baseados em hierarquias conceituais no cenário específico do processamento automático do português; e (ii) multidocumento, posto que um *corpus* desse tipo fornece vários *clusters* de textos que, ao versarem sobre um mesmo assunto ou tópico, tornam-se pertinentes como fontes para a SAM.

Diante de tais características, buscou-se identificar na literatura um *corpus* que as satisfizesse. Nessa investigação, identificou-se o CSTNews (CARDOSO *et al.* 2011), *corpus* multidocumento de notícias jornalísticas em português. O CSTNews possui 50 *clusters* (ou coleções). Cada coleção aborda um assunto distinto, sendo cada texto (notícia) da coleção proveniente de um jornal distinto. No total, o CSTNews possui 140 textos, que somam 2.088 sentenças e 47.240 palavras. Os textos foram coletados dos seguintes jornais online: *Folha de São Paulo, Estadão, O Globo, Jornal do Brasil* e *Gazeta do Povo*. Essas fontes foram escolhidas devido à popularidade e circulação na web, garantindo a coleta de uma mesma notícia veiculada por fontes distintas. Os *clusters* estão organizados em categorias, cujos rótulos indicam a seção do jornal da qual os textos que os constituem foram compilados. Assim, têm-se as categorias "mundo", "política", "cotidiano", "ciência", "dinheiro" e "esporte" (Quadro 5).

Quadro 5 – Distribuição dos *clusters* nas categorias do CSTNews.

| **Categoria** | ***Cluster* (C)** |
|---|---|
| Mundo | C1, C10, C12, C13, C14, C15, C18, C23, C26, C29, C32, C35, C46, C47 |
| Política | C2, C9, C16, C17, C20, C21, C40, C42, C43, C44, C50 |
| Cotidiano | C3, C4, C5, C6, C11, C22, C33, C34, C36, C37, C39, C45, C49 |
| Ciência | C7 |
| Esportes | C8, C19, C24, C25, C27, C28, C31, C38, C41, C48 |

Fonte: Aleixo e Pardo (2008).

Cada *cluster* é constituído por: (i) 2 ou 3 textos-fonte, (ii) sumário manual de cada texto-fonte, (iii) 6 *abstracts* e 6 extratos multidocumento manuais, (iv) sumários automáticos multidocumento, (v) interconexão entre os textos-fonte via CST[27] (RADEV, 2000), (vi) anotação de expressões temporais nos textos-fonte, (vii) etiquetação morfossintática e sintática dos textos-fonte, (viii) anotação dos sentidos dos substantivos e verbos via *synsets* da WN.Pr, (ix) anotação de aspectos informativos (p.ex.: o quê, onde, etc.) de um dos sumários multidocumento de referência (manuais), (x) anotação discursiva de cada texto-fonte via RST (MANN, THOMPSON, 1987) e (xi) anotação de subtópicos informativos dos texto-fonte.

Ressalta-se que todos os sumários multidocumento do CSTNews (manuais e automáticos) possuem taxa de compressão de 70%, ou seja, apresentam 30% do número de palavras do maior texto-fonte de sua coleção correspondente.

Para esta pesquisa, realizou-se um recorte no *corpus*, construindo-se um *subcorpus* do CSTNews. Inicialmente, esse recorte consistiu na seleção de 1 *cluster*, cujo conteúdo dos textos-fonte foi representado em uma estrutura conceitual. No Quadro 6, apresentam-se os 3 textos-fonte e 1 sumário manual (de referência) multidocumento.

Quadro 6 – Textos-fonte e sumário de referência do *cluster* C1 do CSTNews.

| |
|---|
| **Texto 1** (D1_C1_Folha)<br>Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto. Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros. Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas. Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa. Apenas uma manteve a permissão. Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como uma vergonha para o setor. |

[27] O *corpus* CSTNews é assim denominado exatamente porque os textos das coleções estão alinhados por meio das relações estabelecidas pela teoria/modelo linguístico-computacional CST.

| |
|---|
| **Texto 2** (D2_C1_Estadão**)**<br>Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas. As vítimas do acidente foram 14 passageiros e três membros da tripulação. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. "Não houve sobreviventes", disse Okala. O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais. |
| **Texto 3** (D3_C1_JB)<br>Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas. As vítimas do acidente foram 14 passageiros e três membros da tripulação. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu. O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. "Não houve sobreviventes", disse Okala. O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais. Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. |
| **Sumário multidocumento**<br>17 pessoas morreram após a queda de um avião na República Democrática do Congo. 14 dessas vítimas eram passageiros e três membros da tripulação, todos de nacionalidade russa. Nenhuma vítima sobreviveu. O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta após se chocar com uma montanha, prejudicado pelo mau tempo. O avião também levava cargas e minerais. |

Fonte: http://nilc.icmc.usp.br/CSTNews/

O *cluster* C1 foi selecionado de forma aleatória, ou seja, não houve nenhum critério específico para a seleção deste. Esse *cluster* é composto por 3 textos compilados dos jornais *on-line Folha de São Paulo, Estadão* e *Jornal do Brasil*. O C1 possui um total de 24 sentenças e 432 palavras (ALEIXO, PARDO, 2008). Nessa coleção, constam textos do domínio "mundo", sendo que os documentos relatam especificamente um desastre, no caso, a "queda de um avião no Congo". Dessa forma, esse *cluster* pertence ao domínio conceitual "desastre/acidente".

A seguir, descreve-se processo de representação do conteúdo dos 3 textos-fonte do referido *cluster* em uma estrutura léxico-conceitual.

## 4. Representação conceitual do *corpus*

Com base na metodologia do trabalho, a tarefa em questão consistiu em representar o conteúdo dos textos-fonte em uma estrutura conceitual. Para isso, duas estratégias foram escolhidas: (i) construção manual de uma estrutura conceitual que represente o conteúdo subjacente aos textos-fonte e (ii) indexação dos textos-fonte a um recurso léxico-conceitual geral existente e subsequente recorte da parcela pertinente.

A estratégia em (i) é complexa e pouco generalizável. A complexidade está ligada ao fato de a construção de tal objeto de forma manual requerer que os desenvolvedores tenham bastante conhecimento sobre o domínio a ser representado a ponto de delimitar os conceitos e as relações entre eles. E o baixo grau de generalização diz respeito ao fato de que, a cada *corpus* distinto a ser sumarizado, uma estrutura conceitual correspondente precisa ser construída.

A segunda estratégia é um pouco menos complexa e mais generalizável. Diz-se que é menos complexa que (i) porque não é necessário organizar os conceitos, pois as relações entre eles são herdadas do recurso original. Entretanto, essa tarefa ainda é complexa, já que, ao partir dos textos, é preciso desambiguar as palavras a serem indexadas e identificar os conceitos armazenados no recurso original que mais adequadamente expressam o significado das palavras nos textos. Por fim, ressalta-se que a estratégia em (ii) é mais generalizável porque ao se utilizar um recurso previamente construído que seja genérico, é possível indexar textos que tratam de assuntos variados a uma mesma ontologia. Ademais, caso a indexação seja feita de forma bem definida, a modelagem de outras coleções de textos-fonte ou *corpus* pode seguir a mesma metodologia.

Assim, optou-se pela indexação léxico-conceitual (textos-fonte → ontologia), sendo necessário delimitar: (i) as palavras dos textos-fonte a serem indexadas, (ii) o recurso léxico-conceitual ao qual as palavras serão indexadas e, por fim, (iii) o método (manual ou semiautomático) de indexação/construção da representação conceitual. Tais atividades estão descritas nas subseções a seguir.

### 4.1. Seleção das unidades lexicais

Para a seleção das palavras, 2 critérios foram analisados: frequência e classe de palavra.

Considerando-se a frequência, apenas as unidades mais frequentes do *cluster* seriam indexadas, o que restringiria os processos aos nomes e verbos. Tal variedade de classes de palavras tornaria a tarefa de indexação léxico-conceitual mais complexa.

Com o intuito de fazer da indexação um processo mais controlado, optou-se pela classe de palavra como critério de seleção. Assim, seguindo-se a tendência da literatura, somente as palavras da classe dos nomes (comuns) do *cluster* C1 do CSTNews foram submetidas à indexação para a geração da representação conceitual do *cluster*.

Do ponto de vista teórico-metodológico, a ênfase dada aos nomes na literatura pode ser justificada por 2 fatores: (i) os nomes são as palavras mais frequentes de um texto ou *corpus*, expressando boa parte do conteúdo central dos mesmos, e (ii) os nomes lexicalizam conceitos que se organizam primordialmente em uma hierarquia, o que facilita a indexação dos conceitos subjacentes às palavras das indexações.

Uma vez que a classe dos nomes foi selecionada, procedeu-se à identificação das unidades dessa classe nos textos do C1. Para tanto, os textos foram anotados no nível morfossintático por meio do etiquetador[28] (em inglês, *tagger*) LX-Tagger (BRANCO, SILVA, 2004), disponível no portal do LX-Center[29]. O LX-Tagger foi selecionado devido à sua alta precisão (96,87%) e, sobretudo, sua interface *on-line* amigável.

Em (10), ilustra-se a etiquetação da sentença S1 ("*Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo*") do documento D1 do C1 pelo LX-Tagger. Em (10), o etiquetador anotou "pessoas", "queda", "avião" e "passageiros" como nomes comuns (CN (=*commun noun*), os quais foram reduzidos às suas formas canônicas, "pessoa", "queda", "avião" e "passageiro", respectivamente.

(10) a_/LADV1 o/LADV2 menos/LADV3 17/DGT pessoas/PESSOA/CN#fp morreram/MORRER/V#pi-3p após/PREP a/DA#fs queda/QUEDA/CN#fs de/PREP um/UM#ms avião/AVIÃO/CN#ms de/PREP passageiros/PASSAGEIRO/CN#mp em_/PREP a/DA#fs República/PNM Democrática/PNM de_/PREP o/DA#ms Congo/PNM ./PNT

Do conjunto total de 42 nomes de C1, "junho", "março", "quinta-feira" e "sexta-feira" foram excluídos, pois a versão *offline* da WordNet de Princeton (WN.Pr) utilizada para fazer a indexação não continha tais conceitos. No entanto, após as atualizações mais recentes, conceitos referentes a meses e dias da semana foram inclusos na versão *online*.

[28] Ferramenta computacional responsável por associar às palavras de um texto ou sentença uma etiqueta que indica sua correta categoria sintática no contexto.
[29] Disponível em http://lxcenter.di.fc.ul.pt/

Assim, 38 nomes foram selecionados a partir do *cluster* C1 para a representação conceitual do mesmo. No Quadro 7, estão descritas as 38 unidades finais selecionadas.

Quadro 7 – Os nomes constitutivos dos textos-fonte do *cluster* C1.

| | |
|---|---|
| acidente | nacionalidade |
| aeronave | país |
| aeroporto | passageiro |
| aterrissagem | permissão |
| avião | pessoa |
| carga | pista |
| chama | porta-voz |
| cidade | propriedade |
| companhia | queda |
| distância | quilômetro |
| estrada | setor |
| fabricação | sobrevivente |
| floresta | tarde |
| fonte | tempestade |
| leste | tempo |
| localidade | transporte |
| membro | tripulação |
| mineral | tripulante |
| montanha | vítima |

Após a seleção do *corpus* e das unidades lexicais, procedeu-se à seleção da ontologia.

## 4.2. Seleção do recurso léxico-conceitual para indexação

Atualmente, há vários recursos conceituais ou ontologias em português, sendo a maioria disponibilizada no Portal de Ontologias[30]. No entanto, tais recursos, ao organizarem conceitos de áreas bastante específicas (p.ex.: Nanotecnologia, Música, Ecologia, etc.), não são adequados para a tarefa em questão.

Apesar da existência de recursos lexicais desenvolvidos para o português como a Onto.PT (GONÇALO OLIVEIRA *et al.*, 2012), a WordNet.Br (DIAS DA SILVA,

[30] http://ontolp.inf.pucrs.br/

2005) e a MultiWordNet.PT[31], optou-se pela WN.Pr (FELLBAUM, 1998), desenvolvida para inglês norte-americano, devido a três motivos: (i) acessibilidade, posto que seu arquivo-fonte está disponibilizado integralmente via *web*[32]; (ii) pertinência linguística, já que sua estrutura pauta-se em pressupostos da Psicolinguística e seus dados foram manualmente compilados de vários dicionários monolíngues do inglês norte-americano, e (iii) abrangência, já que armazena (versão 3.0) um total de 155.287 unidades lexicais, distribuídas em 117.798 substantivos, 11.529 verbos, 21.479 adjetivos e 4.481 advérbios.

A WN.Pr é o resultado das pesquisas que tiveram início em meados da década de 1980 no Laboratório de Ciência Cognitiva da Universidade de Princeton (EUA). Motivados por pressupostos psicolinguísticos sobre a organização do léxico mental, os pesquisadores desse laboratório construíram uma base lexical em que unidades lexicais e expressões estão organizadas em função do seu significado.

Especificamente, na WN.Pr, as unidades lexicais (palavras ou expressões) do inglês norte-americano estão organizadas em quatro classes de palavras: nome, verbo, adjetivo e advérbio. As unidades de cada classe estão codificadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-sinônimas (p.ex.: {car, auto, automobile, machine, motorcar}). Cada *synset* é construído de modo a representar um único conceito lexicalizado[33] por suas unidades constituintes.

Entre os *synsets*, codificam-se 5 relações lógico-conceituais principais: hiponímia, antonímia, meronímia, acarretamento e causa (CRUSE, 1986; FELLBAUM, 1998):

a) **Hiponímia** (Hiperonímia): relação entre um conceito específico (hipônimo) e um conceito mais genérico (hiperônimo). Por exemplo, o *synset* {cruiser, squad car, patrol car, police car, prowl car} é hipônimo de {car, auto, automobile, machine, motorcar} e {car, auto, automobile, machine, motorcar} é hiperônimo de {cruiser, squad car, patrol car, police car, prowl car}. O termo "hiponímia" (e hiperonímia) é utilizado para denominar a relação *is-a* quando envolve conceitos lexicalizados.

[31] http://mwnpt.di.fc.ul.pt/
[32] http://wordnet.princeton.edu/wordnet/
[33] A lexicalização é o processo pelo qual um conteúdo semântico é expresso por uma unidade lexical, seja ela simples, como "casa", composta, como "guarda-roupa", ou mesmo complexa, como "nota fiscal", resulta da interação entre convenção e motivação (TAYLOR, 1985; LAKOFF, 1987).

b) **Antonímia**: relação que engloba diferentes oposições semânticas. São elas: antonímia complementar: relaciona pares de itens lexicais contraditórios cuja afirmação do primeiro acarreta a negação do segundo e vice-versa (p. ex.: {alive} e {dead}); antonímia gradual, que relaciona itens que denotam valores opostos em uma escala (p.ex.: {small} e {large}) e "antonímia recíproca", que relaciona pares de itens que se pressupõem mutuamente, sendo que a ocorrência do primeiro pressupõe a ocorrência do segundo (p.ex.: {buy, purchase} e {sell}).

c) **Meronímia** (Holonímia): relação entre um *synset* ou mais *synsets* que expressam "partes" (merônimos) e um *synset* que expressa o "todo" (holônimo). Por exemplo, os *synsets* {bumper}, {car door}, {car mirror} e {car window} são merônimos do *synset* {car, auto, automobile, machine, motorcar} e este é holônimo dos primeiros. Os termos "meronímia", para denominar a relação conceitual *has-part*, e "holonímia", para denominar a relação *part-of*, são empregados quando se trata de conectar conceitos lexicalizados.

d) **Acarretamento**: relação entre uma ação A1 e uma ação A2; a ação A1 não pode ser feita sem que A2 também o seja. Esse é o caso, por exemplo, da relação entre o *synset* {snore, saw wood, saw logs}, que representa o conceito "roncar" e {sleep, kip, slumber, log Z's, catch some Z's}, que representa o conceito "dormir", já que a ação denotada pelas unidades lexicais do primeiro *synset* acarreta a ação denotada pelas unidades do segundo *synset*. Os verbos que constituem esse par de *synsets* estão relacionados pelo acarretamento e inclusão temporal.

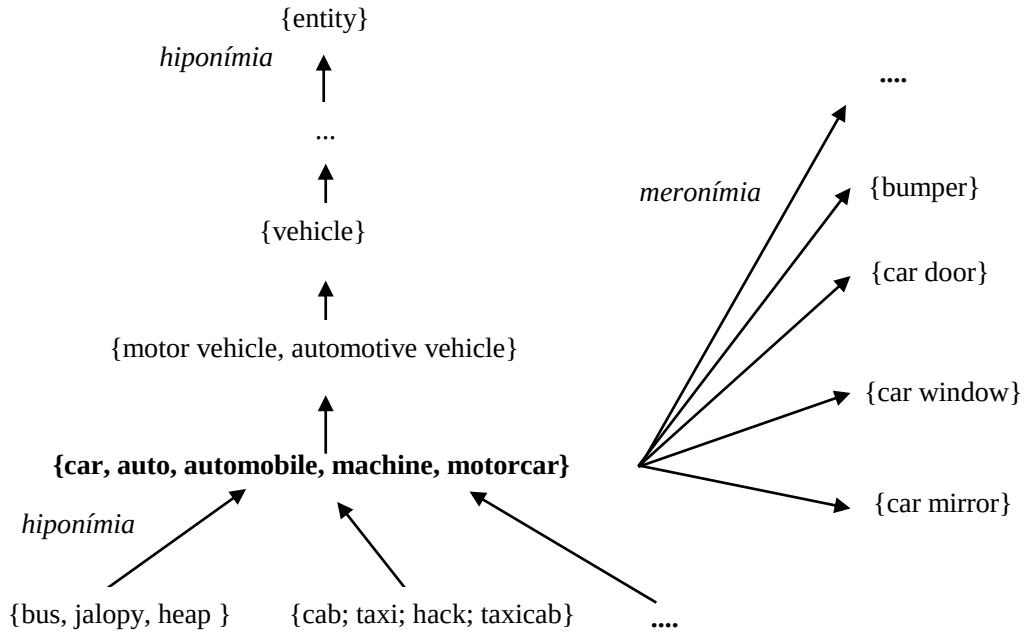
e) **Troponímia:** esse é um tipo de acarretamento entre uma ação A1 e A2, em que *A1 é um A2 de certa maneira*. Esse é o caso, por exemplo, da relação entre os *synsets* {limp, gimp, hobble, hitch}, que representa o conceito "mancar", e {walk}, que codifica "caminhar", pois "mancar" também é um "caminhar" de forma específica. A troponímia, aliás, é conhecida como *manner of* ou *is a way of*. Em outras palavras, "alguém estar mancando" acarreta "alguém está caminhando".

f) **Causa**: relação que se estabelece entre uma ação A1 e uma ação A2 quando a ação A1 denotada pelo verbo *x* causa a ação A2 denotada pelo verbo *y*. Esse é o caso, por exemplo, da relação que se estabelece entre o *synset* {kill} e o *synset* {die, decease, pass away}.

A noção de *synset* e algumas relações lógico-conceituais são ilustrados na Figura 15. Nessa Figura, o *synset* {car, auto, automobile, machine, motorcar} está imediatamente relacionado a:

a) *synset* hiperônimo {motor vehicle, automotive vehicle};
b) *synsets* hipônimos {cruiser, squad car, patrol car, police car, prowl car} e {cab, taxi, hack, taxicab}, etc.;
c) partes que o compõem ou *synsets* merônimos {bumper}, {car door}, {car mirror} e {car window}, etc.

Figura 15 - O construto *synset* e a estruturação dos conceitos nominais.




A WN.Pr também registra outras informações, ditas adicionais, a saber:

a) para cada unidade lexical, há uma frase-exemplo para ilustrar o seu contexto de uso, p.ex.: para "car", no *synset* {car, auto, automobile, machine, motorcar}, há a frase-exemplo "*he needs a car to get to work*" ("ele necessita de um carro para ir trabalhar") (trad. nossa);

b) para cada *synset*, há uma glosa que especifica informalmente o conceito por ele lexicalizado, p.ex.: para o *synset* { car, auto, automobile, machine, motorcar}, há a glosa "*a motor vehicle with four wheels; usually propelled by an internal combustion engine*" ("um veículo com quatro rodas; usualmente impulsionado por um motor de combustão interno") (trad. nossa);
c) para cada synset, há também a especificação do tipo semântico expresso pelo conceito a ele subjacente; p.ex.: o *synset* {bicycle, bike, wheel, cycle} é do tipo semântico <noun.artifact>.

Como modelo de representação do conhecimento, a WN.Pr tem sido considerada uma "ontologia linguística", ou seja, recurso de larga escala que é, por um lado, um estoque de unidades lexicais de uma língua geral e, por outro, um inventário de conceitos lexicalizados compartilhados por uma comunidade linguística (VOSSEN, 1998).

Em outras palavras, as ontologias linguísticas são um tipo especial de ontologia porque elas armazenam conceitos lexicalizados (em dada língua) e não são objetos formais (MAGNINI, SPERANZA, 2002). Esse tipo de ontologia se opõe às formais, cujos conceitos e relações estão explicitamente descritos por um formalismo e cujo inventário de conceitos extrapola a representação de conceitos lexicalizados.

A WN.Pr, enquanto ontologia linguística, tem sido utilizada em várias aplicações de PLN como recuperação e extração de informação, desambiguação lexical de sentido (do inglês, *word sense disambiguation*), categorização e estruturação de documentos, entre outras (MORATO *et al.*, 2004; DI FELIPPO, 2008).

Como dito, as unidades lexicais e expressões do inglês estão organizadas em quatro classes de palavras na WN.Pr. Cada uma delas constitui uma base léxico-conceitual própria, em que os *synsets* estão organizados por relações semântico-conceituais específicas, encarregados da estruturação interna dos conceitos lexicalizados por cada classe. Com base em Fellbaum (1998), o Quadro 8 resume o conjunto principal de relações em função das 4 classes de palavras (nome, verbo, adjetivo e advérbio).

Como mencionado, a representação estruturada do conteúdo dos textos resultará da indexação dos nomes à WN.Pr. Apesar de a classe dos nomes estar estruturada na WN.Pr em função da antonímia, hiponímia e meronímia, somente a hiponímia será considerada na representação dos textos-fonte, tendo em vista a relevância desta relação para a estruturação do conhecimento, já apontada na literatura. Assim, a representação do conteúdo do *cluster* se caracteriza por uma hierarquia conceitual.

Quadro 8 – As relações da WN.Pr em função das classes de palavras.

| Relação | Classe | Exemplo |
|---|---|---|
| Antonímia | N-N<br>V-V<br>Adj-Adj | {man, adult man} → {woman, adult female}<br>{enter, come in, ...} → {exit, go out, ...}<br>{beautiful} → {ugly} |
| Hiponímia (*is-a*) / Hiperonímia | N-N | {bus, jalopy, heap}→{car, auto, automobile, machine, motorcar}<br>{motor vehicle, automotive vehicle}→{car, auto, automobile, machine, motorcar} |
| Meronímia (*has-part*) / Holonymy (*part-of*) | N-N | {car, auto, automobile, machine, motorcar}}→{car mirror}<br>{car mirror}→{car, auto, automobile, machine, motorcar |
| Acarretamento | V-V | {snore, saw wood, ... } → {sleep, kip, slumber, log Z's, ... } |
| Troponímia (*is a way of*) | V-V | {limp, gimp, hobble, hitch}→{walk} |
| Causa | V-V | {kill}→{die, decease, pass away} |
| Legenda: N(ome), V(erbo), Adj(etivo), Adv(érbio) | | |

Fonte: Fellbaum (1998).

A seguir, descreve-se o processo de indexação dos nomes de C1 à hierarquia de *synsets* da WN.Pr e o recorte da parcela da hierarquia que representa o conteúdo de C1.

## 4.3. Indexação léxico-conceitual manual

A indexação léxico-conceitual consistiu na identificação do *synset* que mais adequadamente representa o conceito subjacente a cada nome e na subsequente seleção da hierarquia superior (hiperônimos) do referido *synset.* Ao final, a união das hierarquias parciais resultantes da indexação de cada um dos 38 nomes constitui a representação conceitual estruturada do *cluster* C1.

Para a realização da indexação, optou-se pelo método manual, via consulta ao arquivo *off-line* da WN.Pr (3.0), para buscar os dados conceituais e subsequente armazenamento dos mesmos no editor de planilhas *Microsoft Excel*. Optou-se pelo método manual porque não havia uma ferramenta à época que atendesse às necessidades do projeto[34,35]. Tal método é composto pelos passos descritos e ilustrados na sequência.

[34] À época da modelagem do *cluster* C1, tinha-se disponível o NASP (NÓBREGA, 2013), isto é, um editor de auxílio à anotação semântica dos nomes de uma coleção multidocumento via *synsets* da WN.Pr. Esse editor, no entanto, permite apenas a indexação à WN.Pr dos nomes que compõem o conjunto dos 10% mais frequentes da coleção e não organiza os *synsets* anotados em uma estrutura conceitual.

[35] Ressalta-se que atualmente já está disponível uma extensão do NASP, denominada NASP++ (CABEZUDO, 2015), que (semi)automatiza a anotação dos nomes e verbos de uma coleção multidocumento em português à WN.Pr e também organiza os conceitos/*synsets* anotados em uma estrutura hierárquica. Tal editor foi utilizado para a efetiva proposição dos métodos de SAM, pois essa

Especificamente, cada um dos 38 nomes foi indexado manualmente à WN.Pr por meio da seguinte metodologia:

(a) tradução das unidades extraídas de C1 para a língua inglesa;
(b) busca pelo *synset* da WN.Pr que possui o termo em inglês;
(c) identificação do conceito/*synset* mais adequado, e
(d) seleção de todos os hiperônimos do *synset* escolhido em (iii).

Para a tradução, dois dicionários bilíngues português-inglês foram utilizados, a saber: (i) a versão *online* do "Michaelis: moderno dicionário inglês-português" (WEISZFLOG, 2000), disponível no portal UOL[36], e (ii) *WordReference*[37], dicionário multilíngue disponível *online*. Quando necessário, outros recursos também auxiliaram na tarefa de tradução, como o *Linguee*[38] e o *Google Translator*[39].

Para ilustrar a indexação, toma-se como ponto de partida o nome "acidente" de C1 (cf. Quadro 7, pág. 50). De acordo com a metodologia, o primeiro passo consistiu em traduzir "acidente" para o inglês. Com base nos referidos dicionários e recursos de tradução, seleciou-se a palavra *accident* como equivalência mais adequada.

Após a tradução, buscou-se pela unidade traduzida na interface *online* da WN.Pr. No caso, *accident* é elemento constitutivo de dois *synsets* da WN.Pr: (i) {accident}, cuja glosa é *"a mishap; especially one causing injury or death"*[40], e (ii) {accident, fortuity, chance event}, definido como "*anything that happens by chance without an apparent cause*"[41]. Ao constituir dois *synsets*, *accident* lexicaliza dois conceitos distintos em inglês, sendo necessário identificar o que de fato está expresso nos textos-fonte. Com base nos hiperônimos e nas glosas de cada *synset*, identificou-se {accident} como o *synset* que mais adequadamente representa o conceito em questão.

A seguir, todos os hiperônimos que constituem a hierarquia de {accident} foram selecionados. No sentido *bottom-up*, o conjunto de hiperônimos é composto por: {mishap, misadventure, mischance} → {misfortune, bad luck} → {trouble} → {happening, occurrence, occurrent, natural event} → {event} → {psychological feature} → {abstraction} → {abstract entity} → {entity}. Assim, a hierarquia de {accident} possui no total 10 níveis.

---

tarefa requer a representação conceitual hierárquica de outras coleções do CSTNews. Mais informações sobre esse editor são fornecidas na pág. 84.

