UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Murillo Lagranha Flores

SBI: Um Método de Sumarização Extrativa de Texto Baseado em Instâncias de uma Ontologia

Florianópolis
2019

Murillo Lagranha Flores

SBI: Um Método de Sumarização Extrativa de Texto Baseado em Instâncias de uma Ontologia

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.
Orientador: Prof. Dr. Ricardo Azambuja Silveira
Coorientador: Prof. Dr. Elder Rizzon Santos

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Flores, Murillo Lagranha
SBI: Um Método de Sumarização Extrativa de Texto Baseado em Instâncias de uma Ontologia / Murillo Lagranha Flores ; orientador, Ricardo Azambuja Silveira, coorientador, Elder Rizzon Santos, 2019.
115 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2019.

Inclui referências.

1. Ciência da Computação. 2. Sumarização automática de texto. 3. Ontologias. 4. Processamento de linguagem natural. I. Silveira, Ricardo Azambuja. II. Santos, Elder Rizzon. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Murillo Lagranha Flores

SBI: Um Método de Sumarização Extrativa de Texto Baseado em Instâncias de uma Ontologia

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Thiago Alexandre Salgueiro Pardo, Dr.
Universidade de São Paulo

Prof. Renato Fileto, Dr.
Universidade Federal de Santa Catarina

Prof. Roberto Willrich, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Ciência da Computação.

Digitally signed by Jose Luis Almada Guntzel:53378067004
Date: 2019.09.05 17:26:43 -03'00'

Prof. Dr. José Luís Almada Güntzel
Coordenador do Programa

Assinado de forma digital por Ricardo Azambuja Silveira:23122579049
Dados: 2019.09.05 21:14:30 -03'00'

Prof. Dr. Ricardo Azambuja Silveira
Orientador

Florianópolis, 2 de Setembro de 2019.

Este trabalho é dedicado a todos aqueles que ousam sonhar com
um futuro melhor e que tem a coragem e a persistência
necessárias para fazê-lo acontecer.

AGRADECIMENTOS

Gostaria de agradecer imensamente ao meu orientador, Prof. Dr. Ricardo Azambuja Silveira, e ao meu co-orientador, Prof. Dr. Elder Rizzon Santos. Sem a orientação e sabedoria de ambos este trabalho não teria sido possível.

Agradeço também a minha mãe, professora Carmen, por ter me ensinado a ser um homem livre, em alma e pensamento, e a minha noiva, Dra. Kamilla, por ter me ensinado que a felicidade é um bem que se multiplica ao ser dividido. Vocês duas estiveram sempre ao meu lado e foram o suporte que eu precisava para concluir este trabalho. Sem vocês ele não teria sido possível.

Os meus agradecimentos, enfim, a todos os professores com quem tive aula, por terem sido e continuarem sendo a grande fonte de inspiração da minha jornada.

A capacidade de nos surpreendermos é a única coisa de que precisamos para nos tornarmos bons filósofos. (GAARDER, 1998)

RESUMO

A abundância de documentos de texto disponíveis na web, juntamente com a facilidade de encontrar e recuperar tais documentos trazida pelos buscadores, cria a necessidade de se desenvolverem ferramentas computacionais capazes de criar uma versão resumida destes documentos para que se possa capturar a informação presente nos mesmos sem que para isso haja a necessidade de lê-los na íntegra. Um sumarizador automático de texto cria uma versão resumida de um documento ou de um conjunto de documentos. Sumarizadores extrativos selecionam algumas unidades de texto, como parágrafos ou sentenças, do documento ou dos documentos originais para compor o sumário. Existem diversas técnicas empregadas na seleção e extração de sentenças, dentre elas o uso de medidas baseadas na análise semântica das sentenças. Nestas técnicas, a semântica das sentenças geralmente é representada a partir das formalizações encontradas em uma ontologia. Uma ontologia pode formalizar, entre outros, conceitos e indivíduos, que são instâncias destes conceitos. Os métodos de sumarização extrativa no estado-da-arte exploram somente os conceitos definidos nas ontologias para representar a semântica das sentenças, deixando indivíduos de lado. Desta forma, esta dissertação apresenta uma proposta de método de sumarização extrativa que utiliza as instâncias de uma ontologia para representar a semântica das sentenças, bem como uma série de experimentos realizados para avaliar a relevância dos resultados obtidos pelo mesmo na tarefa de sumarização automática de texto. Os resultados indicam que o método proposto alcança resultados relevantes, revelando que a representação semântica proposta para as sentenças é uma alternativa viável no contexto da sumarização automática

Palavras-chave: Sumarização Automática de Texto. Ontologias. Processamento de Linguagem Natural.

ABSTRACT

The abundance of text documents available on the web, coupled with the ease of finding and retrieving such documents brought by search engines, creates the need to develop computational tools capable of creating summary versions of these documents so that the information present in them can be captured by a reader without them having to read the documents in full. An automatic text summarizer creates a shortened version of a document or set of documents. Extractive summarizers will select textual units, such as paragraphs or sentences, from the original document or documents to compose the summary. There are several techniques employed in sentence selection and extraction, including the ones based on the semantic analysis of sentences. In these techniques, sentence semantics are usually represented using formal descriptions found in an ontology. An ontology can formalize, among others, concepts and individuals, which are instances of the concepts. State-of-the-art extractive summarization methods explore only the concepts defined in ontologies to represent sentence semantics, leaving individuals aside. Thus, this dissertation presents a proposal for an extractive summarization method that uses the instances in an ontology to represent sentence semantics, as well as a series of experiments performed to evaluate the relevance of the results obtained by this method in the automatic text summarization task. The results indicate that the proposed method achieves relevant results, revealing that the proposed semantic representation for sentences is a viable alternative in the context of automatic summarization.

Keywords: Automatic Text Summarization. Ontologies. Natural Language Processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EL	*Entity Linking*.
ILS	*Instances Linking System*.
MMR	*Maximal Marginal Relevance*.
NER	*Named Entity Recognition*.
RSL	Revisão Sistemática da Literatura.
SBI	Sumarização baseada em Instâncias.

SUMÁRIO

1 INTRODUÇÃO

A popularização de computadores e dispositivos móveis junto com a universalização do acesso a web, possibilitaram a geração de um número crescente de documentos de texto disponíveis online. Ferramentas de busca são bastante acessíveis e podem recuperar milhares de documentos sobre virtualmente qualquer assunto. Neste contexto de abundância de documentos e facilidade de busca é improvável que haja tempo suficiente para ler todos os documentos recuperados em sua integridade para, a partir dessa leitura, capturar as informações relevantes neles contidas. A necessidade de se utilizar ferramentas que possam, automaticamente, compreender, classificar, indexar e apresentar esta informação de forma clara e concisa surge deste crescimento na disponibilidade de informações na forma de texto e da impossibilidade de se analisar integralmente o grande volume de informações disponíveis (RIBALDO et al., 2012).

A sumarização automática de texto é uma das muitas técnicas dentro da área de processamento de linguagem natural (PLN) que busca solucionar o problema do gerenciamento de grandes quantidades de informação na forma de texto (ANTIQUEIRA et al., 2009). Um sumarizador automático de texto cria automaticamente uma versão resumida de um documento ou de um conjunto de documentos, enquanto mantém a maioria das informações relevantes presentes nestes documentos (UMBRATH; WETZKER; HENNIG, 2008).

As técnicas de sumarização automática de texto dividem-se em dois grandes grupos: técnicas extrativas e técnicas abstrativas. Sumarizadores baseados em técnicas extrativas selecionam algumas unidades de texto, como parágrafos ou sentenças, do documento ou dos documentos originais para compor o sumário (NENKOVA; MCKEOWN, 2012). As sentenças que compõem um sumário extrativo são portanto um subconjunto das sentenças presentes nos documentos originais. Já sumarizadores baseados em técnicas abstrativas criam o sumário parafraseando os documentos originais, mas não necessariamente usando as mesmas sentenças (NENKOVA; MCKEOWN, 2012). Sumarizadores que utilizam técnicas abstrativas empregam sistemas capazes de gerar texto em língua natural para compor o sumário final, o que demanda o uso de técnicas e recursos computacionais sofisticados. Por outro lado, sumarizadores que utilizam técnicas extrativas selecionam e extraem as informações mais relevantes diretamente dos documentos originais, podendo alcançar resultados tão bons quanto os alcançados por sumarizadores utilizando técnicas abstrativas, porém consumindo menos recursos computacionais. O objetivo geral desta dissertação pode ser alcançado tanto com técnicas de sumarização extrativa quanto com técnicas de sumarização abstrativa. A simplicidade conceitual das técnicas extrativas, a sua menor demanda por recursos, e o fato de que o esforço de implementação associado a estas técnicas é usualmente menor do que o associado a técnicas abstrativas faz com que a sumarização extrativa seja o foco desta dissertação.

A seleção de sentenças nos sumarizadores extrativos é usualmente baseada na avaliação de suas características a partir de medidas de avaliação pré-determinadas (DAS; MARTINS, 2007). Estas medidas podem se basear em diferentes propriedades das sentenças, como propriedades relacionadas a frequência e estrutura, por exemplo (DAS; MARTINS, 2007). A quanti-

dade de vezes que uma palavra aparece em um documento é um exemplo de medida (bastante simples) baseada em frequência. A posição em que uma determinada sentença aparece dentro de um parágrafo é um exemplo de medida baseada em estrutura. Alguns estudos exploram a utilização de medidas baseadas na análise semântica das sentenças, obtendo bons resultados. Essas medidas têm como entrada as sentenças do documento acompanhadas de uma representação formal de suas semânticas. Essas representações comumente são construídas a partir de uma ontologia.

Segundo Gruber (1993), "Uma ontologia é uma especificação formal e explícita de uma conceitualização" e é construída a partir de dois elementos básicos e de relações e asserções sobre eles. Os elementos básicos são classes (ou conceitos) e instâncias (ou indivíduos). A partir destes elementos é possível construir desde ontologias simples até ontologias mais sofisticadas. Uma taxonomia é um exemplo de ontologia simples, onde uma hierarquia de conceitos é formalizada. Em ontologias mais sofisticadas, instâncias e asserções são adicionadas a fim de expressar relações mais complexas e restringir e exemplificar as interpretações pretendidas.

Através de uma revisão sistemática da literatura, verificou-se que métodos de sumarização extrativa no estado da arte que utilizam ontologias como base para representar a semântica das sentenças usam um tipo específico de ontologias simples, as taxonomias ou terminologias. Há, portanto, um elemento básico da construção de ontologias que ainda não foi explorado na sumarização extrativa de texto: as instâncias.

As instâncias explicitam descrições e relações que não são explicitadas pelos conceitos. Se usadas para representar a semântica de uma sentença podem, portanto, representar uma parte desta semântica que não pode ser representada por conceitos. Um método de sumarização extrativa que represente a semântica através de instâncias de uma ontologia obterá uma representação diversa de um que utilize conceitos, podendo portanto obter resultados também diversos.

A proposição de um método de sumarização de texto baseado em instâncias descritas em uma ontologia e nas suas relações abre a possibilidade de se explorar a utilização de ontologias construídas automaticamente a partir de descrições estruturadas de indivíduos do mundo real. Estas ontologias não se restringem a um determinado domínio do conhecimento, como usualmente acontece com ontologias construídas manualmente por especialistas. Portanto, a proposição de um método de sumarização que utiliza instâncias na representação semântica das sentenças é um passo para criar sumarizadores independentes de domínio e que explorem melhor as informações semânticas contidas nas ontologias. Estes sumarizadores poderiam ser usados em documentos de diversas áreas do conhecimento, desde que estas estejam representadas na ontologia. Com uma mesma ontologia seria possível, por exemplo, sumarizar uma página de notícias para que um jornalista pudesse rapidamente saber quais datas e fatos são importantes no documento, ou um capítulo de um livro-texto para um estudante que esteja procurando por uma versão condensada daquele material, ou ainda os artigos publicados em uma conferência para que um professor seja capaz de economizar tempo enquanto se atualiza nos seus interesses de pesquisa, por exemplo.

Esta dissertação apresenta a pesquisa realizada com o objetivo de propor um método de sumarização extrativa de texto a partir da representação semântica de sentenças com base em instâncias de uma ontologia e avaliar a relevância de tal método através da análise dos resultados obtidos pelo mesmo na tarefa de sumarização automática de texto. De maneira mais ampla, a análise da relevância de tal método é um primeiro passo na análise da relevância da utilização de instâncias para a representação semântica de sentenças na sumarização automática de texto.

1.1 OBJETIVO GERAL

Este trabalho tem o objetivo de propor um método de sumarização extrativa de texto que utilize as instâncias definidas em uma ontologia para representar a semântica das sentenças durante o processo de sumarização, analisando a relevância dos resultados obtidos por tal método.

1.2 OBJETIVOS ESPECÍFICOS

O Objetivo Geral pode ser alcançado através dos seguintes objetivos específicos:

- Analisar o estado da arte em métodos de sumarização extrativa que utilizem ontologias para representação semântica.
- Representar a semântica de um documento através de instâncias definidas em uma ontologia.
- Comparar semanticamente diferentes documentos a partir de suas representações através de instâncias.
- Integrar a comparação semântica de documentos a partir de suas representações através de instâncias a um método de sumarização extrativa de texto, propondo assim um novo método.
- Analisar a proposta através de sumários gerados por uma implementação deste método a partir de corpus amplamente utilizado para analisar sumarizadores extrativos de texto.

1.3 METODOLOGIA

A metodologia empregada neste trabalho é inspirada na abordagem de pesquisa construtiva - em inglês *design research* ou *design science*. Esta abordagem busca a resolução de problemas através da execução de duas atividades principais: construir artefatos concretos, como modelos, diagramas, planos, etc, que buscam solucionar um problema e avaliá-los. Peffers et al. (2007) propuseram uma metodologia de processo para a execução desta abordagem composta das seguintes 6 atividades.

1. **Motivação e identificação do problema:** nesta etapa, um problema específico de pesquisa deve ser definido e o valor de uma solução para este problema, justificado.

2. **Definir objetivos de uma solução:** nesta etapa, definem-se os objetivos de uma solução a partir da definição do problema e do conhecimento sobre o que é possível e sobre o que é factível.

3. **Projetar e desenvolver:** nesta etapa, constrói-se o artefato que será a solução ou uma parte da solução para problema identificado. Este artefato pode ser um constructo, um modelo, um método, uma implementação ou "novas propriedades de recursos técnicos, sociais ou informacionais" (PEFFERS et al., 2007; JÄRVINEN, 2007).

4. **Demonstrar:** nesta etapa, deve-se demonstrar o uso do artefato construído para solucionar uma ou mais instâncias do problema definido.

5. **Avaliar:** nesta etapa, observa-se e mede-se quão bem o artefato produzido atinge uma solução para o problema.

6. **Comunicar:** nesta etapa, deve-se comunicar o problema e a sua importância, bem como o artefato produzido, sua utilidade, possíveis inovações, o rigor de seu projeto e a sua efetividade a todas as audiências relevantes.

O problema abordado nessa dissertação, como apresentado anteriormente, tem a seguinte definição: "Como é possível utilizar a representação semântica de sentenças através de instâncias de uma ontologia para propor um método de sumarização extrativa de texto, e quão relevantes podem ser os resultados alcançados por tal método?".

Inspirado na metodologia de processo proposta por Peffers et al. (2007) este trabalho adota uma abordagem iterativa e incremental para investigar o problema apresentado. Cada iteração busca atingir resultados completos ou parciais para cada uma das atividades de 2 a 6, visando construir soluções de modo incremental. A motivação de cada iteração não é necessariamente a mesma por trás dos objetivos definidos para essa dissertação, mas os resultados destas iterações são utilizados para alcançar estes objetivos.

- **Iteração 1**

  - Motivação: compreender o estado da arte do uso de ontologias por métodos de sumarização extrativa de texto.
  - Artefato construído: uma revisão sistemática da literatura seguindo a metodologia proposta por Kitchenham (2004).
  - Resultados: identificou-se como são utilizadas as ontologias no processo de sumarização extrativa de texto, de que maneiras estas ontologias são construídas, que resultados alcançam os sumarizadores utilizando estas ontologias, e quais componentes das ontologias são empregados no processo de sumarização de texto. Concluiu-se

que na maioria dos casos as instâncias definidas em uma ontologia são ignoradas no processo de sumarização, sendo esta uma oportunidade de pesquisa. A revisão sistemática da literatura na integra pode ser encontrada no capítulo 3.

- **Iteração 2**

  - Motivação: criar uma representação da semântica de sentenças usando instâncias definidas em uma ontologia.
  - Artefato construído: uma estratégia de anotação de instâncias em trechos de texto.
  - Resultados: foram obtidos resultados parciais para todas as etapas da metodologia construtiva. Nessa iteração, foram identificados os fatores que contribuem positivamente para que um maior número de sentenças seja representada por instâncias, e uma estratégia de construção de representações focada em maximizar tal número foi especificada. Aumentar o número de sentenças que possuem uma representação não-vazia é importante para garantir que os sumarizadores baseados em tais representações tenham a maior quantidade de informação possível disponível em tempo de sumarização. Os resultados alcançados nessa iteração são apresentados na seção 4.1.1.

- **Iteração 3**

  - Motivação: comparar sentenças a partir de suas representações utilizando instâncias.
  - Artefato construído: uma medida de similaridade entre duas sentenças.
  - Resultados: foram obtidos resultados parciais para todas as etapas da metodologia construtiva. Nessa iteração, buscou-se compreender de que forma duas sentenças distintas poderiam ser comparadas a partir de suas representações como um conjunto de instâncias de uma ontologia. O ponto de partida foi a definição de uma medida de similaridade entre duas instâncias, seguido pela definição de uma medida de similaridade entre sentenças. Testes foram realizados para garantir que a medida de similaridade entre sentenças produzia resultados dentro de um conjunto de expectativas pré-definidas. Os resultados dessa iteração podem ser encontrados nas seções 4.1.2 e 4.1.3.

- **Iteração 4**

  - Motivação: empregar as instâncias definidas em uma ontologia na sumarização extrativa de texto.
  - Artefato construído: um método de sumarização extrativa de texto baseado nas instâncias de uma ontologia.

  - Resultados: foram obtidos resultados parciais para todas as etapas da metodologia construtiva. Nesta iteração, buscou-se definir um método de sumarização extrativa de texto a partir dos resultados obtidos nas iterações anteriores. Tal método integra a medida de similaridade entre sentenças definida na terceira iteração ao algoritmo MMR (CARBONELL; GOLDSTEIN, 1998), utilizando também a estratégia de anotação especificada na segunda iteração. Os resultados dessa iteração podem ser encontrados na seção 4.3.

- **Iteração 5**

  - Motivação: definir a melhor arquitetura de software para o método de sumarização proposto.
  - Artefato construído: uma implementação do método de sumarização proposto.
  - Resultados: foram obtidos resultados parciais para todas as etapas da metodologia construtiva. Buscou-se nesta iteração encontrar a melhor forma de lidar com o grande número de comparações entre instâncias necessárias para computar a similaridade entre sentenças e, por consequência, para se construir o sumário no método proposto na iteração anterior. Buscou-se ainda encontrar a melhor forma de armazenar e consultar tais instâncias, definindo assim a arquitetura de software a ser empregada para uma implementação do modelo que fosse capaz de rodar experimentos com tempo e recursos computacionais limitados. Os resultados dessa iteração podem ser encontrados na seção 5.1.

- **Iteração 6**

  - Motivação: avaliar os resultados obtidos pelo método proposto na tarefa de sumarização de texto.
  - Artefato construído: a execução de testes com o sumarizador implementado em dois conjuntos de dados amplamente utilizados para avaliar sumarizadores extrativos.
  - Resultados: foram obtidos resultados parciais para todas as etapas da metodologia construtiva. O objetivo desta iteração foi avaliar os resultados obtidos pela implementação do método proposto a fim de posicioná-lo frente aos demais sumarizadores extrativos existentes na literatura e entender em que casos utilizar instâncias para descrever as semântica de um documento pode melhorar tais resultados. Testes foram executados em dois conjuntos de dados amplamente utilizados para avaliar sumarizadores extrativos. Os resultados obtidos nessa iteração foram descritos no artigo "Ontology-based Extractive Text Summarization: The Contribution of Instances" aceito para a conferência *"20th International Conference on Computational Linguistics and Intelligent Text Processing"*, ainda a ser publicado, e que pode ser encontrado no anexo A.

A figura 1 ilustra como as atividades da metodologia de processo definida por Peffers et al. (2007) se relacionam, com especial destaque para o processo iterativo que se forma entre as atividades do final de cada iteração e as do início da próxima. Este ciclo é ilustrado por uma linha tracejada na figura. Cada vez que essa linha tracejada é percorrida no processo, uma nova iteração se inicia e por isso essas linhas estão ligadas as linhas de transição entre as iterações executadas durante o desenvolvimento dessa dissertação, numeradas de 1 a 6 e ilustradas logo abaixo.

A figura 2 apresenta um diagrama que busca ilustrar como as iterações se conectam logicamente para alcançar o objetivo da pesquisa, de maneira que os resultados obtidos em uma iteração servem de entrada para uma ou mais das iterações seguintes. Dessa forma, ele representa as dependências de uma iteração em relação as demais, explicitando quando os resultados de uma iteração servem como entrada para mais de uma iteração subsequente e também como os resultados de mais de duas iterações são combinados para formar a entrada de uma terceira.

Figura 1 – Atividades da metodologia e como estas são executadas durante cada iteração

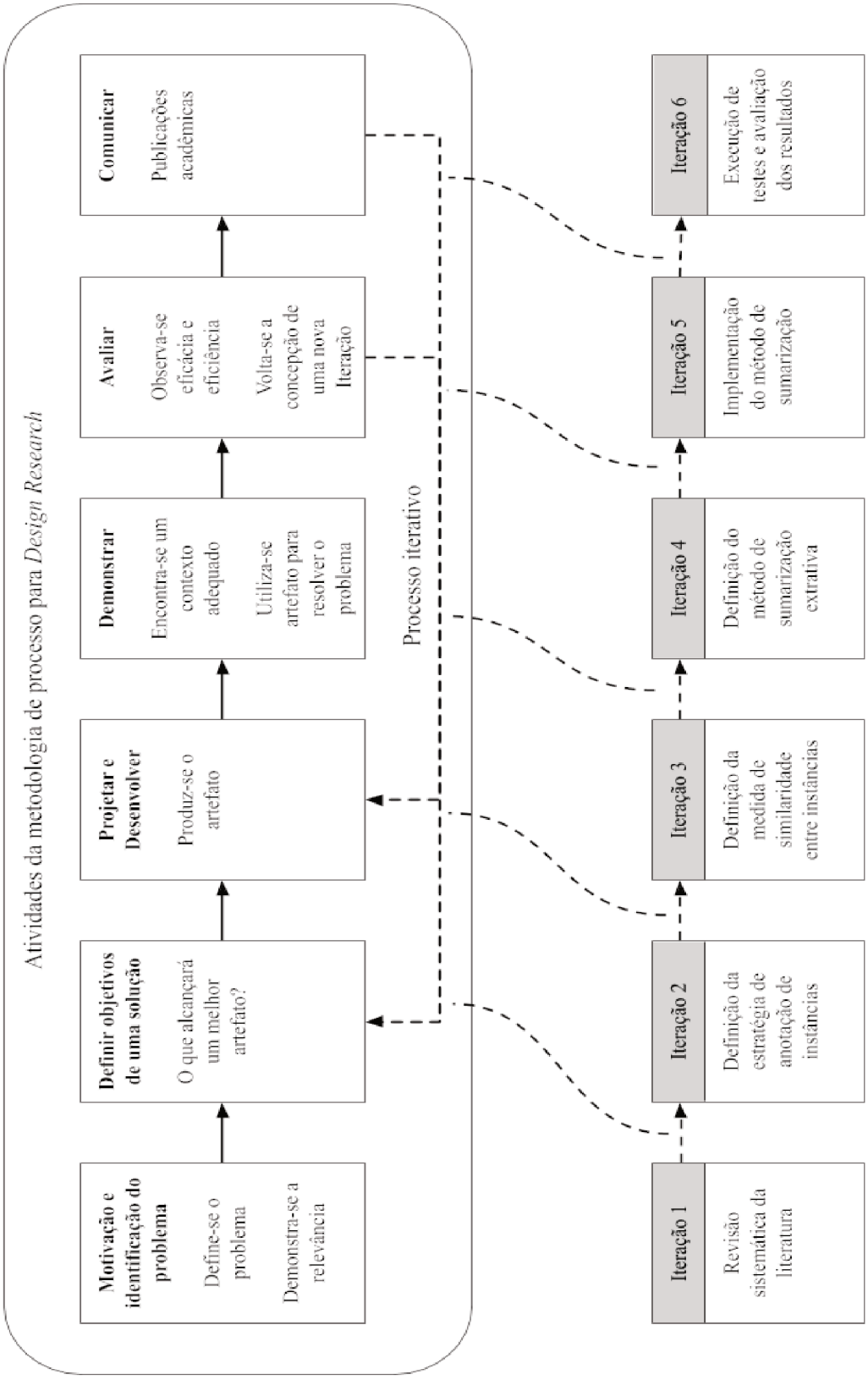
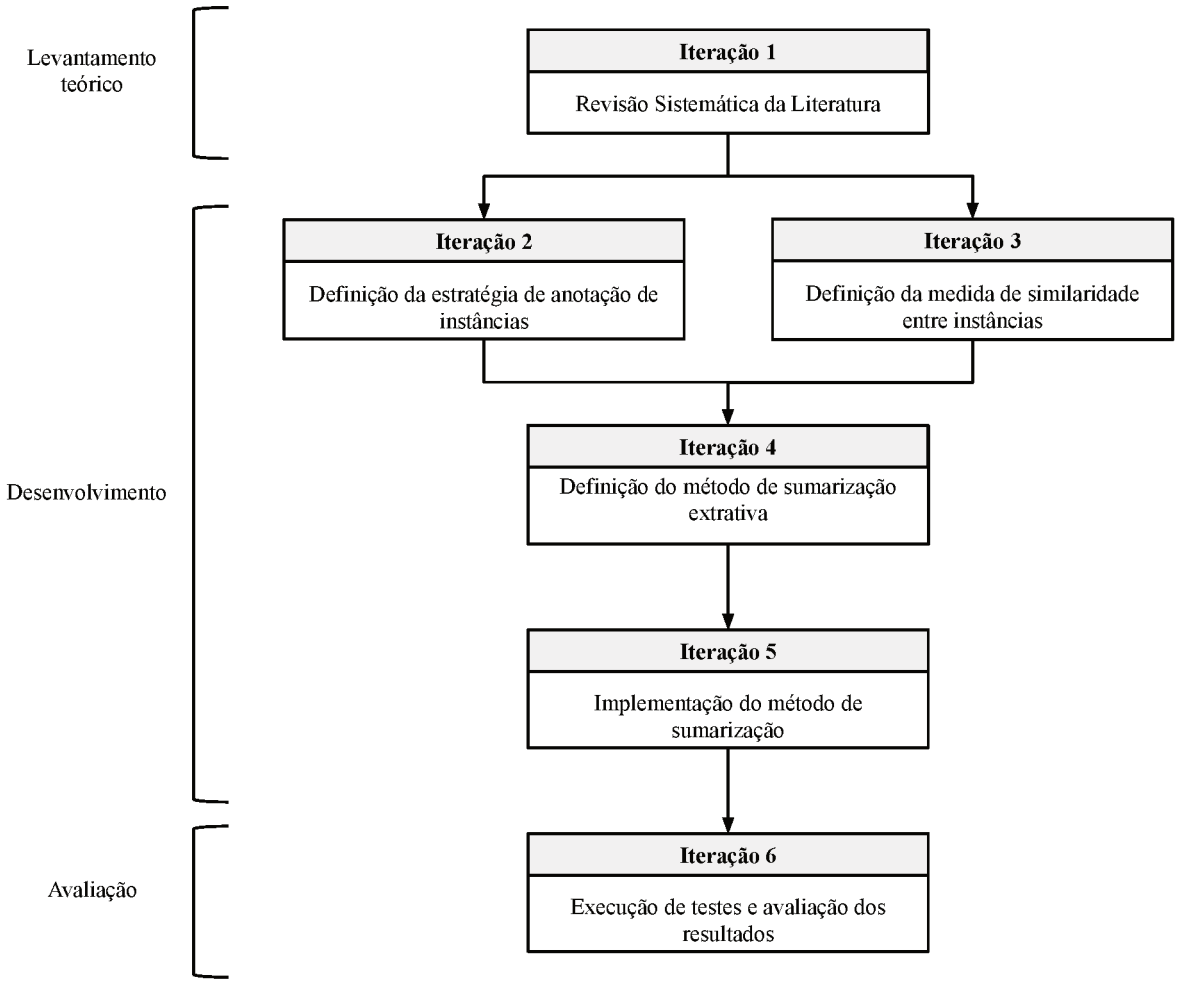


Figura 2 – Diagrama da metodologia - Sequência lógica das iterações.



