Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

RISO-GCT –Determinação do Contexto Temporal de Conceitos em Textos

George Marcelo Rodrigues Alves

Campina Grande, Paraíba, Brasil.
Fevereiro - 2016

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

RISO-GCT –Determinação do Contexto Temporal de Conceitos em Textos

George Marcelo Rodrigues Alves

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Recuperação de Informação

ULRICH SCHIEL
(Orientador)

Campina Grande, Paraíba, Brasil.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

A474r Alves, George Marcelo Rodrigues.

RISO – GCT – Determinação do contexto temporal de conceitos em textos / George Marcelo Rodrigues Alves. – Campina Grande, 2016.

95 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.

"Orientação: Prof. Dr. Ulrich Schiel".

Referências.

1. Processamento de Linguagem Natural. 2. Indexação Temporal. 3. Reconhecimento de Padrões Temporais. 4. Mapa de Tópicos. I. Schiel, Ulrich. III. Título.

CDU 004.43(043)

Resumo

Devido ao crescimento constante da quantidade de textos disponíveis na Web, existe uma necessidade de catalogar estas informações que surgem a cada instante. No entanto, trata-se de uma tarefa árdua e na qual seres humanos são incapazes de realizar esta tarefa de maneira manual, tendo em vista a quantidade incontável de dados que são disponibilizados a cada segundo. Inúmeras pesquisas têm sido realizadas no intuito de automatizar este processo de catalogação.

Uma vertente de grande utilidade para as várias áreas do conhecimento humano é a indexação de documentos com base nos contextos temporais presentes nestes documentos. Esta não é uma tarefa trivial, pois envolve a análise de informações não estruturadas presentes em linguagem natural, disponíveis nos mais diversos idiomas, dentre outras dificuldades.

O objetivo principal deste trabalho é criar uma abordagem capaz de permitir a indexação de documentos, determinando mapas de tópicos enriquecidos com conceitos e as respectivas informações temporais relacionadas. Tal abordagem deu origem ao RISO-GCT (Geração de Contextos Temporais), componente do Projeto RISO (Recuperação da Informação Semântica de Objetos Textuais), que tem como objetivo criar um ambiente de indexação e recuperação semântica de documentos possibilitando uma recuperação mais acurada.

O RISO-GCT utilizou os resultados de um módulo preliminar, o RISO-TT (Temporal Tagger), responsável por etiquetar informações temporais presentes em documentos e realizar o processo de normalização das expressões temporais encontradas. Deste processo foi aperfeiçoada a abordagem responsável pela normalização de expressões temporais, para que estas possam ser manipuladas mais facilmente na determinação dos contextos temporais. .

Foram realizados experimentos para avaliar a eficácia da abordagem proposta nesta pesquisa. O primeiro, com o intuito de verificar se o Topic Map previamente criado pelo RISO-IC (Indexação Conceitual), foi enriquecido com as informações temporais relacionadas aos conceitos de maneira correta e o segundo, para analisar a eficácia da abordagem de normalização das expressões temporais extraídas de documentos. Os experimentos concluíram que tanto o RISO-GCT, quanto o RISO-TT incrementado obtiveram resultados superiores aos concorrentes.

Palavras-chave: indexação temporal, mapa de tópicos, reconhecimento de padrões temporais, processamento de linguagem natural.

Abstract

Due to the constant growth of the number of texts available on the Web, there is a need to catalog that information which appear at every moment. However, it is an arduous task in which humans are unable to perform this task manually, given the increased amount of data available at every second. Numerous studies have been conducted in order to automate the cataloging process.

A research line with utility for various areas of human knowledge is the indexing of documents based on temporal contexts present in these documents. This is not a trivial task, as it involves the analysis of unstructured information present in natural language, available in several languages, among other difficulties.

The main objective of this work is to create a model to allow indexing of documents, creating topic maps enriched with the concepts in text and their related temporal information. This approach led to the RISO-GCT (Temporal Contexts Generation), a part of RISO Project (Semantic Information Retrieval on Text Objects), which aims to create a semantic indexing environment and retrieval of documents, enabling a more accurate recovery.