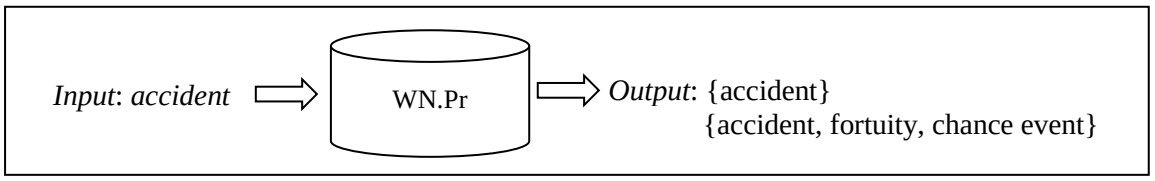
[36] http://michaelis.uol.com.br/
[37] http://www.wordreference.com/
[38] http://www.linguee.com.br
[39] http://translate.google.com/
[40] "um acidente; especialmente um que cause lesão ou morte" (trad. nossa)
[41] "tudo o que acontece por acaso, sem causa aparente" (trad. nossa)

As Figuras 16, 17, 18 e 19 ilustram, respectivamente, os passos (a), (b), (c) e (d).

Figura 16 - Tradução (a)

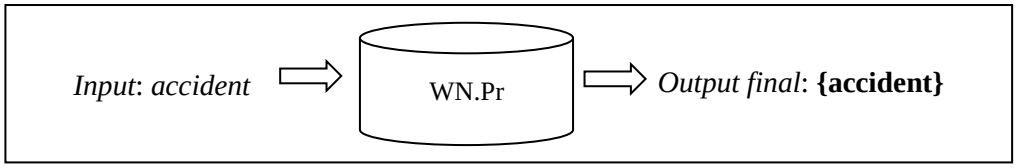

Fonte: autoria própria.

Figura 17 - Busca na WN.Pr (b).




Fonte: autoria própria.

Figura 18 - Identificação do conceito/ *synset* (c).




Fonte: autoria própria.

Figura 19 - Seleção dos hiperônimos do *synset* identificado em (d).

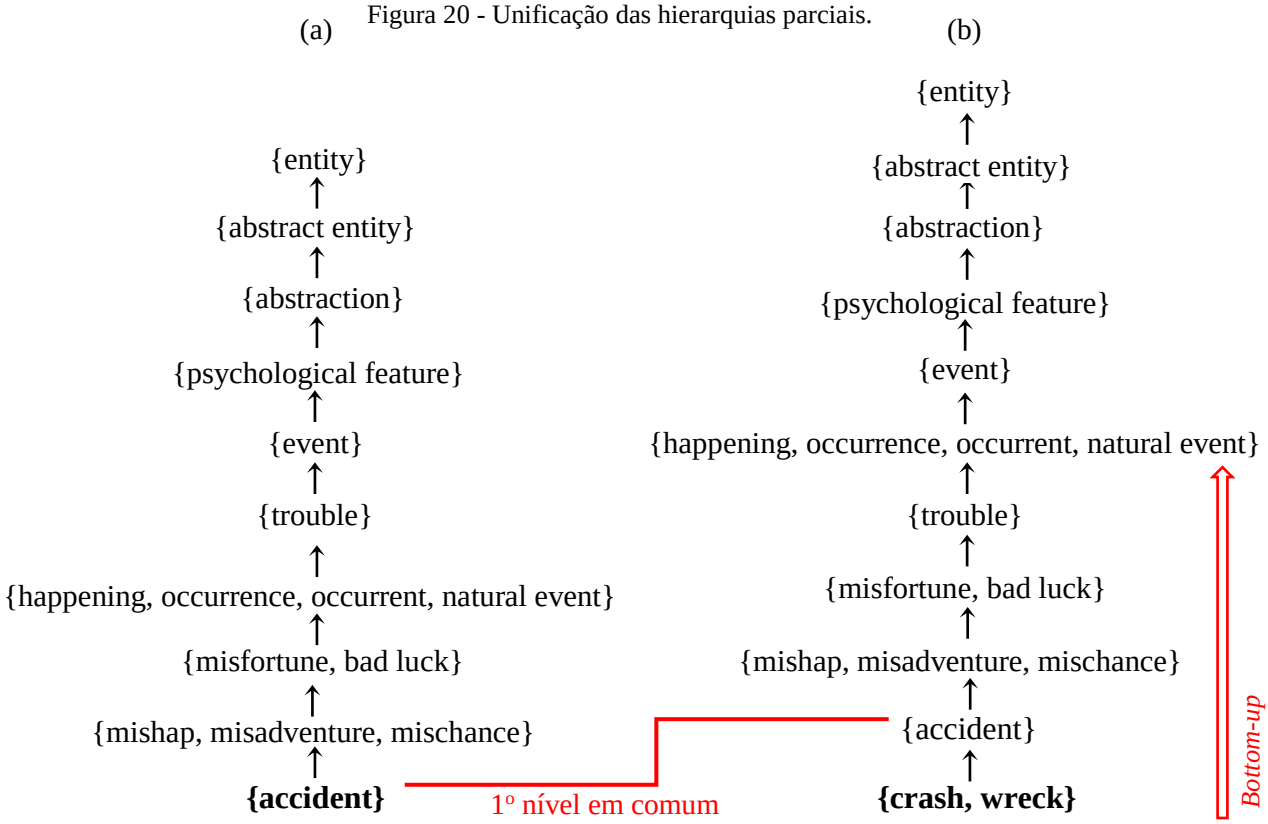
{entity}
↑
{abstract entity}
↑
{abstraction}
↑
{psychological feature}
↑
{event}
↑
{happening, occurrence, occurrent, natural event}
↑
{trouble}
↑
{misfortune, bad luck}
↑
{mishap, misadventure, mischance}
↑
{accident}

Fonte: autoria própria.

Ao final da indexação, obtiveram-se 37 hierarquias parciais, pois dos 38 nomes apenas 2 ("tripulação" e "tripulante") representam conceitos muito similares e, por isso, foram indexados a um mesmo *synset* ({crew} "*the men and women who man a vehicle (ship, aircraft, etc.*"[42]). As hierarquias parciais resultantes da indexação são então compostas por conceitos que ocorreram nos textos-fonte e por outros que foram herdadas da WN.Pr para garantir a organização do conteúdo. Todo conceito que ocorreu nos textos-fonte inicia uma hierarquia própria, como se observa na Figura 19. Nela, tem-se a árvore resultante da indexação do nome "acidente". Para que se obtivesse uma modelagem única da coleção C1, as hierarquias parciais foram unificadas como descrito a seguir.

Uma vez armazenadas em um único arquivo Excel, as hierarquias parciais foram automaticamente unificadas. Tal unificação consistiu em percorrer as hierarquias no sentido *bottom-up* e unificar as árvores ao se identificar o primeiro *synset* em comum entre elas. Na Figura 20, em que se ilustra essa estratégia, vê-se que as hierarquias (a) e (b) são praticamente idênticas, sendo que (b) possui um nível a mais ({crash, wreck}). Entre elas, o primeiro conceito/*synset* em comum no sentido *bottom-up* é {accident}, utilizado como ponto de união das árvores. Em outras palavras, a hierarquia (b) engloba (a) e, por isso, a hierarquia resultante da unificação das árvores de "acidente" e "queda" é, na verdade, a hierarquia (b) (Figura 21).

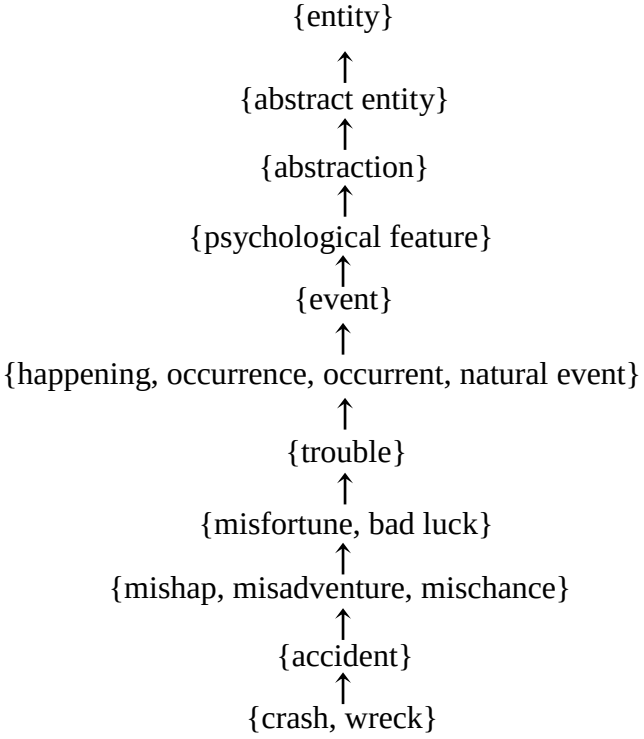
Figura 20 - Unificação das hierarquias parciais.




Fonte: autoria própria.

[42] "Os homens e mulheres que operam um veículo (navio, aviões, etc.)" (tradução nossa).

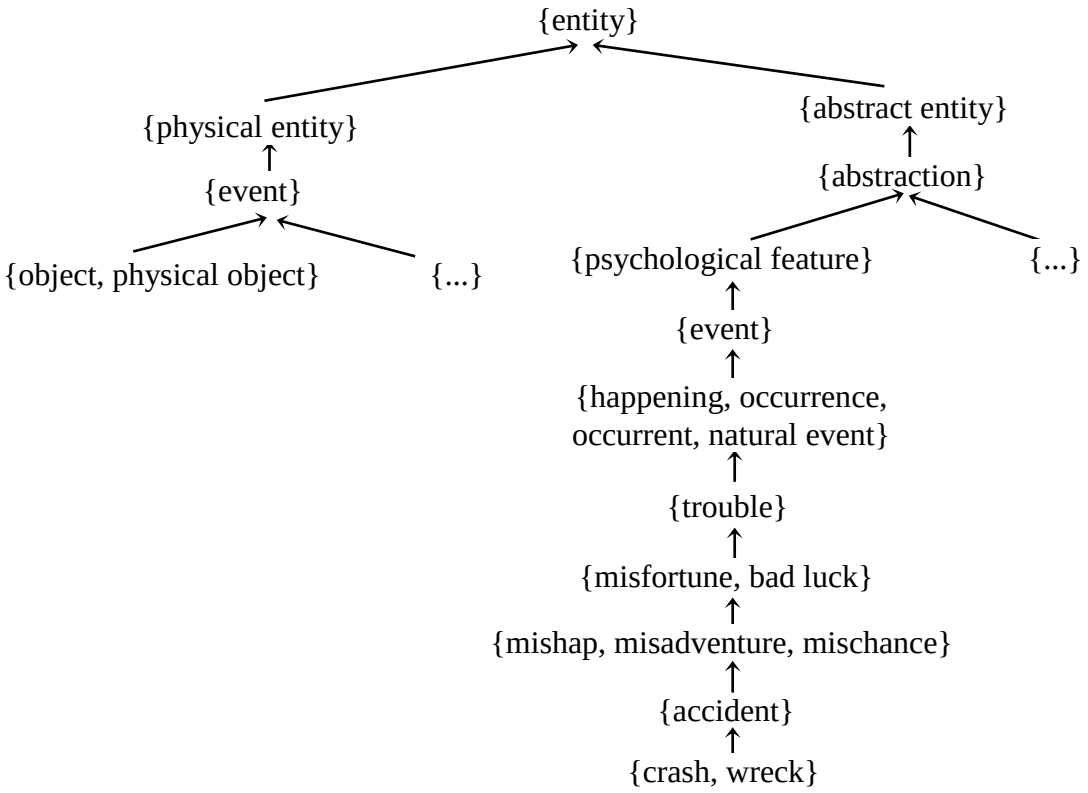
Figura 21 - Resultado da unificação de hierarquias parciais




Fonte: autoria própria.

Na Figura 22, tem-se uma ilustração simplificada da modelagem final do *cluster* C1.

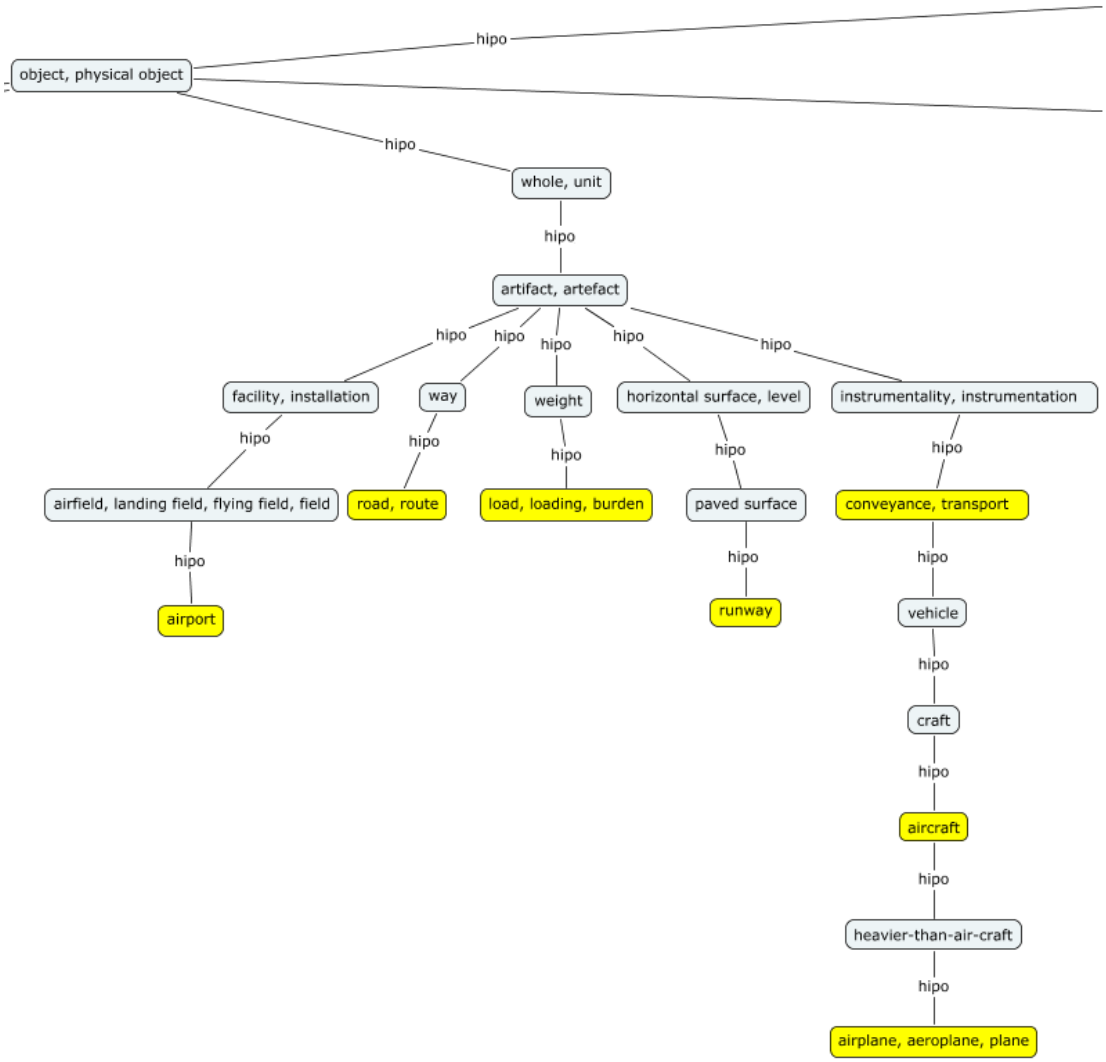
Figura 22 - Hierarquia conceitual simplificada de C1 a partir da WN.Pr.




Fonte: autoria própria.

Após a unificação, a hierarquia foi representada graficamente (em árvore) por meio da ferramenta de visualização e compartilhamento de conhecimento *Cmap Tools*[43]. Na Figura 23, tem-se a árvore de C1 gerada no Cmap, com foco em parte da hierarquia que engloba as entidades da categoria {physical entity}. Nela, os nós em amarelo correspondem aos conceitos provenientes dos textos-fonte e os demais foram herdados da WN.Pr para a construção da hierarquia.

Figura 23 - Hierarquia conceitual de C1 como árvore.




Fonte: autoria própria.

Sobre a hierarquia de C1, cabem aqui alguns destaques. A hierarquia de C1 possui no total 132 conceitos, sendo 37 provenientes dos textos-fonte e 95 herdados da WN.Pr para organização do conteúdo. Quanto às duas grandes categorias de conceitos, as das entidades abstratas e concretas, a hierarquia de C1 é bastante equilibrada, possuindo 65

[43] http://ftp.ihmc.us/

conceitos organizados sob o hiperônimo {abstract entity} e 66 sob o *synset* {physical entity}. Quanto à classificação dos 37 conceitos provenientes dos textos em concretos e abstratos, a hierarquia de C1 possui 22 conceitos concretos e 15 abstratos. Vê-se aí uma ligeira prevalência das entidades concretas dado o assunto das notícias que compõem o referido *cluster*. Ademais, vale ressaltar que a hierarquia possui no total 12 níveis. Por fim, tendo em vista o conteúdo dos textos-fonte, é interessante destacar que no que diz respeito aos conceitos abstratos, a hierarquia de C1 engloba entidades dos seguintes campos semânticos, especificados aqui pelos *synsets*: {communication}, {measure, quality, amount}, {relation}, {psychological feature}, {attribute} e {group, grouping}. Os campos semânticos cobertos pelas entidades concretas são: {substance matter}, {process, physical process}, {phenomenon} e {object, physical object}.

A seguir, descreve-se essa investigação.

## 5. Análise das métricas de grafo como critério de relevância

Para essa investigação, selecionou-se um conjunto inicial de critérios/métricas potencialmente pertinentes para a captura do conteúdo relevante de uma coleção multidocumento. A seguir, apresenta-se a hipótese sobre a pertinência de cada critério e a definição de cada um deles em função do cenário em que este trabalho se enquadra.

### 5.1. Delimitação das métricas

Com base na revisão da literatura, selecionaram-se 5 critérios de relevância, codificados nas medidas: *Simple Frequency*, *Cumulative Frequency, Centrality*, *Closeness* e *Level*. De um modo geral, tais critérios foram selecionados por serem relativamente simples, dada a natureza exploratória e pioneira deste trabalho, e eficientes, posto que foram os mais explorados na literatura, ainda que por meio de denominações diferentes, como é o caso da medida "grau" em Akabane *et AL*. (2011) e da medida *Centrality* em Sousa (2011). A seguir, descrevem-se os critérios/medidas selecionados.

a) *Simple Frequency*

Neste caso, *simple frequency* (em português, frequência simples) consiste no número de ocorrência dos conceitos nos textos-fonte. Esse critério tem sido amplamente utilizado nos trabalhos de SA e SAO para pontuar os conceitos representados em hierarquias ou

ontologias. Essa ampla aplicação na SAM, em especial, pauta-se no fato de que a frequência de ocorrência dos conceitos é uma alternativa eficaz para capturar a redundância, que é o principal critério usado na seleção de conteúdo/sentenças na SAM. A captura da redundância por meio da frequência simples pauta-se na hipótese de que os conceitos mais frequentes (redundantes) são os mais importantes em dada coleção de textos. Assim, a *Simple Frequency* de um conceito *x* em uma hierarquia que modela uma coleção multidocumento equivale ao total de ocorrência das diferentes palavras da coleção que expressam *x.*

b) *Cumulative Frequency*

O outro tipo de frequência, denominada *Cumulative Frequency* (em português, frequência acumulada), também têm sido utilizada em SA. Essa frequência busca privilegiar os conceitos mais genéricos, na medida em que o valor da *Cumulative Frequency* de um conceito *x* equivale à soma da frequência de ocorrência de todos os seus conceitos hipônimos ou subordinados. Tendo em vista a construção de sumários informativos e genéricos, essa medida é relevante porque permite dar maior peso às informações genéricas.

c) *Centrality*

O critério aqui denominado *Centrality* (em português, centralidade) é definido pelo número de ligações que um conceito possui com outros da hierarquia, sendo os relacionamentos codificados em arestas. Selecionou-se esse atributo porque ele busca definir o quão um conceito está relacionado a outros, o que pode ser relevante para selecionar sentenças que veiculam informações relacionadas, contribuindo para a coerência do sumário. Salienta-se que não se considerou a distinção entre arestas de entrada e de saída.

d) *Closeness*

Esse critério não determina a relevância dos conceitos de forma isolada, mas sim em relação aos outros conceitos. Em outras palavras, ele utiliza o relacionamento entre os conceitos de maior importância de uma representação conceitual para determinar a relevância. Essa medida, então, pode ser importante para garantir a identificação dos conceitos de maior importância relacionados entre si.

Especificamente, a medida *Closeness* de um conceito (ou nó) *Cn* considera a distância geodésica[44] entre *Cn* e todos os seus conceitos/nós alcançáveis na árvore (*Cx*). Assim, *Closeness* é menor para conceitos/nós que são mais centrais, pois o caminho é mais curto que a média das distâncias para outros conceitos/nós.

Neste trabalho, a fórmula comumente utilizada na literatura, apresentada em (9), foi adaptada, como evidenciado em (11).

(11) $$Closeness\ (Cn) = \frac{\sum Cx \in (C-Cn)\ (C\ \frac{cumulative\ frequency\ (Cx)}{distance\ (Cn,Cx)}}{\sum Cx \in (C-Cn)\quad 1/distance\ (Cn,Cx)}$$

Como mencionado, a *Closeness* é uma medida que pressupõe o cálculo prévio de um fator de relevância dos conceitos (*relevance(Cx)* em (9)). Segundo a fórmula em (11), destaca-se que a relevância dos conceitos neste trabalho é determinada pela frequência acumulada – *Cumulative Frequency (Cx)*, ou seja, a frequência do próprio conceito somada à frequência de seus conceitos hipônimos ou filhos.

e) *Level*

Esse atributo determina a localização de um conceito *x* em uma representação conceitual. No caso de um modelo hierárquico, o nível expressa a generalidade ou especificidade de *x*. Assim, há conceitos genéricos, intermediários e específicos. Segundo estudos de diferentes áreas, os conceitos intermediários costumam ser os mais representativos do ponto de vista cognitivo, não sendo nem tão genéricos e nem tão específicos.

A seguir, descreve-se como cada medida foi calculada para caracterizar os 37 conceitos constitutivos da hierarquia de C1 que ocorrem nos textos-fonte.

## 5.2. Cálculo das métricas

a) *Simple Frequency*

Como mencionado, a frequência simples consiste no número de ocorrência dos conceitos nos textos-fonte. Como quase todos os nomes de C1 foram indexados a conceitos distintos, a *Simple Frequency* de um conceito/*synset* neste trabalho equivale à

[44] A distância geodésica é a menor distância que une dois pontos ou nós em um grafo.

frequência de ocorrência dos nomes nos textos-fonte da coleção C1. A única exceção é a *Simple Frequency* de {crew}, que é a soma da frequência dos nomes "tripulação" e "tripulante", isto, é, 4+1=5. Assim sendo, a *Simple Frequency* foi manualmente especificada para cada um dos 37 conceitos da hierarquia que ocorreram nos textos-fonte. Em outras palavras, isso quer dizer que somente os conceitos provenientes da indexação dos nomes da coleção possuem *Simple Frequency*, ao passo que os herdados da WN.Pr para organização da árvore não possuem.

Na Tabela 1, tem-se a frequência de ocorrência nos textos-fonte de cada um dos 38 nomes e dos 37 conceitos correspondentes.

Tabela 1 - Especificação da Simple Frequency dos conceitos de C1.

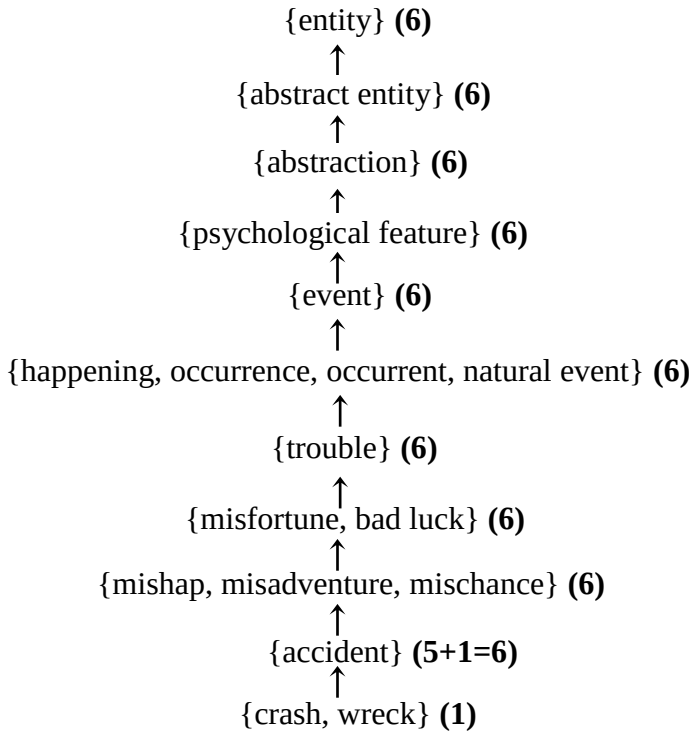
| **Qt.** | **Nome** | **Conceito/*synset* correspondente** | ***Simple Frequency*** |
|---|---|---|---|
| 1 | avião | {airplane, aeroplane, plane} | 11 |
| 2 | porta-voz | {spokesperson, interpreter, representative, voice} | 7 |
| 3 | acidente | {accident} | 5 |
| 4 | passageiro | {passenger, rider} | 5 |
| 5 | aeroporto | {airport} | 4 |
| 6 | companhia | {line} | 4 |
| 7 | membro | {member} | 4 |
| 8 | tripulação/ tripulante | {crew} | 5 |
| 9 | fabricação | {fabrication, manufacture} | 3 |
| 10 | floresta | {forest, wood, woods} | 3 |
| 11 | pessoa | {person, individual, someone, somebody, mortal, soul} | 3 |
| 12 | pista | {runaway} | 3 |
| 13 | quilômetro | {kilometer, kilometre, km, klick} | 3 |
| 14 | aterrissagem | {landing} | 2 |
| 15 | carga | {load, loading, burder} | 2 |
| 16 | distância | {distance} | 2 |
| 17 | fonte | {source} | 2 |
| 18 | leste | {east, due east, eastward, E} | 2 |
| 19 | localidade | {vinicity, locality, neighborhood, neighbourhood, neck of the woods} | 2 |
| 20 | mineral | {mineral} | 2 |
| 21 | nacionalidade | {nationality} | 2 |
| 22 | país | {country, state, land} | 2 |
| 23 | propriedade | {place, property} | 2 |
| 24 | sobrevivente | {survivor, subsister} | 2 |
| 25 | tarde | {afternoon} | 2 |
| 26 | tempo | {weather, weather condition, conditions, atmospheric condition} | 2 |
| 27 | vítima | {victim} | 2 |
| 28 | aeronave | {aircraft} | 1 |
| 29 | chama | {fire, flame, flaming} | 1 |
| 30 | cidade | {city, metroplolis, urban center} | 1 |
| 31 | estrada | {road, route} | 1 |
| 32 | montanha | {mountain, mount} | 1 |
| 33 | permissão | {license, permission} | 1 |
| 34 | queda | {crash, wreck} | 1 |
| 35 | setor | {sector} | 1 |
| 36 | tempestade | {storm, violent storm} | 1 |
| 37 | transporte | {conveyance, transport} | 1 |

b) *Cumulative Frequency*

Como fora definida, o valor da medida *Cumulative Frequency* de um conceito *x* equivale à soma da frequência de ocorrência (*Simple Frequency*) de todos os seus conceitos hipônimos ou subordinados. Sendo assim, os conceitos que estão mais no topo da árvore são aqueles que têm maior frequência acumulada. Ao contrário da *Simple Frequency*, a *Cumulative Frequency* foi manualmente calculada para todos os conceitos/*synsets* da hierarquia de C1, tanto os provenientes do *corpus* quanto os exclusivamente herdados da WN.Pr. O cálculo manual foi feito através da observação dos conceitos devidamente representados via *Cmap*, ou seja, cada conceito recebeu o devido valor de acordo com a herança da frequência dos hipônimos.