1.4 ESTRUTURA DA DISSERTAÇÃO

Os próximos capítulos dessa dissertação estão organizados da seguinte forma:

- **Capítulo 2:** Apresenta e exemplifica os principais conceitos relacionados aos assuntos abordados nessa dissertação, principalmente aqueles ligados a sumarização automática de texto e as ontologias.
- **Capítulo 3:** Traz uma breve apresentação do estado da arte em sumarização automática de texto utilizando ontologias como base semântica, obtido através de uma revisão sistemática da literatura.
- **Capítulo 4:** Descreve o método proposto nessa dissertação através da apresentação de suas partes constituintes. Inicialmente, apresenta-se uma definição de como a similaridade semântica entre as sentenças será medida através de sua representação com instâncias de uma ontologia, para depois apresentar-se a utilização desta definição no método proposto a partir e um método base.
- **Capítulo 5:** Apresenta os experimentos realizados para validar o método proposto bem como os resultados alcançados nestes experimentos e uma breve discussão dos mesmos.
- **Capítulo 6:** Por fim, as considerações finais a respeito do trabalho apresentado nessa dissertação são apresentadas neste último capítulo

Além disso, há ainda no corpo deste documento um anexo (anexo A) que apresenta o artigo intitulado “Ontology-based Extractive Text Summarization: The Contribution of Instances”, fruto do mesmo trabalho que deu origem a essa dissertação.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentadas as principais definições necessárias para se compreender o modelo de sumarização automática de texto proposto nesta dissertação. Para tanto, o capítulo está organizado da seguinte forma: A seção 2.1 apresenta a sumarização automática de texto, sendo que as suas subseções apresentam as categorias nas quais modelos para a mesma podem ser classificados (2.1.1) e como os resultados alcançados por um sumarizador automático de texto podem ser avaliados (2.1.2). A seção 2.2 apresenta a definição de ontologia adotada nesta dissertação, sendo que suas subseções apresentam os componentes que formam as ontologias (2.2.1) assim como uma classificação de categorias de ontologias (2.2.2). A seção 2.3 apresenta a definição de duas tarefas de extração de informação que são importantes para entender o processo de ligação de instâncias de ontologias a trechos de texto. Por fim, a seção 2.4 apresenta as considerações finais sobre este capítulo.

2.1 SUMARIZAÇÃO AUTOMÁTICA DE TEXTO

Segundo Radev, Hovy e McKeown (2002), um sumário pode ser definido como "um texto que é produzido a partir de um ou mais textos originais, que contém as informações importantes presentes nos textos originais, que não é maior do que a metade do tamanho dos textos originais e é normalmente muito menor que isso" (tradução livre). Essa definição é bastante simples e pouco específica, segundo indica o próprio autor. No entanto, tentativas de elaborar uma definição mais refinada tendem a gerar discordância dentro da comunidade (DAS; MARTINS, 2007). Mesmo a partir dessa definição simples, é possível delinear três aspectos importantes que caracterizam a pesquisa em sumarização de texto:

- Sumários podem ser produzidos a partir de um único documento ou de múltiplos documentos.
- Sumários devem preservar informações importantes.
- Sumários devem ser curtos.

Dá-se o nome de sumarizadores automáticos de texto a sistemas computacionais capazes de gerar um sumário automaticamente (sem intervenção humana) a partir de um ou mais textos originais. Sumarização automática de texto (SAT), portanto, é o nome dado a atividade desempenhada por tais sistemas. A SAT tem sido investigada pela comunidade de pesquisadores em processamento de linguagem natural (PLN ou NLP da sigla em inglês para *Natural Language Processing*) já desde o final da década de 1950 (DAS; MARTINS, 2007). Os sumarizadores automáticos de texto podem ser classificados em diversas categorias, dependendo das características da SAT que desempenham. Essas categorias, bem como as características que as definem são apresentadas na subseção seguinte.

2.1.1 Categorias de SAT

Assim como ocorre com outras áreas do conhecimento onde existe pesquisa ativa, a SAT está em constante evolução, por isso a classificação de sumarizadores automáticos de texto quanto as suas características também está. Existem, contudo, algumas distinções que podem ser feitas a partir da análise de estudos previamente publicados e de como estes classificam a SAT. A partir dessa ideia, alguns autores, como Jones (1998), propuseram taxonomias que classificam a SAT a partir de diferentes caraterísticas. A ideia desta seção não é a de apresentar uma descrição detalhada de todas estas características e de como podem ser classificadas, mas sim de apresentar uma visão geral que permita classificar o modelo proposto nesta dissertação e compará-lo aos demais modelos existente. Em especial, apresentam-se a seguir definições que dizem respeito a características relacionadas a entrada, ao propósito e a saída da SAT.

A primeira característica através da qual é possível classificar a SAT diz respeito a entrada recebida pela mesma. Se essa entrada é composta por um único documento, diz-se que a SAT é **monodocumento** (*single-document* em inglês), enquanto que se a entrada é composta por mais de um documento, diz-se que é **multidocumento** (em inglês *multi-document*). Esta característica está relacionada aos documentos que se quer sumarizar e não a construção do sumarizador em si, e cada um dos tipos apresenta um conjunto característico de desafios. Na sumarização multidocumento, o desafio está justamente em produzir um sumário coerente a partir de diversas fontes de informação que sobrepõem-se e complementam-se, sendo por vezes contraditórias. Dessa forma, é necessário não apenas identificar e lidar com a redundância, mas também identificar e assegurar que as informações corretas sejam parte do sumário, tornando o mesmo coerente e completo (DAS; MARTINS, 2007).

A SAT também pode ser classificada de acordo com a finalidade a que se destina o sumário a ser produzido. Existem diversas categorias que englobam diferentes possibilidades para esta característica, conforme apresentado por Lloret e Palomar (2012). Duas categorias bastante comuns são a sumarização **genérica** e a sumarização **focada em consulta**. Na primeira, os sumários tem por finalidade representar todos os fatos relevantes presentes nos documentos originais, podendo ser um substituto dos textos-fonte. Na segunda, os sumários tem por finalidade atender a necessidade de um usuário que é expressa através de uma consulta (LUO et al., 2013a) sendo que usuários diferentes podem (e geralmente tem) necessidades diferentes.

Por fim, uma das características mais importantes para a classificação da SAT diz respeito a forma como o sumário é construído, sendo que a depender desta característica da SAT um sumarizador pode ser classificado como **extrativo** ou **abstrativo**. Diz-se que a SAT é **abstrativa** quando o sumário é produzido a partir da geração de texto novo em língua natural, que não necessariamente está presente ou é diretamente relacionado ao texto-fonte mas que ainda assim constitui um sumário do mesmo (LLORET; PALOMAR, 2012). Por outro lado, diz-se que a SAT é **extrativa** quando o sumário é gerado a partir da identificação e extração de trechos do texto-fonte considerados relevantes, usualmente sentenças (DAS; MARTINS, 2007).

A tabela 1 apresenta um resumo das características apresentadas e das classificações

Característica	Tipos
Entrada da SAT	**Monodocumento**: a entrada do sumarizador é composta por apenas um documento. **Multidocumento**: a entrada do sumarizador é composta por mais de um documento.
Finalidade do sumário	**Genérica**: o sumário tem por finalidade representar todos os fatos relevantes presentes nos documentos originais **Focada em consulta**: o sumário tem por finalidade atender a necessidade de um usuário que é expressada através de uma consulta
Construção do sumário	**Extrativa**: o sumário é gerado a partir da identificação e extração de trechos do texto-fonte considerados relevantes, usualmente sentenças **Abstrativa**: o sumário é produzido a partir da geração de texto novo em língua natural.

Tabela 1 – Resumo de características de um sumarizador automático de texto e de categorias associadas a estas características.

possíveis apresentadas de acordo com estas características.

O modelo proposto e apresentado nessa dissertação é um modelo de sumarização **extrativa, multidocumento e focada em consulta**.

2.1.2 Avaliação de Sumários

Segundo Jones (2007), a avaliação de sumários pode ser classificada, de acordo com a forma como ocorre, em **intrínseca** ou **extrínseca**. Uma avaliação intrínseca avalia o sumário por si mesmo, através da verificação da sua qualidade e informatividade. Uma avaliação extrínseca, por outro lado, busca avaliar o sumário em seu uso em tarefas específicas, diferentes da sumarização automática, como por exemplo na geração de respostas a perguntas específicas, classificação de documentos ou recuperação de informação. Tanto na avaliação intrínseca quanto na avaliação extrínseca podem ser usadas medidas de avaliação calculadas automaticamente ou manualmente (emitida por avaliadores humanos).

Alguns exemplos de critérios de avaliação intrínseca são:

- Qualidade linguística ou legibilidade do sumário.
- Informatividade ou a cobertura de um conteúdo específico no sumário, em relação a uma informação que necessariamente deveria estar presente.
- A concisão ou o quão não redundante é o sumário.

Alguns exemplos de critério de avaliação extrínseca são:

- O quanto a indexação dos sumários, ao invés dos documentos originais, melhora a recuperação de informação.
- Taxa de acerto na classificação de documentos, quando classificados através de seus sumários.
- Número de acertos a perguntas pré definidas, quando as respostas são elaboradas após a leitura do sumário, em comparação a respostas elaboradas após a leitura dos documentos originais.

A avaliação de sumários é uma tarefa que encontra uma série de dificuldades, a começar pela inexistência de uma definição do que seria um sumário ideal para um documento ou conjunto de documentos. Mesmo quando realizada por humanos, tanto a sumarização quanto a avaliação de sumários tendem a ser divergentes (DAS; MARTINS, 2007). Em outras palavras, mesmo quando feita por humanos, as avaliações tendem a ser distintas. A avaliação manual também é bastante custosa. Segundo Lin (2004), a avaliação manual de sumários em larga escala, com as feitas durante as conferência DUC necessitariam de mais de 3000 horas-homem de trabalho. Apesar disso, a comunidade de pesquisadores na área de SAT sempre buscou a definição de medidas e métricas de avaliação que fossem capazes de gerar consenso quanto as avaliações que fariam, visto que a existência de tais medidas e métricas tornaria possível a comparação entre diferentes modelos de sumarização. Além disso, como afirmam Helena e Pardo (2003) "É por meio da avaliação que se torna possível verificar o avanço do estado da arte em sumarização automática".

Neste contexto, Lin (2004) propôs uma família de medidas chamada ROUGE (sigla em inglês para *Recall-Oriented Understudy for Gisting Evaluation*) com o objetivo de avaliar os sumários intrinsecamente e de forma automática. Ao longo dos anos, outras medidas foram propostas, tais como "factoides" (TEUFEL; HALTEREN, 2004), o método da "pirâmide" (NENKOVA; PASSONNEAU, 2004) e o de *Basic Elements* (BE) (HOVY et al., 2006). Apesar disso, desde a sua proposição a família de medidas ROUGE se tornou o padrão *de-facto* para a avaliação automática de sumários (DAS; MARTINS, 2007), tendo este fato sido fundamental para o avanço do estado da arte em SAT. A subseção seguinte apresenta a definição das medidas ROUGE.

2.1.2.1 ROUGE

A família de medidas ROUGE tem como objetivo medir a similaridade entre os sumários gerados automaticamente e sumários de referência, considerados representativos do que seria um "sumário ideal". Essa ideia baseia-se na suposição de que quanto mais similar aos sumários de referência um sumário gerado automaticamente for, melhor este será. As medi-

das ROUGE avaliam a similaridade entre os diferentes sumários através da contagem da co-ocorrência de elementos textuais básicos, tais como *n-gramas*, em ambos (LIN, 2004). Estas medidas apresentam uma boa correlação com a avaliação humana dos mesmos critérios a que se propõe avaliar (LIN, 2004).

Ao todo, existem quatro diferentes tipos de medidas na família: ROUGE-N, que compara os n-gramas nos dois sumários e conta as co-ocorrências; ROUGE-L, para a comparação de sequências de palavras longas; ROUGE-W para a ponderação da subsequência comum mais longa; e ROUGE-S para a comparação de bigramas (*n-gramas* onde $n = 2$) em sequências arbitrárias.

Nas medidas ROUGE-N, *N* refere-se ao tamanho do n-grama para aquela medida específica. Assim, ROUGE-1 mede a co-ocorrência de n-gramas de tamanho 1, enquanto ROUGE-2 mede a co-ocorrência de n-gramas de tamanho 2, e assim sucessivamente. As co-ocorrências são medidas sempre através das medidas de precisão, cobertura e medida-f, sendo definidas da seguinte maneira:

$$Precisão = \frac{|n-gramas \in \{Sum.aut.\} \cap n-gramas \in \{Sum.ref.\}|}{|n-gramas \in Sum.aut.|} \tag{2.1}$$

$$Cobertura = \frac{|n-gramas \in \{Sum.aut.\} \cap n-gramas \in \{Sum.ref.\}|}{|n-gramas \in Sum.ref.|} \tag{2.2}$$

$$Medida-f = \frac{2 \times (Precisão \times Cobertura)}{Precisão + Cobertura} \tag{2.3}$$

Onde $Sum.aut.$ se refere ao sumário automático sendo avaliado e $Sum.ref.$ se refere ao conjunto de sumários de referência que servem como base da avaliação.

Nas medidas ROUGE-L, *L* refere-se a *longest*, ou mais longo, que vem de *Longest common subsequence* ou sequência comum mais longa. Nessa medida entendem-se as sentenças como sequências de palavras e busca-se a subsequência de palavras comum a ambas as sentenças, ou seja, uma sequência de palavras em ordem que for comum a ambas as sentenças, mais longa, como forma de medir a similaridade entre sentenças de dois sumários. É importante salientar que as palavras não precisam ser adjacentes, apenas precisam aparecer uma após a outra na sentença, com possivelmente outras palavras entre elas. A partir desta definição calcula-se precisão, cobertura e medida-f da seguinte maneira:

$$Precisão = \frac{LCS(X,Y)}{n} \tag{2.4}$$

$$Cobertura = \frac{LCS(X,Y)}{m} \tag{2.5}$$

$$Medida-f = \frac{(1+b^2) * Cobertura * Precisão}{Cobertura + b^2 * Precisão} \tag{2.6}$$

Onde $LCS(X,Y)$ é o tamanho da subsequência comum mais longa entre X e Y, X é uma sentença de um sumário de referência, de tamanho m, Y é uma sentença de um sumário automático sendo avaliado, de tamanho n, e b é um parâmetro de ponderação entre precisão e cobertura para o cálculo da medida-f.

A partir das definições anteriores, é possível calcular a medida-f de ROUGE-L para o sumário inteiro sob avaliação através das seguintes definições:

$$Cobertura_{lcs} = \frac{\sum_{i=1}^{u} LCS(r_i, C)}{m} \tag{2.7}$$

$$Precisão_{lcs} = \frac{\sum_{i=1}^{u} LCS(r_i, C)}{n} \tag{2.8}$$

$$Medida - f_{lcs} = \frac{(1+b^2) * Cobertura_{lcs} * Precisão_{lcs}}{Cobertura_{lcs} + b^2 * Precisão_{lcs}} \tag{2.9}$$

Onde C é um sumário candidato contendo v sentenças e um total de n palavras, sendo comparado a um sumário de referência contendo u sentenças e um total de m palavras.

Nas medidas ROUGE-W, W vem de *weighted*. Estas medidas são uma variação das medidas ROUGE-L, onde mais peso é dado a subsequências comuns aonde as palavras que formam as subsequências sejam adjacentes nas sentenças, ou seja, não exista nenhuma outra palavra entre elas. Por exemplo, dada uma sentença de referência $X = [\underline{ABCD}EFG]$ e duas sentenças candidatas $Y_1 = [\underline{ABCD}HIK]$ e $Y_2 = [\underline{A}H\underline{B}K\underline{C}I\underline{D}]$, tanto Y_1 quanto Y_2 teriam o mesmo valor de avaliação nas medidas ROUGE-L apesar de Y_1 ter a subsequência comum a X com as palavras adjacentes, como em X. Entende-se que Y_1 deveria ser melhor avaliada e ROUGE-W tem justamente esta proposta. Lin (2004) apresenta diversas maneiras de se dar mais peso as subsequências adjacentes. Cobertura, precisão e medida-f de ROUGE-W são calculadas da mesma forma como se calculam para ROUGE-L, apenas substituindo-se LCS por $WLCS$ de acordo com uma destas diversas maneiras apresentadas pelo autor.

Nas medidas ROUGE-S, S vem de *skip-bigram*. Um *skip-bigram* é qualquer par de palavras na ordem em que aparecem em um trecho de texto, com qualquer número de palavras entre elas. As estatísticas de co-ocorrência de *skip-bigrams* medem a co-ocorrência de *skip-bigrams* entre um sumário de referência e um sumário sob avaliação. O número de *skip-bigrams* de um sumário pode ser calculado através da função de combinação. Assim, um sumário hipotético formado por 4 palavras terá 6 *skip-bigrams* pois $C(4,2) = 4!\,(2!*2!) = 6$. As co-ocorrências são medidas sempre através das medidas de precisão, cobertura e medida-f, sendo definidas da seguinte maneira:

$$Precisão = \frac{SKIP2(X,Y)}{C(n,2)} \tag{2.10}$$

$$Cobertura = \frac{SKIP2(X,Y)}{C(m,2)} \tag{2.11}$$

$$Medida - f = \frac{(1 + b^2) * Cobertura * Precisão}{Cobertura + b^2 * Precisão} \quad (2.12)$$

Onde X é um sumário de referência de tamanho m, Y é um sumário de referência de tamanho n, $SKIP2(X,Y)$ é o número de co-ocorrência de *skip-bigrams* entre X e Y e b é um parâmetro de ponderação entre precisão e cobertura para o cálculo da medida-f.

As medidas ROUGE foram avaliadas através da sua correlação com avaliações realizadas por avaliadores humanos, sendo que ROUGE-1 e ROUGE-2 tiveram bom desempenho dentre as medidas ROUGE-N (LIN, 2004) e por isso são largamente utilizadas como padrão *de-facto* na avaliação automática de sumários.

2.2 ONTOLOGIAS

Sistemas computacionais baseados em conhecimento são dependentes da existência de uma representação deste conhecimento que possam acessar e manipular. Um conhecimento que está representado de tal maneira é um conhecimento formalizado. Todo corpo de conhecimento que está formalizado é baseado em uma conceitualização, formada por objetos, conceitos e outras entidades que presumivelmente existam em uma área específica de interesse, além das relações existente entre estes (GRUBER, 1993). Uma conceitualização é uma visão simplificada e abstrata do mundo, que se deseja representar para algum propósito. Portanto, todo sistema computacional baseado em conhecimento está apoiado em uma conceitualização, seja ela implícita ou explícita.

Na literatura, existem diversas definições de ontologia, dependendo da área de conhecimento que se estuda e da aplicação da mesmas. A definição adotada neste trabalho, que vem da literatura de Inteligência Artificial e surge da necessidade de sistemas computacionais baseados em conhecimento de apoiar-se sobre uma conceitualização, diz que **uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada** (GRUBER, 1993). Nessa definição, "formal e explícita" refere-se ao fato desta representação existir e ser representada de forma a possibilitar a manipulação por sistemas computacionais, enquanto "compartilhada" refere-se ao fato desta mesma representação poder ser utilizada por diversos sistemas. A partir dessa definição a ontologia pode ser entendida como uma base de conhecimento descrita em uma lógica descritiva específica.

2.2.1 Componentes de uma ontologia

Uma ontologia, segundo a definição apresentada por Gruber (1993) e adotada nesta dissertação, pode ser formada pela definição de **classes, relações, funções, instâncias e axiomas**.

As **classes** ou **conceitos** podem ser entendidos como sendo os conceitos fundamentais sobre os quais versa a ontologia, resultantes da articulação do conhecimento básico sobre

um determinado domínio. Esses conceitos podem ser representados usando um vocabulário especializado.

As **relações** representam as interações que existem entre as classes ou conceitos do domínio. Comumente, relações do tipo é-um (em inglês *is_a*) estão presentes, formando uma hierarquia de conceitos ou taxonomia.

As **funções** são um caso específico das relações, onde a relação entre os elementos se dá de forma única, com mais de um elemento envolvido para formar uma única relação. A definição de função, porém, é uma das menos coesas entre as diversas linguagens de especificação de ontologias, variando bastante (CORCHO; GóMEZ-PéREZ, 2000).

Os **axiomas** são regras que definem e restringem os usos e interpretações das relações, formando sentenças que são sempre verdadeiras.

As **instâncias**, por vezes chamadas de indivíduos, modelam objetos concretos ou fatos-base derivados dos conceitos, bem como seus atributos.

Além disso, uma ontologia, sendo uma base de conhecimento expressa em uma lógica descritiva, é formada por dois componentes tradicionalmente chamados de TBox e ABox (GIACOMO; LENZERINI, 1996). O primeiro, TBox, diz respeito a parte terminológica da ontologia, ou seja, seus conceitos e as relações entre eles. O segundo, ABox, diz respeito as instâncias e as asserções sobre as mesmas. Um exemplo de informação encontrada no TBox, por exemplo, seria a de que um conceito é uma especialização de outro. Já no ABox encontraríamos, por exemplo, a informação de que uma instância é uma instância de um determinado conceito.

A título de exemplo, uma ontologia empregada para modelar o domínio do futebol poderia definir conceitos como *esportista* e *jogador de futebol*, existindo entre eles uma relação do tipo *é-um*: *jogador de futebol é-um esportista*. Na mesma ontologia, poderiam estar definidas as instâncias *Pelé* e *Rivellino*, ambas sendo instâncias do conceito *jogador de futebol*: *Pelé é-um jogador de futebol*, *Rivellino é-um jogador de futebol*. Além disso, ambas as instâncias poderiam estar ligadas por uma relação do tipo *jogou-com*: *Pelé jogou-com Rivellino*. A figura 3 apresenta uma ilustração desta ontologia de exemplo.

2.2.2 Categorias de Ontologias

De acordo com a definição de Guarino (1998) as ontologias podem ser categorizadas, de acordo com com a sua generalidade ou especificidade em: Ontologias de alto nível, Ontologias de domínio, Ontologias de tarefa e Ontologias de aplicação.

As **ontologias de alto nível** são ontologias que tem como propósito armazenar o conhecimento e também noções sobre conceitos fundamentais ou elementares que permeiam diversos domínios. Por isso mesmo, os conceitos representados por tal tipo de ontologia geralmente são independentes de um domínio ou problema particular.

As **ontologias de domínio**, por outro lado, representam o conhecimento sobre um determinado domínio específico. Como exemplo, podem se citar ontologias específicas para o

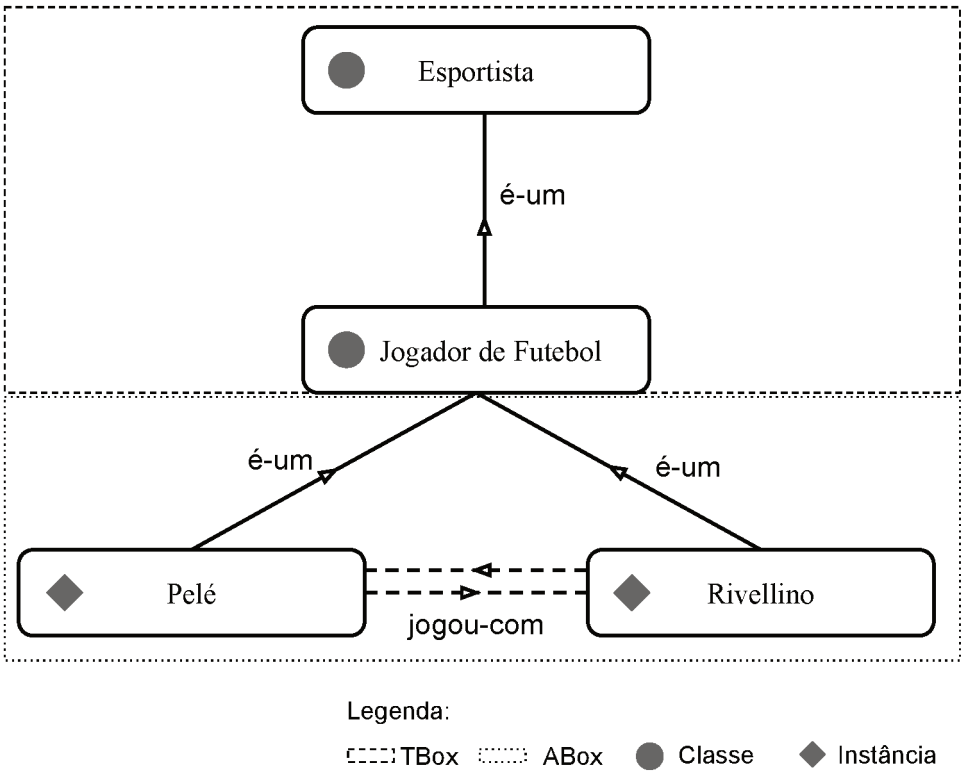


Figura 3 – Exemplo de ontologia com conceitos, relações e instâncias definidas

domínio da medicina, biologia, ciência da computação, gerenciamento de desastres, entre outros. Ontologias desta categoria são as mais comumente usadas por sumarizadores automáticos de texto.

Já as **ontologias de tarefa**, como o nome sugere, representam não o conhecimento de um domínio específico mas sim o conhecimento sobre um conjunto de objetos que seja pertinente a uma determinada tarefa, como a tarefa "coleta de sangue para exames" no domínio de medicina.

Por fim, as **ontologias de aplicação** são as ontologias mais específicas, desenvolvidos para o uso em uma parte específica de uma tarefa ou domínio.

A classificação proposta por Guarino (1998) traz a ideia de que existe uma hierarquia de ontologias, sendo que uma categoria deriva da outra conforme a sua especificidade aumenta. Nesse sentido as ontologias de alto nível são as mais abrangentes, tratando o conhecimento de forma genérica e sugerindo o reuso, enquanto as ontologias de aplicação são as mais específicas, lidando apenas com os objetos diretamente envolvidos na aplicação na qual esta é utilizada. A figura 4 ilustra esta ideia.

As ontologias também podem ser classificadas quanto a forma como são geradas. Neste caso, classificam-se como **ontologias geradas manualmente** aquelas que são geradas inteiramente de forma manual. Estas ontologias usualmente estão limitadas a descrição de um domínio específico feita por um especialista neste domínio. Já as **ontologias geradas automaticamente** são todas aquelas que são geradas, parcial ou inteiramente, de forma automática, ou

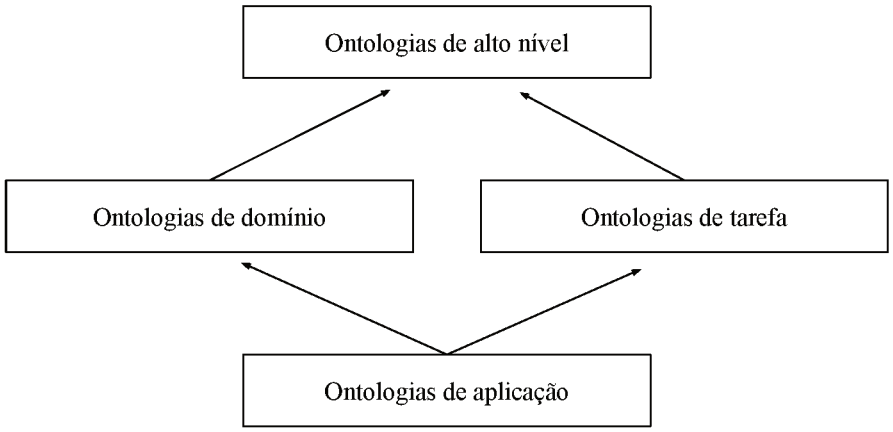


Figura 4 – Hierarquia de categorias de ontologias

seja, sem intervenção humana. Estas ontologias podem ser geradas, por exemplo, a partir de dados estruturados ou semi-estruturados publicados na *web*, usando tecnologias de *web* semântica como *Linked Data*.

A ontologia da DBPedia, que é a ontologia utilizada na implementação do método proposto nesta dissertação utilizada para a realização dos experimentos apresentados também nesta dissertação, é uma ontologia de alto nível, gerada automaticamente.

2.3 TAREFAS DE EXTRAÇÃO DE INFORMAÇÃO

Para que seja possível utilizar as instâncias definidas em uma ontologia para representar a semântica de uma sentença é preciso primeiro identificar quais instâncias estão relacionadas a sentença. Essa identificação se dá através da execução de duas tarefas de extração de informação: a tarefa de Reconhecimento de Entidades Nomeadas e tarefa de Ligação de Entidades. Para defini-las é preciso antes definir o conceito entidade nomeada.