RISO-GCT uses the results of a preliminary module, the RISO-TT (Temporal Tagger) responsible the labeling temporal information contained in documents and carrying out the process of normalization of temporal expressions. Found. In this module the normalization of temporal expressions has been improved, in order allow a richer temporal context determination.

Experiments were conducted to evaluate the effectiveness of the approach proposed a in this research. The first, in order to verify that the topic map previously created by RISO-IC has been correctly enriched with temporal information related to the concepts correctly, and the second, to analyze the effectiveness of the normalization of expressions extracted from documents. The experiments concluded that both the RISO-GCT, as the RISO-TT, which was evolved during this work, obtained better results than similar tools.

Keywords: temporal indexing, topic map, recognition of temporal patterns, natural language processing.

Agradecimentos

Agradeço primeiramente a Deus, que me proveu força, saúde e determinação cruciais para conclusão este trabalho, principalmente nos momentos em que eu mais me desacreditei. Ele sempre esteve presente durante todo os momentos da minha vida, e não seria diferente durante as etapas que compreenderam o mestrado. Obrigado por nunca ter me abandonado, inclusive nos momentos nos quais questionei minha fé.

Agradeço aos meus pais "Seu Gildo" e "Dona Zezinha", que com muito sacrifício, investiram na minha educação, com a qual pude alçar voos cada vez mais altos. Obrigado por todo carinho e amor, principalmente nesses últimos anos onde enfrentei várias dificuldades e nunca deixaram de me estender a mão. Espero um dia ser o filho que vocês sempre mereceram. Saibam que para mim vocês são exemplo e vou tê-los como base na educação dos meus futuros filhos

Agradeço aos meus irmãos Gildo Junior e Giullyane. Apesar de termos nos distanciado nos últimos anos, crescemos juntos compartilhamos muitos momentos felizes e que foram fundamentais para minha formação como ser humano.

Agradeço aos meus avós Seu Manoel Carneiro, Dona Maria Carneiro (*In Memoriam*), Dona Nininha e Seu Serafim (*In Memoriam*), que foram e são os melhores avós que uma pessoa pode ter.

Agradeço aos meus tios, tias, primos e primas, que são pessoas que me enchem de orgulho de fazer parte da família Rodrigues/Alves e que são exemplo de união, carinho e respeito. Obrigado por toda a força e os inúmeros momentos felizes que compartilhamos. Peço desculpas pelas vezes em que precisei me ausentar do convívio de vocês, correndo em busca dos meus objetivos.

Agradeço aos meus sogros "Dona Ana" e "Seu João" que me receberam de braços abertos como filho. Agradeço à minha cunhada Rafaela e seu futuro esposo Jhonnatan, e minhas cunhadas Joanna e Ramilla Morais. Vocês merecem tudo que há de melhor nessa vida.

Agradeço aos meus amigos e colegas de trabalho que foram válvula de escape nos momentos de tensão, me proporcionando momentos de muita alegria e os com os quais sempre pude contar. Espero não estar cometendo nenhuma injustiça, mas existem alguns que não poderia deixar de citar e que representam todos os outros: Braulyo, Elvis, Elymar, Lennerson, Vital, Geverson e Douglas, amigos de todas as horas que não dispensam uma farra. Silas, Marcus e Rael amigos da época do colégio e que tenho o prazer de conviver até hoje. Renally, Taiane, Sabrina e toda turma de Administração e "Medicina". Fernando e Daniel, grandes amigos que fiz na época da graduação. Jailson, Arthur, Alcione e Vicente, representado meus colegas de trabalho

e que representam aqueles que fazem parte da melhor equipe de TI que este mundo já viu.

Agradeço ao meu orientador Ulrich Schiel, um cara fantástico que tive o prazer de trabalhar. Obrigado por ter acreditado em mim, pela paciência e por estar sempre disponível nas horas quem que precisei de um norteamento na minha pesquisa. Seus conselhos foram de fundamental importância para conclusão deste mestrado, e dos quais muitos levarei para a vida inteira.

Agradeço a todos os colegas de mestrado, que contribuíram direta e indiretamente na minha pesquisa, dentre os quais poderia deixar de citar meus companheiros de várias madrugadas de estudo: Anderson Felinto, Arthur Marques, Hamon Barros e Ruan Pierre. Agradeço também aos colegas Adriano Santos, Zé Gildo e Magna Bispo, que participaram anteriormente do Projeto RISO e sempre se mostraram solícitos a sanar todas as dúvidas relacionadas ao projeto que tive durante o mestrado.