Na Figura 24, por exemplo, tem-se a hierarquia resultante da unificação de apenas duas hierarquias parciais, no caso, geradas a partir da indexação de "acidente" e "queda" à WN.Pr. Nela, observa-se que o *synset*-folha {crash, wreck}, que constitui o nível mais inferior, possui *Simple Frequency*=1, a qual é somada à *Simple Frequency* de seu *synset* hiperônimo, {accident}. Assim, o *synset* {accident}, que possui *Simple Frequency*=5, tem uma *Cumulative Frequency* de valor 6. Na sequência, todos os hiperônimos de {accident} (ou seja, de {mishap, misadventure, mischance} à {entity}, no sentido *bottom-up*) herdam o valor 6, até que outras unificações venham a modificar o valor da *Cumulative Frequency* de algum conceito da hierarquia.

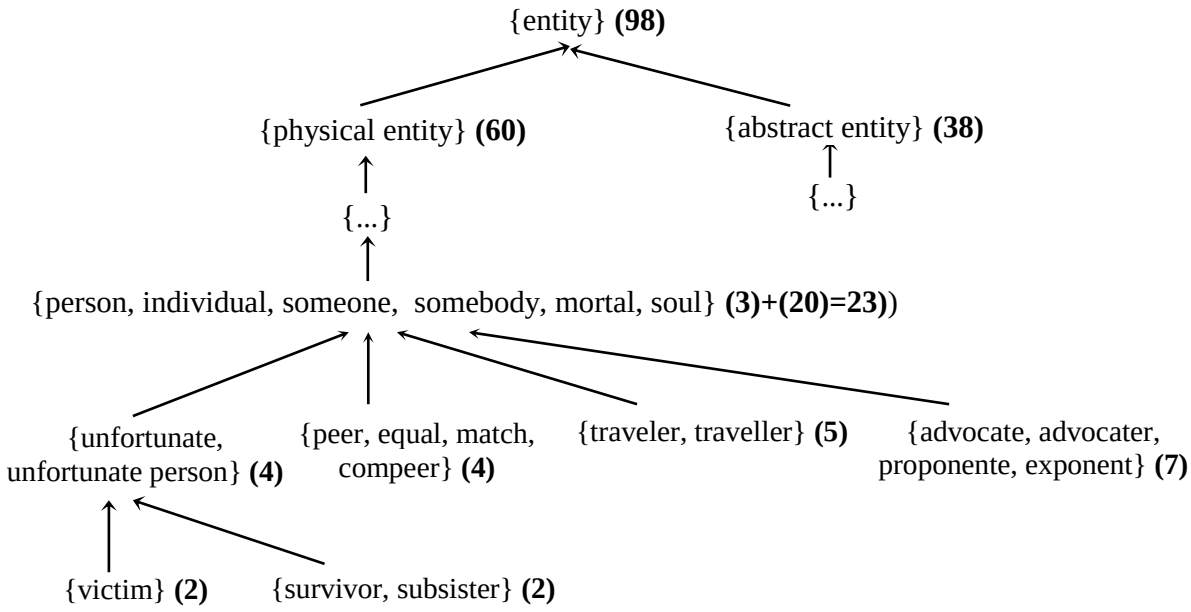
Figura 24 - Cálculo da *Cumulative Frequency*.




Fonte: autoria própria.

Na Figura 25, ilustra-se mais uma vez o cálculo da *Cumulative Frequency*, considerando-se a hierarquia final de C1. Nela, vê-se, por exemplo, que a *Cumulative Frequency*=4 associada ao *synset* {unfortunate, unfortunate, person} resulta da soma da *Simple Frequency* de seus dois hipônimos {victim} (*Simple Frequency*=2) e {survivor, subsister} (*Simple Frequency*=2). Além disso, destaca-se nessa Figura que os conceitos/*synsets* mais superiores possuem *Cumulative Frequency* com valores altos, quando se considera a hierarquia final do *cluster* C1. Por exemplo, o *synset* {entity}, que constitui o *top-concept* da árvore, isto é, o conceito mais genérico e abstrato da modelagem, possui *Cumulative Frequency* de valor 98, já que esta resulta da soma da frequência de seus conceitos hipônimos, {physical entity} e {abstract entity}, isto é, 60 + 38 = 98.

Figura 25 - A *Cumulative Frequency* dos *top*-conceitos.



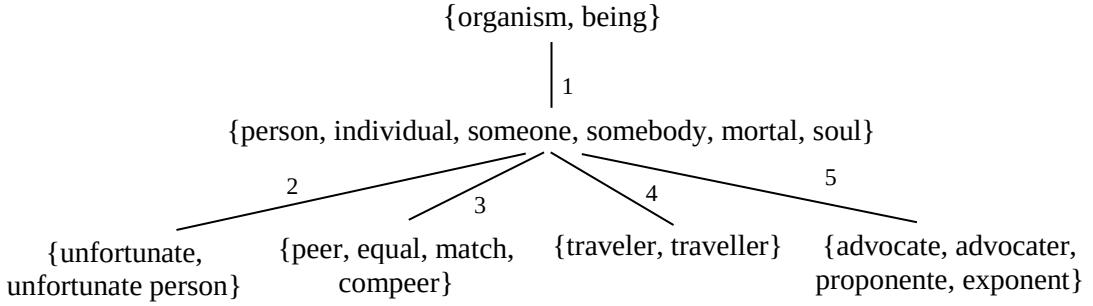

Fonte: autoria própria.

*c) Centrality*

Desconsiderando-se a direção das arestas, a métrica *Centrality* aqui definida consiste no número de ligações que um conceito possui com outros conceitos da hierarquia, o que é codificado pelo número de arestas do nó na árvore conceitual. Por exemplo, a *Centrality* do conceito representado pelo *synset* {person, individual, someone, somebody, mortal,

soul}, cuja glosa é "*a human being*[45]", tem valor 5, como ilustrado na Figura 26. A medida *Centrality* também foi manualmente calculada para todos os conceitos que constituem a hierarquia de C1 com base na observação do número de arestas provenientes de cada conceito via *Cmap*.

Figura 26 - Cálculo da *Centrality*




Fonte: autoria própria.

Nessa Figura, o conceito em questão está relacionado ao hiperônimo direto {organism, being} ("*a living thing that has (or can develop) the ability to act or function independently*"[46]) e a 4 hipônimos (isto é, {unfortunate, unfortunate person} ("*a person who suffers misfortune*"[47]), {peer, equal, match, compeer} ("*a person who is of equal standing with another in a group*"[48]), {traveler, traveller} ("a person who changes location"[49]) e {advocate, advocater, proponente, exponent} ("*a person who pleads for a cause or propounds an idea*"[50])), totalizando 5 arestas.

*d) Closeness*

Na fórmula de *Closeness* adaptada para este trabalho (11), o fator de relevância *relevance(Cx)* da medida original (9) é a *Cumulative Frequency(Cx).* Assim, para

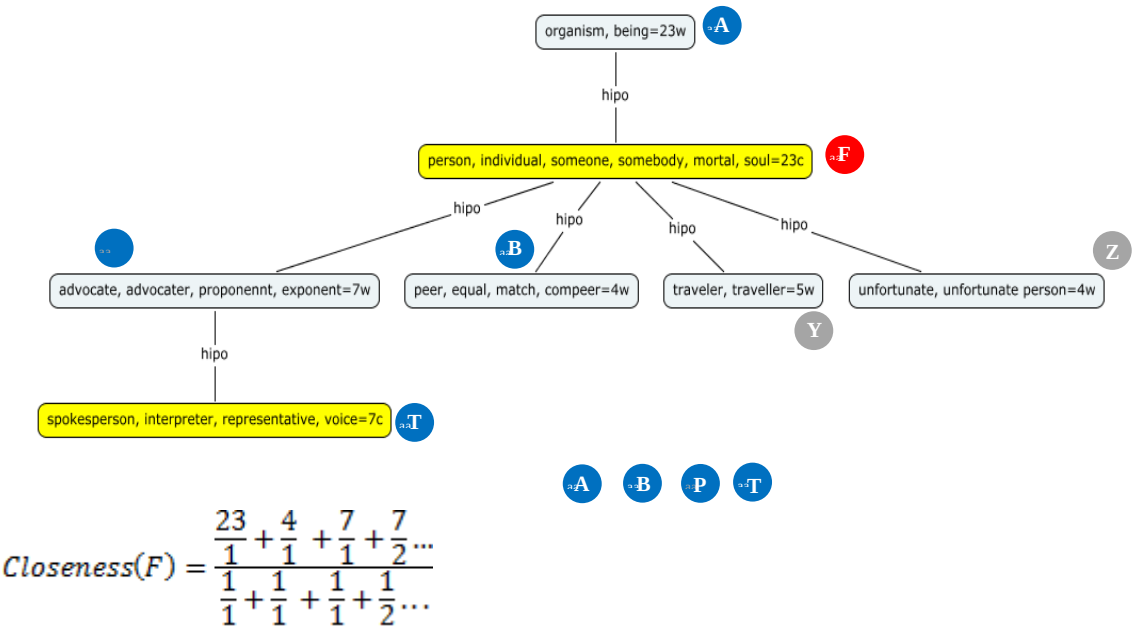
[45] "Um ser humano" (tradução nossa).
[46] "Uma coisa viva que tem (ou pode desenvolver) a capacidade de agir ou funcionar de forma independente" (tradução nossa).
[47] "Uma pessoa que sofre infortúnio" (tradução nossa).
[48] "Uma pessoa que está em pé de igualdade com outro em um grupo" (tradução nossa).
[49] "Uma pessoa que muda de local" (tradução nossa).
[50] "Uma pessoa que defende uma causa ou propõe uma ideia" (tradução nossa).

calcular *Closeness*, partiu-se da especificação prévia da referida frequência. Ademais, *Closeness* foi a única medida calculada de forma automática\[51].

Na Figura 27, ilustra-se especificamente o cálculo de *Closeness* do conceito F (isto é, *Closeness*(F)) da hierarquia, isto é, do *synset* {person, individual, someone, somebody, mortal, soul}. Na fórmula da Figura 27, vê-se que a variável *Cx* da fórmula descrita em (11) representa todos os conceitos da hierarquia, exceto F (isto é, {organism, being}=A, {peer, equal, match, compeer}=B, {advocate, advocater, proponennt, exponent}=P, {spokesperson, interpreter, representative, voice}=T, etc.).

Assim, *cumulative relevance(Cx)* é preenchida pelos valores da *Cumulative Frequency* previamente identificada para cada um dos conceitos. Na Figura 27, esse valor está associado aos *synsets* pelo símbolo de "=". Por exemplo, *cumulative Frequency(A)=23*, *Cumulative Frequency(B)=4*, *Cumulative Frequency(P)=7*, *Cumulative Frequency(T)=7*, etc. A *distance(Cn,Cx)* da fórmula em (11), por sua vez, é preenchida pela distância entre F e os demais conceitos. Por exemplo, *distance(F,A)=1*, *distance(F,B)=1*, *distance(F,T)=2*, etc. Dessa forma, calculou-se o valor de *Closeness* para cada um dos 37 conceitos da hierarquia.

Figura 27 - Cálculo da medida *Closeness*




Fonte: autoria própria.

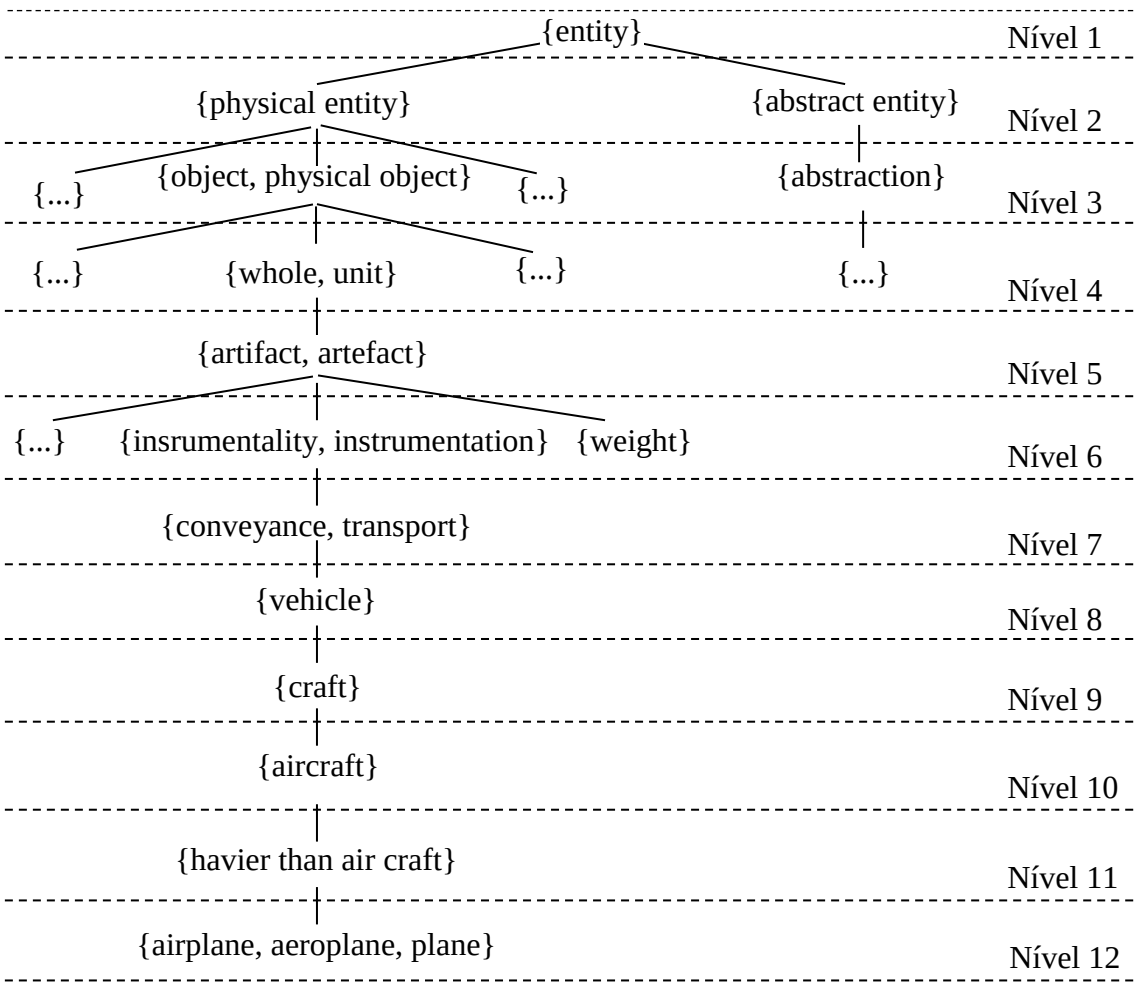
[51] A medida *Closeness* foi calculada por meio de um *script* específico, desenvolvido por Fernando A. A. Nóbrega, doutorando em Ciência da Computação (ICMC/USP) que integra o grupo de SA do NILC, e a quem se agradece muito o suporte.

e) *Level*

Para determinar o nível (do inglês, *level*) do conceito na hierarquia, identificou-se que esta possui no total 12 níveis conceituais. No sentido *bottom-up*, a localização do conceito mais inferior, representado pelo *synset* {airplane, aeroplane, plane} ("*an aircraft that has a fixed wing and is powered by propellers or jets*"[52]), foi especificada como "nível 12". Assim, o *top-concept* da hierarquia, codificado pelo *synset* {entity} ("*that which is perceived or known or inferred to have its own distinct existence (living or nonliving)*"[53]), está localizado no "nível 1". Na Figura 28, destaca-se o ramo mais profundo da árvore referente ao *cluster* C1, em que todos os níveis podem ser observáveis.

Uma vez que os conceitos da árvore referente ao *cluster* C1 foram caracterizados ou especificados pelas 5 medidas selecionadas e adaptadas da literatura, procedeu-se à investigação da pertinência das mesmas para a distinção da relevância dos conceitos.

Figura 28 - Especificação da métrica *Level.*




Fonte: autoria própria.

[52] "Uma aeronave que tem asa fixa e é alimentada por hélices ou jatos" (trad. nossa).
[53] "Aquilo que é percebido, reconhecido ou inferido como tendo existência própria (vivo ou não vivo)" (trad. nossa).

## 5.3. Análise da pertinência das métricas

A pertinência das métricas foi analisada por 2 métodos, considerados complementares. O método manual consistiu em analisar a pertinência das métricas por meio da intuição/olhar do linguista. A análise automática, por sua vez, caracterizou-se pela utilização de algoritmos de Aprendizado de Máquina (AM). Em ambos, os conceitos/*synsets* subjacentes aos 13 nomes constitutivos de um dos sumários humanos do *cluster* C1 foram considerados referência. Isso se deve ao fato de que tais sumários, sendo informativos e genéricos, veiculam o conteúdo principal de C1 e, por isso, os conceitos que neles ocorrem como resultado da seleção humana de conteúdo podem ser considerados os mais relevantes da coleção. A seguir, descreve-se a análise manual.

### *5.3.1. Análise manual*

A análise manual visou à correlação entre as 5 métricas e a relevância dos conceitos.

Para tanto, tomou-se como dado de partida a coleção dos 37 conceitos/*synsets* de C1, devidamente caracterizados pelas 5 medidas ou métricas selecionadas e pelo fator de relevância, isto é, a ocorrência no sumário de referência. A ocorrência dos conceitos no sumário foi especificada por dois valores, que dividem os conceitos em duas classes: "sim" (isto é, conceitos que ocorrem no sumário) e "não" (isto é, conceitos que não ocorrem no sumário). Os dados para a análise manual estão organizados na Tabela 2. Por questão de espaço, os conceitos estão descritos na Tabela 2 pelos nomes que os lexicalizam. Diante de dúvidas sobre a delimitação dos conceitos, estes podem ser encontrados na Tabela 1 codificados em *synsets*.

A partir da Tabela 2, a análise manual teve início com o cálculo da média simples dos valores obtidos por cada métrica. Na sequência, verificou-se o número de conceitos que obteve valor igual ou superior à média em função de cada classe ("sim" e "não"). Por exemplo, a média de *Simple Frequency* para os 37 conceitos foi aproximadamente 2,6. Dos 13 conceitos da classe "sim", 6 apresentam *Simple Frequency* com valor igual/superior à média ("avião", "floresta", "membro", "passageiro", "pessoa" e "tripulação"). Dos 24 conceitos da classe "não", 7 deles possuem *Simple Frequency* com valor igual/superior à média ("acidente", "aeroporto", "companhia", "fabricação", "pista", "posta-voz" e "quilômetro"). A média da *Cumulative Frequency,* por sua vez, foi aproximadamente 3,8. Dos 13 conceitos da classe "sim", 5 apresentam *Simple Frequency* com valor igual/superior à média ("avião", "membro", "passageiro", "pessoa", "tripulação/tripulante"). Dos 24 conceitos

da classe “não”, 6 deles possuem *Simple Frequency* com valor igual/superior à média (“acidente”, “aeronave”, “aeroporto”, “companhia”, “porta-voz” e “transporte”).

A única exceção a esse cálculo com base na média diz respeito à métrica *Level*, pois se identificou, a partir dos dados da Tabela 2, o número de conceitos intermediários pertencente a cada classe. Para determinar o nível intermediário, os conceitos da hierarquia de C1 foram assim classificados em função dos 12 níveis: (i) *top-concepts*, correspondem aos conceitos dos níveis 12, 11, 10 e 9, ou seja, os conceitos do topo da hierarquia, (ii) *middle-concepts*, englobam os conceitos dos níveis 8, 7, 6 e 5, ou seja, os conceitos do meio da hierarquia, e (iii) *lower-concepts*, compreendem os *synsets* dos níveis 4, 3, 2, e 1, ou seja, os conceitos debaixo da hierarquia. Dentre eles, os *middle-concepts* são considerados os de nível intermediários dada a hierarquia de C1. Na Tabela 2, todos os valores iguais ou acima da média estão em destaque.

Tabela 2 - Os conceitos do *cluster* C1 do CSTNews e suas métricas de relevância.

| Qtdade. | Nome/ conceito | Sumário | Métrica | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Simple Frequency* | *Cumulative Frequency* | *Centrality* | *Closeness* | *Level* |
| 1 | avião | Sim | 11 | 11 | 1 | 8,041560382 | 1 |
| 2 | carga | Sim | 2 | 2 | 1 | 7,547655013 | 6 |
| 3 | floresta | Sim | 3 | 3 | 1 | 6,810928806 | 6 |
| 4 | membro | Sim | 4 | 4 | 1 | 7,314982201 | 4 |
| 5 | mineral | Sim | 2 | 2 | 1 | 7,686018373 | 8 |
| 6 | montanha | Sim | 1 | 1 | 1 | 7,538074088 | 7 |
| 7 | nacionalidade | Sim | 2 | 2 | 1 | 6,695537498 | 6 |
| 8 | passageiro | Sim | 5 | 5 | 1 | 7,657742031 | 5 |
| 9 | pessoa | Sim | 3 | 23 | 5 | 8,656908225 | 7 |
| 10 | queda | Sim | 1 | 1 | 1 | 6,26750143 | 2 |
| 11 | tempo | Sim | 2 | 2 | 1 | 6,953761859 | 6 |
| 12 | tripulação, tripulante | Sim | 5 | 5 | 1 | 6,811762991 | 5 |
| 13 | vítima | Sim | 2 | 2 | 1 | 7,424212214 | 5 |
| 14 | acidente | Não | 5 | 6 | 2 | 6,524584767 | 3 |
| 15 | aeronave | Não | 1 | 12 | 2 | 8,274407488 | 3 |
| 16 | aeroporto | Não | 4 | 4 | 1 | 7,403368688 | 5 |
| 17 | aterrissagem | Não | 2 | 2 | 1 | 6,227287261 | 4 |
| 18 | chama | Não | 1 | 1 | 1 | 6,377498738 | 4 |
| 19 | cidade | Não | 1 | 1 | 1 | 6,52026769 | 4 |
| 20 | companhia | Não | 4 | 4 | 1 | 6,500379639 | 3 |
| 21 | distância | Não | 2 | 2 | 1 | 6,437310632 | 5 |
| 22 | estrada | Não | 1 | 1 | 1 | 7,480978846 | 6 |
| 23 | fabricação | Não | 3 | 3 | 1 | 6,22151722 | 3 |
| 24 | fonte | Não | 2 | 2 | 1 | 6,491493504 | 5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 25 | leste | Não | 2 | 2 | 1 | 6,283283198 | 4 |
| 26 | localidade | Não | 2 | 2 | 1 | 6,848025562 | 5 |
| 27 | país | Não | 2 | 2 | 1 | 6,848025562 | 5 |
| 28 | permissão | Não | 1 | 1 | 1 | 5,838955311 | 3 |
| 29 | pista | Não | 3 | 3 | 1 | 7,310985946 | 5 |
| 30 | porta-voz | Não | 7 | 7 | 1 | 7,803125792 | 5 |
| 31 | propriedade | Não | 2 | 2 | 1 | 6,989497846 | 6 |
| 32 | quilômetro | Não | 3 | 3 | 1 | 6,585324996 | 5 |
| 33 | setor | Não | 1 | 1 | 1 | 6,443209937 | 5 |
| 34 | sobrevivente | Não | 2 | 2 | 1 | 7,424212214 | 5 |
| 35 | tarde | Não | 2 | 2 | 1 | 6,473907029 | 5 |
| 36 | tempestade | Não | 1 | 1 | 1 | 6,905572131 | 6 |
| 37 | transporte | Não | 1 | 13 | 2 | 8,452418538 | 6 |
| | | **MÉDIA** | **2,6486** | **3,838** | **1,1892** | **7,028980639** | **---** |

Na Tabela 3, sistematizam-se os resultados da verificação aqui descrita para as 5 métricas quanto às 2 classes (“sim” e “não”). Tendo em vista o desbalanceamento das classes, os resultados estão dispostos em valores absolutos e porcentagem.

Tabela 3 - Análise da correlação entre as métricas e a relevância dos conceitos.

| Métrica | Classe (Quantidade absoluta e porcentagem) | | | |
|---|---|---|---|---|
| | **Sim** | **Não** | **Sim** | **Não** |
| *Simple Frequency* | 6/13 | 7/24 | 46% | 29% |
| *Cumulative Frequency* | 5/13 | 6/24 | 38% | 25% |
| *Centrality* | 1/13 | 3/24 | 7,6% | 12,5% |
| *Closeness* | 8/13 | 7/24 | 61,5% | 29% |
| *Level* | 10/13 | 15/24 | 77% | 62,5% |

Quanto à Tabela 3, observa-se que:

a) As medidas *Simple Frequency* e *Cumulative Frequency* parecem capturar a relevância, pois se destacam em quase metade dos conceitos da categoria “sim” (46% e 38%) e em aproximadamente um quarto (29% e 25%) dos da classe “não”; assim, os conceitos constitutivos do sumário humano de referência parecem ser os que mais se repetem nos textos-fonte;

b) A métrica *Centrality* parece não distinguir os conceitos relevantes dos demais, pois pouco se destaca nas classes “sim” e “não”, isto é, 7,6% e 12,5%, respectivamente; isso pode ser explicado pelo fato de que os 37 conceitos dos textos-fonte estão

conectados na grande maioria das vezes unicamente ao seu hiperônimo, possuindo, assim, apenas 1 relacionamento na hierarquia;

c) A medida *Closeness* aparentemente se caracteriza como um bom indicativo de relevância, já que 61,5% dos conceitos da classe "sim" possuem valor igual ou superior à média, enquanto isso ocorre em apenas 29% dos casos da classe "não"; esse resultado pode indicar que os conceitos do sumário são semanticamente relacionados entre si;

d) A medida *Level* também parece capturar, segundo a análise manual, a relevância, pois 77% dos conceitos da classe "sim" estão localizados em posições ou níveis intermediários na hierarquia ou árvore do *cluster* C1; tal atributo, no entanto, também é relativamente significativo nos casos da classe "não" (62,5%).

Na sequência, descreve-se a análise automática dos dados da Tabela 2.

#### *5.3.2. Análise automática*

Como mencionado, a análise automática da pertinência das medidas foi feita por meio de algoritmos de AM. Especificamente, utilizou-se o paradigma supervisionado de AM, amplamente utilizado no PLN. Os algoritmos supervisionados adquirem conhecimento implícito de exemplos previamente classificados, aos quais estão associados os atributos (no caso, as medidas de relevância), e geram classificadores (p.ex.: conjunto de regras) que relacionam os atributos/valores às classes. A precisão de um classificador para predizer as classes depende diretamente de quanto os atributos são capazes de representar o conhecimento implícito nos exemplos.

Apesar de se ter uma quantidade pequena de dados (isto é, 37 casos (conceitos)), optou-se por essa análise automática porque os algoritmos de AM são capazes de analisar os dados de forma a correlacionar as medidas. Dessa forma, os algoritmos de AM analisam os dados de forma complementar à investigação manual aqui realizada, que só verificou a pertinência das medidas isoladamente. Em outras palavras, o AM permite identificar padrões estatisticamente relevantes que englobam correlações entre os atributos, o que é bastante complexo de se observar manualmente.

a) *Pré-processamento*

Para a aplicação dos algoritmos de AM, os valores absolutos das medidas da Tabela 2 foram normalizados. A normalização ajusta todos os valores das tabelas entre 0 e 1, ou seja, coloca-os em uma escada comum, eliminando valores muito discrepantes que podem influenciar nos resultados. Neste trabalho, a normalização foi realizada

dividindo-se o valor da medida de cada conceito pelo maior valor da medida dentre os 37 conceitos. Por exemplo, a normalização da *Simple Frequency* do conceito subjacente a "avião" foi feita pela divisão do valor de *Simple Frequency* do conceito em questão, 11, pelo maior valor obtido pela medida dentre os 37 conceitos, no caso, 11. Essa divisão resultou na *Simple Frequency* com valor normalizado 1 para "avião". Já a *Simple Frequency* normalizada de "carga" é 0,181818182, pois esta resulta da divisão de 2 por 11. Como resultado do processo de normalização, todas as medidas passaram a ter valores entre 0 e 1.