Na área de recuperação de informação, o termo **entidade nomeada** é usado para se referir a algo do mundo real ou imaginário, geralmente uma instância de um conceito (GRISHMAN; SUNDHEIM, 1996). Entidades nomeadas são referenciadas em texto através de suas formas de superfície, que são sequências de um ou mais termos em língua natural. Uma entidade nomeada pode ter diversos nomes de superfície. Pode-se referenciar, por exemplo, a cidade de *São Paulo* com as formas de superfície: "Sampa", "SP", "Terra da garoa", entre outros. A ocorrência de um nome de superfície em um texto em língua natural é chamada de menção.

A tarefa de **Reconhecimento de Entidades Nomeadas**, ou **NER** da sigla em inglês para *Named Entity Recognition*, é a tarefa de reconhecer menções a entidades nomeadas em um trecho de texto (GRISHMAN; SUNDHEIM, 1996). A partir deste reconhecimento, estas entidades nomeadas podem ser ligadas a recursos externos que descrevem a sua semântica,

definidos em uma base de conhecimento, que pode ser uma ontologia. A tarefa de realizar esta ligação entre as entidades nomeadas reconhecidas em um trecho de texto e as descrições formais de suas semânticas, dá-se o nome de **Ligação de Entidades**, ou **EL** da sigla em inglês para *Entity Linking*.

2.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo apresentou a fundamentação teórica necessária para melhor compreender a proposta apresentada nesta dissertação. Primeiramente foi apresentado o conceito de sumário, a partir do qual define-se a sumarização automática de texto. Em seguida foram apresentadas algumas das possíveis categorias aplicáveis a um sumarizador automático de texto, definindo aquelas nas quais se encaixa o modelo proposto nesta dissertação. Logo depois alguns conceitos ligados a avaliação de sumários foram apresentados. Na sequência apresentou-se o conceito de ontologia utilizado neste trabalho, que vem da literatura de web semântica, bem como as partes que as constituem. Em seguida, algumas das possíveis categorias aplicadas as ontologias foram apresentadas, definindo em quais destas a ontologia utilizada neste trabalho se encaixa. Por fim, duas tarefas de extração de informação importantes para compreender como instâncias de uma ontologia são ligadas a trechos de texto foram apresentadas.

3 ESTADO DA ARTE

Este capítulo apresenta o estado da arte em sumarização de texto utilizando ontologias como base semântica, obtido através de uma revisão sistemática da literatura (RSL) e da análise de alguns outros trabalhos relevantes não incluídos na RSL. A revisão sistemática da literatura foi feita seguindo a metodologia proposta e descrita por Kitchenham (2004). A seção 3.1 apresenta uma descrição do protocolo utilizado na revisão. A seção 3.1.3 apresenta uma análise dos resultados obtidos através da aplicação do protocolo, dividida de acordo com as questões de pesquisa definidas para a revisão. A seção 3.2 apresenta alguns trabalhos conhecidos e relevantes que não foram incluídos na revisão. Por fim, a seção 3.3 apresenta as considerações finais e como a análise do estado da arte ajudou a identificar uma oportunidade de pesquisa que foi o ponto de partida para este trabalho.

3.1 REVISÃO SISTEMÁTICA DA LITERATURA

Para que seja possível analisar e compreender o estado da arte de uma determinada área do conhecimento, é necessário que se defina uma maneira de buscar, selecionar e categorizar evidências sobre a pesquisa nesta área. A análise destas evidências pode levar a identificação de lacunas no conhecimento que podem ajudar a direcionar novas pesquisas na área. Neste trabalho busca-se analisar especificamente o estado da arte em sumarização extrativa utilizando ontologias através de uma revisão sistemática da literatura.

Uma revisão sistemática da literatura (RSL) é uma maneira de avaliar e interpretar todos os estudos relevantes para uma determinada área de pesquisa de uma maneira reproduzível e não tendenciosa (KITCHENHAM, 2004). De acordo com a metodologia descrita por Kitchenham (2004) é necessário que uma revisão sistemática da literatura defina um protocolo de pesquisa para que os resultados alcançados pela mesma possam ser reproduzidos e para que a sua não-tendenciosidade possa ser avaliada. Este protocolo deve ser composto de questões de pesquisa bem definidas, uma estratégia de busca e critérios de inclusão e exclusão. A estratégia de busca deve levar em conta as bases de dados que serão pesquisadas e como a *string* de busca para cada uma das bases será construída. Os critérios de inclusão e exclusão devem levar em conta aspectos diretamente relacionados às questões de pesquisa e aspectos relacionados a qualidade dos estudos.

3.1.1 Questões de pesquisa

Quatro questões de pesquisa foram definidas para guiar a RSL. A motivação da questão de pesquisa 1 (QP1) é identificar para que finalidades as ontologias são empregadas dentro da sumarização extrativa de texto. A segunda questão de pesquisa (QP2) visa identificar quais componentes das ontologias, conforme definição apresentada na seção 2.2, são utilizados. A questão de pesquisa número 3 (QP3) busca identificar como são construídas as ontologias e

qual é o seu tamanho, visto que estes dois fatores influenciam a sua abrangência. A última questão de pesquisa (QP4) tem como objetivo entender os resultados alcançados pelos métodos de sumarização extrativa que utilizam ontologia. As quatro questões de pesquisa são:

1. (**QP1**) Com que objetivo as ontologias são usadas?
2. (**QP2**) Quais componentes de ontologias são usados?
3. (**QP3**) Como as ontologias utilizadas são construídas e qual é o seu tamanho?
4. (**QP4**) Como os autores classificam os resultados alcançados pelos métodos de sumarização de texto que usam ontologias?

3.1.2 Estratégia de pesquisa

A definição da estratégia de pesquisa se inicia pela identificação dos principais termos relacionados as perguntas de pesquisa, bem como termos alternativos a estes e seus sinônimos. A partir desta lista de termos uma string de busca é definida, e as etapas seguintes da RSL podem ser executadas. As definições tanto da *string* de busca quanto das demais etapas da RSL são feitas nas seções seguintes.

3.1.2.1 String de busca

A partir das questões de pesquisa construiu-se a *string* de busca. A definição da *string* de busca iniciou-se com a definição de uma *string* inicial a partir de termos já conhecidos e dos operadores booleanos E/OU. O operador OU foi usado para agrupar todos os sinônimos ou termos alternativos a um termo e o operador E foi usado para unir estes agrupamentos. Esta *string* foi então utilizada para conduzir uma busca inicial em três bases de dados (DBLP, IEEE Xplore, e Science Direct) para verificar se estudos relevantes já conhecidos eram retornados. A partir disso, novos termos foram incorporados a *string* de busca na medida em que apareciam em estudos relevantes retornados.

Analisando os títulos dos resultados retornados nessa busca inicial, verificou-se que o termo “Summarization” parece ser precedido por um conjunto de diferentes modificadores para definir tipos específicos de sumarizadores, por isso na *string* de busca final ele aparece como um termo independente (com duas grafias possíveis) precedido por um conjunto de termos representando todos estes modificadores. Apesar do termo “Extractive” não ter aparecido nos estudos utilizados para aprimorar a *string* de busca, ele foi incluído no conjunto de modificadores por ser um tipo de sumarização conhecido e complementar a “Abstractive”. Para avaliar a qualidade da *string* de busca final, alguns estudos relevantes já conhecidos foram tomados como conjunto mínimo a ser retornado por esta *string*. A *string* de busca final é apresentada abaixo. Houve a necessidade de adaptar esta *string* para cada uma das bases onde a pesquisa foi realizada, devido as particularidades de como a busca é implementada em cada base.

```
(‘‘Automatic’’ OR ‘‘Text’’ OR ‘‘Document’’ OR ‘‘Multi-Document’’ OR
‘‘Abstractive’’ OR ‘‘Extractive’’) AND (‘‘Summarization’’ OR ‘‘Summarisation’’)
AND (‘‘Ontology’’ OR ‘‘Ontologies’’ OR ‘‘Taxonomy’’ OR ‘‘Semantic’’ OR
‘‘Semantics’’).
```

É importante salientar que apesar da *string* de busca ter sido escrita exclusivamente na língua inglesa, isso não limita os possíveis resultados apenas a trabalhos escritos nessa mesma língua, uma vez que as bases de busca selecionadas mantém registro de palavras-chave e resumo em inglês, mesmo para trabalhos escritos em outras línguas.

3.1.2.2 Bases de busca

As seguintes bases de dados foram escolhidas para formar as fontes primárias de busca:

- ACL Anthology (https://aclweb.org/anthology).
- IEEE Xplore (https://ieeexplore.ieee.or).
- Science Direct (https://www.sciencedirect.com).
- Springer Link (https://link.springer.com).
- ACM DL (https://dl.acm.org).
- DBLP (https://dblp.uni-trier.de).

Estas bases foram escolhidas por cobrirem a maior parte das conferências e periódicos em Processamento de Linguagem Natural, Recuperação de Informação e Representação de Conhecimento.

3.1.2.3 Critérios de inclusão e exclusão

Depois de definida a *string* de busca e as bases de dados que seriam utilizadas na revisão, foram definidos os critérios de inclusão e exclusão dos trabalhos. Para ser **incluído** como um estudo a ser analisado, cada trabalho retornado na busca deveria:

1. Descrever um método de Sumarização de Texto.
2. Empregar de alguma maneira uma ou mais ontologias.
3. Reportar resultados experimentais obtidos pelo método apresentado utilizando a família de medidas ROUGE.
4. Descrever como a ontologia utilizada foi construída e qual o seu tamanho.

5. Ter sido escrito em português ou inglês.

Por outro lado, seriam **excluídos** os trabalhos que não atendessem aos critérios de inclusão.

O objetivo do critério de inclusão número 3 foi o de garantir que os resultados alcançados pelo método proposto nesta dissertação pudessem ser comparados diretamente com os resultados alcançados pelos trabalhos analisados na revisão sistemática. Essa comparação direta foi fundamental para o atingimento do objetivo geral da dissertação. As medidas da família ROUGE são o padrão *de-facto* na avaliação automática de sumários e por isso foram selecionadas, uma vez que garantiriam a comparação de resultados com um maior número de trabalhos. Apesar disso, a presença deste critério e a escolha pelas medidas da família ROUGE acabaram por excluir alguns trabalhos conhecidos e relevantes no contexto desta dissertação do conjunto final analisado. Estes trabalhos são apresentados, acompanhados de breve discussão, na seção 3.2.

3.1.2.4 Processo de seleção de trabalhos

O processo de seleção dos trabalhos foi conduzido em três fases. Na primeira fase, os títulos dos trabalhos candidatos foram avaliados e os critérios de inclusão e exclusão aplicados. Na segunda fase, os mesmos critérios de inclusão e exclusão foram aplicados aos resumos (ou *abstracts*) de todos os trabalhos que passaram pela primeira fase da seleção. Por fim, na terceira fase, os critérios de inclusão e exclusão foram aplicados a todos os trabalhos que passaram pela segunda fase, desta vez levando em conta o conteúdo destes trabalhos na íntegra. Os trabalhos que passaram pela terceira fase tiveram ainda a sua qualidade avaliada antes de serem finalmente incluídos na lista final de trabalhos a serem analisados para responder as questões de pesquisa desta RSL. Os critérios de qualidade são apresentados a seguir.

3.1.2.5 Controle de qualidade

Os estudos selecionados tiveram a sua qualidade avaliada antes de serem incluídos na lista final de estudos relevantes para esta RSL. Esta avaliação foi feita a partir de critérios de qualidade, cada um representado por uma pergunta. As respostas possíveis a essas perguntas eram: Sim, Parcialmente e Não. Estas respostas somavam 1 ponto, 0,5 ponto e 0 ponto respectivamente ao total de pontos de cada estudo. O primeiro quartil (5/4 = 1,25) foi definido como ponto de corte. Estudos com pontuação abaixo deste valor não seriam incluídos na lista final de estudos relevantes. As perguntas utilizadas foram:

1. Os objetivos da pesquisa estão claramente definidos?

2. O trabalho foi projetado para atingir estes objetivos?

3. Os pesquisadores discutem os potenciais problemas com a validade/confiabilidade dos resultados?

4. Todas as perguntas de pesquisa foram respondidas adequadamente?

5. As conclusões derivam claramente dos dados coletados?

3.1.2.6 Estratégia para extração dos dados e resumo das conclusões

Para extrair os dados relevantes dos estudos selecionados e permitir a posterior análise destes dados um formulário foi criado. Além de conter campos para identificação do estudo, como título, autor e ano, este formulário continha campos relativos a cada uma das perguntas de pesquisa.

A figura 5 ilustra como se deu a execução da revisão a partir de todas as etapas descritas nesta seção.

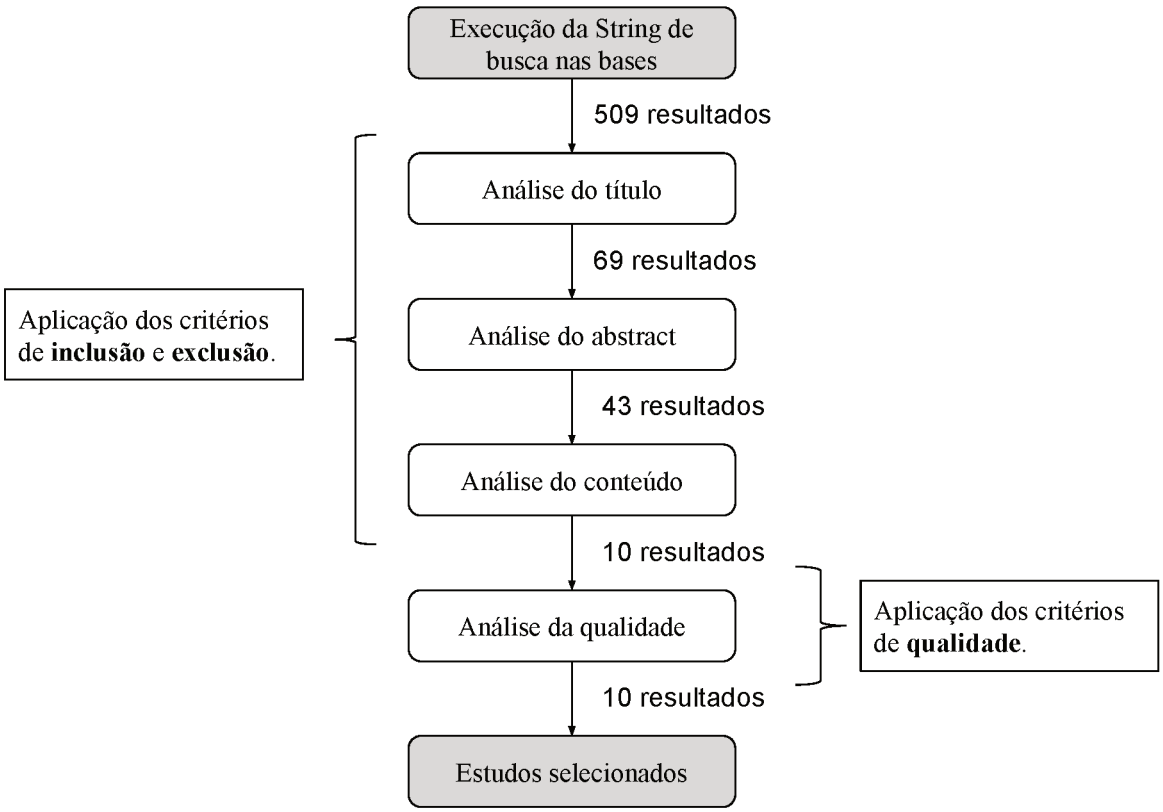


Figura 5 – Etapas da RSL

3.1.3 Análise dos resultados

Aplicando a *string* de busca nas cinco bases de dados definidas, obteve-se um total de 509 resultados. Destes, 7 eram resultados duplicados e foram excluídos. Dos 502 remanescentes, 69 passaram pela primeira fase do processo de seleção, 43 pela segunda e 10 pela terceira - muitos estudos não atendiam ao terceiro critério de inclusão (ROUGE), que só pode ser verificado na leitura integral dos documentos. Ainda na segunda fase um resultado secundário foi incluído, apresentado em Hipola et al. (2014). Este resultado já era conhecido, mas não estava disponível em nenhuma das bases selecionadas. Nenhum estudo foi excluído por não atender aos critérios de qualidade. Um total de 11 estudos foram considerados para responder as perguntas de pesquisa.

Ahmad e Ahmad (2019) descrevem um método de sumarização extrativa multidocumento que se baseia em uma ontologia para criar uma representação intermediária das sentenças. A ontologia utilizada foi construída automaticamente a partir da hierarquia de conceitos da Wikipedia, que tem seis níveis. Nesta ontologia cada conceito está ligado a um artigo da Wikipedia. Para construir a representação das sentenças, a similaridade entre cada sentença e cada artigo é calculada e, a partir dos resultados, a sentença passa a ser representada pelo conceito que estiver ligado ao artigo mais similar a mesma. A partir desta representação os autores utilizam algoritmos da teoria de jogos para ordenar e selecionar as sentenças para extração e criação do sumário. O método proposto é avaliado a partir de experimentos realizados com diferentes conjuntos de dados, tanto direcionados a sumarização focada em consultas (com os corpus DUC2004 e DUC2005), como em sumarização não focada em consultas (com o corpus DUC2002). Os resultados alcançados são descritos como competitivos pelos autores, mas não superam todos os outros resultados com os quais foram comparados.

No trabalho de MacAvaney et al. (2019), duas ontologias especificamente criadas para o domínio do problema em questão são utilizadas para propor um método de sumarização abstrativa baseado em ontologias. Uma delas (UMLS) integra várias outras ontologias menores, portanto parte do seu conteúdo é gerado automaticamente e parte manualmente. A outra (RadLex) também integra outras ontologias menores, mas seu conteúdo foi todo criado manualmente. Os conceitos definidos nas ontologias são utilizados para gerar um vetor de contexto a partir dos documentos sendo sumarizados. A ideia é que apenas substantivos que correspondem a algum conceito definido na ontologia são incluídos no vetor. Este vetor, juntamente com outros, forma a entrada de uma rede neural que gera o sumário abstrativo. O método proposto foi testado com um conjunto de dados específico da área de radiologia. As medidas ROUGE-1, ROUGE-2 e ROUGE-L foram utilizadas para avaliar os resultados alcançados, que foram classificados como excelentes pelos autores visto que o método obteve resultados melhores que todos os demais com os quais estava sendo comparado, especialmente quando a ontologia específica para a área de Radiologia (RadLex) foi utilizada.

O trabalho de Kumar et al. (2018) apresenta um método de geração de respostas para perguntas da área biomédica que utiliza uma ontologia para recuperação de informação a ser su-

marizada. Neste trabalho a ontologia não é aplicada diretamente no processo de sumarização, mas como uma parte dos processos que a antecedem. A ideia dos autores é gerar uma resposta ideal para uma pergunta a partir da sumarização de documentos que são relevantes para aquela pergunta. Inicialmente, o conjunto de documentos disponíveis passa pelos processos de identificação de entidades nomeadas (NER - sigla em inglês para *Named Entity Recognition*) e extração de relações para gerar uma ontologia a partir de conceitos e relações identificados, onde cada conceito guarda a informação de em que documentos o mesmo aparece. Depois a pergunta passa pelo mesmo processo e os conceitos identificados na mesma são então utilizados para construir uma consulta que será executada na ontologia construída a partir dos documentos. Os documentos retornados pela consulta são identificados como sendo os documentos relevantes para a pergunta. Esses documentos são então sumarizados utilizando um método de sumarização abstrativa baseado em redes neurais. O método proposto é avaliado a partir de experimentos em um conjunto de dados especificamente construído para a área de biomedicina (BioASQ). Resultados para diferentes implementações do modelo são reportados. As medidas ROUGE-2 e ROUGE-S são usadas para comparar estas diferentes implementações. Os resultados são descritos pelos autores como promissores, porém os resultados não são comparados com resultados alcançados por nenhum outro modelo.

Chandu et al. (2017) também apresentam um método de sumarização de texto cujo objetivo é servir de mecanismo de geração de resposta para uma pergunta específica na área de biomedicina. Neste trabalho, a ontologia é utilizada para criar uma representação das sentenças através de conceitos que, juntamente com outras representações, serve de base para a sumarização. Os autores utilizam diferentes métodos de sumarização, entre eles MMR, experimentando com cada um a fim de identificar qual gera os melhores resultados. Duas diferentes ontologias são utilizadas para a criação da representação das sentenças: UMLS, que é construída em partes automaticamente e em partes manualmente, e SNOMEDCT que, até onde se pode apurar, também é construída de forma híbrida. Diferentes implementações do método proposto são comparadas entre si, usando as medidas ROUGE-2 e ROUGE-SU4 na análise dos resultados dos experimentos, mas nenhuma é comparada com outros modelos de sumarização que se propõe a mesma tarefa.

No trabalho de Mohamed e Oussalah (2015), os autores utilizam a Wordnet como uma hierarquia de conceitos para construir representações para cada uma das sentenças dos documentos. Essa representação é construída a partir da aplicação de diversas técnicas de pré-processamento que buscam identificar substantivos que melhore representem palavras de outras classes gramaticais. Isso é importante para permitir que as medidas de similaridade semântica definidas para conceitos descritos na Wordnet possam ser empregadas. A Wordnet foi construída automaticamente e possui cerca de duzentos e sete mil pares que ligam conceitos a conjuntos de palavras, cada par podendo ser interpretado como um conceito de uma ontologia. A partir da representação das sentenças através de conceitos da WordNet e de melhorias propostas as medidas de similaridade semântica entre estes conceitos os autores utilizam o algoritmo MMR para produzir sumários extrativos. Eles realizaram experimentos com o corpus

DUC2005 e reportam resultados obtidos na análise com as medidas ROUGE-1, ROUGE-2 e ROUGE-SU4, comparando o modelo proposto com diversos outros. Os resultados alcançados são os melhores dentre aqueles em comparação na medida ROUGE-1, mas não nas demais.

Li e Li (2014) descrevem a construção de um sumarizador automático que utiliza uma ontologia específica para o domínio de engenharia civil costeira e portuária para diversas finalidades, como seleção de documentos e agrupamento de sentenças. O modelo é apresentado através dos diversos sub-agentes que o compõe. O agente de sumarização utiliza a ontologia primeiramente como uma base semântica para identificar conceitos referenciados nas sentenças dos documentos a serem sumarizados. A partir da identificação destes conceitos, as sentenças são agrupadas para formar os vértices de uma árvore que segue a estrutura definida por relações do tipo *is_a* entre os conceitos da ontologia. A partir dessa representação, o sumarizador consegue calcular a similaridade entre sentenças, o que embasa o restante do processo de sumarização. A ontologia utilizada foi gerada manualmente a partir de documentos específicos da área e era composta de instâncias, conceitos e relações *is_a*, tendo por volta de seis mil instâncias e três mil conceitos. Os autores avaliam o modelo proposto e sua implementação a partir de experimentos realizados com um corpus específico da área. Os resultados obtidos foram avaliados através das medidas ROUGE-1 e ROUGE-2 e, segundo os autores, os resultados alcançados podem ser classificados como satisfatórios.

No trabalho apresentado por Hipola et al. (2014) as sentenças dos documentos a serem sumarizados são inicialmente mapeadas para os conceitos descritos em uma ontologia específica para o domínio do problema no qual o trabalho está focado - gerenciamento de desastres. A ontologia utilizada é uma ontologia simples, formada apenas por conceitos e relações do tipo *is_a*, podendo ser entendida com uma taxonomia. Cada sentença é mapeada para exatamente um conceito e a partir desse mapeamento uma representação vetorial das sentenças é construída. Essa representação indica quais conceitos compõe o caminho entre o conceito que representa aquela sentença e conceito raiz da ontologia. A partir dessa representação vetorial, técnicas de clusterização são utilizadas para criar o sumário. Outras representações das sentenças a partir dos conceitos da ontologia também foram exploradas e utilizadas como entrada do processo de clusterização. A ontologia utilizada possuía um total de 109 conceitos e 326 relações, e é a mesma utilizada no trabalho de Wu et al. (2013). Segundo os experimentos realizados pelos autores, todas as combinações de representação das sentenças e técnicas de sumarização baseadas na ontologia obtiveram melhores resultados que as demais técnicas com as quais foram comparadas. Os resultados foram analisados através das medidas ROUGE-2, ROUGE-SU4 e ROUGE-L.

Baralis et al. (2013) utilizam uma ontologia para criar uma representação das sentenças dos múltiplos documentos sendo sumarizados. Essa representação serve de base para a definição de uma medida de similaridade, que é então utilizada numa variante do algoritmo MMR, formando o modelo proposto. A representação das sentenças é formada a partir dos conceitos definidos na ontologia YAGO. Esta ontologia foi gerada automaticamente a partir de documentos e metadados extraídos da Wikipedia. Os autores experimentaram o modelo proposto no

Corpus DUC2004, avaliando os resultados através das medidas ROUGE-2 e ROUGE-4. Os resultados alcançados, segundo os autores, são melhores do que os dos demais modelos com os quais são comparados, mais relevantes e menos redundantes. Além disso os autores realizaram uma investigação qualitativa dos sumários gerados pelo modelo proposto, reportando que os sumários produzidos eram legíveis e faziam sentido, além de serem mais focados nos pontos centrais dos documentos originais dos que os demais com os quais foram comparados.

No trabalho apresentado por Wu et al. (2013), os autores propõem um modelo que utiliza uma ontologia específica para um domínio, construída por especialistas, para sumarizar documentos neste domínio. A ontologia é entendida neste trabalho como uma hierarquia de conceitos ligados por relações do tipo *is_a*. Um conjunto de palavras-chave é definido para cada conceito na ontologia. Estas palavras-chave são então utilizadas para identificar, através de uma busca textual simples, ocorrências destes conceitos nas sentenças a serem sumarizadas. As sentenças são mapeadas para um ou mais conceitos da ontologia e este mapeamento serve de base para calcular o que os autores chamam de "conteúdo de informação" de cada sentença. Esse conteúdo de informação, por sua vez, é usado para selecionar as sentenças que comporão o sumário. A ontologia utilizada foi construída por especialistas no domínio de gerenciamento de desastres e era composta por 109 conceitos e 326 relações, além de um número não citado de indivíduos que não foram utilizados na sumarização. Os autores experimentaram o modelo proposto em diferentes corpus, incluindo o corpus DUC2004 e DUC2005, e analisaram os resultados principalmente através da medidas ROUGE-1, ROUGE-2 e ROUGE-L. Os resultados reportados são comparados com resultados alcançados por implementações de outros modelos, e, segundo os autores, podem ser classificados como bons resultados, sendo que a implementação do modelo proposto atinge os melhores valores nas medidas ROUGE dentre aqueles sob comparação.

Bawakid e Oussalah (2011) propõe um modelo de sumarização extrativa, multi-documento e baseada em consulta que utiliza apenas os conceitos definidos em uma ontologia criada automaticamente a partir da Wikipedia como base semântica da sumarização. No modelo proposto, as sentenças dos documentos sendo sumarizados são representadas pelos conceitos descritos na ontologia que são mencionados nas mesmas e, a partir dessa representação três diferentes características (*features*) que vão avaliar as sentenças são definidas. A partir da avaliação através destas características, as sentenças são ordenadas e depois extraídas em um processo iterativo. O modelo descrito neste trabalho foi proposto para participar da competição TAC2011, e os resultados alcançados no corpus fornecido para esta competição foram avaliados usando, entre outras, as medidas ROUGE-2 e ROUGE-SU4. Os resultados alcançados, segundo os autores, foram competitivos porém passíveis de melhorias.

Umbrath, Wetzker e Hennig (2008) utilizam a ontologia de maneira semelhante a Bawakid e Oussalah (2011). A ontologia utilizada é formada por conceitos e relações do tipo *is_a*, formando assim uma hierarquia de conceitos, ou taxonomia, com 1036 conceitos definidos. Um conjunto de palavras é ligado a cada conceito a partir de uma busca na *web*. Essas palavras passam a representar o conceito e a partir delas calcula-se a similaridade entre uma

sentença e cada um dos conceitos na ontologia. A partir desta similaridade, um classificador seleciona uma sub-árvore da taxonomia para representar cada uma das sentenças, e essa representação é utilizada para definir três características (*features*) que avaliam as sentenças. A partir destas características, um classificador é treinado para selecionar as sentenças que comporão o sumário. O modelo proposto é avaliado usando o corpus DUC2002 e as medidas ROUGE-1 e ROUGE-2. Os resultados alcançados são descritos pelos autores como promissores, embora não tenham atingido os melhores valores para as medidas ROUGE entre os modelos sob comparação.

A partir da leitura e análise dos 11 artigos que atenderam a todos os critérios de seleção definidos para essa revisão sistemática da literatura obtiveram-se respostas para todas as questões de pesquisa. Essas respostas são apresentadas textualmente nas subseções seguintes e na tabela 2.

3.1.3.1 Objetivo do uso de ontologias

Relativo a questão de pesquisa 1 (QP1), com exceção do trabalho de Kumar et al. (2018), onde a ontologia não é utilizada no processo de sumarização em si mas sim em uma etapa anterior de recuperação de informação, todos os modelos apresentados nos trabalhos analisados nessa RSL utilizam a ontologia para criar uma representação da semântica das sentenças presentes nos documentos sob sumarização. Alguns destes trabalhos utilizam esta representação para comparar semanticamente as sentenças; Outros utilizam essa representação para enriquecer a informação disponível para os métodos de extração/abstração.