Por fim, Renata Morais, minha esposa. Aquela que foi a maior incentivadora para que eu realizasse esta etapa dos meus estudos, desde a inscrição para a seleção do mestrado, até as fases finais para conclusão deste estudo. Sei que os esses últimos anos não foram nada fáceis e que em várias vezes não fui tão presente e atencioso quanto gostaria de ser, mas você foi bastante paciente e compreensiva e entendeu que todo esse sacrifício não foi em vão. Saiba que você foi e sempre será minha principal inspiração em tudo que eu fizer nesta vida. Te amo!

Conteúdo

Lista de Siglas

1: PLN – Processamento de Linguagem Natural

2: RISO – Recuperação da Informação Semântica de Objetos Textuais

3: EAT – Extração Automática de Termos

4: EI – Extração da Informação

5: RI – Recuperação da Informação

Lista de Figuras

Lista de Tabelas

Capítulo 1 - Introdução

Devido ao crescimento da quantidade de textos disponíveis na Web, surgiu a necessidade cada vez maior de localizar padrões específicos de informações textuais a fim de extrair estas informações relevantes de maneira estruturada. A extração de maneira automática destas informações deu origem à área de Extração da Informação (EI).

EI possui várias aplicações, por exemplo, a informação disponível em textos não-estruturados pode ser armazenada em bancos de dados tradicionais e usuários podem examiná-las através de consultas padrão. Para isso, há um complexo trabalho de gerenciamento, que é consequência da natureza não estruturada e da difícil análise dos dados (ZAMBENEDETTI, 2002).

Uma sub-área da EI é a Extração Automática de Termos (EAT), cuja extração é focada na em termos em geral. A EAT tem sido muito importante, uma vez que a extração de termos é um trabalho muito difícil de ser realizado manualmente. A EAT é responsável por extrair informações relevantes de um documento, página da Web, entre outros, podendo classifica-la de acordo com algum determinado critério. Estas informações são convertidas em uma estrutura tabular de maneira legível tanto para humanos quanto para uma aplicação. Posteriormente são armazenadas em um banco de dados (SILVA, 2003).

Dentre os profissionais que mais se interessam por EAT em textos destacam-se linguistas computacionais, linguistas aplicados, tradutores, interpretes, jornalistas científicos, entre outros (ZAVAGLIA et al., 2005).

A Recuperação da Informação (RI), frequentemente confundida com EI, diz respeito à tarefa de encontrar documentos a partir de consultas fornecidas como entrada. Esta atividade é avaliada de acordo com a capacidade que o sistema tem de recuperar o maior número de documentos relevantes e excluir o máximo de itens irrelevantes (BISPO, 2013).

Esta dissertação se destina ao processamento de documentos textuais para extração de Contextos Temporais de Conceitos, que consistem na extração de relacionamentos entre os conceitos presentes nos documentos e as expressões (sintagmas) temporais associadas a estes conceitos.

O intuito deste trabalho é prover um processo de RI, onde, durante a fase de indexação, é realizado o enriquecimento dos conceitos presentes no documento com base em dados temporais a eles relacionados. Com isto, pretende-se melhorar a acurácia da recuperação de informação. Especificamente, este indexamento temporal dará suporte ao ambiente de consultas do Projeto RISO (Recuperação da Informação Semântica de Objetos Textuais), sistema em desenvolvimento

pelo grupo de Sistemas de Informação e Bancos de Dados (SINBAD), do DSC/CEEI/UFCG, cujo objetivo é criar um ambiente de indexação e recuperação semântica de documentos.

Identificados os relacionamentos entre conceitos e expressões temporais, podemos ter conceitos cujas expressões temporais associadas, necessitem de um processamento adicional para determinarmos a data real a qual a expressão em questão se refere. Por exemplo a expressão "*two months before*", é necessária a análise de outras expressões temporais presentes ou não no documento.

Existem também casos onde um determinado conceito não possui nenhuma expressão temporal associada a ele presente na mesma frase. Neste caso, é possível a recuperação desta informação em fonte externa ao texto. Para este propósito, utilizamos as informações presentes na base de conhecimento *DBPedia*[1].