Ressalta-se também que, devido ao desbalanceamento dos dados das classes "sim" e "não", as quais são compostas respectivamente por 13 e 24 casos, aplicou-se ao conjunto de 37 conceitos, cujas medidas foram normalizadas, a técnica *oversampling* (em português, sobreamostragem). Por meio dessa técnica, o balanceamento dos dados é é feito pela replicação dos casos da classe minoritária. Com isso, evitam-se discrepâncias entre valores do conjunto inicial de dados, aumentando a acurácia dos algoritmos de AM. Assim, o conjunto de dados submetidos ao AM passou a ser composto por 26 casos da classe "sim" e 24 da classe "não", totalizando 50 conceitos. Na Tabela 4, tem-se o conjunto final de dados, com *oversampling* e normalização.

Tabela 4 - Valores normalizados das medidas dos conceitos de C1 com oversampling.

| Qt. | Nome/ conceito | Ocorrência no sumário (relevância) | Métrica normalizada | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Simple Frequency* | *Cumulative Frequency* | *Centrality* | *Closeness* | *Level* |
| 1 | avião | Sim | 1 | 0,47826087 | 0,200000 | 0,934052258 | 0,125 |
| 2 | carga | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,875258919 | 0,75 |
| 3 | floresta | Sim | 0,272727273 | 0,130434783 | 0,200000 | 0,789452342 | 0,75 |
| 4 | membro | Sim | 0,363636364 | 0,173913043 | 0,200000 | 0,847043349 | 0,5 |
| 5 | mineral | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,893897969 | 1 |
| 6 | montanha | Sim | 0,090909091 | 0,043478261 | 0,200000 | 0,876085787 | 0,875 |
| 7 | nacionalidade | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,776745386 | 0,75 |
| 8 | passageiro | Sim | 0,454545455 | 0,217391304 | 0,200000 | 0,886357139 | 0,625 |
| 9 | pessoa | Sim | 0,272727273 | 1 | 1,000000 | 1 | 0,875 |
| 10 | queda | Sim | 0,090909091 | 0,260869565 | 0,200000 | 0,771228135 | 0,25 |
| 11 | tempo | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,807960251 | 0,75 |
| 12 | tripulação | Sim | 0,363636364 | 0,217391304 | 0,200000 | 0,788619319 | 0,625 |
| 13 | vítima | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,859696692 | 0,625 |
| 14 | avião | Sim | 1 | 0,47826087 | 0,200000 | 0,934052258 | 0,125 |
| 15 | carga | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,875258919 | 0,75 |
| 16 | floresta | Sim | 0,272727273 | 0,130434783 | 0,200000 | 0,789452342 | 0,75 |
| 17 | membro | Sim | 0,363636364 | 0,173913043 | 0,200000 | 0,847043349 | 0,5 |
| 18 | mineral | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,893897969 | 1 |
| 19 | montanha | Sim | 0,090909091 | 0,043478261 | 0,200000 | 0,876085787 | 0,875 |
| 20 | nacionalidade | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,776745386 | 0,75 |
| 21 | passageiro | Sim | 0,454545455 | 0,217391304 | 0,200000 | 0,886357139 | 0,625 |
| 22 | pessoa | Sim | 0,272727273 | 1 | 1,000000 | 1 | 0,875 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 23 | queda | Sim | 0,090909091 | 0,260869565 | 0,200000 | 0,771228135 | 0,25 |
| 24 | tempo | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,807960251 | 0,75 |
| 25 | tripulação | Sim | 0,363636364 | 0,217391304 | 0,200000 | 0,788619319 | 0,625 |
| 26 | vítima | Sim | 0,181818182 | 0,086956522 | 0,200000 | 0,859696692 | 0,625 |
| 27 | acidente | Não | 0,454545455 | 0,260869565 | 0,400000 | 0,774127207 | 0,375 |
| 28 | aeronave | Não | 0,090909091 | 0,52173913 | 0,400000 | 0,960651741 | 0,375 |
| 29 | aeroporto | Não | 0,363636364 | 0,173913043 | 0,200000 | 0,858266296 | 0,625 |
| 30 | aterrissagem | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,722583968 | 0,5 |
| 31 | chama | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,742149824 | 0,5 |
| 32 | cidade | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,759322914 | 0,5 |
| 33 | companhia | Não | 0,363636364 | 0,173913043 | 0,200000 | 0,752786554 | 0,375 |
| 34 | distância | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,746645242 | 0,625 |
| 35 | estrada | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,867692208 | 0,75 |
| 36 | fabricação | Não | 0,272727273 | 0,130434783 | 0,200000 | 0,721653233 | 0,375 |
| 37 | fonte | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,753087539 | 0,625 |
| 38 | leste | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,728802787 | 0,5 |
| 39 | localidade | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,801540285 | 0,625 |
| 40 | país | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,797756975 | 0,625 |
| 41 | permissão | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,677934886 | 0,375 |
| 42 | pista | Não | 0,272727273 | 0,130434783 | 0,200000 | 0,847782287 | 0,625 |
| 43 | porta-voz | Não | 0,636363636 | 0,304347826 | 0,200000 | 0,902855942 | 0,625 |
| 44 | propriedade | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,814938474 | 0,75 |
| 45 | quilômetro | Não | 0,272727273 | 0,130434783 | 0,200000 | 0,763594976 | 0,625 |
| 46 | setor | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,747029368 | 0,625 |
| 47 | sobrevivente | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,859696692 | 0,625 |
| 48 | tarde | Não | 0,181818182 | 0,086956522 | 0,200000 | 0,750950765 | 0,625 |
| 49 | tempestade | Não | 0,090909091 | 0,043478261 | 0,200000 | 0,802491464 | 0,75 |
| 50 | transporte | Não | 0,090909091 | 0,565217391 | 0,400000 | 0,979920812 | 0,75 |

Após a normalização e *oversampling*, os dados referentes aos 50 conceitos foram "discretizados", ou seja, divididos em intervalos iguais para facilitar os cálculos, por meio do ambiente de AM denominado *Weka* (do inglês, *Waikato Environment for Knowledge Analysis*) (HALL *et al.*, 2009). O *Weka,* desenvolvido pela Univesidade de Waikato (Nova Zelândia), é um software de domínio público que consiste em um conjunto de algoritmos de AM. Seguindo-se as opções padrão do *Weka*, os atributos numéricos foram discretizados em um pequeno número de intervalos distintos, os quais são mais adequadamente manipulados com alguns algoritmos. Em seguida, os valores dos atributos foram transformados de numéricos para nominais (isto é, textuais) também por meio da opção padrão do *Weka*. A transformação dos atributos numéricos para nominais *(NumericToNominal)* é necessária pois alguns algoritmos de AM manipulam os dados apenas com atributos nominais como por exemplo algumas implementações do *Naïve* e do *Trees* (cf. pág. 76).

Por fim, para a correta manipulação dos dados, a classe "conceitos" foi removida para que os algoritmos baseados em regras, por exemplo, não criassem regras baseadas

na nomeação dos conceitos e, sim, apenas nas classes numéricas apoiando-se apenas na categoria nominal “sim/não”.

Assim, após a normalização e *oversampling* dos dados iniciais, ou seja, dos 37 conceitos, e da discretização, transformação *NumericToNominal* e exclusão de classe específica (conceitos) dos dados finais (50 conceitos), procedeu-se aos testes das medidas via AM.

b) *Aplicação de diferentes algoritmos: opção Use Training Set*

Para a investigação da pertinência das medidas via algoritmos de AM do *Weka,* utilizou-se primeiramente a opção *Use Training Set,* uma vez que o conjunto de dados é pequeno. Segundo essa opção, os algoritmos utilizam o mesmo conjunto de dados (no caso, os 50 conceitos e suas medidas) para aprender e testar os classificadores. Esse cenário de aprendizado pode gerar resultados superestimados. A precisão das generalizações (que compõem os classificadores), quando testadas no mesmo conjunto de dados do qual foram aprendidas, tende a ser alta, pois elas cobrem os padrões mais recorrentes do conjunto em questão. Assim, para que se tenha uma precisão mais realista sobre classificadores, deve-se utilizar parcelas de dados distintos para aprendizado e teste, como é o caso da validação cruzada (cf. pág. 81).

Ademais, ressalta-se que esta investigação deu ênfase aos algoritmos do paradigma simbólico, pois, além de buscar o classificador de melhor acurácia, buscou-se também explicitar quais medidas capturam a relevância dos conceitos e como isso ocorre. Apesar disso, vários algoritmos dos demais paradigmas (isto é, conexionista, matemático/probabilístico e estatístico) também foram testados para comparação. Na sequência, comentar-se-ão detalhadamente os resultados dos algoritmos simbólicos e de forma mais breve os resultados dos algoritmos dos demais paradigmas.

Dentre os algoritmos conexionistas, utilizou-se o conhecido *Multi-Layer Perceptron* (MLP), com as configurações padrão do *Weka*. No caso, o MLP atingiu a acurácia geral de 98%. Dentre os probabilísticos, foram aplicados os algoritmos Naïve-Bayes e Redes Bayesianas, os quais atingiram 84% e 88% de acurácia, respectivamente. Quanto ao paradigma estatístico, o algoritmo aplicado foi o *Sequential Minimal Optimization* (SMO), que identificou as instâncias das classes “sim” e “não” com 90% de acurácia. Como é possível perceber, a acurácia média dos algoritmos dos paradigmas conexionista, matemático/probabilístico e estatístico foi alta, girando em torno de 90%,

ou seja, os algoritmos deste paradigma classificaram corretamente a maioria dos casos "Sim" e "Não" para os conceitos de C1.

Quanto aos algoritmos do paradigma simbólico, cujos resultados são legíveis por máquinas e humanos, foram testados vários. Dentre os baseados em regras, aplicaram-se os algoritmos básicos *One Rule* (OneR), JRip, e PART, dentre os que representam o conhecimento aprendido na forma de árvores de decisão, testaram-se *Iterative Dichotomiser 3* (ID3) e J48 que é uma extensão do ID3.

O algoritmo ID3 tem como característica priorizar árvores pequenas em detrimento de árvores grandes por meio de generalizações sobre os exemplos de aprendizado. No caso, a árvore de decisão gerada pelo ID3 a partir dos 50 conceitos caracterizados em função das 5 medidas de relevância é composta por um número muito extenso de regras, as quais, no entanto, identificam corretamente as instâncias com 98% de acurácia. O algoritmo J48, que é uma extensão do ID3, gerou uma árvore com menos regras, isto é, 24, mas obteve apenas 86% de acurácia geral, ou seja, o algoritmo classificou erroneamente 14% dos casos.

No paradigma simbólico, o OneR é possivelmente o mais simples, pois ele aposta na hipótese de que basta utilizar apenas um dos atributos para classificar corretamente os exemplos de teste. Assim, esse algoritmo tem a tarefa de encontrar durante o treinamento o atributo que apresenta a menor taxa de erro de classificação. Neste trabalho, OneR gerou 6 regras pautadas na medida *Closeness* (normalizada), as quais obtiveram 78% de acurácia. Tal taxa de acerto é, aliás, surpreendentemente boa, tendo em vista a simplicidade do algoritmo. Outro algoritmo testado foi o JRip, que trabalha com aprendizado de regras na tentativa de minizar erros (*Repeated Incremental Pruning to Produce Error Reduction - RIPPER*). O JRip gerou um conjunto reduzido de apenas 3 regras, pautadas nas medidas *Closeness* e *Level.* Tais regras, no entanto, obtiveram acurácia mais baixa que o OneR, isto é, 74%. O algoritmo PRISM obteve a acurácia mais elevada dentre os baseados em regras, 98%. Entretanto, esse algoritmo gerou um conjunto muito extenso de regras.

O algoritmo PART, que constrói uma árvore de decisão e gera uma lista à parte com a seleção das melhores regras, por sua vez, foi o que gerou um conjunto relativamente pequeno de regras com uma acurácia elevada, no caso, de 90% (isto é, o algoritmo classificou corretamente 45 do total de 50 conceitos). Essa combinação (conjunto de regras gerenciável e mais alta precisão entre as abordagens simbólicas) faz do PART uma boa escolha para os propósitos do trabalho. Na Tabela 5, descrevem-se

as 10 regras sequenciais do PART, que são seguidas pelo número de instâncias (conceitos) classificadas correta e incorretamente, e a precisão das regras, dada pelo número de instâncias corretamente classificadas sobre todos os casos classificados pela regra.

Tabela 5 - Regras do PART com Use Training Set.

| Regra | Acerto/Erro | Precisão |
|---|---|---|
| 1. Se *Closeness* = 0.869794-0.902346, então "sim" | (9.0/1.0) | 90% |
| 2. Senão *Closeness* = 0.739588-0.77214, então "não" | (8.0/0.0) | 100% |
| 3. Senão *Level* = 0.375, então "não" | (3.0/0.0) | 100% |
| 4. Senão *Cumulative Frequency* = inf-0.13913 e *Level* = 0.625, então "não" | (6.0/2.0) | 75,5% |
| 5. Senão *Simple Frequency* = 0.363636-0.454545, então "sim" | (5.0/1.0) | 83.3% |
| 6. Senão *Level* = 0.75 e *Simple Frequency* = 0.181818-0.272727, então "sim" | (5.0/1.0) | 83.3% |
| 7. Senão *Simple Frequency* = 0.272727-0.363636, então "sim" | (4.0/0.0) | 100% |
| 8. Senão *Level* = 0.5, então "não" | (3.0/0.0) | 100% |
| 9. Senão *Level* = 0.75, então "não" | (3.0/0.0) | 100% |
| 10. Senão, "sim" | (4.0/0.0) | 100% |

Para cada conceito, aplicam-se as regras da Tabela 5. Assim, dado um conceito, verifica-se se este satisfaz a condição da Regra 1, que determina *Closeness* com valores entre 0.869794 e 0.902346. Caso o conceito satisfaça essa condição, este é relevante e deve integrar o sumário, o que é indicado pela classe "sim". Caso contrário, aplica-se a Regra 2, segundo a qual o conceito que possui *Closeness* com valor entre 0.739588 e 0.77214 não deve ir para o sumário, que é indicado pela classe "não". Uma vez não satisfeita a Regra 2, aplica-se a Regra 3, segundo a qual um conceito com *Level* igual a 0.375, também não deve ir para o sumário. E assim é seguida a ordem de aplicação das regras até que uma delas seja satisfeita. Caso se atinja a última, a Regra 10, o conceito que não satisfizer nenhuma das 9 regras anteriores, deve compor o sumário ("sim"). E, assim, determina-se a classe (relevante ou não relevate) de cada um dos conceitos.

Ainda na Tabela 5, observa-se que as regras geradas pelo PART se baseiam em 4 das 5 medidas testadas, isto é, *Closeness, Level, Cumulative Frequency* e *Simple Frequency.* Isso quer dizer que, segundo o algoritmo, essas 4 medidas caracterizam os conceitos relevantes ("sim"), ou seja, aqueles que ocorrem no sumário de referência. Das medidas testadas, somente *Centrality* não compõe o conjunto de regras do PART, o que parece corroborar a análise manual da relevância dessa medida.

Dentre as medidas, ressalta-se que as Regras 1 e 2, baseadas exclusivamente em *Closeness,* e a Regra 3, baseada exclusivamente em *Level,* são relativamente

abrangentes, já que juntas englobam 19 do total de 50 conceitos, e bastante precisas, pois a Regra 1 possui precisão de 90% e as Regras 2 e 3 possuem precisão de 100%. A aplicação das Regras 7, 8, 9 e 10 também obteve 100% de precisão, mas essas regras englobam 14 casos do total de 50. Por fim, as Regras 4, 5 e 6 também são abrangentes, totalizando 12 casos, e possuem precisão média, isto é, 75,5%, 83,3% e 83,3%, respectivamente.

Na Tabela 6, a matriz de confusão do PART oferece uma medida efetiva do modelo de classificação ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos. Nessa matriz, observa-se que os acertos e os erros estão relativamente equilibrados entre as classes.

Tabela 6 - Matriz de confusão do PART com *Use Training Set*.

| **Classe** / **Teste** | **Sim (26)** (Ocorrência no sumário) | **Não (24)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 24 (92,3%) | 2 (7,7%) |
| Não | 3 (12,5%) | 21 (87,5%) |

Além de investigar a pertinência das 5 medidas por meio do treinamento e teste de vários algoritmos no mesmo conjunto de dados, testou-se apenas o algoritmo PART, considerado o mais adequado, em função de diferentes opções do *Weka*.

c) *Aplicação do algoritmo PART: opções Use Training Set e Seleção de atributos*

A primeira investigação somente com o PART foi feita a partir da opção *Use Training Set*, já mencionada, considerando apenas as medidas mais discriminativas. Para tanto, realizou-se o processo denominado "seleção de atributos", que é uma técnica de AM amplamente utilizada. Especificamente, a seleção de atributos visa reduzir o conjunto inicial de atributos, eliminando os que são redundantes e pouco discriminativos. Uma das razões para a realização desse processo é o fato de que, diante de um número menor de atributos, as classes a serem induzidas (ou seja, a relevância ou não dos conceitos) pelos classificadores podem ser mais adequadamente compreendidas.

Assim, aplicou-se a técnica de seleção de atributos *InfoGainAttributeEval*, que avalia cada atributo em função do ganho de informação ao se determinar as classes, construindo um ranque. Como resultado, 3 atributos foram efetivamente identificados como discriminativos e assim ordenados: *Closeness* (1º), *Level* (2º) e *Cumulative Frequency* (3º). Dessa forma, destaca-se a ausência da medida *Centrality*, o que mais

uma vez confirma as observações resultantes da análise manual. Ademais, nota-se a ausência da medida *Simple Frequency*, o que pode ser justificado pelo fato de que a frequência isoladamente não é suficiente para distinguir os casos "sim" e "não". Aliás, a presença de medida *Cumulative Frequency* dentre as selecionadas como mais relevantes indica que a frequência acumulada em combinação com nível de especialização/ generalização do conceito é um bom fator discriminativo.

Utilizando-se apenas as medidas relevantes e a opção *Use Training Set*, investigou-se então o desempenho do algoritmo PART. Na Tabela 7, apresentam-se as regras geradas pelo PART somente com base nas medidas *Closeness*, *Level* e *Cumulative Frequency*. No caso, esse conjunto de regras da Tabela 7 obteve acurácia de 84%, que é mais baixa que a obtida com todas as 5 medidas (90%).

Tabela 7 - Regras do PART com *Use Training Set* e Seleção de atributos.

| Regra | Acerto/Erro | Precisão |
|---|---|---|
| 1. Se *Closeness* = 0.869794-0.902346, então "sim" | (9.0/1.0) | 90% |
| 2. Senão *Closeness* = 0.739588-0.77214, então "não" | (8.0/0.0) | 100% |
| 3. Senão *Level* = 0.375, então "não" | (3.0/0.0) | 100% |
| 4. Senão *Cumulative Frequency* = -inf-0.13913 e *Level* = 0.75, então "sim" | (9.0/3.0) | 75% |
| 5. Senão *Cumulative Frequency* = -inf-0.13913 e *Level* = 0.625, então "não" | (6.0/2.0) | 75% |
| 6. Senão *Level* = 0.5 e *Closeness* = 0.707037-0.739588, então "não" | (3.0/0.0) | 100% |
| 7. Senão "sim" | (12.0/2.0) | 85,7% |

Na matriz de atributos da Tabela 8, observa-se que o algoritmo classifica erroneamente 6 casos da classe "não" como sendo da classe "sim" (isto é, 25% do total de 24 exemplos), ao passo que somente 2 casos da classe "sim" foram classificados erroneamente como "não" (isto é, 7,7% do total de 26 exemplos). Ademais, destaca-se a maior porcentagem no acerto da classe "sim" (92,3% do total de 26 conceitos da classe).

Tabela 8 - Matriz de confusão do PART com *Use Training Set* e Seleção de atributos.

| **Classe** / **Teste** | **Sim (26)** (Ocorrência no sumário) | **Não (24)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 24 (92,3%) | 2 (7,7%) |
| Não | 6 (25%) | 18 (75%) |

*d) Aplicação do algoritmo PART: 10-fold cross-validation*

A segunda investigação somente o PART consistiu em aplicar a técnica *10-fold cross validation*[54]. Ao contrário da *Use Training Set*, a técnica *10-fold cross-validation* (em português, "validação cruzada de 10 pastas") faz com que as taxas de acerto e erro dos algoritmos sejam mais próximas dos valores obtidos para a tarefa em uma situação real.

Ressalta-se que, nesse caso, o treinamento e teste do algoritmo foram feitos com base em todos os 5 atributos ou medidas, ou seja, *Centrality, Closeness, Level* e *Simple Frequency* e *Cumulative Frequency*. A acurácia geral obtida pelo algoritmo PART com a técnica *10-fold cross-validation* foi de 56% (isto é, o algoritmo classificou corretamente 28 dos 50 conceitos). Como esperado, essa precisão é menor que a obtida quando todo o conjunto de dados é usado para treinamento e teste (opção *Use Training Set*) (90%). Especificamente, vale ressaltar que as regras geradas nesse teste são as mesmas 10 que compõem a Tabela 5, já que, apesar das 10 iterações, as regras via *10-fold cross-validation* são geradas a partir do conjunto completo de dados ou *corpus.* Assim, as regras geradas pelo PART nessa configuração englobam apenas as medidas *Closeness, Level, Cumulative Frequency* e *Simple Frequency*. A acurácia geral, no entanto, é a média da acurácia de todas as iterações.

Com base na matriz de confusão da Tabela 9, observa-se que o número absoluto de acertos e erros das regras do PART, assim como a respectiva porcentagem, está bastante equilibrada entre as classes.

Tabela 9 - Matriz de confusão do PART com *10-fold cross-validation*.

| **Classe** / **Teste** | **Sim (26)** (Ocorrência no sumário) | **Não (24)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 15 (57,7%) | 11 (42,3%) |
| Não | 11 (45,8%) | 13 (54,2%) |

*e) Aplicação do algoritmo PART: 10-fold cross-validation e Seleção de atributos*

A terceira investigação sobre as medidas de relevância utilizando apenas o PART visou aplicar a técnica *10-fold cross-validation* e utilizar somente medidas resultantes do processo de seleção de atributos, isto é, *Closeness, Level* e *Cumulative Frequency.* No caso, o PART gerou as mesmas 7 regras que compõem a Tabela 7, mas com acurácia

[54] O método de validação cruzada denominado *10-fold cross validation* consiste em dividir o conjunto total de dados em 10 subconjuntos mutuamente exclusivos do mesmo tamanho. A partir deles, 1 subconjunto é utilizado para teste e os 10-1 restantes são utilizados para treinamento. Esse processo é realizado 10 vezes, alternando-se o subconjunto de teste a cada iteração. Ao final das 10 iterações, a acurácia média é calculada.

geral de 50%. Com base na matriz de confusão da Tabela 10, observa-se que o algoritmo classificou corretamente mais casos da classe "não" (13) (isto é, 54,2% do total de 24) que casos da classe "sim" (12) (ou seja, 46,2% do total de 26 conceitos).

Tabela 10 - Matriz de confusão do PART com *10-fold cross-validation* e Seleção de atributos.

| **Classe** / **Teste** | **Sim (26)** (Ocorrência no sumário) | **Não (24)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | **12 (46,2%)** | 14 (53,8%) |
| Não | 11 (45,8%) | **13 (54,2%)** |

Resumindo, 4 testes foram realizados com diferentes configurações do *Weka*, a saber:

(a) treinamento e teste no conjunto total de dados (opção *User training set*) em função de todos os 5 atributos (ou medidas);
(b) treinamento e teste no conjunto total de dados (opção *User training set*) com seleção de atributos (isto é, somente com base em *Closeness*, *Level* e *Cumulative Frequency*);
(c) *10-fold cross-validation* em função dos 5 atributos;
(d) *10-fold cross-validation* com seleção de atributos.

Considerando-se os resultados obtidos em todos os testes para a coleção C1, a medida *Centrality* não capturou a distinção entre as classes "sim" e "não", ou seja, entre os conceitos relevantes (presentes no sumário) e os não-relevantes (não presentes no sumário). Isso pode ser justificado pela natureza da modelagem hierárquica dos textos-fonte, na qual os 37 conceitos que ocorrem nos textos-fonte têm na grande maioria das vezes apenas 1 aresta. Essa característica decorre do fato de que a maioria desses conceitos está no nível mais profundo da hierarquia (12), sendo, no sentido *bottom-up*, iniciadora do próprio ramo da árvore, o que acarreta ter apenas um hiperônimo direto ou aresta.

Inicialmente, pode-se dizer que os testes sobre a pertinência das medidas permitiram excluir a medida *Centrality*, tendo em vista a modelagem adotada. Provavelmente, a medida *Centrality* teria outro comportamento em outro tipo de grafo. Por conseguinte, para a proposição dos métodos de SAM baseados na modelagem hierárquica dos conceitos nominais dos textos-fonte e exploração de medidas de grafo, consideraram-se as medidas mais proeminentes nos testes, isto é, *Closeness*, *Level*, *Cumulative Frequency* e, *Simple Frequency*.

Seguindo a metodologia prevista para este trabalho, as etapas subsequentes à investigação da pertinência das medidas de grafo na SAM foram: (i) a proposição de métodos de SAM e geração de extratos e (ii) a avaliação dos métodos. Tais etapas estão descritas abaixo.

# 6. Aprendizado de critérios de relevância com base em hierarquias conceituais

Essa etapa consistiu em aprender critérios para a seleção de conteúdo, os quais se baseiam nas métricas de grafo identificadas como as mais pertinentes para a detecção dos conceitos relevantes de uma coleção multidocumento. Tais métodos se caracterizam pela representação dos textos-fonte em uma estrutura léxico-conceitual, no caso, hierárquica (ou árvore), e pela pontuação e ranqueamento das sentenças em função dos valores das medidas ou métricas dos seus respectivos conceitos na hierarquia.

Para tanto, o *corpus* utilizado para a investigação da pertinência das medidas foi estendido. Devido à complexidade da tarefa de anotação/indexação, no entanto, tal extensão consistiu na anotação de apenas mais dois *clusters* do CSTNews. Tais *clusters* abordam assuntos distintos e possuem extensão (ou tamanho) reduzida. No caso, os *clusters* selecionadas foram o (i) C31, da categoria "esportes", que possui 2 textos-fonte, totalizando 10 sentenças e 217 palavras, e o C37, da categoria "cotidiano", que também possui 2 textos-fonte, em um total de 27 sentenças e 475 palavras.

A seguir, descreve-se o processo de representação conceitual dos textos dos *clusters* C31 e C37 e as ferramentas utilizadas.

## 6.1. Representação conceitual do *corpus*: método e ferramenta

Tendo em vista a experiência adquirida com a tarefa de investigação da pertinência das métricas, é notório que o método manual de anotação (indexação léxico-conceitual e geração da hierarquia) tende a ser mais preciso, já que não envolve possíveis erros causados pela automação das tarefas. No entanto, sabe-se também que o método manual é bastante custoso, principalmente devido ao tempo despendido na indexação dos nomes à WN.Pr e na geração da árvore conceitual (mesmo via o editor *CMap*). Assim, optou-se por pré-processar os 2 novos *clusters* de forma semiautomática.