3.1.3.2 Componentes utilizados

Com relação a questão de pesquisa 2 (QP2), todos os trabalhos analisados utilizam, de alguma forma, os conceitos definidos na ontologia. Kumar et al. (2018) chamam as definições que usam, descritas na ontologia, de entidades, mas a leitura do artigo deixa claro que tratam-se de conceitos. Alguns dos trabalhos analisados também utilizam, de alguma forma, relações entre conceitos, principalmente relações hierárquicas do tipo *is_a* a fim de formar uma árvore de conceitos. Por isso mesmo, muitos trabalhos limitam-se a entender a ontologia como uma taxonomia. Nenhum dos trabalhos analisados utiliza instâncias dos conceitos (indivíduos) descritos na ontologia, ainda que algumas das ontologias utilizadas englobassem a definição de instâncias.

3.1.3.3 Método de construção e abrangência

A respeito da questão de pesquisa 3 (QP3), dos 11 trabalhos analisados, 7 utilizaram ontologias que foram geradas automaticamente, 2 utilizaram ontologias geradas manualmente e 2 utilizaram ontologias híbridas, com partes geradas automaticamente e partes geradas ma-

nualmente. 9 de 11 ontologias utilizadas eram formadas apenas por conceitos e relações entre conceitos, variando de centenas a milhares de conceitos definidos. As duas restantes também englobavam a definição de instâncias, sendo que ambas possuíam milhares de instâncias definidas.

3.1.3.4 Resultados alcançados

Quanto a questão de pesquisa 4 (QP4), como os trabalhos analisados avaliam os resultados alcançados pelos métodos propostos de maneira distinta, e com objetivos distintos, a avaliação como descrita qualitativamente pelos autores de cada trabalho foi levada em consideração. Foram consideradas três possíveis opções para a avaliação expressa pelos autores: (1) **bons** resultados, (2) resultados **medianos** ou (3) resultados **ruins**. Essa avaliação é, claro, subjetiva. Contudo, lendo e analisando os resultados reportados foi possível identificar claramente a avaliação feita pelos autores dentro dessas três possibilidades em 10 dos 11 trabalhos analisados. Apenas no trabalho de Chandu et al. (2017) não foi possível identificar claramente qual era a avaliação dos autores. Dos 10 trabalhos remanescentes, 7 classificaram os resultados alcançados como medianos, 3 como bons e nenhum como ruins.

3.1.3.5 Comparação entre os resultados selecionados

A tabela 2 apresenta um quadro comparativo, construído a partir das respostas às questões de pesquisa QP2, QP3 e QP4. A questão de pesquisa 1 (QP1) foi excluída da comparação porque o objetivo da utilização das ontologias foi praticamente o mesmo em todos os trabalhos: criar uma representação da semântica das sentenças (a exceção do trabalho de Kumar et al. (2018) como apresentado na seção 3.1.3.1). Os componentes das ontologias foram abreviados como: C para Conceitos, I para Instâncias e R para Relações.

Trabalho	QP2	QP3	QP4
Ahmad e Ahmad (2019)	C	Automática	Medianos
MacAvaney et al. (2019)	C	Híbrida	Bons
Kumar et al. (2018)	C	Automática	Medianos
Chandu et al. (2017)	C	Híbrida	-
Mohamed e Oussalah (2015)	C	Automática	Medianos
Li e Li (2014)	C,R	Automática	Medianos
Hipola et al. (2014)	C,R	Manual	Bons
Baralis et al. (2013)	C	Automática	Bons
Wu et al. (2013)	C,R	Manual	Medianos
Bawakid e Oussalah (2011)	C	Automática	Medianos
Umbrath, Wetzker e Hennig (2008)	C,R	Automática	Medianos

Tabela 2 – Quadro comparativo das respostas obtidas através da RSL as questões de pesquisa: (**QP2**) Quais componentes de ontologias são usados?, (**QP3**) Como as ontologias utilizadas são construídas e qual é o seu tamanho? e (**QP4**) Como os **autores** classificam os resultados alcançados pelos métodos de sumarização de texto que usam ontologias?

3.2 OUTROS TRABALHOS RELEVANTES

Alguns trabalhos conhecidos não foram incluídos no conjunto analisado na RSL por não terem atingido o critério de inclusão número 3. Apesar disso são trabalhos importantes da área de sumarização automática de texto que utilizam técnicas análogas aquelas utilizadas pelos trabalhos analisados na RSL, e que por isso devem também ser considerados partes do estado da arte e formadores do arcabouço de conhecimento que embasa essa dissertação. Por esse motivo, os mesmos são apresentados nesta seção.

Os trabalhos apresentados a seguir utilizam uma linguagem conceitual denominada UNL (sigla em inglês para *Universal Networking Language*. O objetivo desta linguagem é criar uma representação intermediária de trechos de texto (sentenças) que independe da língua natural na qual o mesmo foi escrito. Quando da introdução desta linguagem (UCHIDA, 1996) a ideia era a de que codificadores poderiam ser escritos para transformar trechos de texto em qualquer linguagem natural para trechos em UNL. Estes trechos em UNL formariam então a representação do texto original em língua natural a ser utilizada em diferentes tarefas de PLN. Essa representação é formada por um conjunto de relações binárias entre UWs (da sigla em inglês para *Universal Words*). Cada relação binária é associada a uma etiqueta (*label*) que define a semântica desta relação. É possível traçar um paralelo entre as representações formadas por um conjunto de UWs e relações entre estas e por um conjunto de instâncias ou conceitos e relações entre estes, daí os métodos que utilizam UNL para criar uma representação intermediária das sentenças poderem ser considerados análogos aqueles que utilizam conceitos definidos em uma ontologia e relações entre os mesmos.

No trabalho apresentado por Martins e Rino (2001) os autores propõe um método de sumarização baseado na utilização de UNL para criar uma representação intermediária das sentenças de entrada e de heurísticas de poda que utilizam esta representação. Estas heurísticas

foram utilizadas para podar cada uma das sentenças de entrada representadas em UNL removendo trechos considerados irrelevantes. O sumarizador proposto é chamado de UNLSumm. Um total de 58 heurísticas foram definidas. Estas heurísticas foram divididas em dois grupos: heurísticas de poda simples e encadeada. As primeiras realizavam a poda de uma determinada relação dentro da representação das sentenças baseadas exclusivamente na análise da etiqueta de uma relação binária, enquanto as segundas realizavam a análise de mais de uma relação binária. Por realizar a sumarização removendo trechos das sentenças considerados irrelevantes o método proposto poderia ser considerado um método extrativo. Por outro lado, por utilizar um decodificador de UNL para língua natural, dado que o método proposto só faz a sumarização do texto em UNL, o mesmo poderia ser classificado como abstrativo, sendo difícil avaliar qual das duas classificações é a mais correta. As autoras avaliaram o método proposto em dois corpora, através da taxa de compressão e de uma análise qualitativa dos sumários gerados reportando que os resultados alcançados, apesar de não representativos, são úteis no cenário de utilização de UNL.

No ano seguinte, as mesmas autoras apresentaram um estudo de caso conduzido para identificar heurísticas possivelmente ruins e algumas melhorias feitas ao UNLSumm com base nesse estudo de caso (MARTINS; RINO, 2002). O estudo buscou avaliar cada uma das 84 heurísticas inicialmente propostas nesta iteração do UNLSumm de acordo com quatro restrições que, segundo as autoras, deveriam ser satisfeitas para que uma heurística pudesse ser considerada uma boa heurística. As heurísticas foram divididas em dois grupos (A e B) e o estudo de caso em seis partes, para que fosse possível avaliar heurísticas simples e encadeadas em diferentes contextos. A análise dos resultados alcançados em cada uma das partes do estudo de caso levou a algumas conclusões, entre as quais destaca-se a de que quanto maior o número de heurísticas usadas para comprimir as sentenças de entrada, piores os resultados alcançados na tarefa de sumarização. Quanto mais heurísticas foram usadas, maior foi a compressão e menos legível o sumário final, sendo esta uma característica comum a outros sumarizadores que fazem compressão intra-sentença, segundo as autoras. A partir destas conclusões algumas alterações foram feitas ao UNLSUmm, sendo este chamado após as alterações de UNLSumm 2.0.

No trabalho de Chaud e Felippo (2018) uma investigação acerca de diferentes estratégias de seleção de sentenças é realizada. Estas estratégias são propostas objetivando a sumarização multi-documento multilíngue. A base da construção das estratégias são três medidas de análise léxica-conceitual: (i) a frequência de conceitos; (ii) a frequência de conceitos corrigida pelo inverso da frequência dos documentos e (iii) a frequência de conceitos normalizada pelo número de conceitos em cada sentença. Estas medidas são aplicadas a todas as UWs presentes nas representações das sentenças de entrada em UNL, considerando-se que cada UW equivale a um conceito. As sentenças são então ordenadas de acordo com o somatório dos valores recebidos por cada uma das UWs que as constituem. Três ordens diferentes são produzidas, uma para cada medida utilizada. Estudos foram realizados utilizando o corpus CM2Corpus, e a análise dos resultados revelou que nem sempre sentenças com os termos mais frequentes no corpus são as mais relevantes, reforçando o valor de analisar diferentes medidas de análise conceitual

no contexto da sumarização extrativa. Por fim, os autores sugerem que novos estudos sejam realizados utilizando corpus maiores e com documentos de áreas diferentes da dos documentos usados nos experimentos (notícias), visto que documentos desta área tendem a posicionar as informações mais relevantes nas primeiras sentenças.

3.3 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o estado da arte em sumarização automática de texto utilizando ontologias, obtido através de uma revisão sistemática da literatura e da análise de alguns outros trabalhos relevantes não analisados na RSL. Foi possível observar que as ontologias têm sido empregadas de maneira muito semelhante em diversos trabalhos: como uma base semântica que permite criar uma representação da semântica das sentenças para, a partir dessa representação, definir medidas e métodos que permitam comparar as sentenças entre si ou as manipular com o objetivo de gerar novas sentenças. A forma como as ontologias são construídas e os resultados alcançados variam, mas não aparentam estar diretamente relacionados. O aspecto mais interessante revelado pela RSL, no entanto, se refere a quais componentes das ontologias são efetivamente usados.

Nota-se que a utilização dos componentes das ontologias é bastante limitada. A maioria dos trabalhos analisados durante a RSL utiliza as ontologias como uma taxonomia de conceitos, definida por relações *is_a*, e ignora todo o potencial descritivo das mesmas que está codificado nas instâncias, nas descrições destas instâncias e nas relações entre elas, assim como na descrição dos conceitos e em outros tipos de relação entre eles que não as do tipo *is_a*. Esse potencial descritivo pode contribuir para melhores resultados na sumarização de texto. Esta oportunidade de pesquisa motivou o trabalho que é apresentado nessa dissertação e ajudou a posicionar suas contribuições.

4 SUMARIZAÇÃO BASEADA EM INSTÂNCIAS

Este capítulo está dividido em três seções que apresentam o método de sumarização extrativa baseado em instâncias de uma ontologia proposto neste trabalho. Na primeira seção (4.1), apresentam-se as técnicas utilizadas para construir a representação semântica das sentenças de um documento através de instâncias definidas em uma ontologia, bem como as medidas que embasam o processo de sumarização, definidas a partir desta representação. Na segunda seção (4.2), apresenta-se o modelo teórico que serve de base para a definição do método proposto neste trabalho. Na terceira e última seção (4.3), apresenta-se como o modelo teórico base foi estendido a partir das definições apresentadas na primeira seção, constituindo assim o método proposto neste trabalho.

4.1 SIMILARIDADE SEMÂNTICA ENTRE SENTENÇAS

A construção de uma representação para sentenças que utilize as instâncias definidas em uma ontologia, possibilita que a similaridade semântica entre estas sentenças seja medida. Esta possibilidade de medir a similaridade semântica entre sentenças, por sua vez, pode ser explorada de diferentes maneiras na construção de um método de sumarização extrativa.

A tarefa de sumarização extrativa de texto pode ser entendida como a tarefa de se selecionar quais sentenças deverão ser extraídas de um conjunto de sentenças presentes em documentos originais para formar um sumário destes documentos. Essa seleção pode ser feita de diversas maneiras e usualmente leva em conta a relevância de cada sentença para a construção do sumário de acordo com a finalidade do mesmo e a necessidade de se evitar a redundância no sumário para que seja possível atingir esta finalidade dentro dos limites de tamanho impostos.

Em um sumário cuja finalidade é representar o mais fielmente possível todas as informações presentes nos documentos originais (sumarização genérica), por exemplo, é usualmente necessário avaliar o quanto cada uma das sentenças representa estas informações, através de alguma medida de avaliação. Já em um sumário cujo finalidade é atender a necessidade específica de um usuário expressa, por exemplo, através de uma consulta (sumarização focada em consulta), pode-se avaliar o quanto cada uma das sentenças atende aos objetivos expressos nessa consulta através, também, de uma medida de avaliação. Em ambos os casos a similaridade semântica pode ser usada como esta medida de avaliação capaz de fornecer subsídios para a tomada de decisão sobre a inclusão ou não de uma sentença no sumário. Quanto mais semanticamente similar uma sentença for aos documentos originais, no caso da sumarização genérica, ou à consulta, no caso da sumarização focada em consulta, mais relevante será a inclusão desta sentença para o atingimento da finalidade do sumário.

Também é necessário evitar a redundância dentro do sumário. A redundância ocorre naturalmente em documentos em língua natural sendo muitas vezes um aspecto desejado dos mesmos. Em um sumário, porém, devido a limitação de espaço, é desejável que as sentenças sejam o menos redundantes entre si possível. Desta forma é possível ampliar o número

de assuntos distintos abordados, utilizando o mesmo espaço. Também é possível utilizar a similaridade semântica como uma medida que indica o quanto duas sentenças são redundantes. Quanto mais semanticamente similares forem duas sentenças, mais redundantes tendem a ser. Quanto menos semanticamente similares, menos redundantes tendem a ser.

Nas seções seguintes, apresenta-se a estratégia de construção da representação das sentenças através de instâncias de uma ontologia, e também como esta representação pode ser utilizada para comparar semanticamente quaisquer conjuntos de sentenças.

4.1.1 Representando sentenças através de instâncias

Para poder calcular a similaridade semântica entre sentenças através do uso de instâncias de uma ontologia é necessário primeiramente construir uma representação de ambas que utilize as referidas instâncias. Para tanto, é preciso identificar quais instâncias são mencionadas em cada sentença e construir uma ligação entre estas instâncias e a sentença, o que é feito através do uso um sistema específico para ligar instâncias de uma ontologia a trechos de texto onde as mesmas são mencionadas. Estes sistema é chamado de sistema de ligação de instâncias e o seu funcionamento está baseado na execução de duas tarefas de extração de informação: Reconhecimento de entidades nomeadas, ou NER da sigla em inglês para *Named Entity Recognition*, e ligação de entidades, ou EL da sigla em inglês para *Entity Linking*. O papel de NER e EL, respectivamente, é o de localizar menções a entidades nomeadas em trechos de texto, e ligar essas menções as respectivas entradas em uma base de conhecimento, que no caso desta proposta é uma ontologia.

Um sistema de ligação de instâncias ou ILS, da sigla em inglês para *Instances Linking System*, recebe como entrada um trecho de texto em língua natural e tem como saída um conjunto de instâncias de uma ontologia que são mencionadas naquele trecho de texto, como demonstra a figura 6. Nesta dissertação, o processo de ligação de instâncias a um trecho de texto em língua natural é chamado também de processo de anotação de instâncias. Neste trabalho a entrada do ILS é sempre uma sentença.

A saída do ILS, ou seja, o conjunto de instâncias definidas na ontologia e mencionadas na sentença de entrada, passa a ser a representação interna daquela sentença utilizada no processo de sumarização. A figura 7 apresenta o exemplo de uma sentença e da representação construída para a mesma através do uso o ILS.

Um problema tipicamente enfrentado por um ILS é a ausência de instâncias a serem ligadas a um determinado trecho de texto de entrada devido a ontologia não definir nenhuma instância que represente entidades porventura mencionadas naquele trecho, ou a alguma ineficiência interna do ILS. Para contornar este problema, sistemas ILS podem oferecer um parâmetro que configura a confiança mínima necessária para o ILS ligar uma instância. A confiança expressa o grau de certeza do ILS de que um determinado trecho de texto faz referência a uma instância. Quanto menor for o valor configurado para este parâmetro, maior tenderá a ser o número de instâncias ligadas e por consequência mais trechos de texto terão uma representação

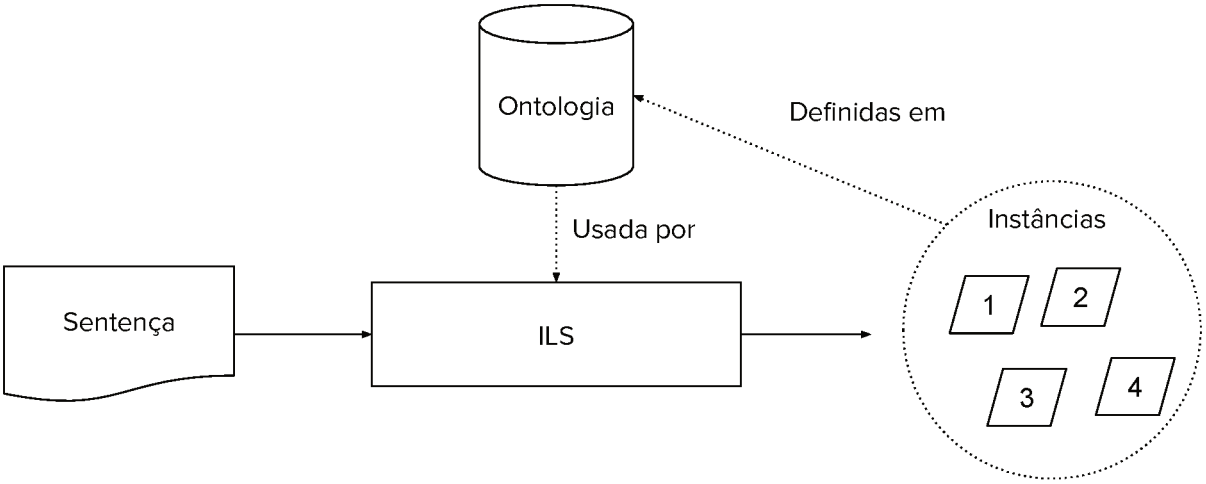
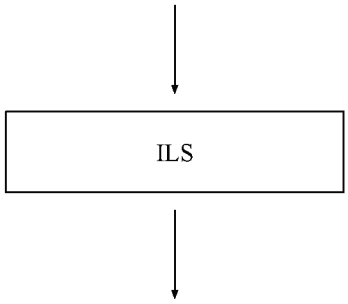


Figura 6 – Funcionamento de um ILS

Entrada - sentença em língua natural:

Está chegando a hora! Nesta quinta-feira (30), Toronto Raptors e Golden State Warriors entram em quadra, em Toronto, às 22h, para começar a decidir o destino do troféu Larry O'Brien. Esse será o primeiro de possivelmente sete jogos.



Saída, lista de instâncias definidas em uma ontologia, passa a ser a representação da sentença no processo de sumarização.

[Toronto Raptors, Golden State Warriors, Toronto, Larry O'Brien]

Figura 7 – Exemplo de sentença com a respectiva representação interna que seria usada no processo de sumarização

válida, ou seja, não-vazia, com pelo menos uma instância ligada. Certificar-se de que o maior número possível de sentenças de um documento tenham uma representação válida através de instâncias é fundamental para garantir o bom funcionamento de um sumarizador baseado em instâncias. O algoritmo 1 apresenta um método para lidar com este problema, baseado na possibilidade de configurar um parâmetro que controla a confiança mínima para anotação de um ILS. Assume-se que este parâmetro pode receber valores entre 0 e 1, onde 0 significa confiança mínima, ou seja, que o ILS pode anotar uma instância ainda que a sua confiança de que aquela instância foi mencionada no texto seja a mínima aceitável definida pelo próprio ILS, e 1

significa confiança máxima.

Algoritmo 1 Ligação de Instâncias com confiança mínima variável

1: **Entrada** T: Texto de uma sentença ou consulta.
2: **Entrada** c: Valor inicial para confiança mínima para ligação.
3: **Método** LINK(T, c)
4: $s \leftarrow ILS(T, c)$
5: **se** $s = \emptyset$ & $c >= 0.1$ **então**
6: $c \leftarrow c - 0.1$
7: $s \leftarrow LINK(T, c)$
retorne s

Ligar instâncias a um trecho de texto em um nível de confiança mais baixo pode significar que instâncias possivelmente menos relevantes serão ligadas. Estas instâncias podem ser consideradas como ruído se não contribuírem para a melhora dos resultados alcançados na sumarização. Contudo, a intuição por trás da proposição do algoritmo 1 é a de que a melhora dos resultados possivelmente gerada por um maior número de sentenças com representações não-vazias compensa qualquer ruído porventura introduzido pelo mesmo.

4.1.2 Similaridade entre conjuntos de instâncias

Para calcular a similaridade semântica entre sentenças utiliza-se a representação das mesmas através das instâncias definidas em uma ontologia. Estas representações são conjuntos de instâncias, portanto a similaridade semântica entre duas sentenças é a similaridade semântica entre os dois conjuntos de instâncias que as representam. Usando esta mesma definição compara-se semanticamente também conjuntos de sentenças, unindo para isso todas as suas representações e utilizando o conjunto de instâncias resultante desta união na comparação.

A similaridade semântica entre dois conjuntos de instâncias é definida como a média das similaridades máximas entre cada instância de cada um dos conjuntos, quando comparada as instâncias do outro conjunto, conforme a definição apresentada na expressão 4.1. Essa definição é uma adaptação da definição de similaridade entre dois conjuntos de palavras apresentada por Mohamed e Oussalah (2015). Enquanto Mohamed e Oussalah (2015) definem uma medida que compara dois conjuntos de palavras que representam conceitos definidos em uma taxonomia, compondo uma sentença e uma consulta, a adaptação apresentada neste trabalho define uma medida que compara dois conjuntos quaisquer de instâncias definidas em uma ontologia, podendo estes conjuntos compor as representações de qualquer unidades textuais. A definição apresentada na expressão 4.1 assume uma contribuição simétrica de cada uma das instâncias nos conjuntos sob comparação.

$$Sim(I_1, I_2) = \frac{1}{2}\left[\frac{\sum_{i_1 \in I_1} \max_{i_2 \in I_2} sim(i_1, i_2)}{|I_1|} + \frac{\sum_{i_2 \in I_2} \max_{i_1 \in I_1} sim(i_2, i_1)}{|I_2|}\right] \tag{4.1}$$

Onde I_1 e I_2 são conjuntos de instâncias e *sim* é uma medida de similaridade semântica entre duas instâncias de uma ontologia. A medida de similaridade entre duas instâncias de uma ontologia utilizada neste trabalho é definida e descrita na seção seguinte (4.1.3).

Esta definição de similaridade semântica entre conjuntos de instâncias garante que a contribuição à medida de similaridade dada por uma instância em particular não será afetada quando novas instâncias forem adicionadas aos demais conjuntos se estas novas instâncias forem menos similares a esta instância em particular do que a instância mais similar já presente. Em outras palavras, o valor máximo possível para a similaridade de uma instância em particular não irá diminuir com a adição de novas instâncias ao conjunto com o qual esta instância está sendo comparada. Quando usada em conjunto com o algoritmo 1 (definido na seção 4.1.1), esta definição garante que instâncias ligadas as sentenças com confiança de ligação menor não diminuam a similaridade entre as instâncias já ligadas anteriormente e os demais conjuntos. É importante salientar, no entanto, que a adição de novas instâncias ao conjunto que representa uma sentença aumenta o número total de instâncias neste conjunto, o que aumenta o valor do denominador em um dos lados da soma da expressão 4.1. Portanto, a adição de novas instâncias aos conjuntos que representam as sentenças pode alterar a similaridade entre as sentenças, mas esta alteração se deve as instâncias adicionadas e não a alteração da similaridade das instâncias previamente presentes nos conjuntos.

4.1.3 Similaridade entre instâncias

Para que seja possível calcular a similaridade entre dois conjuntos de instâncias é preciso primeiro obter a similaridade entre duas instâncias. A medida de similaridade semântica entre duas instâncias utilizada neste trabalho baseia-se na definição teórica de similaridade apresentada por Lin (1998). Segundo Lin (1998) "A similaridade entre A e B é medida pela razão entre a quantidade de informação necessária para expressar tudo que é comum a A e B e a quantidade de informação necessária para descrever totalmente o que A e B são.". Esta definição apresentada pelo autor é baseada na teoria da informação, que diz que "a quantidade de informação contida em uma afirmação é medida pelo logaritmo negativo da probabilidade daquela afirmação" (LIN, 1998). A partir destas definições Lin (1998) apresenta uma série de suposições e a partir dela define a similaridade com a seguinte expressão:

$$sim(A,B) = \frac{logP(comum(A,B))}{logP(descrever(A,B))} \tag{4.2}$$

Portanto, segundo Lin (1998), a similaridade entre A e B é dada pela razão entre o logaritmo da probabilidade de todas as afirmações necessárias para descrever tudo que é comum a A e B e o logaritmo da probabilidade de todas as afirmações necessárias para descrever completamente o que A e B são.

Para poder interpretar essa definição e aplicá-la a este trabalho é preciso saber quais são as afirmações necessárias para descrever uma instância de uma ontologia. Neste trabalho, entende-se que estas afirmações são aquelas necessárias para criar uma descrição de cada uma

destas instâncias. Nesta descrição devem estar presentes aspectos relevantes da semântica de cada instância.

As relações que uma instância de uma ontologia mantém com outras instâncias representam uma parte importante de sua semântica. O tipo de uma instância - o conceito ao qual esta instância pertence - representa também uma parte importante de sua semântica. Desta forma, define-se que a descrição de uma instância é formada por suas relações e tipos, para fins de calcular a sua similaridade semântica com outras instâncias. Além disso, define-se que as relações de uma instância contribuem com duas informações distintas para a descrição da mesma, separadamente: seus tipos e as instâncias as quais estas se conectam. Assim sendo, a descrição de uma instância é formada por três conjuntos distintos de informação: seus **tipos**, os **tipos de suas relações** e as **instâncias de suas relações**. A figura 8 ilustra duas instâncias de conceitos distintos que são relacionadas uma a outra e as suas descrições, tal qual seriam utilizadas para computar a similaridade semântica entre ambas.