Com base nas informações expostas, se fez necessário criar evoluções no sistema RISO (Recuperação da Informação Semântica de Objetos Textuais), possibilitando uma recuperação mais acurada. Os requisitos funcionais para o trabalho desta dissertação são:

a) Estender a função NORM do RISO-TT (Temporal Tagger), para que este fosse capaz de normalizar novos formatos de expressões temporais.
b) Desenvolver uma nova funcionalidade responsável extrair conceitos e seus contextos temporais correspondentes e inclui-las no Mapa de Tópicos (Topic Map) previamente criado pelo RISO-IC (Indexação Conceitual).

Foram encontradas algumas ferramentas com propósito semelhante: *Temporal Fact and Event Extraction from Free Text* e *TM-Gen*. Apesar de contatarmos os seus responsáveis, nenhuma delas pode ser disponibilizada. Entretanto, com base nos trabalhos publicados sobre as ferramentas semelhantes e realizando-se uma análise comparativa entre os resultados obtido entre a ferramenta que implementa a abordagem proposta neste trabalho de dissertação e as ferramentas concorrentes, pode-se afirmar que o RISO-GCT obteve desempenho satisfatório, se saindo melhor que estas.

1.1 Objetivos

O objetivo principal desta dissertação é criar uma abordagem capaz de permitir a indexação de documentos de acordo com os conceitos presentes em documentos textuais e seus

[1] Projeto responsável por extrair informação estruturada da enciclopédia digital Wikipedia e tornar estas informações disponíveis na Web (DBPEDIA, 2015).

respectivos contextos temporais, através da determinação de relacionamentos entre conceitos e contextos temporais e inclusão destas informações no Topic Map[2] do sistema.

1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

1) Desenvolvimento de uma extensão do RISO-TT[3] (Temporal Tagger), para o mapeamento de padrões temporais ainda não considerados.
2) Desenvolvimento do RISO-GCT[4] (Geração de Contextos Temporais), responsável por determinar um possível contexto temporal dos conceitos presentes no texto
3) Atualização do Topic Map gerado pelo RISO-IC[5] com os relacionamentos encontrados.
4) Executar o protótipo e avaliar os resultados.

1.3 Relevância

A relevância deste trabalho está relacionada ao fato de que, com sua incorporação ao Projeto RISO, é possível incrementar a qualidade da recuperação da informação em textos levando em consideração o contexto temporal dos conceitos e, além disso, integrar fontes textuais em Bancos de Dados Temporais.

Durante a revisão da literatura realizada durante a escrita deste trabalho de dissertação não foram encontrados métodos ou ferramentas de construção automática de Topic Maps que envolvesse a determinação do contexto temporal de conceitos extraído tanto do próprio documento, quanto de fontes externas.

Além disto, as ferramentas que mais se aproximaram da abordagem proposta neste trabalho, *Temporal Fact and Event Extraction from Free Text* e *TM-Gen* que serão detalhadas são ferramentas proprietárias e de disponibilidade restrita.

O processo de atribuição de contextos temporais aos conceitos presentes no documento poderá utilizar informações temporais presentes no próprio texto ou em dados obtidos na *DBPedia,* que é uma base de conhecimento constantemente atualizada e que possui informações estruturadas referentes aos artigos presentes na Wikipedia[6].

[2] A definição de Topic Map encontra-se na seção 2.5 deste documento.
[3] Mais detalhes sobre o RISO-TT encontram-se na seção 5.3.2.1 deste documento.
[4] Mais detalhes sobre o RISO-GCT encontram-se na seção 6.2 deste documento.
[5] Mais detalhes sobre o RISO-IC encontram-se na seção 5.3.1 deste documento.
[6] É um projeto de enciclopédia multilíngue de licença livre, baseado na Web e escrito de maneira colaborativa (WIKIPEDIA, 2015).

O acréscimo do contexto temporal aos conceitos possibilitará uma recuperação mais precisa de documentos. Uma pesquisa poderá se concentrar em recuperar documentos que descrevem fatos de uma certa época específica. Assim, todas as informações que se sobrepõe parcial ou totalmente com a época desejada serão recuperadas.