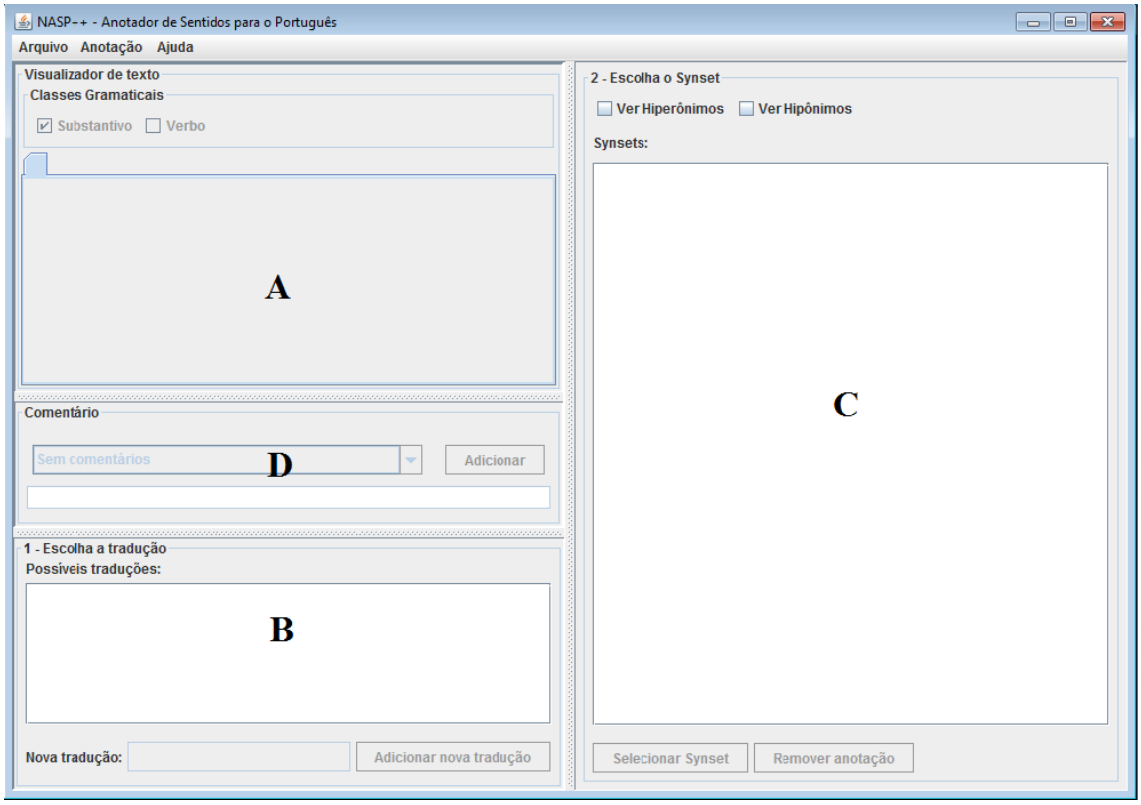
Especificamente, utilizou-se o editor NASP++ (CABEZUDO, 2015), que automatiza (i) a anotação semântica de nomes e verbos de uma coleção multidocumento via indexação dos mesmos ao *synsets* da WN.Pr e (ii) a subsequente construção de uma ontologia dos conceitos/*synsets* indexados.

Para a anotação/indexação dos nomes à WN.Pr e geração da ontologia, o editor realiza o pré-processamento dos textos-fonte, que engloba os processos de (i) tokenização (isto é, delimitação de *tokens*, que comumente são sequências de caracteres

separados por espaços em branco), (ii) etiquetação morfossintática[55] (isto é, identificação das categorias das palavras, como nome, verbo, adjetivo, etc.) e (iii) lematização[56] (isto é, transformação das palavras à sua forma canônica ou básica).

Feito isso, o NASP++ automatiza a anotação semântica por meio das seguintes etapas: (i) seleção do nome *x* a ser anotado; (ii) tradução de *x* para o inglês, o que é feito pelo acesso ao dicionário bilíngue WordReference®[57]; (iii) recuperação dos *synsets* dos quais *x* é elemento constitutivo; (iv) seleção do *synset* que mais adequadamente representa o conceito subjacente a *x*, e (v) anotação de *x* com o *synset* escolhido em (iv). Na Figura 29, tem-se a tela principal do NASP++, composta por vários campos nas quais as etapas (i-v) são realizadas.

Figura 29 - Tela principal do NASP++.



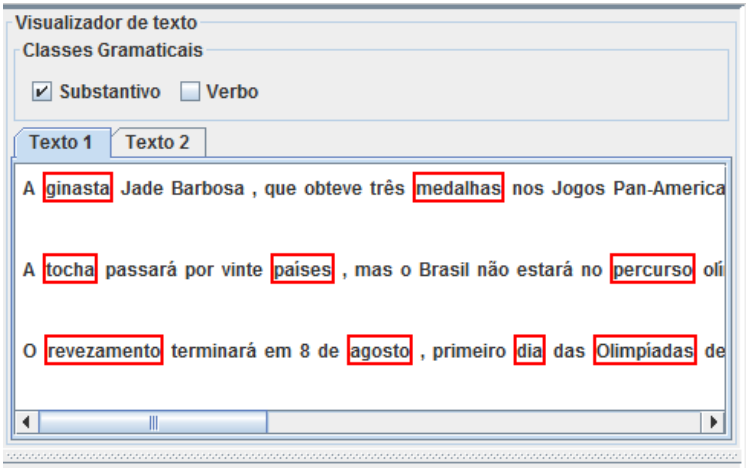

Fonte: autoria própria.

Na Figura 30, tem-se, como ilustração, os dois textos que compõem a coleção C31. Os textos são exibidos e anotados pelos anotadores humanos por meio do campo "visualizador de textos-fonte" (A). Nesse mesmo campo, a ferramenta permite

[55] O NASP utiliza o etiquetador MXPost (RATNAPARKHI, 1996).
[56] Disponível em http://www.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar
[57] http://www.wordreference.com/

especificar a classe das palavras a serem anotadas (substantivo ou verbo); no caso, selecionou-se "substantivo". Nos textos-fonte exibidos, as palavras destacadas em vermelho foram automaticamente identificadas como "nome" (ou substantivo) pelo etiquetador morfossintático MXPost, ou seja, quando a seleção da caixa "substantivo" é feita, a ferramenta destaca em vermelho todas as palavras identificadas como substantivos pelo etiquetador. A partir das palavras em destaque, tem-se início o processo de anotação/indexação. Por exemplo, a anotação do Texto 1 da Figura 30, teve início com o primeiro nome em destaque, no caso, "ginasta".

Figura 30 - Visualizador de textos.

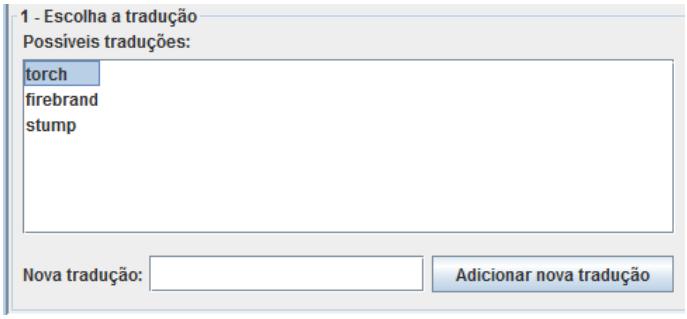



Fonte: autoria própria.

Ainda no campo "visualizador de textos-fonte", ao se clicar na palavra a ser anotada, o editor ativa o campo "comentários" (D)[58] e recupera automaticamente, a partir do acesso ao dicionário online WordReference®, as possíveis traduções em inglês para a palavra em questão, exibindo-as no campo "escolha a tradução" (B). Em outras palavras, o campo "escolha a tradução" permite visualizar todas as possíveis traduções para a palavra em questão e selecionar a mais pertinente. No caso de "tocha", o editor recuperou três equivalentes de tradução, "*torch*", "*firebrand*" e "*stump*" (Figura 31).

[58] O campo "comentários" possibilita adicionar comentários às anotações escolhendo-se uma das opções disponíveis, a saber: (i) sem comentários; (ii) não é verbo, erro de anotação ; (iii) é predicado complexo; (iv) é verbo auxiliar; e (v) outros. As opções (ii), (iii) e (iv) são exclusivas para anotação de sentidos de verbos; a opção (i) é aplicável quando não há observações sobre a anotação; e a opção (v) é aplicável quando existem outros tipos de observação sobre o processo de anotação de uma palavra, incluindo dificuldades de anotação.

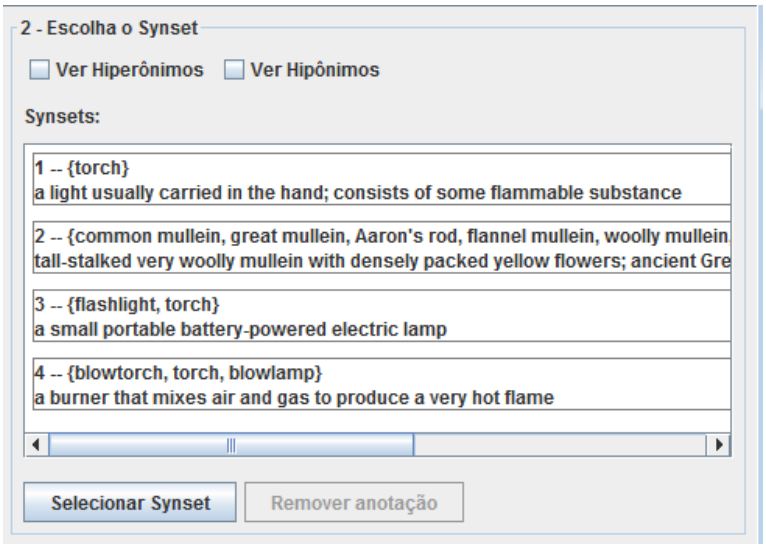
Figura 31 - Tela com “lista de traduções possíveis”




Fonte: autoria própria.

Escolhida a tradução “torch” como a mais adequada, o editor NASP++ recuperou automaticamente todos os *synsets* da WN.Pr que possuem esse nome como um de seus elementos constitutivos. Na Figura 32, observa-se que o editor recuperou 4 *synsets* e suas respectivas glosas e eventuais frases-exemplos. As glosas e as frases-exemplo auxiliam a identificaçãoo do *synset* que melhor codifica o conceito expresso pelo nome original em português.

Figura 32 - Tela de seleção do *synset*.




Fonte: autoria própria.

Dentre os *synsets* recuperados, cabe ao anotador escolher ou selecionar o que mais adequadamente representa o conceito subjacente ao nome "torch" ("tocha"). Caso os *synsets* constituídos pelo equivalente de tradução ("torch"), as glosas e as frases-exemplo não sejam suficientes para se definir a representação mais adequada do conceito do nome em português, o editor oferece a visualização dos hiperônimos e hipônimos dos *synsets* inicialmente recuperados. Essa funcionalidade corresponde aos botões "ver hiperônimos" e "ver hipônimos" do painel C da tela principal.

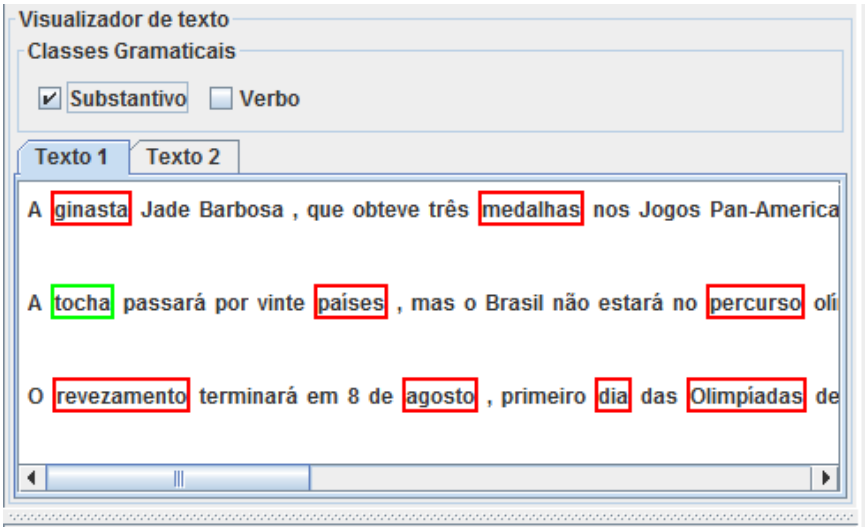
No caso de "*torch*", o primeiro *synset* exibido ao anotador foi selecionado, ou seja, {torch}, cuja glosa é "*a light usually carried in the hand; consists of some flammable substance*"[59], foi considerado adequado para rotular conceitualmente o nome "tocha" na coleção C31. Para selecionar, basta clicar no *synset* em questão e, na sequência, no botão "Selecionar *synset*" (C). O clique no botão "Selecionar *synset*" exibe uma janela de confirmação, cuja opção "sim" finaliza a anotação. Diante de dúvidas, o anotador pode clicar em "Não" e retornar à análise dos *synsets.*

Uma vez selecionado um *synset*, a palavra sob anotação (p.ex.: "tocha") é destacada no campo "visualizador de textos" em "verde", como ilustrado na Figura 33a. O destaque indica que à palavra foi associado um rótulo semântico, ou seja, um *synset.*

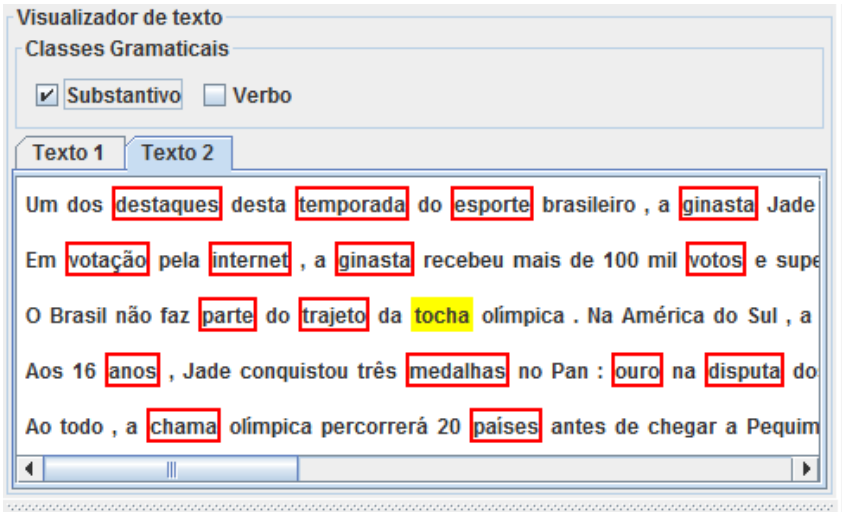
Partindo-se do pressuposto de que as diversas ocorrências de uma palavra em um texto (ou em textos que abordam mesmo assunto) tendem a ter um mesmo sentido, ressalta-se que, uma vez que uma palavra *x* tenha sido anotada com um sentido *y*, todas as demais ocorrências de *x* também são pré-anotados pelo editor com *y*. No NASP++, as demais ocorrências de *x* pré-anotadas com *y* são destacadas de "amarelo". Na Figura 33b, por exemplo, vê-se que outra ocorrência de "tocha" foi pré-anotada com o *synset* selecionado para a anotação da primeira ocorrência de "tocha". Ressalta-se aqui que a pré-anotação semântica é realizada para todas as ocorrências do nome "tocha", independentemente de sua forma flexionada. Assim, caso ocorra o nome "tochas", este também será pré-anotado. Ao anotador humano, cabe a tarefa de verificar se, de fato, o sentido/*synset* pré-anotado é pertinente para as diferentes ocorrências.

[59] Tradução da glosa de {torch}.

Figura 33 - Anotação da 1ª ocorrência de “tocha” e pré-anotação das demais




(a)




(b)

Fonte: autoria própria.

Após a anotação de todos os nomes pertinentes de uma coleção, a ferramenta salva os textos-fonte anotados no formato XML (do inglês, *Extensible Markup Language*), um dos mais utilizados para a tarefa de anotação de *corpus*. Ademais, o NASP++ gera uma estrutura conceitual a partir dos *synsets* utilizados na anotação e das relações herdadas da WN.Pr.

Quanto à estrutura conceitual, especificamente, a cada conceito/*synset* *x* selecionado para a anotação de nome de um *cluster*, o editor obtém da WN.Pr: (i) os hipônimos imediatos de *x*, (ii) os co-hipônimos (isto é, *synsets* do mesmo nível de *x* e filhos do mesmo hiperônimo de *x*) de *x*, (iii) o hiperônimo imediato, os intermediários e o *top-concept* (ou seja, o hiperônimo mais genérico de *x* que inicia a hierarquia de que *x* parte no sentido *top-down*) de *x*. Em outras palavras, o NASP++ recupera toda a

hierarquia conceitual da qual o conceito/*synset x* é parte integrante, gerando uma árvore parcial interna.

Esse processo é repetido a cada conceito/*synset* distinto selecionado para anotar um nome em português. Ao final, as árvores parciais, referentes aos diferentes conceitos/*synsets* de um *cluster*, são unificadas em uma hierarquia final, que representa conceitualmente o conteúdo dos textos-fonte de um *cluster*. O arquivo que contém a hierarquia final é salvo automaticamente em formato XML pelo editor. Na Figura 34, apresenta-se a hierarquia do conceito "medalha" (do inglês, *medal*), que é representado na WN.Pr pelo *synset* {*decoration, laurel wreath*, *medal, medallion, palm, ribbon*}. Os *synsets* que pertencem ao mesmo nível de "medalha", como {*Prix Congourt*}, {*trophy*}, {...} e {*aliyah*}, são seus co-hiperônimos. O *synset* que pertence ao nível imediatamente superior a "medalha" (e a seus co-hipônimos), no caso, {*award, accolade, honor, honour, laurels, symbol*}, é chamado de hiperônimo imediato. Os *synsets* que compõem os demais níveis, desde {*symbol*} até {*entity*} são também hiperônimos de "medalha". Vale lembrar que os hiperônimos são itens lexicais que expressam conceitos mais genéricos ou amplos. Por fim, os *synsets* de nível inferior a "medalha", como {*Medal of Honor, Congressional Medal of Honor*}, {N*avy Cross*} e {*Air Medal*} são hipônimos, expressando conceitos mais específicos que (ou tipos de) "medalha".

Figura 34 - Parte da hierarquia conceitual de C31 viaNasp++


Fonte: autoria própria.

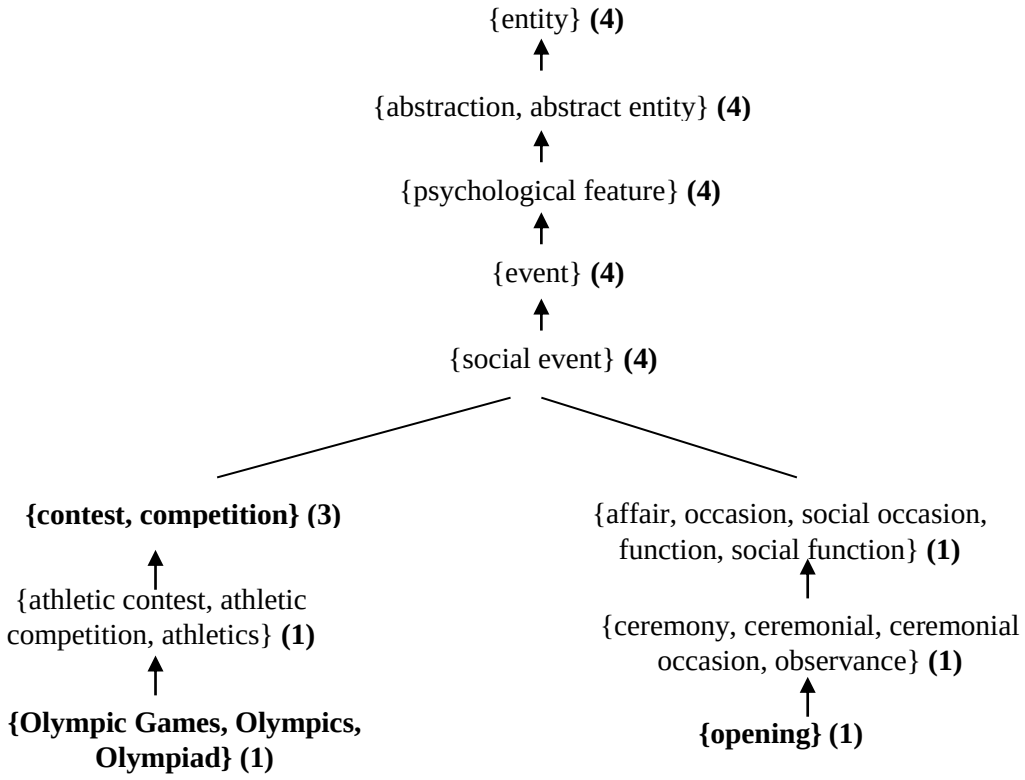
A ferramenta herda todos os hipônimos e co-hipônimos (conceitos de mesmo nível hierárquico) do conceito anotado. Com base na contagem da categoria *nSelect*, que expressa a *Simple Frequency*, identificaram-se os conceitos efetivamente ativados. Por exemplo, se o atributo *nSelect* de um conceito *x* tem valor "0", isso significa que este não ocorreu no texto, pois tal atributo expressa a frequência de anotação de um conceito. O atributo *countSub* funciona como um contador da hierarquia, pois é responsável por somar a frequência de todos os conceitos hipônimos, sendo que a frequência final é herdada pelos hiperônimos. Sendo assim, o conceito que está no topo terá o maior valor associado a *countSub* da hierarquia.

Ainda na Figura 34, nota-se que "*decoration, laurel wreath, medal, medallion, palm, ribbon"* é um conceito ativado com uma *Simple Frequency* de valor "2", o que é indicado pelo atributo-valor *countSub*="2". Esse par atributo-valor é herdado por todos os seus hiperônimos, por exemplo, *"award, accolade, honor, honour, laurels"*, que tem então uma *Cumulative Frequency* de valor 2.

Para evitar possíveis erros do editor, todos os nomes das coleções C31 e C37 foram contabilizados manualmente com base na observação dos dados inseridos via ferramenta *Cmap*.

Por exemplo, na Figura 35, os conceitos que estão em negrito são os que aconteceram nos textos e, ao lado deles, tem-se a frequência de ocorrência dos mesmos. Os conceitos restantes são os que foram herdados da WN.Pr. Sendo assim, conforme a hierarquia cresce e recebe os conceitos herdados, estes recebem também o peso de seus conceitos hipônimos. A distribuição dos pesos pode ser visualmente verificada na estrutura gráfica elaborada por meio da ferramenta *Cmap*.

Figura 35 - Contabilização das frequências via Cmap




Fonte: autoria própria.

## 6.2. Cálculo das métricas: métodos e ferramenta

A caracterização dos conceitos da hierarquia do *cluster* C1 em função das 5 métricas inicialmente selecionadas da literatura para serem analisadas no cenário da SAM foi feita majoritariamente de forma manual, sendo *Closeness* a única exceção.

Para a caracterização dos conceitos das coleções C31 e C37 do CSTNews em função das 4 métricas mais relevantes, *Simple Frequency*, *Cumulative Frequency*, *Closeness* e *Level*, automatizou-se também o cálculo de *Simple Frequency, Cumulative Frequency* e *Level*. Vale ressaltar que a medida *Centraliy* foi descartada devido a não-pertinência da mesma descrita anteiormente.

Para a caracterização dos *clusters* adicionais quanto às métricas de frequência, salienta-se que o editor NASP++ salva automaticamente o número de ocorrências de um nome *x* anotadas com um *synset y* como a *Simple Frequency* de *x.* E, uma vez que a *Simple Frequency* tenha sido especificada, a *Cumulative Frequency* é automaticamente calculada em função da organização hierárquica dos conceitos/*synset* gerada pela ferramenta.

Com base na organização hierárquica dos *synsets* via *Cmap*, é possível calcular os valores da métrica *Level*, cujo único requisito é a posição dos conceitos/*synsets* na hierarquia. Para o cálculo automático da métrica *Closeness,* utilizou-se o mesmo *script* aplicado para a caracterização de C1

Na sequência, descreve-se a proposição dos métodos de SAM a partir da representação das coleções C1, C31 e C37 em herarquias conceituais e na caracterização dos conceitos em função de métricas de grafo.

Para tanto, vale ressaltar que a classe "Sim" (isto é, "ocorrência do conceito no sumário") e "Não" (isto é, "não-ocorrência do conceito no sumário") (Tabela 11) da coleção C1 foram redimensionadas. Tendo em vista a expansão do *corpus* CSTNews pela inclusão de 5 novos sumários humanos a cada coleção, o critério utilizado para a descrição das classes "Sim" e "Não" passou a ser a ocorrência dos conceitos nos 6 sumários humanos disponíveis por coleção. Especificamente, se um conceito dos textos-fonte ocorre em pelo menos um dos seis sumários humanos, então esse conceito pertence à classe "Sim"; caso contrário, se tal conceito não ocorre em nenhum dos seis sumários humanos, então ele pertence à classe "Não".

Considerando-se que os humanos tendem a produzir, por várias razões, sumários relativamente diferentes dada uma mesma coleção de textos-fonte, a alteração no critério de classificação dos conceitos em "Sim" e "Não" feita com base em uma quantidade maior de sumários humanos busca garantir a efetiva relevância dos conceitos. Por exemplo, quando da classificação com base em apenas 1 sumário humano, a coleção C1 tinha inicialmente apenas 13 conceitos classificados como "Sim" e 24 conceitos classificados como "Não". Desses 13, um deles era o conceito "acidente", que tem frequência relativamente alta em relação aos outros conceitos. Com a nova classificação, segundo os 6 sumários humanos do CSTNews, a coleção C1 passou a ter 26 conceitos classificados como "Sim" e 11 conceitos classificados como "Não". Sendo assim, a possibilidade de um conceito importante ser classificado como não-relevante é menor, o que pode proporcionar a elaboração de sumários mais informativos.

Tabela 11 - O *cluster* C1 e suas métricas de relevância redimensionadas

| Qt. | Nome/ conceito | Sumários | Métrica | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Simple Frequency* | *Cumulative Frequency* | *Centrality* | *Closeness* | *Level* |
| 1 | acidente | Sim | 5 | 6 | 2 | 6,524584767 | 3 |
| 2 | aeronave | Sim | 1 | 12 | 2 | 8,274407488 | 3 |
| 3 | aeroporto | Sim | 4 | 4 | 1 | 7,403368688 | 5 |
| 4 | aterrissagem | Sim | 2 | 2 | 1 | 6,227287261 | 4 |
| 5 | avião | Sim | 11 | 11 | 1 | 8,041560382 | 1 |
| 6 | carga | Sim | 2 | 2 | 1 | 7,547655013 | 6 |
| 7 | cidade | Sim | 1 | 1 | 1 | 6,52026769 | 4 |
| 8 | companhia | Sim | 4 | 4 | 1 | 6,500379639 | 3 |
| 9 | fabricação | Sim | 3 | 3 | 1 | 6,22151722 | 3 |
| 10 | floresta | Sim | 3 | 3 | 1 | 6,810928806 | 6 |
| 11 | leste | Sim | 2 | 2 | 1 | 6,283283198 | 4 |
| 12 | localidade | Sim | 2 | 2 | 1 | 6,848025562 | 5 |
| 13 | membro | Sim | 4 | 4 | 1 | 7,314982201 | 4 |
| 14 | mineral | Sim | 2 | 2 | 1 | 7,686018373 | 8 |
| 15 | montanha | Sim | 1 | 1 | 1 | 7,538074088 | 7 |
| 16 | nacionalidade | Sim | 2 | 2 | 1 | 6,695537498 | 6 |
| 17 | passageiro | Sim | 5 | 5 | 1 | 7,657742031 | 5 |
| 18 | pessoa | Sim | 3 | 23 | 5 | 8,656908225 | 7 |
| 19 | pista | Sim | 3 | 3 | 1 | 7,310985946 | 5 |
| 20 | porta-voz | Sim | 7 | 7 | 1 | 7,803125792 | 5 |
| 21 | propriedade | Sim | 2 | 2 | 1 | 6,989497846 | 6 |
| 22 | queda | Sim | 1 | 1 | 1 | 6,26750143 | 2 |
| 23 | tempestade | Sim | 1 | 1 | 1 | 6,905572131 | 6 |
| 24 | tempo | Sim | 2 | 2 | 1 | 6,953761859 | 6 |
| 25 | tripulação, tripulante | Sim | 5 | 5 | 1 | 6,811762991 | 5 |
| 26 | vítima | Sim | 2 | 2 | 1 | 7,424212214 | 5 |
| 27 | chama | Não | 1 | 4 | 1 | 6,377498738 | 4 |
| 28 | distância | Não | 2 | 5 | 1 | 6,437310632 | 5 |
| 29 | estrada | Não | 1 | 6 | 1 | 7,480978846 | 6 |
| 30 | fonte | Não | 2 | 5 | 1 | 6,491493504 | 5 |
| 31 | país | Não | 2 | 5 | 1 | 6,848025562 | 5 |
| 32 | permissão | Não | 1 | 3 | 1 | 5,838955311 | 3 |
| 33 | quilômetro | Não | 3 | 5 | 1 | 6,585324996 | 5 |
| 34 | setor | Não | 1 | 5 | 1 | 6,443209937 | 5 |
| 35 | sobrevivente | Não | 2 | 5 | 1 | 7,424212214 | 5 |
| 36 | tarde | Não | 2 | 5 | 1 | 6,473907029 | 5 |
| 37 | transporte | Não | 1 | 6 | 2 | 8,452418538 | 6 |

A seguir, nas Tabelas 12 e 13 aprentam-se os conceitos do *cluster* C31 e C37 respectivamente, com suas métricas e classificações especificadas de acordo o novo critério de relevância, ou seja, presença ou não dos conceitos nos 6 sumários humanos.