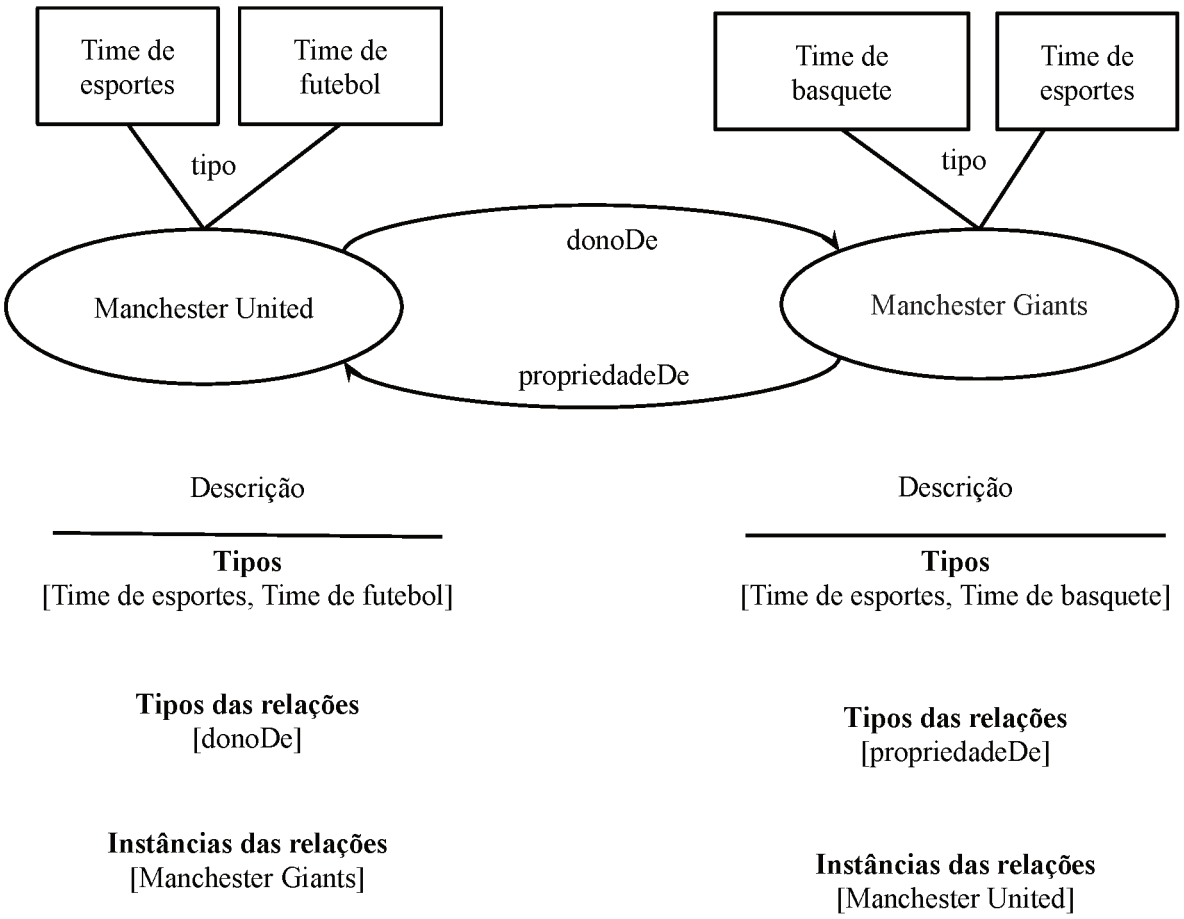


Figura 8 – Exemplo de descrição de instâncias

Visando interpretar a definição de similaridade apresentada por Lin (1998) a este trabalho, é preciso ainda conhecer a probabilidade associada a cada um dos componentes da des-

crição de uma instância. Define-se que a probabilidade associada a um componente qualquer da descrição de uma instância é igual a probabilidade deste componente estar presente em uma instância qualquer da ontologia selecionada ao acaso, como ilustrado na equação 4.3. A probabilidade associada a um tipo de relação, por exemplo, é a probabilidade daquele tipo de relação estar presente em uma instância qualquer da ontologia selecionada ao acaso.

$$P(componente) = \frac{contagem(\text{instâncias com componente})}{contagem(\text{número total de instâncias})} \tag{4.3}$$

Conhecendo a expressão 4.2, as categorias de componentes que formam a descrição de cada instância e a probabilidade associada a cada componente, conforme definida na expressão 4.3, define-se que neste trabalho a similaridade entre duas instâncias é então calculada indiretamente através da similaridade entre cada uma das categorias de suas descrições. A similaridade entre cada uma das categorias das descrições de duas instâncias (tipos, tipos das relações e instâncias das relações) é dada pela razão entre o dobro do somatório do logaritmo das probabilidades de todos os componentes daquela categoria comuns a ambas as instâncias e o somatório do logaritmo das probabilidades de todos os componentes daquela categoria que descrevem cada uma das instâncias. Esta definição deriva diretamente da expressão 4.2 e da interpretação da mesma oferecida por Lin (1998) para o cálculo de similaridade semântica, e é apresentada na expressão 4.4 onde *cat* é uma função que retorna todos os componentes de uma categoria da descrição de uma instância.

$$Sim_{categoria}(A,B) = \frac{2*\left(\sum_{c\in(cat(A)\cap cat(B))} logP(c)\right)}{\left(\sum_{c\in cat(A)} logP(c)\right)+\left(\sum_{c\in cat(B)} logP(c)\right)} \tag{4.4}$$

Por fim, a similaridade entre duas instâncias é definida como a média das similaridades entre cada uma das categorias da sua descrição:

$$sim(A,B) = \frac{1}{3}\left[Sim_{tipos}(A,B) + Sim_{tiposRel}(A,B) + Sim_{instanciasRel}(A,B)\right] \tag{4.5}$$

A tabela 3 apresenta a similaridade entre algumas instâncias definidas na ontologia da DBPedia de 2014, calculadas seguindo as definições apresentadas nesta seção. Os resultados apresentados nesta tabela estão alinhados com os resultados esperados de uma medida de similaridade que segue as premissas definidas por Lin (1998), o que é evidenciado pelas seguintes observações:

- Similaridade máxima é atingida quando uma instância é comparada com ela mesma (linha 1)
- Times na mesma liga - *Los Angeles Lakers* e *Golden State Warriors* estão na mesma liga, assim como *New England Patriots* e *Seattle Seahawks* - tem similaridade mais alta quando comparados entre si do que quando comparados com times em outras ligas (linhas 2, 4 e 3)

- Os valores das similaridade para times na mesma liga são próximos (linhas 2 e 4)
- A similaridade entre um time de esportes e um personagem de ficção é uma ordem de magnitude menor do que entre dois times de esportes de ligas diferentes (linhas 3 e 5)

#	Medida	Valor
1	*Similaridade*(L.A. Lakers, L.A. Lakers)	1
2	*Similaridade*(L.A. Lakers, G.S. Warriors)	0.6248
3	*Similaridade*(L.A. Lakers, N.E. Patriots)	0.3958
4	*Similaridade*(N.E. Patriots, S. Seahawks)	0.6301
5	*Similaridade*(L.A. Lakers, Spider Man)	0.0363

Tabela 3 – Similaridade entre instâncias da ontologia da DBPedia de 2014

4.2 MÉTODO BASE

Diversos métodos de sumarização extrativa existentes na literatura utilizam uma medida de similaridade entre sentenças para comparar as mesmas e definir, a partir desta comparação, quais comporão o sumário. Com uma medida de similaridade semântica entre sentenças representadas através de instâncias de uma ontologia definida, foi possível estender um destes métodos para que utilizasse a referida medida. O método selecionado, e que portanto serve de base para o método proposto neste trabalho, foi definido por Carbonell e Goldstein (1998) e implementado através do algoritmo MMR. Este método já era conhecido pelo autor dessa dissertação e foi selecionado sem que outros métodos tenham sido considerados, uma vez que este atendia a todos os critérios necessários para garantir o atingimento dos objetivos deste trabalho (a saber: (i) método de sumarização extrativa e (ii) extensível através da utilização de uma medida de similaridade entre as sentenças) além de ser conceitualmente simples, portanto fácil de estender, bastante flexível quanto a entrada da sumarização e quanto a finalidade do sumário, e também bastante utilizado na literatura. Apesar de ter sido originalmente proposto para a tarefa de sumarização extrativa focada em consultas, este método já foi utilizado também para a tarefa de sumarização genérica em trabalhos como os de Murray, Renals e Carletta (2005) e Gong e Liu (2001).

O algoritmo MMR funciona iterativamente, e a cada iteração seleciona uma sentença para ser extraída e incluída no sumário, até que o tamanho desejado seja atingido. As sentenças são avaliadas através de sua (i) similaridade à uma consulta que expressa uma necessidade de um usuários específico e (ii) diversidade adicionada ao sumário quando comparadas as sentenças previamente selecionadas. A sentença selecionada para extração é sempre aquela que melhor combinar estas características, ponderadas pelo parâmetro α. Quanto maior o valor de α, mais peso é dado a similaridade com a consulta. Quanto menor o valor de α mais peso é dado a diversidade adicionada. A expressão 4.6 define o passo iterativo do algoritmo:

$$MMR \stackrel{def}{=} \max_{D_i \in R/S} \left[\alpha(sim_1(D_i, Q)) - (1-\alpha) \max_{D_j \in S} sim_2(D_i, D_j) \right] \quad (4.6)$$

Onde Q é uma consulta que expressa as necessidades de um usuário específico, R é uma coleção de sentenças, S é um subconjunto das sentenças em R já selecionadas, R/S é o conjunto das sentenças ainda não selecionados e sim_1 e sim_2 são medidas de similaridade.

4.3 MÉTODO PROPOSTO

A partir da construção da representação semântica das sentenças utilizando instâncias de uma ontologia, propõe-se a extensão do método base - MMR - através da realização da comparação semântica entre sentenças utilizando a medida de similaridade descrita na seção 4.1.2, definindo assim o que chama-se nesta dissertação de método de sumarização baseado em instâncias ou SBI. O SBI, por se basear no algoritmo MMR, segue a ideia de seleção iterativa de sentenças, sendo que seu passo iterativo é definido pela expressão 4.7, onde S_Q são as sentenças da consulta, S_D são as sentenças do documento ou documentos originais, S_S é o subconjunto das sentenças selecionadas, S_D/S_S são as sentenças ainda não selecionadas, *sim* é a medida de similaridade definida na seção 4.1.2 e α é um parâmetro que controla a ponderação entre similaridade a consulta ou diversidade adicionada ao sumário na hora de selecionar uma sentença.

$$SBI \stackrel{def}{=} \max_{S_i \in S_D/S_S} \left[\alpha(Sim(S_i, S_Q)) - (1-\alpha) \max_{S_j \in S} Sim(S_i, S_j) \right] \quad (4.7)$$

De maneira mais ampla, SBI é composto de dois processos principais: (i) Ligação de instâncias às sentenças de entrada e (ii) seleção e extração destas sentenças. A figura 9 apresenta uma representação do método, com destaque para estes dois principais processos e suas respectivas entradas e saídas. Primeiramente, instâncias são ligadas as sentenças de entrada, através de um sistema de anotação de instâncias. Esta ligação de instâncias pode acontecer com um valor fixo para a confiança necessária para a anotação configurado ou utilizando a estratégia de confiança variável discutida na seção 4.1.1. Depois disso as sentenças são selecionadas e extraídas iterativamente seguindo a definição apresentada na expressão 4.7.

Assim como acontece com o algoritmo MMR, o método proposto pode ser usado tanto para sumarização baseada em consulta quanto para sumarização genérica. Neste último caso os próprios documentos de entrada são utilizados como consulta, ou seja, a união de todas as sentenças que compõe os documentos de entrada representa as sentenças da consulta. Como citado na seção 4.2 essa mesma abordagem já foi empregada na utilização do algoritmo MMR para sumarização genérica. Esta abordagem está baseada no entendimento de que um sumário genérico deverá ser semanticamente similar aos documentos originais e também, por definição, à consulta, sendo portanto os próprios documentos originais uma boa representação de tal consulta.

Ainda no que tange as entradas aceitas pelo SBI, o mesmo pode ser usado tanto para a sumarização monodocumento quando para a sumarização multidocumento. No caso desta

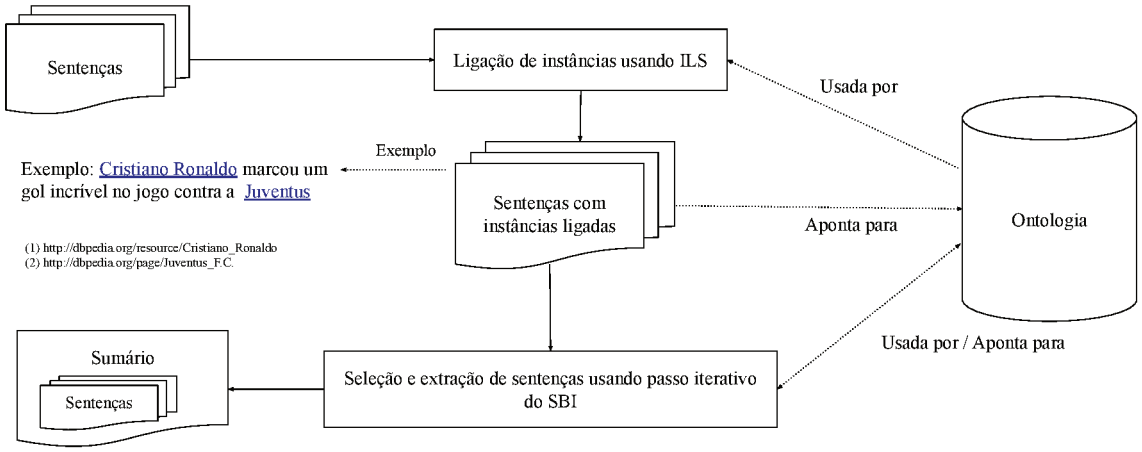


Figura 9 – Estrutura do método proposto nesta dissertação

última, os documentos de entradas devem ser concatenados e tratados como um único documento. É fato que esta não é uma abordagem para a sumarização multidocumento que leve em conta as suas particularidades e desafios específicos, mas também é fato que a seleção iterativa focando no aumento da relevância e diminuição da redundância aliada a análise semântica das sentenças garante que as múltiplas entradas possam ser concatenadas, em qualquer ordem, sem afetar a seleção e extração das sentenças que formarão o sumário.

4.4 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado o método de sumarização extrativa baseado em instâncias de uma ontologia, denominado SBI, que é a principal contribuição e o principal resultado alcançado durante o desenvolvimento deste mestrado. Este método baseia-se na construção de uma representação para as sentenças dos documentos a serem sumarizados que utiliza as instâncias definidas em uma ontologia, possibilitando a comparação semântica entre as sentenças. A estratégia de construção da representação baseada em instâncias assim como a definição e apresentação da medida de similaridade que possibilita a comparação semântica são apresentadas no começo do capítulo.

A partir da definição da estratégia de construção da representação das sentenças e da medida de similaridade, estende-se o método base, representado pelo algoritmo MMR para formar o método proposto neste trabalho (SBI). Tanto o método base quanto o método proposto são apresentados neste capítulo.

A partir da apresentação e proposição do SBI feitas neste capítulo, segue-se para a apresentação dos experimentos realizados para analisar a relevância dos resultados obtidos pelo mesmo na tarefa de sumarização extrativa de texto, em diferentes configurações da mesma, as quais serão apresentadas no capítulo seguinte (capítulo 5), para que se cumpra o objetivo geral deste trabalho de propor um método de sumarização extrativa baseado em instâncias de uma

ontologia, analisando a relevância dos resultados obtidos pelo mesmo.

5 EXPERIMENTOS E RESULTADOS

Apresentam-se neste capítulo os experimentos realizados a fim de analisar a relevância dos resultados obtidos pelo SBI em diferentes configurações da tarefa de sumarização extrativa, possibilitando o cumprimento do objetivo geral deste trabalho. Os experimentos foram definidos de forma a possibilitar a análise da utilização do SBI tanto para sumarização genérica quanto para sumarização focada em consulta. Além disso, o corpora selecionado abrange tanto a sumarização monodocumento, no caso da sumarização genérica, quando multidocumento, no caso da sumarização focada em consulta. Desta forma, as possibilidades de utilização oferecidas pelo método podem ser exploradas em diferentes contextos, possibilitando uma avaliação mais abrangente do mesmo assim como a sua validação através da análise dos resultados alcançados nestes diferentes contextos.

O restante deste capítulo está assim organizado: Inicialmente são descritos detalhes da implementação do método utilizada na realização dos experimentos (seção 5.1). Em seguida apresentam-se os critérios de avaliação a serem usados na avaliação dos resultados alcançados, bem como a medida de avaliação adotada (seção 5.2). Em seguida, apresentam-se os experimentos realizados a fim de avaliar a relevância dos resultados alcançados pelo SBI na tarefa de sumarização extrativa genérica (seção 5.4), seguidos dos experimentos realizados a fim de avaliar a mesma relevância dos resultados alcançados na tarefa de sumarização extrativa focada em consulta (seção 5.3). Por fim, apresenta-se uma discussão destes resultados assim como as considerações finais deste capítulo (seção 5.5).

5.1 IMPLEMENTAÇÃO

Para realizar os experimentos descritos nas seções seguintes implementou-se o método proposto neste trabalho selecionando a versão de 2014 da ontologia da DBPedia para servir de base de conhecimento. Esta ontologia é construída a partir do mapeamento de caixas de informação da Wikipedia para uma única ontologia compartilhada. Além disso, esta ontologia possui links RDF que apontam para pelo menos 30 fontes de dados externas o que possibilita que dados destas fontes sejam utilizados em conjunto com dados da DBPedia (LEHMANN et al., 2014). A versão de 2014 da ontologia da DBPedia consiste de 685 classes, 2795 propriedades e mais de 4 milhões de instâncias. Estas instâncias descrevem o conhecimento de diversos domínios diferentes tornando esta ontologia extremamente genérica e abrangente. Os componentes TBox e ABox da ontologia, contendo respectivamente conceitos e instâncias, são disponibilizados separadamente.

A utilização das instâncias definidas na ontologia da DBPedia no processo de sumarização depende da identificação de referências às mesmas nos documentos a serem sumarizados, no processo de ligação de instâncias, que é realizado através de um ILS. Na implementação do SBI utilizada nos experimentos esse processo é realizado através do sistema *DBPedia Spotlight*, que é um sistema para ligação automática de instâncias da DBPedia a textos em língua natural

(DAIBER et al., 2013). *DBPedia Spotlight* identifica quais instâncias da ontologia da DBPedia são mencionadas em um trecho de texto através de alguma de suas possíveis formas de superfície, que são diferentes formas escritas de se referenciar a mesma instância, e as liga a este trecho de texto. O sistema garante que apenas instâncias são ligadas a este trecho de texto de entrada porque utiliza apenas o componente ABox da ontologia da DBPedia. Além disso, este sistema permite a configuração de alguns parâmetros da ligação das instâncias, tais como proeminência, pertinência ao tópico, ambiguidade contextual e confiança na anotação (MENDES et al., 2011), sendo este último o mais relevante por influenciar diretamente o número de ligações realizadas, afetando assim a proporção de sentenças que possuem uma representação válida, ou seja, que tem ao menos uma instância ligada. A existência deste parâmetro de configuração também é fundamental para permitir a implementação da estratégia de ligação de instâncias com confiança variável definida na seção 4.1.1. O desempenho do *DBPedia Spotlight* na tarefa de ligação de instâncias foi analisado por Mendes et al. (2011), sendo que os resultados alcançados pelo mesmo, segundo os autores, estão dentro de um intervalo "competitivo". A maioria dos resultados alcançados por outros sistemas com os quais foi comparado, para medida-f, mostrando-se piores em todos os valores de confiança mínima para anotação testados.

Para cada um dos experimentos, duas implementações diferentes do SBI foram usadas. Uma delas utilizou um valor fixo para a confiança da ligação de instâncias às sentenças de entrada, enquanto a outra utilizou um valor variável para esta confiança, conforme descrito na seção 4.1.1. A nomenclatura utilizada na apresentação dos resultados nas seções seguinte é um reflexo da utilização destas duas versões do SBI. Assim, por exemplo, o sistema nomeado *VAR-0.6* faz referência a versão da implementação que utilizou confiança na anotação de instâncias variável começando com o valor 0.6, enquanto o sistema nomeado *FIX-0.7* faz referência a versão da implementação que utilizou o valor fixo 0.7 para o parâmetro de confiança na anotação de instâncias. Cada uma das duas versões diferentes da implementação do SBI (VAR e FIX) foi executada com três diferentes valores para a confiança na anotação de instâncias - 0,3; 0,6 e 0,9 - (esses valores referem-se a confiança inicial na implementação que utiliza confiança variável) e cinco valores diferentes para o parâmetro α - 0,3; 0,4; 0,5; 0,6 e 0,7 - totalizando 15 configurações diferentes por implementação disponíveis para os experimentos, perfazendo um total de 30 diferentes versões do método proposto.

Durante a apresentação dos resultados dos experimentos descritos neste capítulo, os resultados alcançados pelos mesmos são agrupados em *sistemas*. Cada sistema corresponde a uma versão específica da implementação do SBI com uma configuração específica para o parâmetro de confiança da ligação de instâncias. A notação usada para descrever o nome dos sistemas apresentados nos resultados segue a seguinte convenção: Cada nome de sistema é formado por um sufixo e um prefixo separados por um traço ("-"). O prefixo indica se aquela versão específica da implementação do modelo utilizou um valor fixo (FIX) ou variável (VAR) para o parâmetro de confiança na anotação de instâncias (feita pelo ILS *DBPedia Spotlight*), conforme descrito na seção 4.1.1. O sufixo indica qual era o valor inicial da confiança na ligação de instâncias no caso de confiança na ligação de instâncias variável, ou qual era o único valor no

caso de confiança na ligação de instâncias fixa. O sistema nomeado *VAR-0.6*, por exemplo, faz referência a implementação do modelo que usou confiança na anotação de instâncias variável iniciando com o valor 0,6.

5.2 CRITÉRIOS DE AVALIAÇÃO

Neste trabalho, adotaremos a avaliação intrínseca usando como critério de avaliação a informatividade ou cobertura de conteúdo, embasando-se no fato deste ser o critério de avaliação mais amplamente adotado na literatura, possibilitando que os resultados obtidos sejam comparados com resultados obtidos em trabalhos já publicados.

5.2.1 Medidas de avaliação

A avaliação da informatividade dos sumários gerados pela implementação do SBI foi feita através da comparação destes sumários com os sumários de referência presentes no corpora selecionado. Esta comparação se deu através da família de medidas ROUGE, que são o padrão *de-facto* para avaliação de sumários extrativos.

Para computar as medidas ROUGE, utilizou-se o script *Perl* ROUGE-1.5.5.pl que foi originalmente usado para computar estas mesmas medidas na competição DUC2005. Os parâmetros usados para calcular os resultados também foram os mesmos usados na competição DUC2005[1], sendo os principais descritos na tabela 4.

-n 2	calcular ROUGE-1 e ROUGE-2
-x	não calcular ROUGE-L
-m	aplicar *stemmer*[2]Porter nos sumários gerados automaticamente e nos sumários de referência
-c 95	usar intervalo de confiança de 95%
-r 1000	número de pontos de amostragem (1000)
-f A	seleciona a fórmula de escore, onde 'A' significa média do modelo considerando múltiplos sumários de referência
-p 0.5	razão entre a importância relativa de precisão e cobertura das medidas ROUGE

Tabela 4 – Descrição do significado dos principais parâmetros utilizados no cálculo das medidas ROUGE para o resultado dos experimentos, utilizando o script *Perl* ROUGE-1.5.5.pl

5.3 SUMARIZAÇÃO FOCADA EM CONSULTA COM SBI

Apresentam-se a seguir os experimentos realizados a fim de avaliar a relevância dos resultados obtidos pela implementação do SBI quando utilizada para a sumarização extrativa focada em consulta, ou seja, aquela na qual a finalidade do sumário é atender a uma necessidade

[1] ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d

[2] Sistema capaz de reduzir uma palavra ao seu radical

específica de um usuário expressada através de uma consulta. Diferentemente do que acontece quando o SBI é utilizado para a sumarização genérica, neste caso não é preciso definir uma consulta artificialmente, uma vez que os corpus específicos para esta finalidade já as definem para cada conjunto de documentos de entrada, a fim de guiar a construção do sumário.

5.3.1 Corpus

O corpus selecionado para a realização dos experimentos apresentados nesta seção foi o corpus DUC2005. Este corpus foi construído para uma conferência de mesmo nome, onde diferentes sumarizadores automáticos competiram pelos melhores resultados, e é apresentado por Dang (2005). O corpus é formado por 50 conjuntos de documentos, cada um contendo entre 25 e 50 documentos diferentes sobre um mesmo tópico, sendo portanto destinado a sumarização multidocumento. Cada conjunto de documentos tem em média 31 documentos e 20.236 palavras.

O número de palavras nos sumários gerados deve ser de 250. Para cada conjunto de documentos estão disponíveis entre 4 e 10 sumários modelo, produzidos por humanos, que são os sumários de referência utilizados para a avaliação dos sumários gerados pelo SBI.

Como este conjunto de dados foi criado especificamente para analisar sumarizadores focados em consulta, cada conjunto de documentos tem uma consulta específica definida que deve ser levada em consideração na hora de gerar sumários para aquele conjunto.

5.3.2 Definição dos Experimentos

Os experimentos realizados consistiram em utilizar uma implementação do SBI para gerar diferentes sumários para cada um dos 50 conjuntos de documentos presentes neste corpus e comparar a avaliação destes sumários através das medidas ROUGE com avaliações obtidas através das mesmas medidas por implementações de outros modelos.

Os modelos selecionados para comparação foram aqueles que mais se aproximavam do método proposto neste trabalho, conforme levantado na revisão sistemática da literatura apresentada no capítulo 3.

Os diferentes sumários gerados pela implementação do modelo apresentado neste trabalho dizem respeito às diferentes configurações aplicadas ao mesmo, conforme descrito na seção 5.1.

5.3.3 Resultados

Apresentam-se a seguir os resultados obtidos nos experimentos realizados para avaliar os sumários gerados pela implementação do SBI na tarefa de sumarização extrativa focada em consulta. Os sumários foram avaliados através das medidas-f de ROUGE-1 e ROUGE-2, pois estas apresentam grande correlação com a avaliação humana (LIN, 2004) sendo padrão

de-facto para avaliação de sumários extrativos. A notação usada para nomear as diferentes versões e configurações do SBI, denominadas sistemas, para as quais se reportam resultados é apresentada na seção 5.1.

A figura 10 apresenta os resultados obtidos para a medida-f de ROUGE-1 por todos os sistemas, com cinco valores diferentes configurados para o parâmetro α (um em cada rodada de experimentos). Este parâmetro controla o equilíbrio entre relevância para a consulta e diversidade do sumário no momento da seleção de novas sentenças para compor o sumário. Observando a figura, é possível perceber que todas as versões do sistema obtiveram resultados melhores conforme a confiança na anotação de indivíduos diminuía, para valores de α maiores ou iguais a 0,5. Com esse valor de α a relevância para a consulta possuía um peso maior no momento de selecionar novas sentenças para compor o sumário. Com o valor de α configurado para valores menores que 0,5 o inverso ocorreu: os resultados pioraram conforme o valor da confiança na anotação diminuiu, com uma queda particularmente aguda entre os sistemas com confiança na anotação configurados em 0,6 e 0,3. É importante notar também que a exceção do sistema FIX-0.9 todos os sistemas atingiram seus melhores resultados com α configurado para 0,7. Esses resultados indicam que quanto mais peso é atribuído a relevância para a consulta, maior a quantidade de instâncias ligadas, melhorando os resultados. Se mais peso é dado a diversidade no sumário, no entanto, mais instâncias ligadas pode levar a piores resultados.

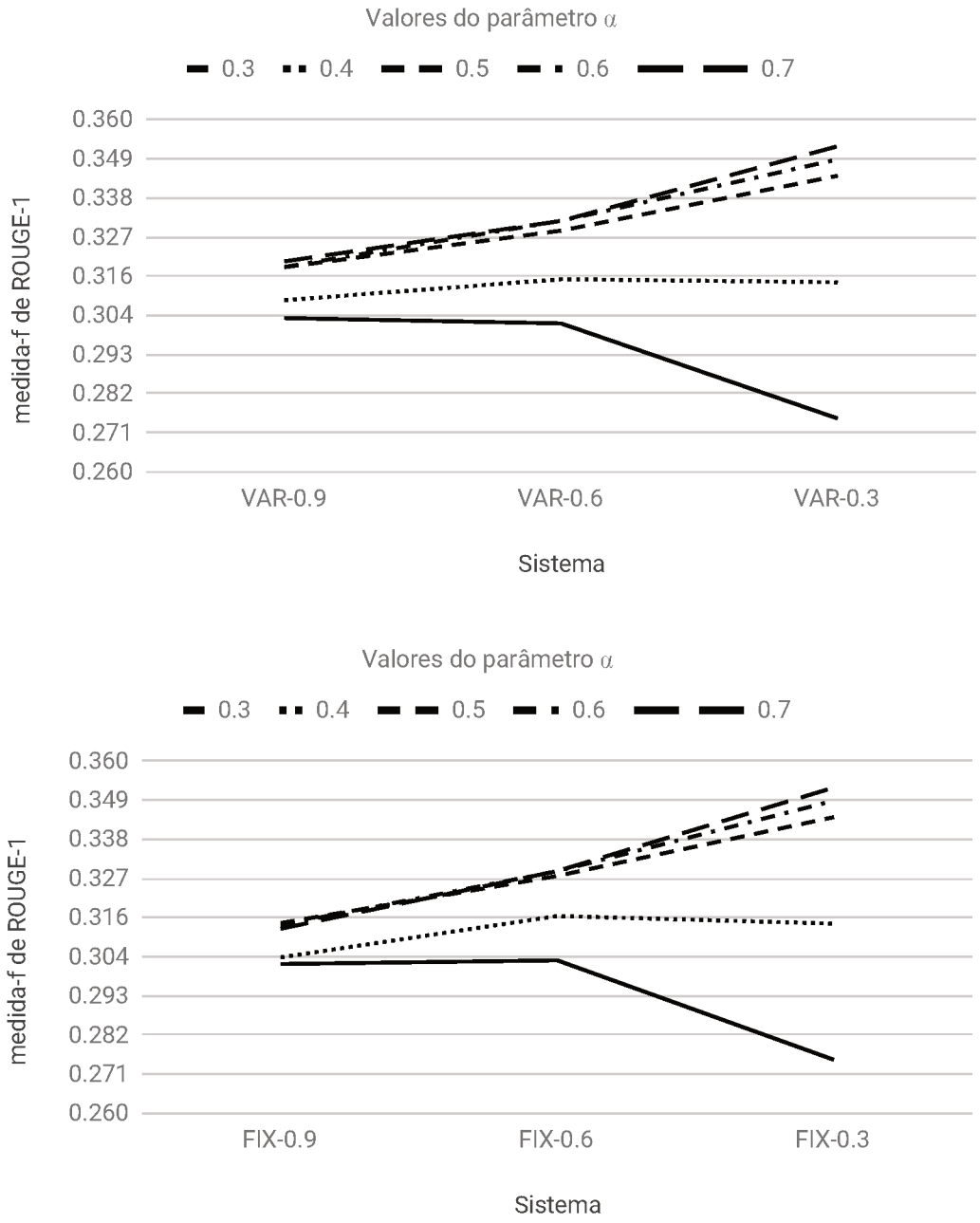


Figura 10 – Resultados obtidos na medida-f de ROUGE-1 por sistema, com cinco valores diferentes de α no corpus DUC2005

A figura 11 apresenta uma comparação entre os sistemas (versões da implementação do método proposto neste trabalho) com confiança na anotação de instâncias fixa e variável através dos valores obtidos para a medida-f de ROUGE-1, para três valores iniciais de confiança de anotação e um valor fixo para o parâmetro α de 0,7. Os sistemas com confiança na anotação variável atingiram melhores resultados dos que os sistemas com confiança na anotação fixa em duas ocasiões.