1.4 Estrutura da Dissertação

No Capítulo 2 é apresentado o referencial teórico sobre os principais assuntos abordados neste trabalho, como Extração Automática de Termos (EAT), Extrações de Termos Temporais e Extração de Conceitos presentes em documentos, Lógica Temporal de ALLEN (1984) e Topic Maps.

No Capítulo 3 é apresentada a metodologia utilizada nesta pesquisa.

No Capítulo 4 são apresentados os trabalhos relacionados à área de recuperação de informações temporais e geração automática de Topic Maps.

No Capítulo 5 é apresentada uma visão geral do projeto RISO.

No Capítulo 6 é apresentado a arquitetura e a abordagem criada para a extração de contextos temporais relacionados aos conceitos presentes em documentos.

No Capítulo 7 é apresentada a validação do trabalho em comparação com ferramentas semelhantes.

Por fim, as conclusões obtidas com os resultados dos experimentos, bem como os trabalhos futuros, são apresentados no Capítulo 8.

Capítulo 2 - Fundamentação Teórica

Este capítulo apresenta o resultado do estudo realizado na área na qual este trabalho está inserido. Desta forma, são apresentados uma introdução aos temas Extração Automática de Termos (seção 2.1), Extração de Termos Temporais (seção 2.2), Contexto Temporal de Conceitos (seção 2.3), Lógica Temporal (seção 2.4) e Topic Maps (seção 2.5). Estes temas possibilitam um melhor entendimento da área pesquisada e motivam o desenvolvimento a pesquisa.

2.1 Extração Automática de Termos (EAT)

Extração de Informação (EI) é o processo de identificar automaticamente tipos específicos de entidades, conceitos, relações ou eventos em textos livres e armazenar esta informação de uma forma estruturada (Yangarber et. Al., 2000). Sistemas de Extração de Informação são úteis na execução de diversas tarefas como, por exemplo, identificação e classificação de nomes próprios, extração de eventos e relações típicas de um domínio de conhecimento, extração de multi-palavras e extração de termos (ZAVAGLIA et al., 2005). Se a extração é focada em termos em geral, fala-se em Extração Automática de Termos (EAT).

De acordo com LAGUNA (2014), ainda não existe um consenso sobre a definição correta de "*termo*". A autora cita definições como as de BARROS (2004) que classifica "*termos*" como sendo uma unidade léxica com um conteúdo específico dentro de um domínio específico e define *"unidade léxica*" como sendo "*símbolo linguístico, composto de expressão e de conteúdo, que pertence a uma das grandes classes gramaticais (substantivo, verbo, adjetivo ou advérbio)*". LAGUNA (2014) também cita a definição dada pela norma internacional ISO 1087, segundo a qual "*termos*" são determinações, por meio de uma unidade linguística, de um conceito definido em uma língua de especialidade. De acordo com o dicionário online DICIO[7] um termo significa "*uma expressão própria de uma área de conhecimento*". TELINE (2004) considera que "*termos*" são, ao mesmo tempo, unidades lexicais (vocábulos) e unidades de conhecimento, consequentemente objetos para estudos científicos.

O gargalo da EAT é a avaliação da sua eficácia, pois exige a opinião de especialistas, sendo esse processo caro e demorado. Por outro lado, contar com recursos como glossários ou dicionários, isto é, com listas de referências, também traz seus riscos, uma vez que tais recursos são incompletos, dada a constante produção de novos termos (ZAVAGLIA et al., 2005). A EAT é

[7] DICIO - http://www.dicio.com.br/termo/

baseada, tradicionalmente, em três abordagens: estatística, linguística e híbrida, que serão detalhadas nas seções 2.1.1, 2.1.2 e 2.1.3.