O *cluster* C31 tem 39 conceitos no total, com 20 conceitos classificados como “Sim” e 19 conceitos para a categoria “Não” e o *cluster* C37 possui 76 conceitos no total, com 50 conceitos para a categoria “Sim” e 26 conceitos para a categoria “Não”.

Tabela 12 - O *cluster* C31 e suas métricas de relevância.

| Qt. | Nome/ conceito | Sumários | Métrica | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Simple Frequency* | *Cumulative Frequency* | *Centrality* | *Closeness* | *Level* |
| 1 | abertura | Sim | 1 | 1 | 7 | 3,796232031 | 7 |
| 2 | abril | Sim | 1 | 1 | 7 | 3,731727305 | 7 |
| 3 | agosto | Sim | 2 | 2 | 7 | 3,776722461 | 7 |
| 4 | chama | Sim | 2 | 2 | 7 | 3,785290102 | 7 |
| 5 | cidade | Sim | 1 | 1 | 8 | 3,582626117 | 8 |
| 6 | competição | Sim | 1 | 3 | 5 | 4,148480188 | 5 |
| 7 | dia | Sim | 3 | 3 | 4 | 4,323130451 | 4 |
| 8 | evento | Sim | 1 | 18 | 3 | 5,02917651 | 3 |
| 9 | ginasta | Sim | 3 | 3 | 9 | 3,858066261 | 9 |
| 10 | internet | Sim | 2 | 2 | 9 | 3,70822816 | 9 |
| 11 | jogo | Sim | 4 | 4 | 6 | 4,179517921 | 6 |
| 12 | país | Sim | 2 | 2 | 6 | 3,87483925 | 6 |
| 13 | parte | Sim | 1 | 1 | 3 | 4,393134881 | 3 |
| 14 | percurso | Sim | 1 | 1 | 5 | 3,971045437 | 5 |
| 15 | representante | Sim | 2 | 2 | 9 | 3,759743977 | 9 |
| 16 | revezamento | Sim | 4 | 4 | 11 | 3,92592477 | 11 |
| 17 | símbolo | Sim | 1 | 1 | 7 | 3,69991571 | 7 |
| 18 | tocha | Sim | 4 | 4 | 9 | 3,950818581 | 9 |
| 19 | trajeto | Sim | 1 | 1 | 6 | 3,920611422 | 6 |
| 20 | votação | Sim | 2 | 2 | 7 | 3,977594852 | 7 |
| 21 | ano | Não | 1 | 1 | 1 | 3,818305183 | 6 |
| 22 | apresentação | Não | 1 | 1 | 1 | 3,616138776 | 9 |
| 23 | bronze | Não | 1 | 1 | 1 | 3,614122198 | 7 |
| 24 | destaque | Não | 1 | 1 | 1 | 3,792704263 | 6 |
| 25 | disputa | Não | 1 | 3 | 2 | 4,148480188 | 5 |
| 26 | equipe | Não | 1 | 1 | 1 | 3,792704263 | 6 |
| 27 | esporte | Não | 1 | 1 | 1 | 3,844831188 | 7 |
| 28 | julho | Não | 1 | 1 | 1 | 3,731727305 | 7 |
| 29 | medalha | Não | 2 | 2 | 1 | 3,910822915 | 6 |
| 30 | nadador | Não | 1 | 1 | 1 | 3,754417247 | 9 |
| 31 | noite | Não | 1 | 1 | 1 | 3,972959737 | 5 |
| 32 | olimpíada | Não | 1 | 1 | 1 | 3,814372078 | 7 |
| 33 | ouro | Não | 2 | 2 | 1 | 3,564658011 | 10 |
| 34 | prata | Não | 1 | 1 | 1 | 3,614122198 | 7 |
| 35 | salto | Não | 1 | 1 | 1 | 4,027637396 | 6 |
| 36 | solo | Não | 1 | 1 | 1 | 3,69991571 | 7 |
| 37 | temporada | Não | 1 | 1 | 1 | 3,972959737 | 5 |
| 38 | terça-feira | Não | 1 | 1 | 1 | 3,587013752 | 8 |
| 39 | voto | Não | 1 | 1 | 1 | 3,493255717 | 9 |

Tabela 13 - O *cluster* C37 e suas métricas de relevância.

| Qt. | Nome/ conceito | Sumários | Métrica | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Simple Frequency* | *Cumulative Frequency* | *Centrality* | *Closeness* | *Level* |
| 1 | adulto | Sim | 1 | 1 | 1 | 7,502319676 | 7 |
| 2 | água | Sim | 2 | 2 | 1 | 6,608509897 | 6 |
| 3 | arma | Sim | 2 | 3 | 2 | 6,34201678 | 9 |
| 4 | atendimento | Sim | 1 | 1 | 1 | 5,875043108 | 7 |
| 5 | cadeia | Sim | 1 | 13 | 1 | 7,515044598 | 10 |
| 6 | capacidade | Sim | 2 | 2 | 1 | 6,737649737 | 4 |
| 7 | choque | Sim | 2 | 2 | 1 | 5,882311817 | 8 |
| 8 | começo | Sim | 1 | 1 | 1 | 6,675161112 | 4 |
| 9 | começo | Sim | 1 | 3 | 1 | 6,525120708 | 4 |
| 10 | compromisso | Sim | 1 | 1 | 1 | 6,316598904 | 5 |
| 11 | criança | Sim | 5 | 6 | 1 | 7,307362439 | 8 |
| 12 | dia | Sim | 2 | 2 | 1 | 6,252221491 | 6 |
| 13 | diretor | Sim | 3 | 3 | 1 | 6,643480438 | 10 |
| 14 | estado | Sim | 1 | 1 | 1 | 6,558842133 | 7 |
| 15 | festa | Sim | 3 | 3 | 1 | 6,324823384 | 6 |
| 16 | fuga | Sim | 1 | 1 | 1 | 6,126180883 | 6 |
| 17 | garantia | Sim | 2 | 2 | 1 | 6,244375861 | 6 |
| 18 | greve | Sim | 1 | 1 | 1 | 5,599877635 | 9 |
| 19 | homem | Sim | 1 | 1 | 1 | 7,058438583 | 8 |
| 20 | hora | Sim | 1 | 1 | 1 | 6,680484889 | 4 |
| 21 | início | Sim | 2 | 3 | 1 | 6,525120708 | 5 |
| 22 | líder | Sim | 1 | 5 | 2 | 7,500209914 | 7 |
| 23 | luz | Sim | 2 | 2 | 1 | 6,692229203 | 8 |
| 24 | manhã | Sim | 2 | 2 | 1 | 6,440509348 | 5 |
| 25 | massa | Sim | 2 | 2 | 1 | 7,163129091 | 3 |
| 26 | menor | Sim | 1 | 6 | 1 | 7,307362439 | 8 |
| 27 | mês | Sim | 2 | 2 | 1 | 6,707103772 | 4 |
| 28 | motim | Sim | 4 | 4 | 1 | 6,385244559 | 5 |
| 29 | negociação | Sim | 2 | 2 | 1 | 6,269585261 | 6 |
| 30 | objeto | Sim | 1 | 65 | 3 | 10,26876341 | 2 |
| 31 | pavilhão | Sim | 1 | 1 | 1 | 7,43302097 | 6 |
| 32 | pessoa | Sim | 2 | 40 | 11 | 8,329734983 | 6 |
| 33 | plano | Sim | 1 | 1 | 1 | 6,216490074 | 6 |
| 34 | polícia | Sim | 5 | 5 | 1 | 6,383286545 | 6 |
| 35 | policial | Sim | 1 | 1 | 1 | 6,511216177 | 10 |
| 36 | presídio | Sim | 6 | 13 | 1 | 7,515044598 | 10 |
| 37 | preso | Sim | 12 | 14 | 1 | 7,68685439 | 9 |
| 38 | quarta-feira | Sim | 3 | 3 | 1 | 6,016404706 | 8 |
| 39 | rebelião | Sim | 7 | 7 | 1 | 6,191557502 | 7 |
| 40 | refém | Sim | 2 | 16 | 2 | 7,901593571 | 8 |
| 41 | represália | Sim | 1 | 1 | 1 | 5,957598677 | 7 |
| 42 | revólver | Sim | 1 | 1 | 1 | 5,941168976 | 12 |
| 43 | secretaria | Sim | 1 | 1 | 1 | 5,907165904 | 7 |
| 44 | segurança | Sim | 1 | 1 | 1 | 5,692164205 | 9 |
| 45 | semana | Sim | 1 | 1 | 1 | 6,412250681 | 5 |
| 46 | tensão | Sim | 1 | 1 | 1 | 6,175962003 | 5 |
| 47 | terça-feira | Sim | 1 | 1 | 1 | 5,943544686 | 8 |
| 48 | transferência | Sim | 2 | 2 | 1 | 7,530379159 | 7 |
| 49 | tropa | Sim | 2 | 2 | 1 | 5,899581985 | 8 |
| 50 | túnel | Sim | 1 | 1 | 1 | 6,660710999 | 8 |
| 51 | administração | Não | 1 | 1 | 1 | 6,345790635 | 5 |
| 52 | agente | Não | 3 | 3 | 1 | 6,808078524 | 9 |
| 53 | auxiliar | Não | 1 | 1 | 1 | 6,741262987 | 9 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 54 | braço | Não | 1 | 1 | 1 | 6,128626958 | 8 |
| 55 | buraco | Não | 1 | 1 | 1 | 5,945722489 | 7 |
| 56 | cabeça | Não | 1 | 1 | 1 | 6,472978982 | 6 |
| 57 | capital | Não | 1 | 1 | 1 | 6,342241126 | 8 |
| 58 | cela | Não | 1 | 1 | 1 | 6,737729296 | 8 |
| 59 | detento | Não | 2 | 14 | 1 | 7,68685439 | 9 |
| 60 | enfermagem | Não | 1 | 1 | 1 | 5,844566832 | 8 |
| 61 | esposa | Não | 1 | 1 | 1 | 6,741262987 | 9 |
| 62 | fim | Não | 1 | 1 | 1 | 6,675161112 | 4 |
| 63 | forma | Não | 1 | 1 | 1 | 6,316598904 | 5 |
| 64 | informação | Não | 2 | 2 | 1 | 6,658876939 | 4 |
| 65 | lanche | Não | 1 | 1 | 1 | 6,320540399 | 7 |
| 66 | libertação | Não | 1 | 1 | 1 | 5,954134905 | 7 |
| 67 | muro | Não | 1 | 1 | 1 | 7,038613525 | 7 |
| 68 | oportunidade | Não | 1 | 1 | 1 | 6,067223743 | 6 |
| 69 | parte | Não | 1 | 1 | 1 | 5,529036987 | 10 |
| 70 | revista | Não | 1 | 1 | 1 | 5,913885767 | 7 |
| 71 | secretário-adjunto | Não | 1 | 1 | 1 | 6,759366219 | 9 |
| 72 | sindicância | Não | 1 | 1 | 1 | 5,77093685 | 8 |
| 73 | suspeito | Não | 1 | 1 | 1 | 7,502319676 | 7 |
| 74 | tarde | Não | 1 | 1 | 1 | 6,201511072 | 6 |
| 75 | termo | Não | 1 | 1 | 1 | 6,383490708 | 5 |
| 76 | unidade | Não | 6 | 13 | 1 | 7,515044598 | 10 |

## 6.3. Aprendizado de critérios de relevância via AM

A delimitação dos critérios de relevância baseados nas métricas de grafo foi feita de forma automática, isto é, com base nos resultados do aprendizado de algoritmos de AM, em um processo similar ao realizado para o estudo da pertinência das métricas.

Uma vez que as métricas *Simple Frequency, Cumulative Frequency, Closeness* e *Level* foram calculadas para as coleções C1, C31 e C37 do CSTNews, estas foram submetidas em conjunto a algoritmos do paradigma supervisionado de AM. No caso, vários testes e treinamentos foram feitos em função de diferentes configurações fornecidas pelo ambiente *Weka*, como *Use Training Set* com seleção de atributos e *Supplied Test Set* com seleção de atributos.

Com ênfase nos algoritmos simbólicos, buscou-se identificar, com base nas diferentes configurações do *Weka*, o menor conjunto de regras que obtivesse a mais alta precisão na tarefa de determinar os conceitos relevantes (e não relevantes) de dada coleção multidocumento. As 3 coleções selecionadas, com todas as medidas calculadas e classificações de relevância (“Sim” foi para os sumários, e “Não” não foi para os sumários) foram submetidas a algoritmos de aprendizado de máquina do *Weka*.

A seguir, apresentam-se as classificações e treinamentos/testes realizados com algoritmos de AM para todos os 3 *clusters* selecionados.

### *6.3.1. Pré-processamento do corpus*

Primeiramente, aplicou-se a técnica de *oversampling* aos *clusters* C1 e C37 para garantir o balanceamento dos dados das classes “Sim” e “Não”. Especificamente, os 11 conceitos da classe “Não”, que compõem o total de 37 conceitos do *cluster* C1, foram duplicados, totalizando 22 conceitos. Logo, o *cluster* C1 passou a ter 48 conceitos no total (isto é, 26 da classe “Sim” e 22 da classe “Não”). Quanto ao *cluster* C37, os 26 conceitos da classe “Não”, que compõem o total de 76 conceitos, foram duplicados, somando 52 conceitos. Por conseguinte, o *cluster* C37 passou a ter 102 conceitos no total (isto é, 50 da classe “Sim” e 52 da classe “Não”). O *cluster* C31 já estava balanceado com 20 conceitos para a classe “Sim” e 19 conceitos para a classe “Não”, totalizando 39 conceitos.

Uma vez que os *clusters* estavam balanceados, passou-se ao pré-processamento padrão do ambiente *Weka*. Sendo assim, aplicou-se a opção padrão “normalização”, para que as métricas apresentassem valores entre 0 e 1.

Em seguida, utilizou-se a opção padrão “discretizar”, para que os todos valores, além de terem intervalos entre 0 e 1, apresentassem intervalos numéricos iguais de acordo com as métricas.

É importante ressaltar que para a manipulação dos dados de treino e teste, todos os atributos devem estar dispostos de maneira igual e na mesma ordem, caso contrário, o *Weka* apresentará erros. Neste trabalho, os atributos são os nomes das métricas e os conceitos e valores são as instâncias. Após a discretização e normalização dos dados foi importante conferir, num editor de texto, se a classificação do atributo “Foi para o sumário?” para “Sim” e “Não” estava na ordem correta. Na Figura 36, tem-se os atributos “frequência texto” (*simple frequency*), “proximidade” (*closeness*), “frequência hierarquia” (*cumulative frequency*), “nível hierarquia” *(level*), e “foi para o sumário”. O atributo “foi para o sumário”, aliás, representa as classes a serem aprendidas e tem os valores “sim, não” (isto é, relevante, não-relevante). Em todos os conjuntos, a ordem dos valores deve ser sempre a mesma (ou seja, “sim” seguido de “não), caso contrário, o *Weka* não consegue testar os dados.

Figura 36 - Classificação do atributo “Foi para o sumário?”

```
3  @attribute 'Frequência texto'
   {'\'(-inf-0.1]\'','\'(0.1-0.2]\'','\'(0.2-0.3]\'
   }
4  @attribute Proximidade
   {'\'(-inf-0.1]\'','\'(0.1-0.2]\'','\'(0.2-0.3]\'
   }
5  @attribute 'Frequência hierarquia'
   {'\'(-inf-0.1]\'','\'(0.1-0.2]\'','\'(0.2-0.3]\'
   }
6  @attribute 'Nível hierarquia'
   {'\'(-inf-0.1]\'','\'(0.1-0.2]\'','\'(0.2-0.3]\'
   }
7  @attribute 'Foi para o sumário?' {SIM,NÃO}
8
```

Fonte: autoria própria.

Na Figura 36, por exemplo, o atributo “foi para o sumário?”, referente ao *corpus* de treinamento, está na ordem “sim, não”, fazendo com que o mesmo ocorra com o *corpus* de teste para que não haja erros de manipulação dos dados.

### *6.3.2. Treinamento e teste*

Após o pré-processamento, deu-se início aos treinos e testes com o *corpus* composto pelos *clusters* C1, C31 e C37. Para tanto, realizaram-se 3 rodadas de treino/teste. Em cada rodada, utilizou-se uma combinação distinta de 2 *clusters* para treino e 1 *cluster* para teste, a saber: (a) treino com C1+C31 e teste em C37; (b)treino com C1+C37 e teste em C31 e (c) treino com C31+C37 e teste em C1.

Nas 3 rodadas, os treinamentos foram feitos por meio da opção do *Weka* “*Use Training Set*” e os testes foram feitos por meio da opção “*Supplied Test Set*”. Além disso, ressalta-se que não foi feita a “seleção de atributos” clássica do *Weka* tendo em vista que esta foi feita, de certa forma, quando da investigação da pertinência das medidas (Seção 5.3), que resultou na exclusão da métrica *Centrality* desta fase do trabalho. Assim, nas 3 rodadas de treino e teste, utilizaram-se as 4 medidas tidas como pertinentes (*Simple Frequency*, *Cumulative Frequency*, *Closeness* e *Level*).

*a) Treino com C1 e C31 e teste em C37*

Nas primeiras investigações com os algoritmos simbólicos, os valores mais altos de precisão foram obtidos pelos algoritmos *Non-Nested Generalized Exemplars* (NNGe) (89%), que trabalha com regras “se-então”, *Decision Table/Naïve Bayes* (DTNB) (77%) algoritmo híbrido que divide os atributos em 2 subconjuntos disjuntos, Prism (75%), e

JRip (68%), os quais geraram conjuntos relativamente extensos de regras, com exceção de JRip e DTNB, que geraram apenas 3 e 1 regras, respectivamente. O algoritmo PART foi o que apresentou a melhor combinação dos critérios precisão e número de regras. No caso, ele obteve precisão relativamente alta de 77% e produziu um conjunto de 11 regras, o que é computacionalmente tratável (Tabela 14).

Tabela 14 - Regras obtidas pelo PART a partir do treino em C1 e C31.

| Regra | Acerto/Erro | Precisão |
|---|---|---|
| 1. Se *Cumulative Frequency* = -inf-0.1 e *Closeness* = 0.2-0.3 e *Simple Frequency* = 0.1-0.2, então "não" | (15.0/2.0) | 86,7% |
| 2. Senão *Cumulative Frequency* = -inf-0.1 e *Closeness* = 0.3-0.4, então "sim" | (14.0/6.0) | 57,1% |
| 3. Senão *Cumulative Frequency* = 0.1-0.2, então "sim" | (12.0/1.0) | 91,7% |
| 4. Senão *Closeness* = 0.5-0.6, então "sim" | (6.0/2.0) | 66,7% |
| 5. Senão *Cumulative Frequency* = -inf-0.1 e *Closeness* = 0.2-0.3 e *Simple Frequency* = 0.4-0.5, então "sim" | (4.0/0.0) | 100% |
| 6. Senão *Cumulative Frequency* = -inf-0.1 e *Closeness* = 0.2-0.3, então "não" | (5.0/1.0) | 80% |
| 7. Senão *Cumulative Frequency* = -inf-0.1 e *Closeness* = 0.1-0.2 e *Level* = 0.5-0.6, então "não" | (4.0/1.0) | 75% |
| 8. Senão *Closeness* = 0.6-0.7, então "sim" | (5.0/0.0) | 100% |
| 9. Senão *Cumulative Frequency* = -inf-0.1e *Closeness* = -inf-0.1, então "não" | (3.0/0.0) | 100% |
| 10. Senão *Cumulative Frequency* = -inf-0.1 e *Level* = 0.7-0.8, então "sim" | (3.0/1.0) | 66,7% |
| 11. Senão "sim" | (16.0/6.0) | 62,5% |

Com base na análise da matriz de confusão gerada pelo PART (Tabela 15), nota-se que o algoritmo classifica corretamente as instâncias da classe "Sim" com precisão relativamente alta, de 87%, as da classe "Não" com precisão de 65,9%.

Tabela 15 - Matriz de confusão do PART a partir do treino em C1 e C31.

| **Classe** / **Teste** | **Sim (46)** (Ocorrência no sumário) | **Não (41)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 40 (87%) | 6 (15%) |
| Não | 14 (34,1%) | 27(65,9%) |

A acurácia das regras aplicadas ao *cluster* C37 foi de 51%, ou seja, o PART acerta um pouco mais da metade dos casos "Sim" (54%) e exatamente a metade dos casos "Não" (50%), o que pode ser visto na matriz de confusão da Tabela 16.

Tabela 16 - Matriz de confusão do PART a partir do teste em C37.

| **Classe** / **Teste** | **Sim (50)** (Ocorrência no sumário) | **Não (52)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 27 (54%) | 23 (46%) |
| Não | 26 (50%) | 26 (50%) |

*Aplicação do algoritmo PART:10-fold cross –validation*

A aplicação da técnica *10-fold cross-validation* para o algoritmo PART gerou as mesmas 11 regras do conjunto de treino com precisão geral de 60%. Com base na matriz de confusão, observa-se que o algoritmo classificou corretamente 67,4% dos casos da classe "Sim" e 53,6% dos casos da classe "Não" (Tabela 17).

Tabela 17 - Matriz de confusão do PART com *10-fold cross-validation* para C1 e C31.

| **Classe** / **Teste** | **Sim (46)** (Ocorrência no sumário) | **Não (41)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 31 (67,4%) | 15 (32,6%) |
| Não | 19 (46,3%) | 22 (53,6%) |

*b) Treino com C1 e C37 e teste em C31 e validação cruzada*

O segundo treinamento realizado se deu com os *clusters* C1 e C37. Dos algoritmos simbólicos, os que obtiveram as melhores acurácias foram: NNGe (80%), Prism (72%) e PART (61%). O NNGe e o Prism geraram conjuntos de regras muito extensos. Em compensação, o PART gerou um conjunto de 6 regras, as quais estão na Tabela 18.

Tabela 18 - Regras obtidas pelo PART a partir do treino em C1 e C37.

| Regra | Acerto/Erro | Precisão |
|---|---|---|
| 1. Se *Simple Frequency* = 0.2-0.3, então "sim" | (7.0/2.0) | 71,4% |
| 2. Senão *Simple Frequency* = 0.3-0.4, então "sim" | (6.0/0.0) | 100% |
| 3. Senão *Level* = 0.5-0.6, então "não" | (36.0/12.0) | 66,7% |
| 4. Senão *Closeness* = 0.5-0.6, então "sim" | (6.0/0.0) | 100% |
| 5. Senão *Level* = 0.4-0.5, então "sim" | (19.0/8.0) | 57,9% |
| 6. Senão "não" | (76.0/36.0) | 52,6% |

A partir da matriz de confusão do PART (Tabela 19), observa-se que o algoritmo classifica corretamente 36,8% dos casos de "Sim" e 86,5% dos casos de "Não".

Tabela 19 - Matriz de confusão do PART a partir do treino em C1 e C37.

| **Classe** / **Teste** | **Sim (76)** (Ocorrência no sumário) | **Não (74)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 28 (36,8%) | 48 (63,2%) |
| Não | 10 (13,5%) | 64 (86,5%) |

A aplicação das regras em C31 alcançou a marca dos 56% de precisão, o que também pode ser observado na matriz de confusão (Tabela 20), pois o algoritmo erra e acerta praticamente a metade dos casos da classe "Sim" e dos casos da classe "Não".

Tabela 20 - Matriz de confusão do PART a partir do teste em C31.

| **Classe** / **Teste** | **Sim (20)** (Ocorrência no sumário) | Não (19) (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 11 (55%) | 9 (45%) |
| Não | 8 (42,1%) | 11 (57,9%) |

*Aplicação do algoritmo PART:10-fold cross –validation*

Para o conjunto de treino C1 e C31, o algoritmo de validação cruzada (*10-fold*) também gerou as mesmas 6 regras com 58% de precisão. No caso, o algoritmo classificou corretamente 42,1% dos casos da classe "Sim" e 72,3% dos casos da classe "Não"(Tabela 21).

Tabela 21 - Matriz de confusão do PART com *10-fold cross-validation* para C1 e C37.

| **Classe** / **Teste** | **Sim (76)** (Ocorrência no sumário) | **Não (74)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 32 (42,1%) | 44 (58%) |
| Não | 19 (25,7%) | 55 (72,3%) |

*c) Treino com C31 e C37 e teste em C1*

O último treinamento foi feito com os *clusters* C31 e C37. Os algoritmos simbólicos que apresentaram as melhores acurácias foram: NNGe (73%), DNTB (64%), Prism (63%) e PART (63%). Os algoritmos NNGe e Prism geraram um conjunto muito extenso de regras, o DTNB gerou apenas uma regra e o PART gerou 7 regras. Os resultados com o PART são apresentados na Tabela 22.

Tabela 22 - Regras obtidas pelo PART a partir do treino em C31 e C37.

| Regra | Acerto/Erro | Precisão |
|---|---|---|
| 1. Se *Simple Frequency* = 0.9-inf, então "sim" | (4.0/0.0) | 100% |
| 2. Senão *Simple Frequency* = 0.4-0.5 e *Level* = 0.5-0.6, então "sim" | (3.0/0.0) | 100% |
| 3. Senão *Simple Frequency* = 0.1-0.2, então "não" | (31.0/12.0) | 61,3% |
| 4. Senão *Simple Frequency* = -inf-0.1 e *Level* = 0.6-0.7 , então "não" | (29.0/11.0) | 62,1% |
| 5. Senão *Simple Frequency* = -inf-0.1 e *Closeness* = 0.1-0.2, então "não" | (23.0/9.0) | 60,9% |
| 6. Senão *Level* = 0.7-0.8 e *Closeness* = 0.4-0.5, então "sim" | (7.0/2.0) | 71,4% |
| 7. Senão "sim" | (44.0/18.0) | 59,1% |

Pela matriz de confusão gerada pelo PART para C31 e C37 (Tabela 23), observa-se que o algoritmo acerta mais de 70% dos casos da classe "Não" e acerta um pouco mais da metade dos casos da classe "Sim".

Tabela 23 - Matriz de confusão do PART a partir do treino em C31 e C37.

| **Classe** / **Teste** | **Sim (70)** (Ocorrência no sumário) | **Não (71)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 38 (54,3%) | 32 (45,7%) |
| Não | 20 (28,2%) | 51 (71,8%) |

O último teste realizado foi a aplicação das regras apreendidas com base nos *clusters* C31 e C37 ao *cluster* C1. Tal teste, realizado com o algoritmo PART, atingiu 58% de precisão com as mesmas 7 regras do conjunto de treinamento. Pela matriz de confusão apresentada na Tabela 24, observa-se que o PART testado em C1 acertou 61,5% dos casos da classe "Sim" e 54,5% dos casos da classe "Não", ou seja, ele consegue acertar pouco mais da metade dos casos para todo o conjuntos de teste.