A tabela 5 apresenta uma comparação entre a média dos resultados obtidos pelos sistemas que participaram da competição da conferência DUC2005, sistema apresentados em traba-

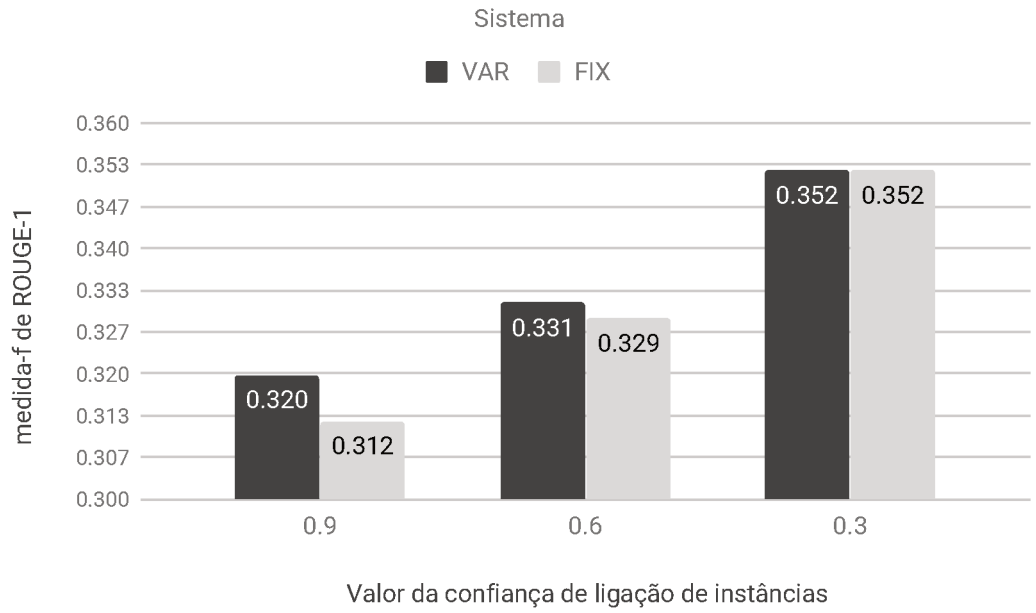


Figura 11 – Resultados obtidos na medida-f de ROUGE-1, por sistema, com um valor fixo para o parâmetro α de 0,7, para o corpus DUC2005

lhos correlatos a este e a melhor versão da implementação do método proposto neste trabalho, o sistema VAR-0.3 com α configurado para 0,7. Nosso sistema é melhor que a média dos resultados do sistemas que participaram da DUC2005 mas também é pior que os demais, em ambas as métricas.

Sistema	Medida-f	
	ROUGE-1	ROUGE-2
Média sistemas DUC2005	0.3434	0.0602
Luo et al. (2013b)	0.3728	**0.0807**
Canhasi e Kononenko (2014)	0.3945	0.0797
Mohamed e Oussalah (2015)	**0.3949**	0.0693
Este trabalho	0.3524	0.0639

Tabela 5 – Comparação entre a média dos resultados obtidos por sistemas que participaram da DUC2005, resultados obtidos por implementações de modelos de trabalhos correlatos e pela implementação do método proposto neste trabalho, para o corpus DUC2005.

A tabela 6 apresenta todos os resultados obtidos durante os experimentos no corpus DUC2005. Alguns destes resultados foram apresentados nas figuras e tabelas anteriores. Os resultados foram obtidos por diferentes sistemas, utilizando as duas diferentes estratégias de anotação de instâncias definidas na seção 4.1.1 (VAR e FIX), diferentes valores iniciais para a confiança na anotação (0,3; 0,6; 0,9) e diferentes valores para o parâmetro α. Observando a

ROUGE-1						
α	VAR-0.9	VAR-0.6	VAR-0.3	FIX-0.9	FIX-0.6	FIX-0.3
0.3	0.30367	0.30215	0.27519	0.30236	0.3034	0.27519
0.4	0.30875	0.31471	0.3138	0.30426	0.31599	0.31380
0.5	0.31811	0.32848	0.34404	0.31399	0.32736	0.34404
0.6	0.31818	0.33128	0.34870	0.31342	0.32871	0.34870
0.7	0.31976	0.33125	0.35242	0.31238	0.32874	0.35242
ROUGE-2						
α	VAR-0.9	VAR-0.6	VAR-0.3	FIX-0.9	FIX-0.6	FIX-0.3
0.3	0.03987	0.03711	0.03538	0.03991	0.03850	0.03538
0.5	0.04834	0.05317	0.05877	0.04634	0.05361	0.05877
0.7	0.04977	0.05356	0.06390	0.04590	0.05280	0.06390

Tabela 6 – Resultados obtidos para a medida-f de ROUGE-1 e ROUGE-2 por todos os sistemas nos experimentos realizados no corpus DUC2005.

tabela é possível constatar que os valores seguem, como esperado, as tendências apresentadas na descrição das figuras 10 e 11, em especial a tendência de resultados melhores para menores valores de confiança inicial, sempre que α é igual ou maior a 0,5.

5.3.4 Exemplo de sumário gerado

A fim de tornar possível também uma avaliação qualitativa dos sumários gerados pela implementação do SBI no contexto da sumarização extrativa focada em consulta apresenta-se a seguir um exemplo destes sumários, gerado pelo sistema VAR-0.3 com α configurado para 0,7, visto que esta foi a configuração que atingiu os melhores resultados neste contexto.

Sumário para tópico 683j do corpus DUC2005

Consulta

Discuss the events leading to the breakup of Czechoslovakia, the formation of the Czech Republic and Slovakia, and the outlook for these nations after one year.

Sumário de referência

[1] On New Year's day, 1993, Slovaks sloughed off a thousand year subservience to Hungary and seven decades as the junior partner in Czechoslovakia and celebrated the birth of a sovereign, independent republic.la
[2] But only six months earlier, most Slovaks went into the crucial general elections of June 1992 intending to negotiate a new and looser union with their richer Czech cousins, but not to be divorced from them.

[3] Opinion polls showed that independence was only sought by a small minority in both the Czech lands and Slovakia. By 1992, the polarization of politics between nationalists in Slovakia and market reformers in the Czech lands, as well as the defeat of Vaclav Havel, a strong advocate for the union, had put the federation of Czechs and Slovaks at risk.
[4] Before the elections, Vaclav Klaus, the Czech leader, rejected any "soft options".
[5] He turned down the looser federation proposed by Vladimir Meciar, the Slovak leader.
[6] He rejected Slovak demands for a central bank and requests for federal funds.
[7] Mr. Klaus argued that both sides should either agree on a smaller but more effective federal government or a quick divorce.
[8] Meciar agreed to the lalatextter.
[9] It appeared that Klaus and many Czechs were relieved.
[10] A year after independence, Slovakia's economy was deeply in recession, and political infighting threatened to bring down the government.
[11] Democracy suffered in both countries as evidenced by growing disillusionment with politicians in Slovakia and widespread cynicism in the Czech Republic over corruption among business and political elites.

Sumário gerado pela implementação do método proposto neste trabalho

[1] The Slovak leader acknowledges that 'founding the new state brought many problems.'
[2] The dissolution of the post-communist state of Czechoslovakia, which was regarded as perhaps the most promising candidate for full European Community membership, will have repercussions well beyond its borders.
[3] We did not want to be cut off from Prague and the Czech part of our country.
[4] In the 1960s the the East Slovakian Steel Works (VSZ) was built outside Kosice by Czech engineers with Czech equipment.
[5] The spiritual fathers of the Czechoslovak state, mainly Czech and Slovak emigres in Pittsburgh and other American industrial towns during the first world war, sold the idea of the new dual nation to President Woodrow Wilson as a Slav bulwark to reduce German influence in postwar Europe.
[6] Slovakia has always felt like a poor relation of the Czech republic.
[7] New political forces arising from the collapse of the Soviet bloc, together with the deep-rooted Slovak nationalism, are threatening to split the country again into its Czech and Slovak components.
[8] Slovakia has no so such self-confident past.
[9] The communists who ran Czechoslovakia in the Stalinist years and after the Soviet invasion of 1968 were, almost to a man, grey mediocrities who showed an insensitivity bred of ignorance to tradition and theenvironment.

[10] Thus Mr Milos Zeman, the Czech Social Democrat leading the opposition to the break-up of Czechoslovakia, made a shrewd move last week when he proposed to transform the present federation into a Czecho-Slovak Union on the Maastricht model, which would come to

5.4 SUMARIZAÇÃO GENÉRICA COM SBI

Apresentam-se a seguir os experimentos realizados a fim de avaliar a relevância dos resultados obtidos pela implementação do SBI quando utilizada para a sumarização extrativa genérica, ou seja, aquela na qual a finalidade do sumário é representar todos os fatos relevantes presentes nos documentos originais. Como discutido na seção 4.3 quando o método é utilizado para esta finalidade a união de todas as sentenças dos documentos de entrada é utilizada como consulta, uma vez que neste caso não há uma consulta definida no corpus para guiar a construção do sumário.

5.4.1 Corpus

O corpus selecionado para a realização dos experimentos descritos nesta seção foi o CNN/DailyMail. Este corpus foi apresentado por Hermann et al. (2015) e é composto de 312.085 artigos extraídos dos sites da CNN (canal de notícias da TV a cabo estado-unidense) e do DailyMail (jornal britânico), separados de maneira que cada documento corresponda a um artigo.

A intenção deste conjunto quando da sua apresentação era avaliar algoritmos de sumarização baseados em técnicas de aprendizagem supervisionada, e por isso os autores que o apresentaram sugeriram que o mesmo fosse dividido em três conjuntos menores contendo 287.227/13.368/11.490 documentos para treino, validação e teste respectivamente. Como o método proposto neste trabalho baseia-se em aprendizagem não supervisionada, somente o subconjunto de teste será considerado. Isso é também conveniente uma vez que todos os resultados publicados obtidos com este corpus utilizam também o conjunto de testes.

Assim como em estudos anteriores (NARAYAN; COHEN; LAPATA, 2018; TAN; WAN; XIAO, 2017; SEE; LIU; MANNING, 2017) neste trabalho os sumários de referência para este corpus, necessários para a avaliação dos sumários gerados, foram construídos a partir dos destaques de cada artigo, grifados pelos próprios autores. Cada artigo tem em média 3,88 sentenças no seu sumário de referência, por isso escolhemos definir como 4 o número de sentenças que cada sumário gerado para este corpus deveria ter. Como cada artigo gera, a partir dos destaques, um sumário de referência, este é um corpus para sumarização monodocumento.

5.4.2 Definição dos Experimentos

Os experimentos realizados consistiram na utilização da implementação do SBI para geração de sumários para cada um dos 11490 documentos no conjunto de teste do corpus CNN/-DailyMail. Posteriormente, estes sumários foram avaliados através das medidas ROUGE e os resultados obtidos comparados com resultados obtidos por implementações de outros modelos.

A seleção de modelos para comparação se baseou na proximidade conceitual com o método proposto neste trabalho conforme discutido na apresentação da revisão sistemática da literatura. Por fim, diferentes sumários foram gerados para diferentes configurações da implementação do SBI.

Como este corpus não é específico para a avaliação de sumarizadores focados em consultas, os próprios documentos foram usados como consulta.

5.4.3 Resultados

Apresentam-se a seguir os resultados obtidos nos experimentos realizados para avaliar os sumários gerados pela implementação do SBI na tarefa de sumarização extrativa genérica. Os sumários foram avaliados através das medidas-f de ROUGE-1 e ROUGE-2, pois estas apresentam grande correlação com a avaliação humana (LIN, 2004) sendo padrão *de-facto* para avaliação de sumários extrativos. A notação usada para nomear as diferentes versões e configurações do SBI, denominadas sistemas, para as quais se reportam resultados é apresentada na seção 5.1.

A figura 12 apresenta os valores obtidos para a medida-f de ROUGE-1 por todos os sistemas, com três valores diferentes configurados para o parâmetro α.

Diferentemente do que aconteceu no corpus DUC2005, reduzir a confiança na ligação de instâncias não se traduziu em melhores resultados na maioria dos sistemas. Os sistemas que utilizaram confiança de ligação de instâncias variável (VAR), obtiveram resultados melhores conforme esta confiança diminui, para valores de α menores que 0,5, porém os resultados foram próximos entre si. Por outro lado, é bastante claro que os resultados alcançados pelos sistemas que utilizaram confiança na ligação de instâncias fixa (FIX) pioraram conforme a confiança na ligação de instâncias diminuía. Também chama a atenção o fato de que o parâmetro que controla a confiança de ligação de instâncias influenciou os resultados também de maneira diferente do que aconteceu no corpus DUC2005. Para um mesmo valor inicial para este parâmetro foram os sistemas que mantém esse valor fixo (FIX) e não os que o variam (VAR) que obtiveram os melhores resultados. Como aconteceu no corpus DUC2005, todas as versões do sistema, exceto uma, atingiram seus melhores resultados quando α era igual a 0,7. Como neste corpus o documento inteiro estava sendo usado como consulta, estes resultados parecem indicar que nestes casos, com consultas maiores, ter mais instâncias anotadas com uma confiança mais baixa nestas anotações tende a produzir resultados piores.

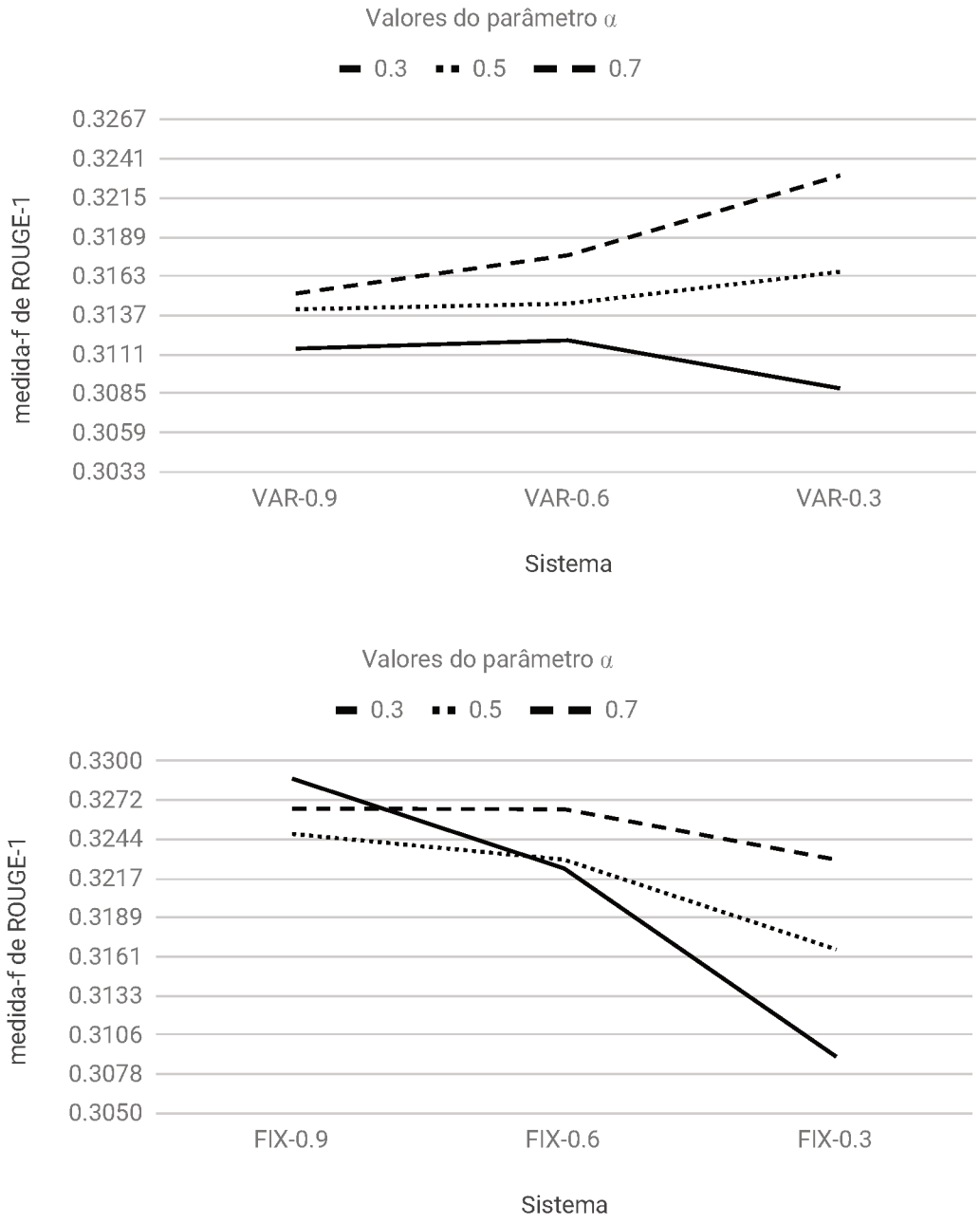


Figura 12 – Resultados obtidos na medida-f de ROUGE-1 por sistema, com três valores diferentes de α no corpus CNN/DailyMail

A figura 13 apresenta uma comparação entre os sistemas com confiança na anotação de instâncias fixa e variável através dos valores obtidos para a medida-f de ROUGE-1, para três valores iniciais de confiança e um valor fixo para o parâmetro α de 0,7. Os sistemas com confiança na anotação fixa atingiram melhores resultados em duas ocasiões.

A tabela 7 apresenta uma comparação entre os resultados alcançados por modelos de sumarização extrativa publicados recentemente e os melhores resultados obtidos pela implementação do método proposto neste trabalho, alcançados pelo sistema FIX-0.9 com α 0,3. O modelo apresentado por Amplayo, Lim e Hwang (2018) é o mais próximo conceitualmente ao

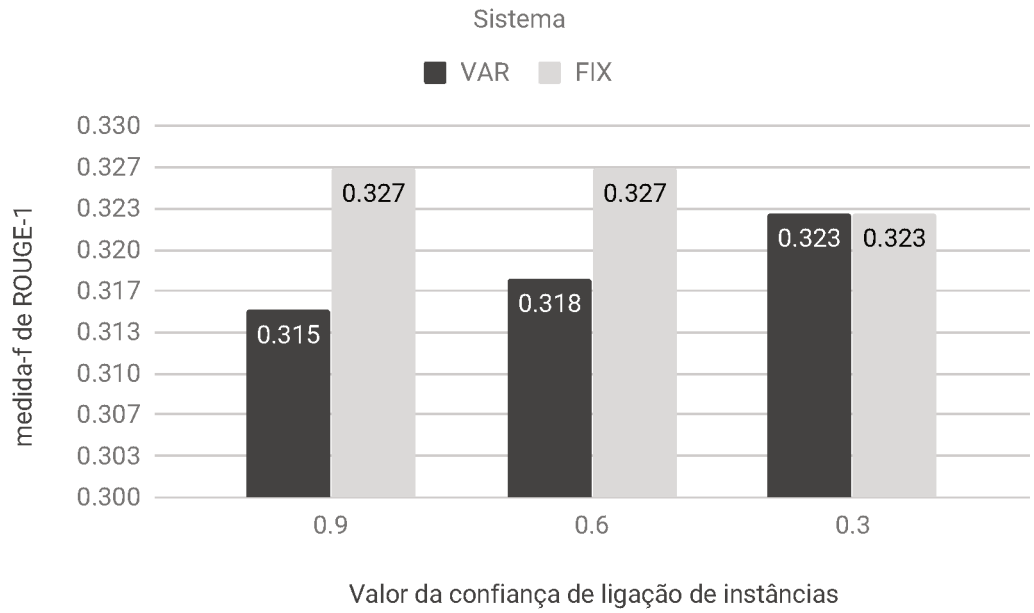


Figura 13 – Resultados obtidos na medida-f de ROUGE-1, por sistema, com um valor fixo para o parâmetro α de 0,7, para o corpus CNN/DailyMail

apresentado neste trabalho, uma vez que os autores exploraram o uso de entidades ligadas para melhorar os resultados atingidos no processo de sumarização. A implementação do método proposto neste trabalho atingiu melhores resultados se comparado com este trabalho, mas estes mesmos resultados foram piores que os atingidos pelos demais trabalhos, que representam o estado da arte de modelos de sumarização extrativa com resultados publicados para este corpus.

Sistema	Medida-f	
	ROUGE-1	ROUGE-2
Zhou et al. (2018)	**0.4159**	**0.1901**
Dong et al. (2018)	0.4150	0.1870
Zhang et al. (2018)	0.4105	0.1877
Amplayo, Lim e Hwang (2018)	0.3190	0.1010
Este trabalho	0.3280	0.1200

Tabela 7 – Comparação entre os resultados obtidos por implementações de modelos de trabalhos correlatos e pela implementação do método proposto neste trabalho, para o corpus CNN/DailyMail.

A tabela 8 apresenta todos os resultados obtidos durante os experimentos no corpus CNN/DailyMail. Alguns destes resultados foram apresentados nas figuras e tabelas anteriores. Os resultados foram obtidos por diferentes sistemas, utilizando as duas diferentes estratégias de anotação de instâncias definidas na seção 4.1.1 (VAR e FIX), diferentes valores iniciais para a confiança na anotação (0,3; 0,6; 0,9) e diferentes valores para o parâmetro α (0,3; 0,6; 0,9).

Observando a tabela é possível notar que os números apresentam, como esperado, as mesmas tendências apontadas na descrição das figuras 12 e 13, em especial a destacada diferença entre o comportamento dos sistemas com confiança na anotação fixa e variável com relação a diminuição da confiança inicial desta anotação.

ROUGE-1						
α	VAR-0.9	VAR-0.6	VAR-0.3	FIX-0.9	FIX-0.6	FIX-0.3
0.3	0.3115	0.3120	0.3089	0.3287	0.3224	0.3090
0.5	0.3141	0.3145	0.3166	0.3248	0.3230	0.3166
0.7	0.3152	0.3177	0.3230	0.3266	0.3266	0.3230
ROUGE-2						
α	VAR-0.9	VAR-0.6	VAR-0.3	FIX-0.9	FIX-0.6	FIX-0.3
0.3	0.1015	0.1015	0.0989	0.1195	0.1133	0.0990
0.5	0.1032	0.1035	0.1046	0.1175	0.1148	0.1046
0.7	0.1041	0.1066	0.1163	0.1205	0.1204	0.1163

Tabela 8 – Resultados obtidos para a medida-f de ROUGE-1 e ROUGE-2 por todas as implementações do método proposto neste trabalho nos experimentos realizados no corpus CNN/DailyMail.

5.4.4 Exemplo de sumário gerado

A fim de tornar possível também uma avaliação qualitativa dos sumários gerados pela implementação do SBI no contexto da sumarização extrativa genérica apresenta-se a seguir um exemplo destes sumários, gerado pelo sistema FIX-0.9 com α configurado para 0,3, que foi a configuração que atingiu os melhores resultados neste corpus.

Sumário para um dos documentos do corpus CNN/DailyMail

Documento original, também usado como consulta

-lrb- cnn -rrb- those poor fish must have been wondering what the heck was happening to them . the oregon parks and recreation department has reported that a section of a fiberglass boat 20 or 30 feet long was spotted off the state 's coast this week and has been towed into harbor . the debris is suspected to be from the earthquake and tsunami that hit japan on march 11 , 2011 . the boat fragment was found this week and towed to newport , oregon , where it is moored at a marina . inside were found – more than four years and 4,000 miles later , if officials ' suspicions are correct – some specimens of a variety of yellowtail jack fish normally found in japanese waters . biologists with the oregon coast aquarium and oregon state university 's hatfield marine science center inspected the debris while it was still at sea and determined that the ecological threat posed by invasive species was small . the remnants of the boat will be dried out , inspected further and taken to a landfill . but

for the yellowtail jack fish , the journey is not over . they 'll be taken to the oregon coast aquarium .

Sumário de referência

[1] debris from boat to be dried , inspected and taken to landfill .
[2] the debris contained fish normally found in japanese waters .
[3] the earthquake and tsunami hit japan in march 2011 .

Sumário gerado pela implementação do método proposto neste trabalho

[1] biologists with the oregon coast aquarium and oregon state university 's hatfield marine science center inspected the debris while it was still at sea and determined that the ecological threat posed by invasive species was small .
[2] inside were found – more than four years and 4,000 miles later , if officials ' suspicions are correct – some specimens of a variety of yellowtail jack fish normally found in japanese waters .
[3] -lrb- cnn -rrb- those poor fish must have been wondering what the heck was happening to them .
[4] the debris is suspected to be from the earthquake and tsunami that hit japan on march 11 , 2011 .

5.5 DISCUSSÃO SOBRE OS RESULTADOS

Os resultados obtidos no corpus DUC2005 se mostraram inversamente proporcionais a confiança na anotação das instâncias e diretamente proporcionais ao valor do parâmetro α. Ou seja, quanto menor o valor do parâmetro que controla a confiança na anotação melhores os resultados, e quanto maior o valor do parâmetro α melhores os resultados. Além disso, os sistemas com confiança de anotação variável (VAR) obtiveram melhores resultados quando comparados com os sistemas com confiança na anotação fixa (FIX), para o mesmo nível de confiança inicial - VAR-0.9 teve melhor desempenho que FIX-0.9, por exemplo. Os sistemas com confiança de anotação variável ligam instâncias às sentenças com nível de confiança na anotação mais baixos (ver seção 4.1.1).

Uma possível explicação para esta observação está na diferença de tamanho entre os documentos sendo sumarizados e a consulta. Como a consulta é menor do que os documentos, com níveis de confiança na anotação mais baixos as instâncias extras anotadas na consulta compensam as instâncias extras e irrelevantes que porventura sejam anotadas nos documentos. Além disso, quanto mais alto o valor do parâmetro α mais importância é dada a relevância para consulta e menos a diversidade do sumário no momento de selecionar que sentenças serão extraídas. Por isso instâncias adicionadas na consulta são mais relevantes e melhoram os

resultados obtidos com valores mais altos para o parâmetro α.

No corpus CNN/DailyMail os sistemas com confiança na anotação fixa (FIX) obtiveram melhores resultados quando comparados com os sistemas com confiança de anotação variável (VAR), para o mesmo nível de confiança inicial - FIX-0.9 obteve melhor desempenho que VAR-0.9, por exemplo, ao contrário do que aconteceu no corpus DUC2005. Neste corpus os próprios documentos foram usados como consulta e neste caso, aparentemente, as instâncias extras anotadas nas consultas com valores mais baixos de confiança na anotação aparentemente não compensaram as instâncias extras e irrelevantes anotadas nos documentos.

As figuras 11 e 13 suportam os resultados apresentados anteriormente, ao apresentar os valores da medida-f de ROUGE-1, por sistema, com um valor fixo para o parâmetro α.

O uso de instâncias permite uma representação semântica das sentenças dos documentos a serem sumarizados que reflete mais fielmente as diferenças reais entre estas sentenças, o que pode impactar a avaliação da similaridade semântica entre sentenças ajudando a identificar sentenças mais relevantes a consulta.

Os experimentos realizados demonstraram que a utilização da estratégia de anotação de instâncias com confiança variável na anotação pode melhorar os resultados obtidos por sumarizadores extrativos de texto focados em consulta. Dois fatos serviram como evidência empírica desse argumento. Primeiro o fato de que com consultas curtas (DUC2005) os resultados foram melhores quando foi adotada a estratégia de confiança variável na anotação. Isto é, melhorar a representação leva a melhores resultados quando a consulta é curta. Segundo o fato de que com consultas longas (CNN/DailyMail) os resultados não melhoraram quando a estratégia de confiança variável na anotação foi adotada. Isto é, melhorar a representação não leva a melhores resultados quando a consulta é longa. Em consultas curtas, as chances de uma instância ser anotada com um nível de confiança alto é reduzida pela extensão do texto ser pequena. Com consultas mais longas, no entanto, estas chances aumentam.