2.1.1 Abordagem Estatística

A abordagem estatística utiliza o conhecimento obtido por meio da aplicação de medidas estatísticas. Para tanto, o *corpus*[8] passa por um pré-processamento, que comumente envolve a identificação de *tokens*[9], a remoção de *stopwords*[10] e a representação dos textos em uma tabela. Nessa tabela, cada linha representa um documento (D_1) e cada coluna representa um n-grama do documento (N_1), sendo que a célula D_1N_1 pode ser preenchida com alguma medida, por exemplo, pela frequência absoluta do n-grama N_1 no documento D_1. A essa representação dos textos é dada a denominação *bag-of-words* (BOW). Nesse sentido, a aplicação de medidas estatísticas por meio de uma BOW ignora qualquer informação estrutural sobre as sentenças dos textos, como a ordem em que os n-gramas ocorrem. A partir dos valores obtidos pela medida escolhida, os candidatos a termos são ranqueados. Nesse ranqueamento, considera-se que os candidatos com pontuação mais elevada têm maior probabilidade de serem termos representativos do corpus (Pazienza et. al., 2005).

As medidas comumente utilizadas no desenvolvimento de extratores semi-automáticos[11] segundo a abordagem estatística, são independentes de língua. A independência de língua é uma característica vantajosa do ponto de vista computacional, pois a aplicação das medidas não requer a especificação (manual ou automática) de qualquer tipo de conhecimento sobre a língua dos textos em processamento, o que torna a extração mais simples e rápida. Em comparação à extração humana, a independência de língua não reflete o processo realizado pelos especialistas do domínio, já que estes utilizam o conhecimento linguístico para identificar termos. Um tipo de conhecimento linguístico é o morfológico, utilizado, por exemplo, para identificar termos compostos por morfemas greco-latinos (TELINE, 2004).

O principal problema dos extratores de termos desenvolvidos segundo a abordagem estatística é o fato deles não identificarem e selecionarem termos reais com baixa frequência em um texto ou *corpus*. Um exemplo para este problema é quando a medida escolhida utiliza a frequência de cada termo no *corpus* como base para a extração de termos e, por isso, um

[8] Conjunto de documentos textuais sobre determinado tema (DICIONÁRIO INFORMAL, 2009).
[9] Conjunto de caracteres que possuem um significado coletivo e que pode ser manipulado por um analisador sintático (NUNES *et. al*., 2012).
[10] http://www.agenciamestre.com/seo/stop-words-como-funcionam-palavras-de-parada/
[11] Nesta variação da abordagem estatística, o usuário define as regras de mapeamento e valida as informações extraídas

determinado termo (p. ex., "*polinização*", do domínio de ecologia) não é extraído por ele ter sido citado poucas vezes no texto (TELINE, 2004).

As medidas estatísticas buscam identificar duas propriedades terminológicas: *unithood*e *termhood*. As medidas estatísticas que expressam *unithood* revelam a força ou estabilidade de expressões complexas (isto é, formadas por dois ou mais elementos separados por espaços em branco). As medidas que expressam *termhood* revelam, por sua vez, o grau de relação entre uma expressão linguística e um domínio do conhecimento. Em outras palavras, *termhood* expressa o quanto uma expressão linguística (seja ela simples, como "*polaridade*", ou complexa, como "*molécula orgânica*" e "*molécula de água*") está relacionada a um domínio (KAGEURA et. al., 1996).

2.1.2 Abordagem Linguística

Neste tipo de abordagem os candidatos a termos são identificados e extraídos de um conjunto de textos sobre determinado tema com base em suas características ou propriedades linguísticas, as quais podem ser de diferentes tipos ou níveis (LAGUNA, 2014). Essas informações linguísticas dizem respeito a: informações lexicográficas – dicionários de termos e lista de palavras auxiliares ("*stopwords*"); informações morfológicas – padrões de estrutura interna da palavra; informações morfossintáticas – categorias morfossintáticas e funções sintáticas; informações semânticas – classificações semânticas; informações pragmáticas – representações tipográficas e informações de disposição do termo no texto (TELINE, 2004).

A abordagem linguística baseia-se em uma hierarquia de diferentes tipos de conhecimentos linguísticos, de acordo com a complexidade da modelagem e o tratamento computacional do conhecimento.

O conhecimento de nível morfológico se refere basicamente à morfemas greco-latinos ou típicos do domínio que indicam a ocorrência de um possível termo (Almeida et. al., 2008). LAGUNA (2014) cita como exemplo os termos utilizados na área da medicina e de domínios correlatos, por exemplo, que geralmente buscam morfemas de origem greco-latina, como *artr(i/o)* (do grego *árthron*) em "*atrite*" (Vivaldi et. Al., 2007).