Tabela 24 - Matriz de confusão do PART a partir do teste em C1.

| **Classe** / **Teste** | **Sim (26)** (Ocorrência no sumário) | **Não (22)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 16 (61,5%) | 10 (38,5%) |
| Não | 10 (45,5%) | 12 (54,5%) |

*Aplicação do algoritmo PART:10-fold cross –validation*

A validação cruzada, para o conjunto de treino C31 e C37 classificou os atributos "Sim" e "Não" com 58% de acurácia também, e gerou as mesmas 7 regras. Na tabela 25

observa-se que o algoritmo classificou corretamente 47,1% dos casos da classe “Sim” e boa parte dos casos da classe “Não” (69%).

Tabela 25 - Matriz de confusão do PART com *10-fold cross-validation* para C31 e C37.

| **Classe** / **Teste** | **Sim (70)** (Ocorrência no sumário) | **Não (71)** (Não-ocorrência no sumário) |
|---|---|---|
| Sim | 33 (47,1%) | 37 (52,8%) |
| Não | 22 (31%) | 49 (69%) |

Diante dos resultados apresentados pelos grupos de treinamento/teste, julgou-se o melhor conjunto de regras para a detecção dos conceitos relevantes com base na acurácia e no número de regras para a construção dos sumários.

Os resultados apresentados em (a) atestaram que o treinamento do algoritmo PART obteve a mais alta precisão (77%) em relação aos outros treinamentos, e além disso, o PART aprendeu pelo menos 51% das regras testadas para o *cluster* C37.

Sendo assim, em comparação aos cenários (b) e (c), pode-se dizer que os melhores resultados foram obtidos no cenário (a) e, por isso, o conjunto de regras desse cenário foi o escolhido para ser utilizado para ranquear e pontuar as sentenças em um dos métodos de SAM. A seguir, apresentam-se os métodos de SAM extrativos cujas abordagens de seleção de conteúdo baseiam-se em medidas de grafo.

# 7. Proposição e Avaliação de métodos de SAM

Como visto na literatura, os métodos de seleção do conteúdo que irá compor o sumário são aplicados durante a transformação, um dos processos que compõem a arquitetura dos métodos/sistemas de SA. A seguir, apresentam-se os dois métodos de SAM extrativa propostos neste trabalho, cujos critérios de seleção de conteúdo pautam-se em medidas de relevância de grafo.

## 7.1. Descrição do métodos de SAM

O primeiro método de SAM extrativa foi denominado CFSumm (do inglês, *Concept Frequency Summarization*), pois as sentenças mais relevantes de uma coleção são selecionadas exclusivamente com base na frequência de ocorrência dos conceitos. Como visto, a frequência dos conceitos já foi explorada na literatura como critério de relevância, mostrando-se pertinente em tal tarefa. No método aqui proposto, a

frequência de ocorrência é capturada pela métrica *Simple Frequency*. No Quadro 9, descreve-se o algoritmo do método CFSumm.

Quadro 9 - Algoritmo do Método CFSumm

| | |
|---|---|
| Análise | 1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os substantivos (nomes comuns) com os conceitos/synsets da WordNet de Princeton. |
| Transformação | 2. Calcular a taxa de compressão em 70%<br>3. Pontuar as sentenças com base no valor da medida *Simple Frequency* de seus conceitos/*synsets* constitutivos na coleção<br>4. Ranquear as sentenças em função da pontuação dos conceitos<br>5. Selecionar a 1ª sentença do ranque<br>6. Caso a taxa de compressão não tenha sido atingida:<br>6.a. Selecionar a próxima sentença do ranque<br>6.b. Verificar a redundância da sentença em questão com a já selecionada<br>6.c. Eleger a sentença somente se não for redundante<br>6.d. Em caso de empate, seleciona-se a sentença na ordem em que seguem os os documentos (Documento 1, Documento 2...etc.)<br>7. Repetir o passo 6 até que a taxa de compressão seja atingida |
| Síntese | 8. Justapor as sentenças na ordem em que foram selecionadas<br>9. Ordenar os segmentos/sentenças pela ordem de ocorrência nos textos-fonte. |

O segundo método de SAM extrativa foi denominado LCHSumm (do inglês, *Lexical Conceptual Hierarchy Summarization*), pois a seleção de conteúdo se baseia nas regras aprendidas pelo PART a partir das métricas *Simple Frequency*, *Cumulative Frequency*, *Closeness* e *Level* utilizadas como atributos. Tal método é descrito pelo algoritmo apresentado no Quadro 10.

Quadro 10 - Algoritmo do Método LCHSumm

| | |
|---|---|
| Análise | 1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os substantivos (nomes comuns) com os conceitos/synsets da WordNet de Princeton. |
| Transformação | 2. Calcular a taxa de compressão em 70%<br>3. Pontuar as sentenças em função da precisão da regra de AM aplicada aos *synsets*/conceitos na coleção classificados como "Sim" pela regra<br>4. Ranquear as sentenças em função da pontuação dos conceitos<br>5. Selecionar a 1ª sentença do ranque<br>6. Caso a taxa de compressão não tenha sido atingida:<br>6.a. Selecionar a próxima sentença do ranque<br>6.b. Verificar a redundância da sentença em questão com a já selecionada<br>6.c. Eleger a sentença somente se não for redundante<br>6.d. Em caso de empate, seleciona-se a sentença na ordem em que seguem os os documentos (Documento 1, Documento 2...etc.)<br>7. Repetir o passo 6 até que a taxa de compressão seja atingida |
| Síntese | 8. Justapor as sentenças na ordem em que foram selecionadas<br>9. Ordenar os segmentos/sentenças pela ordem de ocorrência nos textos-fonte. |

## 7.2. Geração de extratos

Quanto à geração dos extratos, ressalta-se que, diante da complexidade de implementação das regras obtidas pelo AM em um sistema de sumarização multidocumento durante a realização deste trabalho de mestrado, os extratos foram construídos de forma manual. Ademais, utilizou-se uma taxa de compressão de 70%, (calculada a partir do número de palavras do maior texto da coleção) a mesma utilizada para a construção dos demais sumários do CSTNews. Para eliminar a redundância na aplicação de ambos os métodos, utilizou-se a anotação CST existente no *corpus*. Especificamente, descartou-se a sentença candidata que possuía relações de "redundância" com as já selecionadas para o sumário. Por exemplo, se duas sentenças estivessem relacionadas via *Identity*, a segunda sentença candidata era ignorada; se duas sentenças estivessem relacionadas via *Equivalence*, devia-se eliminar a sentença mais longa (em número de palavras); se duas sentenças estivessem relacionadas via *Subsumption*, a sentença que é subsumida deveria ser excluída.

### *7.2.1. Aplicando o método CFSumm*

Seguindo-se o algoritmo do método CFSumm, cada sentença S dos textos-fonte de um *cluster* foi pontuada e ranqueada em função da soma da frequência simples dos conceitos relevantes e não-relevantes de S na coleção. Em outras palavras, a soma da

frequência de todos os conceitos de uma S resulta no peso de S. Por exemplo, a primeira sentença do Documento 1 da coleção C1 obteve peso igual a 20, já que esse valor resultou da soma da frequência de seus conceitos constitutivos, "pessoa", "queda", "avião" e "passageiro" (3+1+11+5=20), na coleção. Na Tabela 26, os conceitos estão sublinhados e os valores da medida *Simple Frequency* estão associadas a cada conceito entre parênteses.

Tabela 26 - Cálculo do peso das sentenças segundo o método CFSumm.

| S1_D1_C1 | Peso |
|---|---|
| Ao menos 17 pessoas **(3)** morreram após a queda **(1)** de um avião **(11)** de passageiros **(5)** na República Democrática do Congo. | 20 |

Na Tabela 27, apresentam-se todas as sentenças do *cluster* C1 e seus respectivos pesos.

Tabela 27 - Peso das sentenças do *cluster* C1 segundo o método CFSumm.

| **Sentença/ Documento** | **Sentença** | **Peso** |
|---|---|---|
| S1_D1 | Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. | 20 |
| S2_D1 | Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. | 26 |
| S3_D1 | A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto. | 18 |
| S4_D1 | Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. | 20 |
| S5_D1 | O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. | 17 |
| S6_D1 | Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros. | 6 |
| S7_D1 | Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas. | 15 |
| S8_D1 | Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa. Apenas uma manteve a permissão. | 5 |
| S9_D1 | Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como "uma vergonha" para o setor. | 3 |
| S1_D2 | Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas. | 21 |
| S2_D2 | As vítimas do acidente foram 14 passageiros e três membros da tripulação. | 20 |
| S3_D2 | Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. | 28 |

| | | |
|---|---|---|
| S4_D2 | Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. | 12 |
| S5_D2 | O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. | 18 |
| S6_D2 | "Não houve sobreviventes", disse Okala. | 2 |
| S7_D2 | O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais. | 31 |
| S1_D3 | Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas. | 21 |
| S2_D3 | As vítimas do acidente foram 14 passageiros e três membros da tripulação. | 20 |
| S3_D3 | Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu. | 28 |
| S4_D3 | O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. | 18 |
| S5_D3 | "Não houve sobreviventes", disse Okala. | 2 |
| S6_D3 | O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais. | 31 |
| S7_D3 | Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. | 12 |

Uma vez que os pesos de todas as sentenças de um *cluster* tenham sido calculados, seleciona-se a sentença mais bem ranqueada para iniciar o sumário. No caso de C1, selecionou-se a sentença S7_D2, cujo peso é 31. Seguindo-se o algoritmo, a segunda sentença a ser selecionada é S6_D3. Tal sentença, no entanto, é idêntica a S7_D2, ou seja, S7_D2 e S6_D3 são totalmente redundantes e, por essa razão, S6_D3 não é selecionada para o sumário e a próxima sentença mais bem ranqueada passa a ser a nova candidata a compor o sumário. A próxima sentença do ranque mais bem pontuada é S3_D3 (peso é 28) e, por não possuir relação CST de redundância com S7_D2, esta é então selecionada para compor o sumário. Tendo em vista a taxa de compressão de 70%, o tamanho previsto para o sumário (54 palavras) é atingido após a seleção de S3_D3. Justapondo as 2 sentenças selecionadas, tem-se o sumário de C1 gerado segundo o método CFSumm, o qual está descrito no Quadro 11.

Quadro 11 - Sumário de C1 com base no método CFSumm.

| |
|---|
| O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.<br>Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. |

Para as outras coleções anotadas, aplicou-se o mesmo algoritmo. No caso do *cluster* C31, a S1 do Documento 1 foi selecionada para compor o sumário porque tinha o maior peso (22) em relação às outras sentenças. A seleção dessa sentença foi o suficiente para atingir a taxa de compressão de 70% prevista para o sumário de C31, já que o maior texto-fonte dessa coleção é relativamente pequeno com 134 palavras. No Quadro 12, tem-se o sumário resultante para esse *cluster* segundo o método CFSumm .

Quadro 12 - Sumário de C31 com base no método CFSumm

| |
|---|
| A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. |

Para a coleção C37, o método CFSumm gerou um sumário composto pelas seguintes sentenças-fonte: (i) S1_D1, com peso igual a 41, (ii) S10_D2, com peso igual a 34 e (iii) S4D1, com peso igual a 30. No Quadro 13, tem-se o sumário resultante para esse *cluster* segundo o método CFSumm .

Quadro 13 – Sumário de C37 com base no método CFSumm

| |
|---|
| 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.<br>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.<br>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. |

### *7.2.2 Aplicando o método LCHSumm*

De acordo com o método LCHSumm, o ranque das sentenças é elaborado com base na soma dos pesos associados somente aos conceitos relevantes. Tendo em vista que os conceitos relevantes (isto é, classificados como "Sim" pelas regras) e não-relevantes (isto é, classificados como "Não" pelas regras) da coleção já tenham sido identificados pela aplicação das regras do AM (cf. Tabela 14), utiliza-se a precisão das regras que identificam os conceitos relevantes para pontuar e ranqueá-los. No caso, os conceitos classificados como relevantes pelas regras de mais alta precisão recebem maior peso. Por exemplo, seguindo a ordem de aplicação das regras pelo algoritmo, a Regra 2 é a primeira que identifica conceitos relevantes (cf. Tabela 14). No caso, a Regra 2 obteve

precisão de 57,1% e, portanto, o conceito que a satisfaz é associado ao peso 0,57. Os conceitos que satisfazem a Regra 3, que identifica os conceitos relevantes com 91,7% de precisão, são associados ao peso 0,92, e assim por diante. Caso um conceito satistizesse mais de uma regra, sua pontuação final seria a soma da pontuação referente a cada uma das regras que satisfaz. Com o ranque dos conceitos, as sentenças dos textos-fonte são pontuadas e ranqueadas em função do peso dos seus conceitos/*synsets* relevantes constitutivos. Assim, o topo do ranque é composto pelas sentenças que contém os conceitos mais relevantes, dentre os relevantes.

Na Tabela 28, apresenta-se o método de pontuação e ranqueamento de uma sentença do *cluster* C1 com base na precisão das regras do AM. Nessa Tabela, os conceitos estão sublinhados e os pesos referentes às regras estão entre parênteses.

Tabela 28 - Cálculo do peso das sentenças segundo o método LCHSumm.

| S1_D1_C1 | **Peso** |
|---|---|
| O porta-voz **(1)** informou que o avião **(0,63),** um Soviet Antonov-28 de fabricação **(0,63)** ucraniana e propriedade **(0,67)** de uma companhia **(0,92)** congolesa, a Trasept Congo, também levava uma carga **(1,67)** de minerais **(1)** | 6,52 |

Uma vez que as sentenças tenham sido ranqueadas, seleciona-se a mais bem pontuada para iniciar o sumário. No caso do C1, por exemplo, a sentença no topo do ranque é a S1_D1, descrita na Tabela 25. A próxima sentença do ranque é S6_D3, a qual foi descartada por ser idêntica a S1_D1. Assim, a próxima sentença candidata a compor o sumário é S3_D2, a qual, por não possuir relação CST de redundância com S1_D1, foi incluída no sumário. Com a seleção da sentença S3_D2, atingiu-se a taxa de compressão de 70% prevista para o sumário de C1. Tal sumário é apresentado no Quadro 14.

Quadro 14 - Sumário de C1 com base no método LCHSumm

| |
|---|
| O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.<br>Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. |

Para o *cluster* C31, a seleção da primeira sentença do ranque (S1_D1), que possui peso de 7,16, foi suficiente para atingir a taxa de compressão de 70% prevista para o sumário (Quadro 15).

Quadro 15 - Sumário de C31 com base no método LCHSumm

| |
|---|
| A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. |

O *cluster* C37, por ser o maior de todos os 3 *clusters*, resultou em um sumário mais extenso, contendo 3 sentenças, já que a taxa de compressão era de 70%. As sentenças selecionadas foram: (i) . S11_D1, com peso de 8,64, (ii) S10_D2, com peso 6,76, e (iii) S4_D1, com peso 6,40. No Quadro 16, apresenta-se o sumário gerado para C37.

Quadro 16 - Sumário de C37 com base no método LCHSumm

| |
|---|
| A cadeia abriga 203 detentos, mas só tem capacidade para 80. Neste mesmo pavilhão, no início deste mês, os presos quebraram objetos e fizeram um túnel para tentar uma fuga em massa.<br>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.<br>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. |

A seguir, descreve-se a avaliação dos métodos CFSumm e LCHSumm.

## 7.3. Avaliação dos métodos

Essa etapa consistiu em avaliar os sumários gerados a partir da aplicação dos métodos propostos neste trabalho. Visto que a avaliação extrínseca é demorada e cara (VAN-HALTEREN; TEUFEL, 2003), optou-se, nesta investigação, pela avaliação intrínseca, que tem como foco a informatividade dos sumários extrativos (MANI, 2001).

Para a avalição da informatividade, utilizou-se o pacote de medidas ROUGE (LIN, 2004) descrita na Seção 2.1.3. Especificamente, utilizaram-se duas medidas: (i) ROUGE-1, que calcula a informatividade pela sobreposição de unigramas entre o sumário automático e o de referência, e (ii) ROUGE-2, que se baseiam na sobreposição

de bigramas entre o sumário automático e o de referência. A aplicação desse pacote é automática e os resultados são fornecidos em termos de precisão, cobertura e medida-f. Para a aplicação do pacote ROUGE, ressalta-se a necessidade de ao menos 1 sumário de referência (humano) para comparação. Nesse caso, o *corpus* CSTNews disponibiliza os dados necessários, já que possui 6 sumários de referência por *cluster*. Sendo assim, neste trabalho, cada sumário gerado de acordo com o método proposto foi avaliado em comparação aos 6 sumários humanos (abstracts) disponíveis por coleção do CSTNews. Como exemplo pode-se citar o sumário obtido por meio do método CFSumm para a coleção C1 que foi avaliado em comparação aos 6 sumários humanos disponíveis na mesma coleção.

Para fins de comparação, utilizou-se como *baseline* o sumarizador extrativo multidocumento GistSumm (PARDO, 2005). O GistSumm foi o primeiro sistema sumarizador automático multidocumento produzido para o Português e segue uma abordagem muito simples, justapondo todos os textos e selecionando as sentenças de acordo com a frequença das palavras. Além disso, o GistSumm é um aplicativo de fácil instalação e que possui uma interface amigável para manipulação dos textos a serem sumarizados. Especificamente, o método superficial do GistSumm pontua e ranqueia as sentenças dos textos de uma coleção com base na frequência de ocorrência de suas palavras na coleção. A sentença de maior pontuação é considerada a *gist sentence* (isto é, sentença que expressa o conteúdo principal da coleção) e selecionada para iniciar o sumário. As demais sentenças que compõem o sumário satisfazem a dois critérios: (i) conter pelo menos um radical em comum com a *gist sentence* e (ii) ter pontuação maior que a média das pontuações de todas as sentenças.

No Quadro 17, exibem-se os sumários gerados pelo GistSumm para os *clusters* C1, C31 e C37, seguindo a mesma taxa de compressão (70%) utilizada nos métodos CFSumm e LCHSumm.

Quadro 17 - Sumários de C1, C31 e C37 gerados pelo GistSumm

| **Sumário C1** |
| --- |
| Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.<br>Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas. |

| **Sumário C31** |
|---|
| A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. |
| **Sumário C37** |
| 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.<br>A unidade ficou sem luz e água e as negociações para a libertação dos reféns foi retomada na manhã desta quarta-feira.<br>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. |

Nas Tabelas 29, 30 e 31, tem-se os resultados da ROUGE para os três metodos em questão, isto é, CFSumm, LCHSumm e GistSumm, respectivamente. Na Tabela 32, apresenta-se a média das medidas ROUGE 1 e ROUGE 2 para os três métodos.

Tabela 29 - Resultado da ROUGE: CFSumm.

| Coleção | ROUGE-1 | | | ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| C1 | 0.40851 | 0.41026 | 0.40938 | 0.17467 | 0.17544 | 0.17505 |
| C31 | 0.42365 | 0.62319 | 0.50440 | 0.27919 | 0.41667 | 0.33435 |
| C37 | 0.37101 | 0.33862 | 0.35408 | 0.20649 | 0.18817 | 0.19690 |

Tabela 30 - Resultado da ROUGE: LCHSumm.

| Coleção | ROUGE-1 | | | ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| C1 | 0.40851 | 0.41026 | 0.40938 | 0.17467 | 0.17544 | 0.17505 |
| C31 | 0.42365 | 0.62319 | 0.50440 | 0.27919 | 0.41667 | 0.33435 |
| C37 | 0.31304 | 0.30508 | 0.30901 | 0.11504 | 0.11207 | 0.11354 |

Tabela 31 - Resultado da ROUGE: GistSumm.

| Coleção | ROUGE-1 | | | ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| C1 | 0.31304 | 0.30508 | 0.30901 | 0.11504 | 0.11207 | 0.11354 |
| C31 | 0.42365 | 0.62319 | 0.50440 | 0.27919 | 0.41667 | 0.33435 |
| C37 | 0.41449 | 0.42560 | 0.41997 | 0.22124 | 0.22727 | 0.22421 |

Tabela 32 - Média da ROUGE para CFSumm, LCHSumm e GistSumm.

| Método | Média ROUGE-1 | | | Média ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| **CFSumm** | 0.40106 | 0.45736 | 0.42262 | 0.22012 | 0.26009 | 0.23543 |
| **LCHSumm** | 0.38173 | 0.44618 | 0.40760 | 0.18963 | 0.23473 | 0.20765 |
| **GistSumm** | 0.38373 | 0.45129 | 0.41113 | 0.20516 | 0.25200 | 0.22403 |

Com base nos resultados apresentados na Tabela 32, nota-se que o método CFSumm tem resultados superiores em relação aos outros dois métodos. Em princípio, isso pode ser explicado pelo fato de que a medida ROUGE apenas compara n-gramas, ou seja, compara as palavras existentes nos sumários humanos com os sumários gerados pelos métodos. Sendo assim, de acordo com a ROUGE, o método CFSumm é o que contém mais informações provenientes dos sumários humanos a nível de palavras. Em contrapartida, observando-se os dados apresentados nas Tabelas 29, 30 e 31, os métodos CFSumm e LCHSumm apresentaram resultados iguais para os *clusters* C1 e C31, e o *baseline* GistSumm apesar de também possuir resultados iguais para o *cluster* C31, obteve valores inferiores tratando-se do *cluster* C1.

Apesar de o método LCHSumm apresentar resultado inferior quando comparado ao GistSumm, isso pode ser justificado pelo fato de que o *cluster* C37, por exemplo, continha mais informações em relação aos outros *clusters* o que permitiu a escolha de informações que não estavam presentes nos sumários humanos no que diz respeito ao número de palavras. Sendo assim, é provável que os conceitos selecionados pelo método em questão apresentem informações distintas quando comparados as sumários humanos.

A informatividade de um sumário é comumente avaliada pela medida ROUGE, no entanto, há outras propriedades textuais que a ROUGE não é capaz de julgar, as quais influenciam a qualidade dos sumários.

De acordo com a TAC, a qualidade de um sumário pode ser avaliada em função das seguintes propriedades: (i) gramaticalidade, que diz respeito à ausência de erros de ortografia, pontuação e sintaxe, (ii) não redundância, que se refere à ausência de informações repetidas, (iii) clareza referencial, que diz respeito à clara identificação dos componentes da superfície textual que fazem remissão a outro(s) elemento(s) do universo textual, (iv) foco, que se refere ao fato de que as informações de uma sentença devem se relacionar com as informações do restante do sumário, e (v) estrutura e coerência, que diz respeito à organização do sumário considerando sua textualidade.

Com o intuito de avaliar a qualidade, os 9 sumários gerados pelos métodos CFSumm, LCHSumm e pelo *baseline* GistSumm foram avaliados manualmente quanto às propriedades especificadas pela TAC.

A avaliação das propriedades relativas à qualidade foi realizada por 9 linguistas computacionais. Para cada um dos 9 sumários selecionados para a avaliação, os juízes pontuaram cada uma das 5 propriedades textuais por meio de um formulário *online*. Para todas as propriedades, os juízes dispunham de uma escala de 1 a 5 pontos, a qual está descrita no Quadro 18.

Quadro 18 - Pontuações e níveis para a avaliação da qualidade linguística.

| **Pontuação** | **Nível** |
|---|---|
| 1 | Péssimo |
| 2 | Ruim |
| 3 | Regular |
| 4 | Bom |
| 5 | Excelente |

Os resultados da avaliação manual da gramaticalidade, não redundância, clareza referencial, foco e estrutura/coerência são apresentados nas Tabelas 30, 31, 32, 33 e 34, respectivamente.

O valor de cada célula das tabelas indica a quantidade de avaliações que cada nível recebeu (cf. Quadro 18) conforme a propriedade em questão, para os três métodos. Para tanto, apresentam-se os valores de duas formas: (i) absoluto, e (ii) porcentagem. Para cada propriedade, calculou-se a média ponderada das avaliações, isto é, o nível "péssimo" recebeu peso 1, para o nível "ruim" considerou-se o peso 2, o nível "regular" recebeu peso equivalente a 3, o nível "bom" passou a ter peso 4 e o nível "excelente" obteve peso equivalente a 5. Portanto, quanto mais próxima de 5 for a média, melhor o resultado, e, quanto mais próxima de 1, pior.

Na Tabela 33, por exemplo, observa-se que a gramaticalidade dos 3 sumários gerados pelo método CFSumm: (i) não recebeu as pontuações 1 ("péssimo") e 2 ("ruim"), (ii) recebeu 2 vezes a pontuação 3 ("regular"), isto é, 7,4% do total; (iii) recebeu 1 vez a pontuação 4 ("bom"), ou seja, 3,7% do total, (iv) recebeu 24 vezes a pontuação 5 ("excelente"), equivalente a 88,9% do total. Com isso, a gramaticalidade teve média ponderada de 4,8, revelando que os juízes a consideram de nível "excelente" na média, já que a pontuação 4,8 está mais próxima de 5.

Tabela 33 - Pontuações dos métodos: critério de “gramaticalidade”

| | **Gramaticalidade** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Péssimo (1) | | Ruim (2) | | Regular (3) | | Bom (4) | | Excelente (5) | | Média |
| **CFSumm** | 0 | 0% | 0 | 0% | 2 | 7,4% | 1 | 3,7% | 24 | 88,9% | 4,8 (**excelente**) |
| **LCHSumm** | 0 | 0% | 1 | 3,7% | 1 | 3,7% | 3 | 11,1% | 22 | 81,5% | 4,7 (**excelente**) |
| **GistSumm** | 0 | 0% | 0 | 0% | 1 | 3,7% | 4 | 14,8% | 22 | 81,5% | 4,8 (**excelente**) |

Na média, a gramaticalidade dos sumários gerados pelo métodos CFSumm e LCHSumm e pelo *baseline* GistSumm foram identificadas como “excelente”, posto que eles receberam a pontuação média de 4,8, 4,7 e 4,8 respectivamente. Tal fato significa que tanto os métodos CFSumm e LCHSumm, bem como o *baseline* GistSumm geraram sumários extrativos com poucos problemas de ortografia, pontuação e sintaxe. Entretanto, a gramaticalidade poderia ser desconsiderada na avaliação, pois os problemas presentes nos extratos são integralmente advindos dos textos-fonte, já que nenhum dos métodos gera abstratos.

Tabela 34 - Pontuações dos métodos: critério de “não-redundância”

| | **Não-redundância** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Péssimo (1) | | Ruim (2) | | Regular (3) | | Bom (4) | | Excelente (5) | | Média |
| **CFSumm** | 0 | 0% | 0 | 0% | 3 | 11,1% | 3 | 11,1% | 21 | 77,8% | 4,7 (**excelente**) |
| **LCHSumm** | 0 | 0% | 1 | 3,7% | 2 | 7,4% | 3 | 11,1% | 21 | 77,8% | 4,6 (**excelente**) |
| **GistSumm** | 0 | 0% | 7 | 25,9% | 2 | 7,4% | 3 | 11,1% | 15 | 55,5% | 3,6 (**bom**) |

Quanto ao critério “não redundância” (Tabela 34), os métodos baseados em conhecimento léxico-conceitual, apresentam resultados bem distintos dos produzidos pelo *baseline.* No caso, a “não redundância” dos sumários gerados pelos métodos CFSumm e LCHSumm receberam a pontuação média de 4,7 e 4,6, respectivamente, ou seja, “excelente”. Os sumários gerados pelo *baseline* GistSumm, por sua vez, receberam a pontuação média de 3,6 (“bom”). Essa diferença pode ser justificada pelo fato de que os métodos CFSumm e LCHSumm englobam um processo de eliminação da redundância baseado na CST. Em contrapartida, o GistSumm não incorpora um processo de remoção de redundância. Castro Jorge (2010), aliás, comprova que métodos

que se baseiam nas relações CST para tratar a redundância dos sumários melhoram significativamente os resultados que se referem a essa propriedade.