Essas observações levam a crer que com consultas curtas, variar a confiança de anotação pode garantir que pelo menos uma instância será anotada na consulta, isto é, que a consulta terá uma representação semântica válida, enquanto que com consultas longas mais instâncias podem ser anotadas com níveis mais altos de confiança na anotação e, por isso, variar esta confiança não surte o mesmo efeito. Portanto, utilizar a estratégia de variar a confiança mínima para anotação na hora de construir a representação semântica através de instâncias de uma ontologia impacta positivamente os resultados do processo de sumarização.

A performance de um sumarizador baseado em instâncias de uma ontologia é altamente dependente da quantidade de instâncias na ontologia, da cobertura semântica destas instâncias e da qualidade do processo de reconhecimento destas instâncias para construção da representação semântica das sentenças. É extremamente importante criar mecanismos que assegurem a possibilidade de criar uma representação semântica através de instâncias para as sentenças. Melhores algoritmos e medidas de similaridade podem remediar um possível excesso de instâncias nas representações da consulta e dos documentos, mas não podem remediar a falta de instâncias.

Quando a estratégia mais relevante para a tarefa específica de sumarização sob teste é utilizada, os resultados alcançados pelo método proposto demonstram-se relevantes. Na sumarização focada em consulta, a utilização da estratégia de anotação de instâncias com confiança variável garante uma maior chance de criar representações válidas para toda as sentenças, incluindo as sentenças da consulta, conduzindo a resultados que, quando comparados com os resultados alcançados por outros trabalhos importantes da área, demonstram-se relevantes. Na sumarização genérica o mesmo acontece quando se usa a estratégia de anotação de instâncias com confiança fixa. É fato que o método proposto não conseguiu obter resultados melhores do que os resultados alcançados por sumarizadores do estado da arte quando utilizados nos mesmos corpus, como pode ser observado nas tabelas 5 e 7. Estes sumarizadores, no entanto, utilizam abordagens específicas de aprendizado de máquina, com menos capacidade de generalização. Também é fato que os resultados alcançados pelo SBI demonstram o potencial descritivo das instâncias para a representação semântica, uma vez que estes são melhores do que os resultados alcançados por outros métodos já publicados, os quais contribuíram significativamente para a exploração das diversas possibilidades para a sumarização extrativa.

Assim sendo, considera-se que os resultados alcançados pelo SBI são relevantes uma vez que são no mínimo tão bons quanto os resultados alcançados por outros métodos de sumarização extrativa que utilizam representações e estratégias de sumarização diversas, cujos resultados são também considerados relevantes por serviram para validar estas representações e apresentar alternativas possíveis para a sumarização extrativa.

Ademais, contribui para a conclusão de que os resultados são relevantes o fato de que utilizar as instâncias de uma ontologia para representar as sentenças de um documento é uma estratégia que possibilita a utilização de ontologias construídas automaticamente, a partir de descrições estruturadas de coisas do mundo real, na sumarização extrativa de texto. Ontologias construídas desta forma estão menos limitadas à representação de um domínio específico do conhecimento, pois não dependem da intervenção de especialistas neste domínio durante a sua construção. Esta é a abordagem de construção de ontologias na *Linked Data*, como a DBPedia e outras grandes ontologias que funcionam como pontos centrais e interligação de diversas outras ontologias. Portanto, o método proposto possibilita a criação de sumarizadores baseados em ontologias construídas automaticamente e que usualmente são independentes de domínio, alcançando resultados relevantes e tão bons quanto os alcançados por métodos que requerem um esforço manual maior. Esta característica é particularmente interessante para a construção de sumarizadores para uso real em aplicações destinadas a usuários finais.

Por fim, para além da análise quantitativa, um outro aspecto pelo qual se pode analisar a relevância dos resultados alcançados pelo método proposto diz respeito a legibilidade e coerência dos sumários gerados. Nos dois exemplos de sumários gerados pelo SBI incluídos nesta dissertação, nas seções 5.3.4 e 5.4.4, é possível observar três aspectos. A primeira observação é a de que as sentenças que compõe os sumários gerados são bastante coerentes e legíveis. Este fato, porém, era esperado uma vez que o sumário é extrativo e por isso a coerência e legibilidade das sentenças do sumário derivam diretamente da coerência e legibilidade das

sentenças dos documentos originais. Em segundo lugar observa-se que o sumário construído para o corpus DUC2005, portanto multidocumento e focado em consulta, é aparentemente menos coerente e 'fluído', quando comparado ao sumário gerado para o corpus CNN/DailyMail, portanto genérico e monodocumento. Esta 'fluidez' na leitura do texto parece surgir do encadeamento das ideias apresentadas nas sentenças de forma mais suave e menos abrupta, propiciando uma leitura com menos esforço, como se a ideia apresentada em uma sentença levasse naturalmente a ideia apresentada na próxima. Nesse sentido a existência de apenas um documento de entrada propicia maiores chances de que um sumário extrativo guarde essa característica se a mesma estiver presente no documento original, principalmente se as sentenças aparecerem no sumário na mesma ordem na qual aparecem no texto, enquanto que a presença de múltiplos documentos de entrada diminui as chances de que essa mesma coerência e encadeamento de ideias nas sentenças apareça no sumário. Por fim, é nota-se que ambos os sumários gerados são pouco redundantes, com cada uma de suas sentenças constituintes trazendo uma ideia central substancialmente distinta das demais e que traz uma informação relevante.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Diversos trabalhos apresentados na literatura buscam analisar a semântica das sentenças presentes em documentos originais a fim de construir um sumário dos mesmos a partir da extração das sentenças consideradas mais relevantes. Para possibilitar esta análise semântica, e também a avaliação e comparação das sentenças, é necessário construir uma representação da semântica dessas sentenças a partir de uma base de conhecimento que formalize os elementos componentes da mesma. As ontologias são bastante utilizadas como base de conhecimento para este fim. As ontologias são formadas, entre outros, por conceitos e instâncias e ambos podem ser utilizados para a construção de uma representação semântica. Através de uma revisão sistemática da literatura, contudo, verificou-se que apenas os conceitos haviam sidos explorados para tal finalidade nos trabalhos já publicados sobre sumarização extrativa. O objetivo geral desta dissertação e do trabalho de mestrado que a gerou foi o de preencher esta lacuna, propondo um método de sumarização extrativa de texto que utilizasse as instâncias de uma ontologia para representar a semântica das sentenças e analisando a relevância dos resultados alcançados pelo mesmo.

O objetivo geral deste trabalho foi alcançado, assim como seus objetivos específicos, através da proposição de um método de sumarização extrativa de texto baseado no uso de uma ontologia para representar a semântica das sentenças, denominado SBI, e de experimentos realizados para avaliar a relevância dos resultados alcançados pelo mesmo.

O método proposto primeiro utiliza a saída de um sistema de ligação de instâncias à trechos de texto para construir a representação semântica das sentenças com instâncias definidas na ontologia, para em seguida selecionar e extrair as sentenças consideradas mais relevantes. A seleção e extração das sentenças está baseada na adaptação do algoritmo MMR, proposto por Carbonell e Goldstein (1998), e em uma medida de similaridade entre sentenças que utiliza as suas representações e, por sua vez, baseia-se em uma medida de similaridade semântica entre instâncias de uma ontologia. Essa medida de similaridade semântica entre instâncias utiliza os componentes das descrições das instâncias para compará-las.

Os experimentos realizados para analisar a relevância dos resultados alcançados pelo método proposto dividiram-se em dois, levando em consideração a configuração específica da tarefa de sumarização que abordaram. Isso ocorreu porque o método proposto possibilitava tanto a sumarização genérica quanto a sumarização focada em consulta - assim como tanto a sumarização monodocumento quanto a sumarização multidocumento, embora esta característica não tenha guiado diretamente a condução dos experimentos.

Os experimentos realizados para analisar a relevância dos resultados alcançados pelo SBI na tarefa de sumarização focada em consulta empregaram o corpus DUC2005 e demonstraram que este alcança resultados relevantes, apesar de não superarem os melhores resultados já alcançados para este corpus. Além disso, os resultados também demonstraram que a construção da representação das sentenças influencia nestes resultados a medida que pode garantir que estas possuam uma representação não-vazia, o que parece ser especialmente importante neste

caso da sumarização focada em consulta quando as consultas são significativamente menores que os documentos.

Da mesma forma, os experimentos realizados para analisar a relevância dos resultados alcançados pelo SBI na tarefa de sumarização genérica demonstraram que o mesmo alcança resultados que podem ser considerados relevantes, apesar de também não serem os melhores já alcançados. O corpus utilizado nestes experimentos foi o CNN/DailyMail. Assim como aconteceu com experimentos realizados com o corpus DUC2005, os resultados demonstraram que a construção da representação das sentenças influencia os resultados alcançados. Ao contrário do que ocorreu com o corpus DUC2005, no entanto, a tentativa de garantir que as sentenças possuam uma representação não-vazia não pareceu especialmente importante neste caso. Especula-se que isso se deva ao fato de que as "consultas" utilizadas neste caso, que eram os próprios documentos, eram mais longas e, portanto, tinham maiores chances de ter instâncias anotadas com níveis de confiança mais altos.

Nas próximas seções, apresentam-se as principais contribuições alcançadas por este trabalho, juntamente com as principais limitações do mesmo (seção 6.1), seguidas de sugestões para trabalhos futuros (seção 6.2).

6.1 PRINCIPAIS CONTRIBUIÇÕES E LIMITAÇÕES

As principais **contribuições** apresentadas nesta dissertação foram:

- Estratégia de ligação de instâncias com confiança variável

  A estratégia de ligação de instâncias de uma ontologia à trechos de texto com confiança variável na ligação, apresentada na seção 4.1.1 demonstrou-se fundamental para a construção de representações não vazias das sentenças, que por sua vez foram importantes para os resultados atingidos nos experimentos realizados na tarefa de sumarização baseada em consulta.

- Medida de similaridade semântica entre instâncias de uma ontologia

  A interpretação da medida de similaridade semântica baseada na teoria da informação e apresentada por Lin (1998), e que neste trabalho encontra-se na seção 4.1.3, alcançou resultados empíricos alinhados a expectativas pré-definidas sobre valores esperados de uma medida de similaridade semântica e serviu de base para as comparações semânticas definidas no método de sumarização extrativa proposto.

- Método de sumarização extrativa baseado em instâncias de uma ontologia

  A integração da definição de como construir uma representação da semântica das sentenças através de instâncias de uma ontologia, com a medida de similaridade entre instâncias que propicia a comparação semântica entre sentenças e a adaptação de um método de sumarização extrativa existente, para a proposição de um método de sumarização extrativa baseado em instâncias de uma ontologia foi a principal contribuição deste trabalho.

Por outro lado, as principais **limitações** do mesmo são:

- Avaliação empírica da medida de similaridade entre instâncias

  A medida de similaridade entre instâncias definida e utilizada neste trabalho baseia-se em uma medida teórica com fundamentos sólidos. Essa, contudo, só foi avaliada empiricamente. Seria interessante que a mesma fosse provada formalmente, ou validada experimentalmente.

- Baixa transparência do processo de ligação de instâncias às sentenças de entrada

  O método proposto apoia-se na utilização de um sistema específico para realizar a ligação de instâncias as sentenças de entrada. Essa ligação é o fundamento da construção da representação das sentenças através de instâncias da ontologia. O sistema utilizado para fazer a ligação, contudo, é à parte do método proposto neste trabalho e não é discutido no mesmo, tornando este método menos transparente, e por consequência a sua avaliação mais difícil.

- Uso de uma única ontologia nos experimentos

  Os experimentos realizados limitaram-se a utilizar as instâncias definidas na ontologia da DBPedia de 2014 como base de conhecimento. Apesar desta ontologia ser bastante abrangente seria interessante que os mesmos experimentos tivessem sido realizados com outras ontologias, a fim de analisar o impacto do uso destas nos resultados alcançados. Realizar os experimentos com outras ontologias permitiria também entender o impacto nos resultados de acordo com as características específicas de cada uma, como o fato de ser específica de um domínio ou independente de domínio.

- Os resultados alcançados estão distantes do estado da arte.

  Apesar de os resultados alcançados poderem ser considerados relevantes para o atingimento dos objetivos propostos nessa dissertação, estes estão distantes do estado da arte. Isto posiciona a representação das sentenças e o método proposto como alternativas viáveis, úteis principalmente por representar a possibilidade de uso de ontologias construídas automaticamente dispensando a necessidade de especialistas de domínio, porém que não apontam para uma melhora significativa da informatividade dos sumários no estado da arte.

6.2 TRABALHOS FUTUROS

A partir dos resultados alcançados no trabalho de mestrado desenvolvido, que são apresentados nesta dissertação, e das principais contribuições e limitações do mesmo, algumas possibilidades de trabalhos futuros são apresentadas a seguir.

- Explorar usos diversos das relações entre instâncias

Neste trabalho as relações das instâncias das ontologias foram utilizadas para criar as suas descrições, que serviram de base para a comparação semântica entre sentenças. Essas mesmas relações poderiam ser utilizadas de outras formas para esta mesma comparação semântica. A existência de uma relação entre as instâncias, por exemplo, seja ela direta ou indireta, poderia ser considerada de alguma forma no cômputo da similaridade.

- Explorar dados ligados na web

  A ontologia da DBPedia, assim como outras ontologias, aponta não apenas para instâncias definidas nela mesma, mas também para instâncias e conceitos definidos em outras ontologias. Esta forma de ligação de instâncias é bastante comum entre ontologias publicadas na web, no contexto da web semântica e dos dados ligados. O trabalho apresentado nesta dissertação não explorou estas ligações à instâncias e conceitos definidos em outras ontologias, mas acredita-se que fazê-lo pode ter um impacto positivo sobre os resultados por aumentar a quantidade de informação disponível para o sumarizador.

- Explorar outras medidas de similaridade

  A medida de similaridade semântica utilizada neste trabalho para comparar instâncias de uma ontologia só foi validada empiricamente e, portanto, é possível que existam medidas mais adequadas à utilização na sumarização extrativa de texto. Além de validar melhor a medida utilizada nesta proposta, testar o uso de outras medidas pode levar ao encontro de uma medida que traga melhores resultados.

- Utilizar outro sistema de ligação de instâncias

  Assim como o que acontece com a medida de similaridade semântica entre instâncias, não foi considerada, neste trabalho, a possibilidade de utilização de um sistema de ligação de instâncias à trechos de texto diferente daquele que selecionou-se para a proposta. Apesar de o ILS selecionado ter servido de base para a entrega de resultados relevantes pelo método proposto, é possível que outros sistemas consigam realizar uma anotação de instâncias mais eficaz, levando a melhores resultados.

- Utilizar outro método base

  Como o foco desta proposta era avaliar a relevância dos resultados obtidos por um sumarizador automático que utilizasse as instâncias de uma ontologia para representar a semântica das sentenças, selecionou-se o MMR como método base sem considerar outras opções. Porém, outros métodos de sumarização extrativa poderiam servir tão bem como métodos base, a serem estendidos pela utilização da medida de similaridade definida nesta proposta, quanto o MMR. Seria interessante, portanto, considerar a utilização da medida de similaridade em outros métodos, realizando experimentos que comparassem esta extensão aquela feita ao MMR.

- Usar outros corpora para validação do método

A validação do método proposto aconteceu através da realização de experimentos utilizando dois corpora: DUC2005 e CNN/DailyMail. Ambos os corpora foram construídos a partir de textos jornalísticos em inglês. Seria interessante utilizar outros corpora, construídos com textos vindo de outros domínios para validar a relevância dos resultados obtidos pelo método proposto neste trabalho. Da mesma maneira, seria interessante utilizar corpora em outras línguas, para comparar os resultados alcançados em diferentes línguas. A utilização de corpora em outras línguas, claro, depende da existência de uma ontologia que defina conceitos e instâncias nesta outra língua.

- Penalizar instâncias ligadas com nível de confiança mais baixo

  É possível que a utilização do algoritmo 1 nos sistemas VAR tenha gerado alguma espécie de ruído no processo de sumarização ao anotar instâncias a um nível de confiança mais baixo que o nível inicial configurado. Para tentar medir essa possibilidade e, ao mesmo tempo, contorná-la, seria interessante penalizar de alguma forma a similaridade entre estas instâncias anotadas a níveis mais baixos de confiança e as instâncias anotadas no nível inicial. A ideia é a de que quanto menor o nível de confiança utilizado para anotar uma instância, menos a sua similaridade com outras instâncias contribuiria para o cômputo da similaridade entre sentenças.

REFERÊNCIAS

AHMAD, A.; AHMAD, T. A game theory approach for multi-document summarization. **Arabian Journal for Science and Engineering**, v. 44, n. 4, p. 3655–3667, Apr 2019.

AMPLAYO, R. K.; LIM, S.; HWANG, S.-w. Entity commonsense representation for neural abstractive summarization. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. Association for Computational Linguistics, 2018. p. 697–707. Disponível em: http://aclweb.org/anthology/N18-1064.

ANTIQUEIRA, L. et al. A complex network approach to text summarization. **Information Sciences**, v. 179, n. 5, p. 584 – 599, 2009. ISSN 0020-0255. Special Section - Quantum Structures: Theory and Applications. Disponível em: http://www.sciencedirect.com/science/article/pii/S0020025508004520.

BARALIS, E. et al. Multi-document summarization based on the yago ontology. **Expert Syst. Appl.**, v. 40, n. 17, p. 6976–6984, 2013.

BAWAKID, A.; OUSSALAH, M. Using SSM for enhancing summarization. In: **Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011**. [s.n.], 2011. Disponível em: http://www.nist.gov/tac/publications/2011/participant.papers/seme1.proceedings.pdf.

CANHASI, E.; KONONENKO, I. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. **Expert Syst. Appl.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 41, n. 2, p. 535–543, fev. 2014. ISSN 0957-4174. Disponível em: http://dx.doi.org/10.1016/j.eswa.2013.07.079.

CARBONELL, J.; GOLDSTEIN, J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: **Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 1998. (SIGIR '98), p. 335–336. ISBN 1-58113-015-5. Disponível em: http://doi.acm.org/10.1145/290941.291025.

CHANDU, K. et al. Tackling biomedical text summarization: OAQA at BioASQ 5B. In: **BioNLP 2017**. Vancouver, Canada,: Association for Computational Linguistics, 2017. p. 58–66. Disponível em: https://www.aclweb.org/anthology/W17-2307.

CHAUD, M. R.; FELIPPO, A. D. Exploring content selection strategies for multilingual multi-document summarization based on the universal network language (unl). **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 26, n. 1, p. 45–71, 2018. ISSN 2237-2083.

CORCHO, O.; GóMEZ-PéREZ, A. Evaluating knowledge representation and reasoning capabilities of ontology specification languages. In: **In Proceedings of the ECAI 2000 Workshop on Application of Ontologies and Problem-Solving Methods**. [S.l.: s.n.], 2000.

DAIBER, J. et al. Improving efficiency and accuracy in multilingual entity extraction. In: **Proceedings of the 9th International Conference on Semantic Systems**. New York, NY, USA: ACM, 2013. (I-SEMANTICS '13), p. 121–124. ISBN 978-1-4503-1972-0. Disponível em: http://doi.acm.org/10.1145/2506182.2506198.

DANG, H. T. Overview of duc 2005. In: **In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP**. [S.l.: s.n.], 2005.

DAS, D.; MARTINS, A. F. T. **A Survey on Automatic Text Summarization**. [S.l.], 2007.

DONG, Y. et al. Banditsum: Extractive summarization as a contextual bandit. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2018. p. 3739–3748. Disponível em: http://aclweb.org/anthology/D18-1409.

GAARDER, J. **Mundo de Sofia, O**. [S.l.]: Cia. Das Letras, 1998. ISBN 8571644756.

GIACOMO, G. D.; LENZERINI, M. Tbox and abox reasoning in expressive description logics. In: **Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996. (KR'96), p. 316–327. ISBN 1-55860-421-9. Disponível em: http://dl.acm.org/citation.cfm?id=3087368.3087406.

GONG, Y.; LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In: **Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2001. (SIGIR '01), p. 19–25. ISBN 1-58113-331-6. Disponível em: http://doi.acm.org/10.1145/383952.383955.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **Proceedings of the 16th Conference on Computational Linguistics - Volume 1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996. (COLING '96), p. 466–471. Disponível em: https://doi.org/10.3115/992628.992709.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, Academic Press Ltd., London, UK, UK, v. 5, n. 2, p. 199–220, jun. 1993. ISSN 1042-8143. Disponível em: http://dx.doi.org/10.1006/knac.1993.1008.

GUARINO, N. **Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy**. 1st. ed. Amsterdam, The Netherlands, The Netherlands: IOS Press, 1998. ISBN 9051993994.

HELENA, L. M. R.; PARDO, A. T. S. A sumarização automática de textos: principais características e metodologias. In: **JAIA - Jornada de Atualização em Inteligência Artificial.** Campinas: [s.n.], 2003. VIII, p. 203–245.

HERMANN, K. M. et al. Teaching machines to read and comprehend. In: **Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1**. Cambridge, MA, USA: MIT Press, 2015. (NIPS'15), p. 1693–1701. Disponível em: http://dl.acm.org/citation.cfm?id=2969239.2969428.

HIPOLA, P. et al. Ontology-based text summarization. the case of texminer. **Library Hi Tech**, 2014. ISSN 0737-8831.

HOVY, E. H. et al. Automated summarization evaluation with basic elements. In: **Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.** [s.n.], 2006. p. 899–902. Disponível em: http://www.lrec-conf.org/proceedings/lrec2006/summaries/438.html.

JÄRVINEN, P. Action research is similar to design science. **Quality & Quantity**, v. 41, n. 1, p. 37–54, Feb 2007. ISSN 1573-7845. Disponível em: https://doi.org/10.1007/s11135-005-5427-1.

JONES, K. S. Automatic summarising: Factors and directions. In: **Advances in Automatic Text Summarization**. [S.l.]: MIT Press, 1998. p. 1–12.

JONES, K. S. Automatic summarising: The state of the art. **Information Processing and management**, v. 43, n. 6, p. 1449 – 1481, 2007. ISSN 0306-4573. Text Summarization. Disponível em: http://www.sciencedirect.com/science/article/pii/S0306457307000878.

KITCHENHAM, B. **Procedures for Performing Systematic Reviews**. Department of Computer Science, Keele University, UK, 2004.

KUMAR, A. N. et al. Ontology-based retrieval & neural approaches for BioASQ ideal answer generation. In: **Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 79–89. Disponível em: https://www.aclweb.org/anthology/W18-5310.

LEHMANN, J. et al. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. **Semantic Web Journal**, IOS Press, 2014.

LI, L.; LI, T. An empirical study of ontology-based multi-document summarization in disaster management. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 44, n. 2, p. 162–171, Feb 2014. ISSN 2168-2216.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: **Proc. ACL workshop on Text Summarization Branches Out**. [S.l.: s.n.], 2004. p. 10.

LIN, D. An information-theoretic definition of similarity. In: **Proceedings of the Fifteenth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. (ICML '98), p. 296–304. ISBN 1-55860-556-8. Disponível em: http://dl.acm.org/citation.cfm?id=645527.657297.

LLORET, E.; PALOMAR, M. Text summarisation in progress: a literature review. **Artificial Intelligence Review**, v. 37, n. 1, p. 1–41, Jan 2012. ISSN 1573-7462. Disponível em: https://doi.org/10.1007/s10462-011-9216-z.

LUO, W. et al. Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. **Know.-Based Syst.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 46, p. 33–42, jul. 2013. ISSN 0950-7051. Disponível em: http://dx.doi.org/10.1016/j.knosys.2013.02.015.

LUO, W. et al. Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. **Know.-Based Syst.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 46, p. 33–42, jul. 2013. ISSN 0950-7051. Disponível em: http://dx.doi.org/10.1016/j.knosys.2013.02.015.

MACAVANEY, S. et al. Ontology-aware clinical abstractive summarization. **CoRR**, abs/1905.05818, 2019. Disponível em: http://arxiv.org/abs/1905.05818.

MARTINS, C. B.; RINO, L. H. M. Pruning unl texts for summarizing purposes. In: **Proc. of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS'2001)**. [S.l.: s.n.], 2001. (6th Natural Language Processing Pacific Rim Symposium, v. 1), p. 539–544.

MARTINS, C. B.; RINO, L. H. M. Revisiting unlsumm: Improvement through a case study. v. 1, p. 71–79, 11 2002.

MENDES, P. N. et al. Dbpedia spotlight: Shedding light on the web of documents. In: **Proceedings of the 7th International Conference on Semantic Systems**. New York, NY, USA: ACM, 2011. (I-Semantics '11), p. 1–8. ISBN 978-1-4503-0621-8. Disponível em: http://doi.acm.org/10.1145/2063518.2063519.

MOHAMED, M. A.; OUSSALAH, M. Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources. In: **2015 IEEE Trustcom/BigDataSE/ISPA**. [S.l.: s.n.], 2015. v. 2, p. 80–87.

MURRAY, G.; RENALS, S.; CARLETTA, J. Extractive summarization of meeting recordings. In: **in Proceedings of the 9th European Conference on Speech Communication and Technology**. [S.l.: s.n.], 2005. p. 593–596.

NARAYAN, S.; COHEN, S. B.; LAPATA, M. Ranking sentences for extractive summarization with reinforcement learning. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. Association for Computational Linguistics, 2018. p. 1747–1759. Disponível em: http://aclweb.org/anthology/N18-1158.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: ____. **Mining Text Data**. Boston, MA: Springer US, 2012. p. 43–76. ISBN 978-1-4614-3223-4.

NENKOVA, A.; PASSONNEAU, R. J. Evaluating content selection in summarization: The pyramid method. In: **Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004**. [s.n.], 2004. p. 145–152. Disponível em: http://aclweb.org/anthology/N/N04/N04-1019.pdf.

PEFFERS, K. et al. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, p. 45–77, 01 2007.

RADEV, D. R.; HOVY, E.; MCKEOWN, K. Introduction to the special issue on summarization. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 28, n. 4, p. 399–408, dez. 2002. ISSN 0891-2017. Disponível em: http://dx.doi.org/10.1162/089120102762671927.

RIBALDO, R. et al. Graph-based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. In: CASELI, H. et al. (Ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 260–271. ISBN 978-3-642-28885-2.

SEE, A.; LIU, P. J.; MANNING, C. D. Get to the point: Summarization with pointer-generator networks. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Association for Computational Linguistics, 2017. p. 1073–1083. Disponível em: http://aclweb.org/anthology/P17-1099.

TAN, J.; WAN, X.; XIAO, J. Abstractive document summarization with a graph-based attentional neural model. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Association for Computational Linguistics, 2017. p. 1171–1181. Disponível em: http://aclweb.org/anthology/P17-1108.

TEUFEL, S.; HALTEREN, H. van. Evaluating information content by factoid analysis: Human annotation and stability. In: **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain**. [s.n.], 2004. p. 419–426. Disponível em: http://www.aclweb.org/anthology/W04-3254.

UCHIDA, H. O. shi. Unl: Universal networking language-an electronic language for communication. In: . [S.l.: s.n.], 1996.

UMBRATH, W.; WETZKER, R.; HENNIG, L. An ontology-based approach to text summarization. **Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on**, IEEE Computer Society, Los Alamitos, CA, USA, v. 03, n. undefined, p. 291–294, 2008.

WU, K. et al. Ontology-enriched multi-document summarization in disaster management using submodular function. **Information Sciences**, v. 224, p. 118 – 129, 2013. ISSN 0020-0255.

ZHANG, X. et al. Neural latent extractive document summarization. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2018. p. 779–784. Disponível em: http://aclweb.org/anthology/D18-1088.

ZHOU, Q. et al. Neural document summarization by jointly learning to score and select sentences. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Association for Computational Linguistics, 2018. p. 654–663. Disponível em: http://aclweb.org/anthology/P18-1061.

ANEXO A – ARTIGO ACEITO PARA PUBLICAÇÃO

Ontology-based Extractive Text Summarization: The Contribution of Instances

Murillo Lagranha Flores[1], Elder Rizzon Santos[1], Ricardo Azambuja Silveira[1]

[1] Department of Informatics and Statistics, Federal University of Santa Catarina, Florianópolis Brazil

murillo.flores@posgrad.ufsc.br, {elder.santos,ricardo.silveira}@ufsc.br

Abstract. In this paper, we present a text summarization approach focusing on multi-document, extractive and query-focused summarization that relies on an ontology-based semantic similarity measure, that specifically explores ontology instances. We employ the DBpedia Ontology and a theoretical definition of similarity to determine query-sentence and sentence-sentence similarity. Furthermore, we define an instance-linking strategy that builds the most accurate sentence representation possible while achieving a better coverage of sentences that can be represented by ontology instances. Using primarily this instances linking strategy, the semantic similarity measure and the Maximal Marginal Relevance Algorithm - MMR - we propose a summarization model that is capable of avoiding redundancy from a more fine-grained representation of sentences, due to their representation as ontology instances. We demonstrate that our summarizer is capable of achieving compelling results when compared with relevant DUC systems and recently published related studies using ROUGE metrics. Moreover, our experiments lead us to a better understanding of how ontology instances can be used to represent sentences and what is the role of said instances in this process.

Keywords. Extractive text summarization, Ontologies, Ontological instances.

Introduction

Text Summarization is the task of creating a shorter version of a document or a set of documents while keeping most of the informational content present in these documents. Automatic Text Summarizers are usually classified with regards to how they construct the final summary as either extractive or abstractive [12]. In extractive summarization, the summary is built by concatenating textual units (usually paragraphs or sentences) extracted from the original documents. Due to its conceptual simplicity and the guarantee that the sentences used in summary will be at least as legible as the sentences in the original documents extractive summarization has been a very prominent approach in automatic text summarization for the last decades.

Constructing an extractive summary that covers most of the information present in the original documents while achieving a significant reduction in length is essential to avoid redundancy. Ontology-based summarizers proposed so far explored the use of concepts as a proxy to represent the semantics of sentences, successfully avoiding redundancy and therefore achieving great results in generating extractive summaries. However, due to their inability to distinguish between different references to the same concept, which reduces their ability to evaluate the semantics of sentences, ontology-based extractive summarizers that only explore concepts have the tendency to leave relevant sentences out of the summary for considering them redundant, when in fact they hold references to different instances of the same concept. The use of manually built ontologies makes the problem more severe, due to their reduced number of concepts.

In this paper we propose to use ontology instances to represent the semantics of sentences, attacking the problem mentioned above.

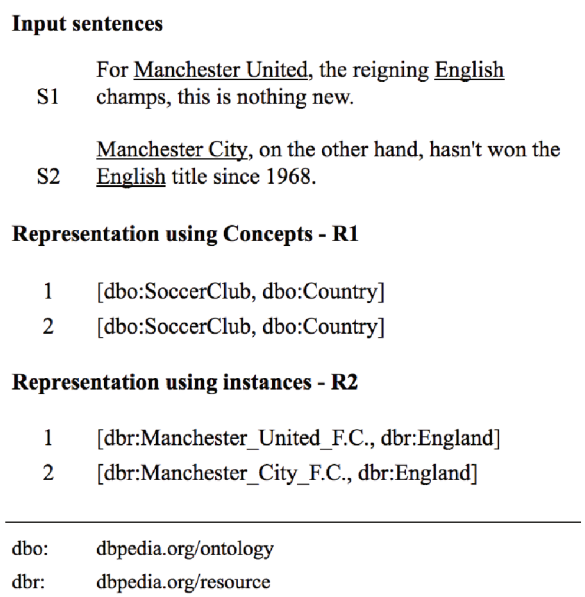


Fig. 1. Representation of sentences using concepts and instances defined in a ontology

Figure 1 depicts the advantages of using instances to represent the semantics of a sentence. Sentences S1 and S2 are referencing two distinct football teams from the same city, and positioning them with regards to their past performance on the English football championship. In conjunction, they are comparing and making an argument about both teams performances. When these sentences are represented as a vector of concepts, as can be seen in R1, their representations are identical. When they are represented as a vector of instances instead, as seen in R2, their representation changes and comes closer to the real semantic differences between them.

With the goal of improving the quality of summaries built by extractive query-focused text summarizers in mind, we present an ontology-instances based summarization model. Our model uses an automatic annotation tool to link sentences to instances defined in an ontology, then uses these instances to represent the sentences and finally a semantic similarity measure to calculate the similarity between two sets of instances. We experiment on the DUC2005 dataset.

Representing Sentences as Concepts

To evaluate the impact that representing sentences as concepts have on the process of detecting redundant sentences, we calculated the overlap between the concepts representing distinct sentences of a summary from two summary sets.

We used the DUC2004 task 2 dataset on this evaluation. In this dataset model summaries are manually created summaries, and peer summaries have been limited to include only summaries created by participating systems.

We started by employing a system for automatic annotation of DBpedia instances on text, DBpedia spotlight [5], to identify references to instances on each summary. Because the instances described in the DBpedia ontology are linked to other ontologies, we then grouped the concepts that appeared in the *rdf:type* of each instance by ontology. After that, we selected the first concept that was listed as the *rdf:type* of each group as the concept that best represented that mention on the text. With that, we created vectors of concepts that appeared in each sentence of each summary, per ontology. Those lists were considered the representation of each sentence as a vector of concepts. We consider the final result of this process to be similar to what an ontology-based summarizer that only employ concepts to represent sentences would achieve.

We calculated the intersection between the vectors of concepts representing the sentences of each summary. We classified the results in **total intersection** when the same vector of concepts represented at least two sentences of the same summary, and **partial intersection** when at least one concept appeared in two distinct vectors representing sentences of the same summary. Table 1 shows the final results. We only included results for the DBpedia and the Schema.org ontologies, as those are the larger ones linked to DBpedia instances and therefore can generate a more diverse representation of sentences.

We found that peer summaries tend to have a lower total and partial intersections than model summaries as can be seen in table 1.

Documents set	Total intersection	Partial intersection
Peer		
dbpedia	0.346	0.747
schema.org	0.343	0.705
Model		
dbpedia	0.601	0.870
schema.org	0.626	0.845

Table 1. Percentage of documents with two or more sentence representations matching totally or partially on DUC 2004.

These results demonstrate that it is common to reference the same concept more than once in model summaries, created by humans. Therefore, using only concepts to detect and avoid redundancy has the potential to remove sentences that appear to be important in human-created summaries. It appears that a more granular semantic representation, that can compare the differences between sentences more precisely can achieve better results. Corroborates to this idea the fact that peer summaries have lower total and partial intersections.

Base Model

We use the MMR algorithm as our base model [3]. At each iteration, the algorithm selects a sentence to extract and include in the summary, until the desired length is reached. The sentence selected is always the one that is (i) more similar to the query and (ii) less redundant when compared with the previously selected sentences. We choose the MMR algorithm as our base model because it is specifically tailored for extractive query-focused summarization and can easily be extended to incorporate the advantages of a better similarity measure capable of comparing sentences and query. MMR is defined as in expression 1.

$$MMR \stackrel{def}{=} \max_{D_i \in R/S} \Big[\alpha(sim_1(D_i, Q)) - (1-\alpha) \max_{D_j \in S} sim_2(D_i, D_j) \Big] \quad (1)$$

Where Q is a query, R is a documents collection (cluster), S is the subset of documents in R already selected, R/S is the set of yet unselected documents and sim_1 and sim_2 are similarity metrics.

Semantic Similarity Using Instances

We extend MMR through the definition of a semantic similarity measure capable of calculating query-sentence and sentence-sentence similarity that can be used by that algorithm.

Representing sentences as instances

In order to determine a sentence's relevance to a specific query using ontology instances a representation of each one of them using said instances must be constructed. To this end, we employ an Instances Linking System (ILS). Instances linking systems will take snippets of text, in our case a sentence, as input and output a list of ontology instances that are mentioned in the input.

One typical problem faced by instances linking systems is the absence of detected mentions due to either a reduced number of instances defined in the underlying ontology or some inefficiency of the ILS. Instances linking systems might allow the configuration of a confidence parameter, that determines the minimum level of confidence that the ILS must have in order to link a mention, to address the problem of ILS inefficiency. Ensuring that all sentences have a valid representation is fundamental to guarantee that an instances based summarizer will be able to operate correctly, therefore we devise an approach to deal with the problem mentioned above based on the possibility of configuring a confidence parameter, as shown in algorithm 1, where $ILSLink$ is a function that links instances at a given level of confidence using the ILS.

Algorithm 1 IL with variable confidence

Input T: Text from a sentence or query.
Input c: Initial confidence value.
procedure LINK(T, c)
    $s \leftarrow ILSLink(T, c)$
    if $s = \emptyset \ \& \ c >= 0.1$ **then**
        $c \leftarrow c - 0.1$
        $s \leftarrow Link(T, c)$
    return s

Similarity between sets of instances

We defined sentence-sentence and query-sentence semantic similarity using their representations as ontology instances. Inspired by the work of [11] the semantic similarity between two sets of instances is defined as the average of the maximal similarity between the instances representing each one of the sets, as shown in expression 2

$$Sim(S_1, S_2) = \frac{1}{2}\left[\frac{\sum_{i_1 \in S_1} \max_{i_2 \in S_2} Sim(i_1, i_2)}{|S_1|} + \frac{\sum_{i_2 \in S_2} \max_{i_2 \in S_1} Sim(i_1, i_1)}{|S_2|}\right] \quad (2)$$

Where Sim is a semantic similarity measure between two ontology instances. We define and describe the measure used in this work in section 4.3. This definition assumes a symmetrical contribution of each one of the instance sets under comparison.

When used in conjunction with the algorithm defined in section 4.1 this definition ensures that the contribution to the overall similarity added by instances linked at a given level of confidence will not decrease with lower confidence values. In other words, its maximal similarity will not decrease with the addition of instances that are less strongly related to the sentence in question. It is worth noting, however, that the addition of said instances does increase the number of instances representing each sentence which might increase the denominator in each side of expression's 2 sum.

Similarity between instances

We use the theoretical definition of similarity presented by [7] to define the semantic similarity measure used in this work. According to [7] "The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are" and can be expressed by the following equation.

$$sim(A, B) = \frac{logP(common(A, B))}{logP(description(A, B))} \quad (3)$$

We believe that the relations that an ontology instance holds with other instances contain valuable semantic information about it. We also believe that an instance's types - the concept that they are an instance of - hold a significant amount of semantic information about it. We define that the description of an instance is formed by its types and relations with other instances, for the sake of computing its semantic similarity with other instances. Furthermore, we define that each relation will add two pieces of information to the description, separately: its type and the instance it connects to. Therefore, the description of an instance will be formed by three distinct categories of information: its **types**, its **relation types** and its **relation instances**. Figure 2 depicts two instances of different concepts that are related to each other and their description as it would be used to compute their semantic similarity.

To apply Lin's [7] definition of similarity to this work, we first define that the probability of any given component of the description of an instance is given by the probability of that component being present at a randomly selected instance of the ontology, as shown in equation 4. The probability of a relation type, for instance, is the probability of a relation of that type being present on a randomly selected instance of the ontology.

$$P(component) = \frac{count(\text{instances with component})}{count(\text{total number of instances})} \quad (4)$$

We expand the definition presented in equation 4 to define the similarity of a description category

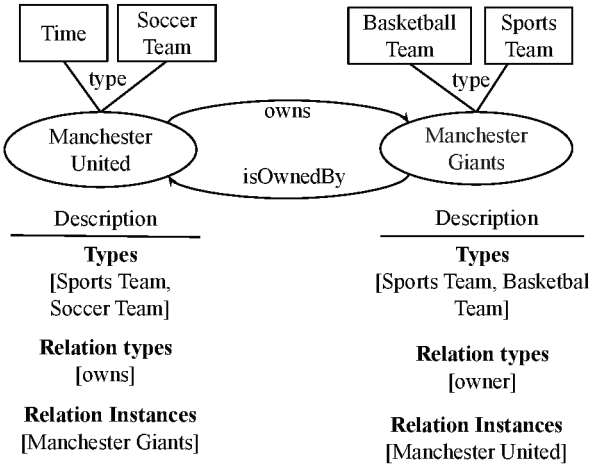


Fig. 2. Instances description example

(types, relation types and relation instances) as in equation 5, where cat is a function that returns all components of one of the categories of information in the description of the instance passed as a parameter.

$$\frac{2*\left(-1*\sum_{c\in(cat(A)\cap cat(B))} logP(c)\right)}{\left(-1*\sum_{c\in cat(A)} logP(c)\right)+\left(-1*\sum_{c\in cat(B)} logP(c)\right)} \quad (5)$$

Finally, the overall similarity is defined as the average of category similarities, as expressed in equation 6, where Sim_{types} is the **types** similarity, $Sim_{relTypes}$ is the **relation types** similarity and $Sim_{relInst}$ is the **relation instances** similarity.

$$sim(A,B) = \frac{1}{3}\Big[Sim_{types} + Sim_{relTypes} + Sim_{relInst}\Big] \quad (6)$$

As an example, table 2 presents the similarity between instances of the 2014 DBpedia Ontology calculated following the previous definitions.

The results in table 2 align well with the values expected from a similarity measure that follows the assumptions defined in [7]. Maximum similarity is achieved when an instance is compared against itself (line 1). Teams of the same league - Los Angeles Lakers and Golden State Warriors are in the same league, as well as New England Patriots and Seattle Seahawks - have a higher similarity when compared against each other than when compared against a team in the other league (lines 2, 4 and 3). The similarities between teams in the same league have close values for both leagues (lines 2 and 4). Moreover, the similarity between a sports team and a fictional character is an order of magnitude smaller than between two sports teams on different leagues (lines 3 and 5).

#	Measure	Value
1	sim(L.A. Lakers, L.A. Lakers)	1
2	sim(L.A. Lakers, G.S. Warriors)	0.6248
3	sim(L.A. Lakers, N.E. Patriots)	0.3958
4	sim(N.E. Patriots, S. Seahawks)	0.6301
5	sim(L.A. Lakers, Spider Man)	0.0363

Table 2. Similarity between 2014 DBPedia Ontology Instances

Our Model

We extend the MMR algorithm by employing the instances linking strategy and the similarity measures defined in the previous section, as shown in figure 3.

First, instances are linked to the input documents and the query. When linking instances to the query either a fixed minimum confidence or the strategy discussed in section 4.1 are used. After that, the input documents and the query are segmented into sentences. The MMR algorithm then uses these sentences and the instances linked to them to extract sentences and build the summary following the definition in expression 7, where S_Q are the query sentences, S_D are the documents sentences, S_S are the subset of the document sentences already selected, S_D/S_S are the yet unselected document sentences and sim is the similarity measure defined in section 4.2.

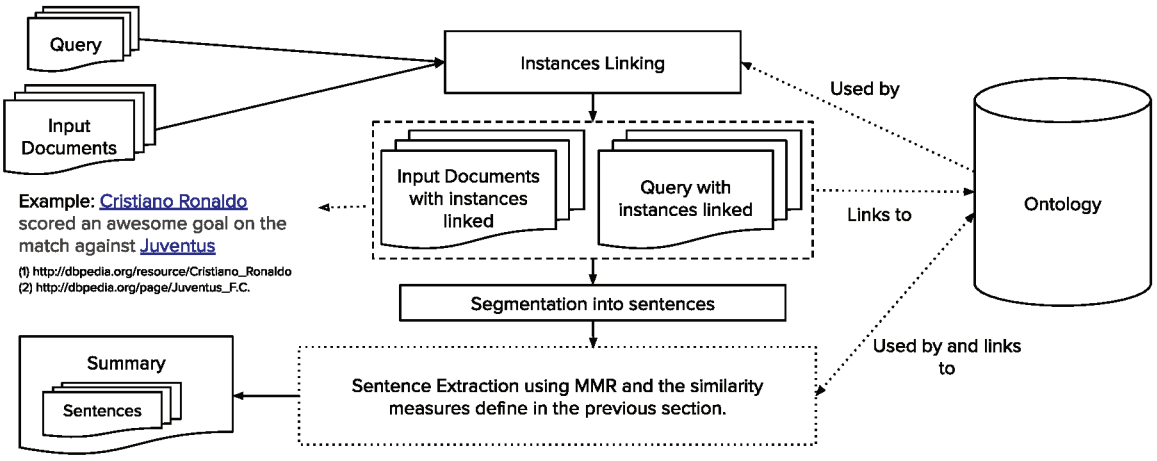


Fig. 3. The full architecture of our model

$$MMR \stackrel{def}{=} \max_{S_i \in S_D / S_S} \Big[\alpha(sim(S_i, S_Q)) - (1-\alpha) \max_{S_j \in S} sim(S_i, S_j) \Big] \quad (7)$$

Related Works

Different authors have used ontologies in numerous approaches to address the extractive summarization problem.

[13] used an ontology to create a graph where each ontological concept of the document becomes a vertex and every relation between concepts becomes an edge. The most "central" sentences on that graph are extracted and establish the summary. [1] used the YAGO ontology to evaluate sentences considering a feature that expresses the sentence's popularity and pertinence, called *entityRank*. Later, sentences are extracted using a variation of MMR strategy [3]. [4] described the adopted techniques and the design of a system called Texminer, which uses ontologies in a very similar approach to the one followed by [14]. [15], heavily influenced by [8] applies an ontology to represent sentences as sets of concepts and to compute the similarity between the sentences. Closely related to our work [11] derived an approach for extractive multi-document query-focused summarization based on a semantic similarity measure that employed the WordNet Taxonomy as its knowledge-based. The authors enhanced the similarity measure with named entity semantic relatedness inferred from Wikipedia.

Different approaches that did not employ ontologies also addressed the multi-document extractive summarization problem. [2] proposed a query-focused approach based on a weighted archetypal analysis (wAA), a multivariate data representation using matrix factorization and clustering. [9] also proposed a query-focused approach suggesting to focus on three different considerations: 1) relevance, 2) coverage and 3) novelty in a probabilistic modeling framework.

Previous studies on extractive summarization have only used ontologies to capture the hierarchy of concepts in a specific domain, effectively using them as a taxonomy. Ontology instances have not been explored so far, and we are the first ones to use them to represent sentences as a way to compare these sentences semantically and enhance the summarizer's performance.

Experiments

In this section, we report the experiments conducted to evaluate the effectiveness of our

proposed model in multi-document query-focused extractive summarization.

Experimental settings

We used the DUC 2005 dataset for evaluation. The DUC 2005 dataset is formed by 50 document clusters, each containing between 25 and 50 different documents on a specific topic. Each cluster has on average 31 documents and 20,236 words. The desired number of words in the summaries is 250. For each document set, between four and ten model summaries are available. This dataset was specifically created for the evaluation of multi-document query-focused summarizers.

To evaluate the summaries generated by our system quantitatively and compare them against baselines summaries as well as against summaries generated by closely related systems we use the Rouge family of metrics [6]. Rouge metrics are the *de-facto* standard in extractive summary evaluation, being widely used in the existing literature. The assessment of the quality of a summary carried out by Rouge-n metrics is based on existing model summaries (usually generated by humans) and the co-occurrence of n-grams between those model summaries and the summaries under evaluation. The evaluation follows the definition in expression 8.

$$ROUGE - N = \frac{\sum_{S \in \{Ref.Summ.\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Ref.Summ.\}} \sum_{gram_n \in S} Count(gram_n)} \quad (8)$$

Where N is the length of the N-gram, $Count(gram_n)$ is the number of n-grams in the reference summary and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the summary being evaluated and a set of reference summaries (Ref. Summ.).

Implementation

To conduct the experiments we implemented our proposed model selecting the DBpedia Ontology as our base ontology. The 2014 DBpedia Ontology was built by knowledge extracted from Wikipedia and has more than four million instances defined in it [5].

We used DBpedia Spotlight to link DBpedia ontology instances to text. DBpedia Spotlight is capable of linking instances through different surface forms and with a configurable disambiguation confidence [10].

We experimented with two different variations of our system, one using a fixed value for instances linking disambiguation confidence on both documents and query and one using a variable value for the query, as described in section 4.1. We ran each variation with three initial confidence values - 0.3, 0.6, 0.9 - and three MMR α values - 0.3, 0.5, 0.7 - totalizing 18 different experimental runs.

As for computing Rouge metrics, we used the same ROUGE-1.5.5.pl Perl script used to compute the scores in the original DUC2005 competition. The parameters used were also the same ones used by DUC2005 [1].

Results

We evaluated the quality of the summaries generated by our systems using Rouge-1 and Rouge-2 as these perform better in multi-document summarization evaluation [6].

The systems name notation used in the figures describing results is defined as follows: Each system name is formed by a prefix and a suffix, separated by a dash ("-"). The prefix indicates whether that version of the system used a fixed (FIX) or a variable (VAR) confidence value when linking instances to the query. The suffix indicates the initial confidence value of the system. As an example, VAR-0.6 indicates that that version of the system ran with a variable confidence value (to link instances to the query, as described in section 4.1) starting from 0.6.

[1] ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d

Figure 4 shows the ROUGE-1 scores obtained by all systems, with three different values of the MMR parameter α configured. This parameter controls the balance between query relevance and summary diversity when selecting sentences. The figure shows that all systems presented better results as the instances annotation confidence decreased for values of α greater than or equal to 0.5 when query relevance had a more significant impact on sentence selection. With the α value set to 0.3 the opposite occurred - the results decreased as the confidence decreased, with a particularly acute drop between systems with confidence configured to 0.6 and 0.3. It is also worthwhile to note that except for FIX-0.9 all systems achieved their best results with α set to 0.7. These results indicate that if more relevance is given to the query, the more instances are annotated in the documents, the better. If summary diversity is given more importance when selecting sentences, more instances annotated in the documents might lead to worst results.

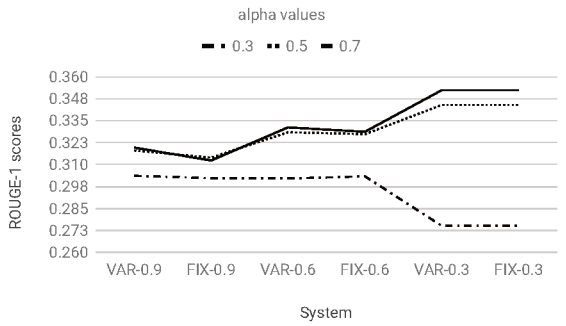


Fig. 4. Rouge-1 scores per system, with three different values of alpha

One possible explanation for achieving better performance with lower confidence and higher alpha values lives on the length difference between the query and the documents. Because the query is very short when compared to the documents, the extra instances it gets with lower values of confidence compensates the noise introduced by the extra, possibly irrelevant, instances linked to the documents. This extra instances will increase sentence extraction performance significantly when alpha is configured to a value that gives query relevance more importance than summary diversity - or at least equals to.

Figure 5 shows that the systems with variable decreasing confidence in instances annotation on the query achieved better results than the versions with fixed confidence at the same starting level of confidence, in two occasions for a fixed α of 0.7. That corroborates with the explanation that more instances annotated on the query are more relevant to increase performance with higher values of α.

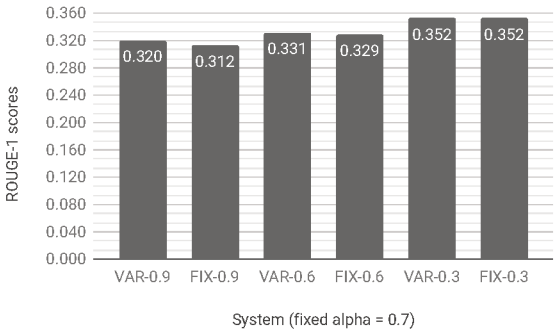


Fig. 5. Rouge-1 scores per system, with a fixed value of alpha (0.7)

Table 3 presents a comparison between the average of DUC2005 systems, closely related works and the results obtained by the best variant of our systems, VAR-0.3 system with α set to 0.7. All systems were experimented in the DUC2005 dataset. Our system outperforms the average DUC2005 systems in both compared metrics, but it also falls behind all the other systems under comparison in both metrics.

System	Rouge-1	Rouge-2
Avg. DUC2005 Systems	0.3434	0.0602
Luo et al. [9]	0.3728	0.0807
Canhasi et al. [2]	0.3945	0.0797
Mohamed et al. [11]	0.3949	0.0693
This work	0.3524	0.0639

Table 3. Comparison between the average of DUC2005 systems, closely related works and our results.

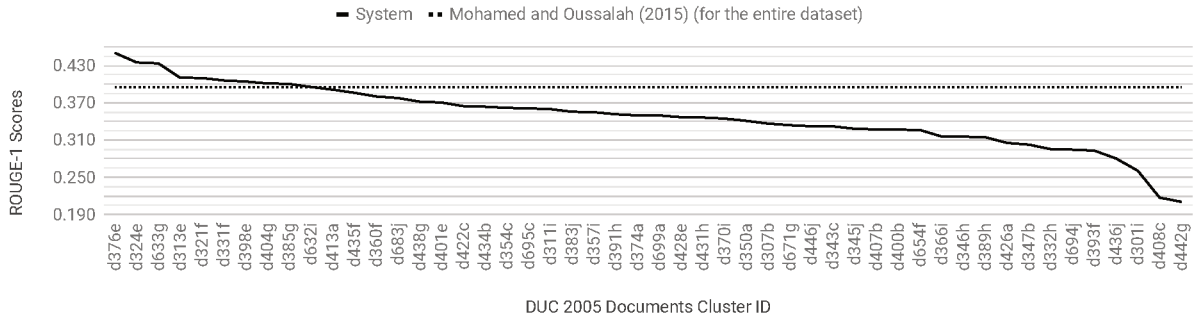


Fig. 6. Rouge-1 scores per DUC2005 document cluster, obtained by the best variant of our system.

We also analyzed the Rouge-1 Scores obtained by the best variant of our system in all 50 DUC2005 document clusters. The results are shown in figure 6, ordered by decreasing Rouge-1 scores from left to right. To help visualize the quality of the results we also plot a line representing the best result shown in table 3 [11] for the entire dataset. As can be seen in the figure, at first the results achieved are above that line. They then decrease in a way that resembles a linear descent with a sudden drop at the end. These results indicate that it is possible to achieve great results using instances to represent sentences and the techniques described in section 5, but further analysis is required to understand what's preventing the system from performing better on the clusters where performance is falling above the compared best.

We can draw from the conducted experiments that ontology instances can contribute to boosting the performance of extractive multi-document query-focused summarizers, by enhancing sentence-query similarity comparison and therefore helping identify sentences that are more relevant to the query. The fact that all versions of our summarizer presented better (or at least equal) results when an effort to enhance the query representation was made, by varying the instances linking confidence parameter as described in section 4.1, is an empirical evidenced of that. Following the same idea, we can also understand that the performance of summarizers based on ontology instances is highly dependent on the quantity and semantic coverage of the instances defined in the ontology and on the quality of the instances linking process. Better algorithms and similarity metrics can remediate an excess of irrelevant instances linked to the query and the documents, but they cannot remediate a lack of instances.

Conclusion

We proposed to use ontology instances to build an extractive query-focused multi-document summarization model, as a way to achieve a more fine-grained representation of the semantics of sentences, and avoid the problem of over-pruning sentences due to a limited semantic representation. We showed that when ontology concepts are used to represent the semantics of sentences, human-created summaries have more sentences with overlapping representations than automatically generated ones. We extended the MMR algorithm to build our model, through an instance linking strategy with variable linking confidence and a similarity measure based on ontology instances. We experimented on the DUC2005 dataset and concluded that although representing sentences as ontology instances can help boost summarization performance further analysis is still needed to achieve better results.

References

1. **Baralis, E., Cagliero, L., Jabeen, S., Fiori, A., & Shah, S. (2013).** Multi-document summarization based on the yago ontology. *Expert Syst. Appl.*, Vol. 40, No. 17, pp. 6976–6984.
2. **Canhasi, E. & Kononenko, I. (2014).** Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Syst. Appl.*, Vol. 41, No. 2, pp. 535–543.
3. **Carbonell, J. & Goldstein, J. (1998).** The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, ACM, New York, NY, USA, pp. 335–336.
4. **Hipola, P., Senso, J. A., Leiva-Mederos, A., & Dominguez-Velasco, S. (2014).** Ontology-based text summarization. the case of texminer. *Library Hi Tech*.
5. **Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2014).** DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.
6. **Lin, C.-Y. (2004).** Rouge: A package for automatic evaluation of summaries. *Proc. ACL workshop on Text Summarization Branches Out*, pp. 10.
7. **Lin, D. (1998).** An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 296–304.
8. **Lin, H. & Bilmes, J. (2010).** Multi-document summarization via budgeted maximization of submodular functions. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 912–920.
9. **Luo, W., Zhuang, F., He, Q., & Shi, Z. (2013).** Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. *Know.-Based Syst.*, Vol. 46, pp. 33–42.
10. **Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011).** Dbpedia spotlight: Shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, ACM, New York, NY, USA, pp. 1–8.
11. **Mohamed, M. A. & Oussalah, M. (2015).** Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources. *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 2, pp. 80–87.
12. **Nenkova, A. & McKeown, K. (2012).** *A Survey of Text Summarization Techniques*. Springer US, Boston, MA, pp. 43–76.
13. **Ramezani, M. & Feizi-Derakhshi, M.-R. (2015).** Ontology-based automatic text summarization using farsnet. *Advances in Computer Science : an International Journal*, Vol. 4, No. 2, pp. 88–96.
14. **Umbrath, W., Wetzker, R., & Hennig, L. (2008).** An ontology-based approach to text summarization. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, Vol. 03, No. undefined, pp. 291–294.
15. **Wu, K., Li, L., Li, J., & Li, T. (2013).** Ontology-enriched multi-document summarization in disaster management using submodular function. *Information Sciences*, Vol. 224, pp. 118 – 129.

Article received on __/__/____ accepted on __/__/____.
Corresponding author is Murillo Flores.