Tabela 35 - Pontuações dos métodos: critério de “clareza referencial”

| | **Clareza referencial** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Péssimo (1) | | Ruim (2) | | Regular (3) | | Bom (4) | | Excelente (5) | | Média |
| **CFSumm** | 0 | 0% | 1 | 3,7% | 3 | 11,1% | 8 | 29,6% | 15 | 55,5% | 4,4 (**bom**) |
| **LCHSumm** | 0 | 0% | 2 | 7,4% | 4 | 14,8% | 7 | 25,9% | 14 | 51,8% | 4,2 (**bom**) |
| **GistSumm** | 0 | 0% | 4 | 14,8% | 4 | 14,8% | 8 | 29,6% | 11 | 40,7% | 4,0 (**bom**) |

Quanto à propriedade “clareza referencial” (Tabela 35), os sumários gerados pelos métodos CFSumm, LCHSumm e GistSumm receberam pontuações médias similares, a saber, 4,4, 4,2 e 4,0 respectivamente. Essas pontuações indicam que os sumários apresentam “bom” nível de “clareza referencial”, apesar de os métodos não realizarem qualquer processo de resolução de correferência.

Tabela 36 - Pontuações dos métodos: critério de “foco”

| | **Foco** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Péssimo (1) | | Ruim (2) | | Regular (3) | | Bom (4) | | Excelente (5) | | Média |
| **CFSumm** | 0 | 0% | 1 | 3,7% | 4 | 14,8% | 10 | 37,3% | 12 | 44,4% | 4,3 (**bom**) |
| **LCHSumm** | 0 | 0% | 2 | 7,4% | 8 | 29,6% | 6 | 22,2% | 11 | 40,7% | 4,0 (**bom**) |
| **GistSumm** | 0 | 0% | 3 | 11,1% | 7 | 25,9% | 4 | 14,8% | 13 | 48,1% | 4,0 (**bom**) |

Com base na Tabela 36, observa-se que o “foco” dos sumários gerados pelos métodos CFSumm e LCHSumm receberam a pontuação média de 4,3 e 4,0, respectivamente, sendo, portanto, considerado de nível “bom” por apresentarem sentenças moderadamente relacionadas. O mesmo se dá em relação ao *baseline* GistSumm, que recebeu pontuação 4,0. Vale ressaltar que, quanto maior forem os resultados obtidos no quesito foco, menor a probabilidade de os sumários apresentarem sentenças com conteúdo disperso ou pouco relacionado.

Tabela 37 - Pontuações dos métodos: critério de “estrutura e coerência”

| | **Estrutura e coerência** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Péssimo (1) | | Ruim (2) | | Regular (3) | | Bom (4) | | Excelente (5) | | Média |
| **CFSumm** | 0 | 0% | 1 | 3,7% | 5 | 18,5% | 5 | 18,5% | 16 | 59,2% | 4,2 (**bom**) |
| **LCHSumm** | 0 | 0% | 4 | 14,8% | 5 | 18,5% | 9 | 33,3% | 9 | 33,3% | 3,9 (**bom**) |
| **GistSumm** | 2 | 7,4% | 7 | 25,9% | 3 | 11,1% | 4 | 14,8% | 11 | 40,7% | 3,6 (**bom**) |

Na Tabela 37, observa-se que a propriedade “estrutura e coerência” dos sumários gerados pelos métodos CFSumm e LCHSumm receberam as pontuações médias de 4,2 e 3,9, respectivamente. Os sumários gerados pelo *baseline* GistSumm receberam a pontuação média de 3,6. Nesses casos, a estrutura e a coerência foram consideradas de nível “bom” para todos os sumários, porém os métodos CFSumm e LCHSumm obtiveram resultados superiores em relação aos sumários gerados pelo *baseline*. Por se tratar de sumários extrativos, isto é, compostos por sentenças extraídas na íntegra dos textos-fonte, a ausência de reescrita pode ter prejudicado a estrutura dos sumários, os quais, por isso, apresentam sentenças desconexas umas com as outras.

No geral, os sumários gerados pelos métodos CFSumm e LCHSumm apresentam melhor qualidade em relação aos sumários gerados pelo *baseline* GistSumm, e especificamente, comparando-se os dois métodos desenvolvidos nesta pesquisa, o CFSumm obteve resultado superior.

## 8. Considerações finais

Neste trabalho, propuseram-se métodos baseados em hierarquias conceituais para a seleção de conteúdo na SAM extrativa. Tais métodos contribuiram para o avanço da SAM, pois, até o momento, poucos trabalhos para o português envolvem esse tipo de conhecimento profundo.

Quanto à hipótese (a), isto é, a de que a organização léxico-conceitual de uma coleção de textos-fonte sobre um mesmo assunto reflete de fato o fenômeno multidocumento da redundância, ressalta-se que tal fenômeno é o único que se manifesta na modelagem hierárquica e é capturado através do peso que é dado aos conceitos por meio das métricas de relevância. Apesar da particularidade de cada métrica, todas elas têm como objetivo final dar peso aos conceitos, e desta forma, o

resultado é a busca pelas informações mais importantes que são transmitidas pela redundância.

Com relação à hipótese (b), de que as métricas relativas às estruturas léxico-conceituais, representadas em grafos, permitem de fato identificar o conteúdo principal da coleção dos textos-fonte, ressalta-se que as métricas permitiram no geral identificar os conceitos mais importantes das coleções, considerando fatores como a frequência simples e acumulada dos conceitos, a especificidade/generalidade dos conceitos e a proximidade de conceitos relevantes.

Sobre a hipótese (c), a de que as propriedades da estrutura léxico-conceitual de fato possibilitam a seleção de conteúdo mais importante da coleção, destaca-se que esta se confirmou, o que foi comprovado pela medida ROUGE. Além de produzir sumários informativos, as propriedades codificadas pelas métricas também são capazes de produzir sumários com boa qualidade linguística.

Por fim, pode-se afirmar que a aplicação do método profundo CFSumm apresentou resultados satisfatórios quando comparados ao *baseline* GistSumm. Portanto, o conhecimento léxico-conceitual aplicado em conjunto com a medida de frequência de ocorrência dos conceitos nos textos (*Simple Frequency*) extrai informações relevantes para a construção dos sumários, em função da relação semântica de sinonímia proveniente da WN.Pr., isto é, as avaliações demonstraram que o método CFSumm consegue superar o *baseline* pois agrega não somente informações superficiais como a frequência, mas também as relações semânticas entre as palavras com mesmo sentido, dada uma coleção de textos que tratam de um mesmo assunto. Além disso, a utilização das métricas agregadas ao conhecimento léxico-conceitual, permitiu o desenvolvimento do método LCHSumm que vai além do conhecimento superficial da frequência e do conhecimento profundo da sinonímia. O LCHSumm se difere dos outros métodos pelo fato de que trata informações semânticas e profundas como a especifidade/generalidade dos conceitos e a proximidade dos conceitos com outros conceitos considerados importantes, através de métricas de grafos aplicadas a textos, sobretudo, em português.

## 8.1. Limitações

Destaca-se aqui a complexa tarefa de modelagem léxico-conceitual dos textos-fonte em formato de árvore. A complexidade resultou de vários fatores. Um deles foi lidar com o conhecimento léxico-conceitual, que envolveu as tarefas subjetivas de tradução e

indexação léxico-conceitual. Outro fator foi a necessidade de lidar com ferramentas computacionais distintas para a realização dos processos intermediários, cujos formatos de entrada e saída não eram totalmente compatíveis, requerendo certo pré-processamento dos dados a cada etapa.

### 8.2. Contribuições

Neste trabalho, realizou-se a primeira pesquisa envolvendo a utilização de medidas de grafo e conhecimento léxico-conceitual na SAM para o português. Além disso, a investigação de tais métodos envolvendo conhecimento léxico-conceitual contribuiu para a extensão das pesquisas na área, que ainda são incipientes.

Contudo, a indexação dos nomes de 3 coleções do CSTNews aos *synsets* da WN.Pr permite que a anotação final possa ser utilizada em outros trabalhos para a futura exploração de diferentes medidas de grafo.

Por fim, pode-se afirmar que a frequência de ocorrência dos conceitos, já investigada por Tosta (2014), é eficiente para a seleção de sentenças a compor extratos multidocumento.

### 8.3. Trabalhos Futuros

Quanto aos trabalhos futuros, ressalta-se a importância de se estender o *corpus* de estudo por meio da indexação léxico-conceitual e representação em árvores de outros *clusters* do CSTNews, tendo em vista o reduzido tamanho do *corpus* utilizado neste trabalho. Além disso, ressalta-se o interesse em investigar a pertinência das medidas em função do assunto, tamanho ou complexidade dos textos-fonte. Em outras palavras, tem-se o interesse em investigar se um conjunto de medidas pode ser determinado como relevantes considerando o assunto, tamanho ou complexidade dos textos-fonte.

## Referências bibliográficas

AKABANE, A. T.; PARDO, T. A. S.; RINO, L. H. M. **Explorando medidas de redes complexas para sumarização multidocumento.** In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2, 2011, Cuiabá. **Proceedings…**Cuiabá, MT, 2011, p.1-3.

ALEIXO, P. PARDO, T. A. S. CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST. **Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação**, NILC-TR-08-05, n. 326. São Carlos, SP, junho, 2008, 15p.

ALENCAR

, V. B. **Uma ferramenta para Sumarização de Ontologias**. Recife, PE, 2008. Monografia de Conclusão de Graduação – Centro de Informática, Universidade Federal de Pernambuco (Cin/UFPE), 31p., 2008.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American,** may, 2001. Disponível em: <http http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html >. Acesso em: 15 ago. 2015.

BLACKBURN, S. **Dicionário Oxford de Filosofia**. Rio de Janeiro: Jorge Zahar, 1997, 438 p. ISBN 85-711-0402-6.

BONDY, J. A.; MURTY, U. S. R. **Graph Theory with applications**. Elsevier Science Ltd/North-Holland, 1976, 264 p.

BORGATTI, S. P.; EVERETT, M. G. A Graph-theoretic perspective on *Centrality*. **Social Networks**, v. 28, Issue 4, p.466–484, 2006

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. Enschede, Netherlands, 1997. Thesis (PhD in Information and Knowledge Systems) – University of Twente, 243 p., 1997.

BRANCO, A; SILVA, J. **Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese.** In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4, 2004, Lisbon. **Proceedings...** Lisbon, Portugal, 2004, p. 507-510.

CABEZUDO, M. A. S. **Investigação de métodos de desambiguação lexical de sentidos de verbos do português do Brasil**. São Carlos, SP, 2015. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (ICMC/USP), São Carlos, SP, 134p., 2015

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. São Carlos, SP, 2013. Dissertação de Mestrado, Programa de Pós-Graduação em Linguística - Universidade Federal de São Carlos (PPGL/UFSCar), São Carlos, SP, 133 p., 2013.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. São Carlos, SP 2014, Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP), São Carlos, SP, 182 p., 2014.

CARDOSO, P. C. F.; MAZIERO, E. G.; CASTRO JORGE, M. L. R.; SENO, E. M. R.; DI-FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. **CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese.** In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…** Cuiabá, MT, 2011a, p. 88-105.

CARDOSO, P. C. F.; PARDO, T. A. S.; NUNES, M. G. V. **Métodos para sumarização automática multidocumento usando modelos semântico-discursivos.** In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings...**Cuiabá, MT, 2011b, p. 59-74.

CASTRO JORGE, M. L. R.; AGOSTINI, V.; PARDO, T. A. S. **Multi-document summarization using complex and rich features.** In: ENCONTRO NACIONAL DE

INTELIGÊNCIA ARTIFICIAL, 8, 2011, Natal. **Proceedings…** Natal, RN, 2011, p. 1-12.

CASTRO JORGE, M. L. R. **Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST** (*Cross-document Structure Theory*). São Carlos, SP, 2010. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (ICMC/USP), São Carlos, SP, 86 p., 2010.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. **A Generative approach for multi-document summarization using the Noisy Channel model.** In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…** Cuiabá, MT, 2011, p. 75-87.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. **Experiments with CST-based Multi-document summarization.** In: ACL WORKSHOP TEXTGRAPHS-5: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010, Uppsala. **Proceedings…**Uppsala, Sweden, 2010, p. 74-82.

CASTRO JORGE, M. L. R. ***Modelagem gerativa para sumarização automática multidocumento*.** São Carlos, SP, 2015. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP). São Carlos, SP, 151 p., 2015.

CHENG, G., GE, W.; QU, Y. **Generating summaries for ontology search.** In: THE INTERNATIONAL CONFERENCE COMPANION ON WORLD WIDE WEB – WWW'11, 20, 2011, New York. **Proceedings…**New York, New York, USA: ACM Press, 2011, p. 27.

CREMMINS, E. T. **The art of abstracting.** Arlington: Information Resources Press, 1996, 230 p.

CRUSE, A. **Lexical semantics.** Cambridge: Cambridge University Press, 1986, 310 p.

DANG, H. T. **Overview of DUC 2005.** In: Proceedings of the Document Understanding Conference, 2005.

DIAS DA SILVA, B. C.; **A construção da base wordnet.br: Conquistas e desafios.** In: WORKSHOP IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 3, 2005. **Proceedings…**Sociedade Brasileira de Computação, São Leopoldo, RS, Brasil, 2005, p. 2238-2247.

DIESTEL, R. **Graph Theory** (Graduate Texts in Mathematics), Springer, 3rd edition, 2005, 415 p.

DI-FELIPPO, A. **Delimitação e alinhamento de conceitos lexicalizados no Inglês Norte-americano e no Português Brasileiro.** Araraquara, SP, 2008. Tese de Doutorado (Doutorado em Linguística), Faculdade de Ciências e Letras - Universidade Estadual Paulista (Unesp/FCLAr), Araraquara, SP, 253 p., 2008.

ENDRES-NIGGEMEYER, B. **Summarization Information**. Berlin: Springer, 1998.

FELLBAUM, C. (Ed.). **WordNet: an electronic lexical database.** Cambridge, MA: MIT Press, 1998, 423 p.

FREEMAN, L. C. *Centrality* in social networks conceptual clarification. **Social Networks**, v. 1, Issue 3, p. 215–239, 1978.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Human-Computer Studies**, v. 43, n. 5-6, p. 907-928, 1995.

GONÇALO OLIVEIRA, H.; ANTÓN PÉREZ, L,; GOMES, P. **Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese**. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE PROCESSIN AND INFORMATION, 17 **Proceedings…**Springer-Verlag, Berlin, Alemanha, 2012, p. 210-215.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H., The *WEKA* Data Mining Software: An Update, **SIGKDD Explorations**, v. 11, Issue 1, p. 1-9, 2009.

HALTEREN, H.; TEUFEL, S. **Examining the consensus between human summaries: initial experiments with factoid analysis.** In: HLTNAACL DUC WORKSHOP, 2003, Edmonton. **Proceedings…** Edmonton, AB, Canada, 2003, 1-8 p.

HEARST, M. **TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages**, Computational Linguistics, 1997, v. 23, n. 1, p. 33-64, 1997.

HENNIG, L., UMBRATH, W., WETZKER, R. **An ontology-based approach to Text Summarization.** In: WORKSHOP ON NATURAL LANGUAGE PROCESSING AND ONTOLOGY ENGINEERING (NLPOE 2008), 3, 2008, Toronto. **Proceedings…**Toronto, Canada, 2008, p. 291-294.

HORRIDGE, M.; HOLGER, K.; RECTOR, K; STEVENS, R.; WROE, C.; JUPP, S.; MOULTON, G.; DRUMMONT, N.; BRANDT, S. **A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools.** The University Of Manchester, 2004. Disponível em: <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf>. Acesso em: 14 ago. 2015.

KNUBLAUCH, H.; FERGERSON, R.; NOY, N. F.; MUSEN M. A. **The protégé OWL plugin: an open development environment for semantic web applications.** In: INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC2004), 3, 2004, Hiroshima. **Proceedings...** (Lecture Notes in Computer Science, 3298), Hiroshima, Japan, Springer, 2004, p. 229-243.

KUMAR, Y. J.; SALIM, N. Automatic multi-document summarization approaches. **Journal of Computer. Science**, Malaysia, 8, 2012, p. 133-140.

LI, L., WANG, D., SHEN, C., LI, T. **Ontology-enriched multi-document summarization in disaster management.** In: ACM SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR), 2010, Geneva. **Proceedings…** Geneva, Switzerland, 2010, p. 819-820.

LI, L.; WANG, D.; SHEN, C.; LI, T. **Ontology-enriched multi-document summarization in disaster management.** In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 33, 2010, Geneva. **Proceedings**... New York, NY: ACM, 2010. p. 819-820.

LI, N., MOTTA; E.; AQUIN, M. **Ontology Summarization: an analysis and an evaluation.** In: INTERNATIONAL WORKSHOP ON EVALUATION OF SEMANTIC TECHNOLOGIES (IWEST 2010), 9, 2010, Shanghai. **Proceedings**...Shangai, China, 2010, p. 1–12.

LIN, C. Y.; HOVY, E. **Automated multi-document summarization in NeATS.** In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2, 2002, San Diego. **Proceedings…** San Diego, California, San Francisco: Morgan Kaufmann Publishers Inc., 2002, p. 59-62.

LIN, C. Y. **ROUGE: a Package for Automatic Evaluation of Summaries.** In: Workshop ON TEXT SUMMARIZATION BRANCHES OUT (WAS 2004), 8, 2004, Barcelona. **Proceedings…** Barcelona, Spain, 2004, p. 74-81.

LOUIS, A. NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, Cambridge, MA, v. 39, n. 2, p. 267-300, 2013.

MAGNINI, B., SPERANZA, M. **Merging global and specialized linguistic ontologies.** In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION - LREC, 3, 2002, Las Palmas. **Proceedings...** Las Palmas: University of Las Palmas, Espanha, 2002, 1-6 p.

MANI, I. **Automatic summarization.** Amsterdam: John Benjamins Publishing Co., 2001, 286 p.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. Cambridge, MA: MIT Press, 1999.

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: a theory of text organization. **Technical Report ISI/RS-87-190**, 1987.

MARCU, D. **The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.** Toronto, Canada, 1997. Thesis (PhD in Philosophy) – University of Toronto – Graduate Department of Computer Science, 351 p., 1997.

MAZIERO, E. G., PARDO, T. A. S. **Multi-document discourse parsing using traditional and hierarchical machine learning.** In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8, 2011, Cuiabá **Proceedings…** Cuiabá, MT, 2011, p. 1-10.

METZ, J.; CALVO, R.; SENO, E. R. M.; ROMERO, R. A. F.; LIANG, Z. Redes Complexas: conceitos e aplicações. **Relatório Técnico do Instituto de Ciências Matemáticas e de Computação**, n. 290. ISSN - 0103-2569, 2007, janeiro, 2007.

MIHALCEA, R.; TARAU, P. **An algorithm for language independent single and multiple document Summarization.** In: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP), 2, 2005, Korea. **Proceedings…** Korea, 2005, p.19-21.

MORATO, J.; MARZAL, M. A.; LLORENS, J.; MOREIRO, J. **WordNet Applications.** In: GLOBAL WORDNET CONFERENCE, 2, 2004, Brno. **Proceedings...** Brno, Czech Republic, 2004. p. 270-278.

NENKOVA, A.; PASSONNEAU, R. **Evaluating content selection in summarization: The pyramid method.** In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (HLT/NAACL), 2004, Boston. **Proceedings...** Boston, MA, 2004, 1-8 p.

NEWMAN, M. **Networks: An Introduction** (1st ed.), Oxford, UK: Oxford University Press, 2010, 784p.

NÓBREGA, F. A. A. **Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento.** São Carlos, SP, 2013, Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP) São Carlos, SP, 126 p., 2013.

OPSAHL, T., AGNEESSENS, F.; SKVORETZ, J. Node *Centrality* in weighted networks: Generalizing degree and shortest paths. **Social Networks**, v. 32, Issue 3, p. 245– 251, 2010.

ORĂSAN, C. **Automatic summarization in the informational age.** In: RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING – INTERNATIONAL CONFERENCE (RANLP - 2009), 7, 2009, Borovets. **Proceedings...** Stroudsburg, PA: Association on Computational Linguistics, Borovets, Bulgaria, 2009.

OTTERBACHER, J. C.; RADEV, D. R.; LUO, A. **Revisions that improve cohesion in multi-document summaries: a preliminary study.** In: WORKSHOP ON AUTOMATIC SUMMARIZATION, 2002, Stroudsburg. **Proceedings…** Stroudsburg, Pennsylvania, 2002, p 27-36.

PARDO, T. A. S. GistSumm - GIST SUMMarizer: extensões e novas funcionalidades. **Série de Relatórios do NILC.** NILC-TR-05-05, São Carlos, SP, p.8, fevereiro, 2005.

PARDO, T. A. S.; NUNES, M. G. V. On the development and Evaluation of a Brazilian Portuguese Discourse Parser. **Journal of Theoretical and Applied Computing**, v. 15, n. 2, p. 43-44, 2008.

PERONI, S.; MOTTA, E; AQUIN, M. **Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures.** In: The SEMANTIC WEB (ISWC-2008). **Proceedings…** Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 242–256.

PIRES, C. E. S. **Um Sistema P2P de Gerenciamento de Dados com Conectividade Baseada em Semântica.** Recife, PE, 2007, Exame de Qualificação e Proposta de Tese, Pós-Graduação em Ciências da Computação – Universidade Federal de Pernambuco - Centro de Informática (UFPE/Cin). Recife, PE, 123 p., 2007.

PIRES, C. E.; SOUSA, P.; KEDAD, Z; SALGADO, A. C. **Summarizing ontology-based schemas in PDMS.** In: 26th INTERNATIONAL CONFERENCE ON DATA ENGINEERING WORKSHOPS (ICDEW 2010). Long Beach, CA: IEEE, 2010, p. 239–244.

RADEV, D. R. **A common theory of information fusion from multiple text sources, step one: Cross-document structure.** In: ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1, 2000, Hong Kong. **Proceedings…**Hong Kong, 2000, p. 74-83.

RADEV, D.; ALLISON , T.; BLAIR-GOLDENSOHN, S.; BLITZER, J.; CELEBI, A.; DIMITROV, S.; DRABEK, E.; HAKIM, A.; LAM, W.; LIU, D.; OTTERBACHER, J.; QI, H.; SAGGION, H.; TEUFEL, S.; TOPPER, M.; WINKEL, A.; ZHANG, Z. **MEAD - A Platform for Multi Document Multilingual Text Summarization.** In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2004), 4, 2004, Lisbon. **Proceedings…** Lisbon, Portugal, 2004, p. 1-4.

RATNAPARKHI, A. **A maximum entropy part-of-speech tagger.** In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING CONFERENCE, 1, 1996, Philadelphia. **Proceedings…** Philadelphia, Pennsylvania, 1996. p. 133-142.

REIMER, U.; HAHN, U. 1988. **Text Condensation as Knowledge Base Abstraction.** In: Conference on Artificial Intelligence Applications, 4, 1988, Washington. **Proocedings...**Washington: IEEE Computer Society Press, 1988, p. 338-344. (1988, apud MANI, 2001).

RIBALDO, R. **Investigação de Mapas de Relacionamento para Sumarização Multidocumento.** São Carlos, SP. 2013. Monografia de Conclusão de Curso, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP). São Carlos, SP, 61 p., 2013.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. **Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information.** In: CONFERÊNCIA INTERNACIONAL DE PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA (PROPOR-2012), 11, 2012, Coimbra. **Proceedings…** Coimbra, Portugal, 2013 p. 1-12.

RIBALDO, R.; PARDO, T. A. S.; RINO, L. H. M. **Sumarização automática multidocumento com mapas de relacionamento.** In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2, 2011, Cuiabá. **Proceedings…**Cuiabá-MT, 2011, p. 1-3.

ROSCH, E, C. B. MERVIS, W. D. GRAY, D. M. JOHNSON; P. BOYES-BRAEM. Basic objects in natural categories. **Cognitive Psychology**, vol. 8, n. 3, p. 382-439, 1976.

RUOHONEN, K. **Graph theory.** Tampere University of Technology, Tampere, Finland, J. Tamminen, K.-C. Lee, & R. Pich, eds., 114 p., 2013.

SALTON, G.; SINGHAL A.; MITRA, M; BUCKLEY C. Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n. 2, p. 193-207, 1997.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, K. **Experiments in multi-document summarization.** In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2, 2002, San Francisco. **Proceedings…** San Francisco, CA, USA, 2002, p. 52-58.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, 27, p. 379–423, 1948.

SILVA, P. P. **ExtraWeb: um sumarizador de documentos web baseados em etiquetas html e ontologia.** São Carlos, SP, 2006. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de São Carlos (PPGCC/UFSCar), São Carlos, SP, 168 p., 2006.

SILVEIRA, S. B.; BRANCO, A. **Combining a Double *Cluster*ing Approach with Sentence Simplification to Produce Highly Informative Multi-document Summaries.** In: INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION (IRI), 13, 2012, Las Vegas. **Proceedings…** Las Vegas: IEEE, NV, USA, 2012.

SINCLAIR, J. Corpus and text: basic principles. In: WYNNE, M. (Ed.). **Developing linguistic corpora: a guide to good practice.** Oxford: Oxbow Books, 2005. p. 1-16. Disponível em: <www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>. Acesso em: 14 ago. 2015.

SOUSA, P. O. V. Q. **Otimização de uma Ferramenta para Sumarização de Ontologias.** Recife, PE, 2011, Monografia de Conclusão de Curso, Centro de Informática – Universidade Federal de Pernambuco (Cin/UFPE), Recife, PE, 43 p., 2011.

SOUSA, P. O. V. Q. **OWLSumBRP: Um Método de Sumarização de Ontologias.** Recife, PE, 2014, Dissertação de Mestrado, Centro de Informática - Universidade Federal de Pernambuco (Cin/UFPE), Recife, PE, 93 p., 2014.

SOUSA, P. O.; PIRES, C. E., SALGADO, A. C. A Method for Building Personalized Ontology Summaries. **Journal of Information and Data Management** (JIDM), v. 4, n. 3, p. 236-250, 2013.

SPARCK JONES, K.; GALLIERS, J. R. **Evaluating Natural Language Processing systems: An analysis and review.** Berlin: Springer-Verlag, 1996.

STAAB, S., STUDER, R. (Eds.). **Handbook on ontologies.** International Handbooks on Information Systems. Berlin, Heidelberg: Springer-Verlag, 2004, p. 275-298.

STUDER, R., BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, Issues 1-2, p.161–197, 1998.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M. S. **Fuzzy Genetic Semantic Based Text Summarization.** In: DEPENDABLE, AUTONOMIC AND SECURE COMPUTING (DASC), IEEE NINTH INTERNATIONAL CONFERENCE ON, 9, 2011, Sydney. **Proceedings...** Washington, DC: IEEE Computer Society, 2011. p. 1184-1191.

TOSTA, F. E. S. **Aplicação de conhecimento léxico-conceitual na Sumarização Automática Multidocumento Multilíngue.** São Carlos, SP, 2014. Dissertação de Mestrado – Programa de Pós-Graduação em Linguística - Universidade Federal de São Carlos (PPGL/UFSCar), São Carlos, SP, 116 p., 2014.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities**, Dordrecht: Kluwer Academic Publichers, v.32, Issue 2-3, p. 73-89, 1998.

WEISZFLOG, W. **Michaelis: moderno dicionário inglês**: inglês-português/ português-inglês. Editora Melhoramentos, 2000. Disponível em <http://michaelis.uol.com.br/moderno/ingles/index.php>. Acesso em: 14 ago. 2015.

WU, C.-W.; LIU, C.-L. Ontology-based Text Summarization for Business News Articles. **Computers and Their Applications**, v. 2003, p. 389-392, 2003.

W3C. **OWL 2 Web Ontology Language Document Overview.** Second Edition, p. 7, 2012. Disponível em: < http://www.w3.org/TR/owl2-overview/>. Acesso em: 14 ago. 2015.

ZHANG, X., CHENG, G., GE, W., QU, Y.: Summarizing Vocabularies in the Global Semantic Web. **Journal of Computer Science and Technology**, v. 24, Issue 1, p.165-174, 2009.

ZHANG, X., CHENG, G., QU, Y.: **Ontology Summarization Based on RDF Sentence Graph.** In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 16, 2007, Banff, **Proceedings...** Banff, Alberta, Canada, 2007, p. 1-9.