A extração com base em conhecimento sintático utiliza a estrutura sintática das sentenças para identificar candidatos a termos. Por exemplo na sentença "*Padrões são unidades de informação que se repetem*", "*padrões*", é identificada como núcleo de um sintagma nominal e, por isso, é selecionada como um candidato a termo (LAGUNA, 2014).

A extração com base em conhecimento semântico utiliza o sentido das palavras que foram reagrupadas pelo analisador sintático para identificar candidatos a termos e os conceitos subjacentes.

2.1.3 Abordagem Híbrida

Sistemas com abordagem híbrida utilizam o conhecimento estatístico em conjunto com o conhecimento linguístico (TELINE et al., 2004), o que torna o sistema mais eficiente, uma vez que ele coleciona resultados. Existem dois tipos de métodos híbridos: aqueles que utilizam o conhecimento estatístico primeiro e o linguístico posteriormente, e aqueles que utilizam o conhecimento estatístico apenas como complemento para o complemento linguístico. No primeiro caso, os problemas de falsos negativos que ocorrem nos sistemas que utilizam a abordagem estatística e para o segundo caso, os resultados são melhores em razão de a estatística auxiliar no momento do processo de detecção.

2.2 Extração de Termos Temporais

Timeline Recognition ou Temporal Information Extraction é o reconhecimento de expressões temporais presente em textos. Expressões estas que podem ser absolutas (E.g., "*17 de junho de 1999 às 12:00h*", "*de 01 de janeiro de 2013 à 05 de outubro de 2013*") ou expressões relativas (E.g., "*ontem*", "*dois dias antes*") ou podem representar a frequência na qual um determinado evento ocorre (E.g., "*mensalmente*", "*por semana*"). A extração automática de expressões temporais é útil para construir uma linha do tempo dos fatos e das entidades descritas em textos, realizando assim a sumarização de informações presentes no texto (NEVES et al., 2013).

Uma expressão temporal pode ser qualquer expressão que denomina informações do tipo: datas de calendário, horários do dia, períodos de tempo, durações, intervalos.

Expressões temporais possuem uma vasta gama de relações gramaticais geralmente sinalizadas por uma ou mais palavras de tempo, os chamados Gatilhos Lexicais (MARSIC, 2011). São eles:

- Substantivos: século, ano, mês, dia, fim de semana, minutos, futuro, passado;
- Adjetivos: próximo, medieval, antigo, mensal;
- Advérbios: Atualmente, então, semanal, hoje, ontem, amanhã, esta noite.
- Padrões de Tempo: 09h00min, 26/12/2002.
- Números: 4hrs.

Para o reconhecimento de expressões temporais em textos, SANTOS (2013) utiliza **padrões** e **regras**.

Padrões são conjuntos de termos agrupados semanticamente de maneira que possam determinar o valor de uma expressão temporal. Estes termos gramaticais podem ser preposições,

advérbios, estações do ano, datas, horas e expressões regulares. A Tabela 1 ilustra exemplo de padrões definidos no RISO-TT.

Os padrões por sua vez, podem ser sequencialmente agrupados de acordo com **regras** (SANTOS 2013). Uma regra considera a posição dos termos que formam uma expressão temporal. Por exemplo, a regra Dia Mês Ano é diferente da regra Mês Dia Ano. Desta maneira, é possível reconhecer expressões temporais presentes em textos. A Tabela 2 lista exemplos de regras definidas no RISO-TT compostas por padrões agrupados e que formam expressões temporais.

Categoria de Padrão	**Exemplo**
Dia	1st, 2nd, 3rd, 4th, ..., 31th... 1,2,3,4...31 One, two, three,..., thirty-one
Ano	d{4} (expressão regular de 4 dígitos)
Estação do Ano	Spring, Summer, Fall, Autumn, Winter
Estrutura Pré-Temporal	In the beginning, In the start of, in early, entre outros.
Advérbios	Early, later, late
Unidade Temporal	Day, month, year, week
Estrutura Básica Temporal	September,7, 1822

Tabela 1 – Exemplos de padrões utilizados no RISO-TT.

A Tabela 2 ilustra exemplos de regras definidos pelo RISO-TT através do agrupamento dos padrões.

Regra	Exemplo
Preposição + Estrutura Básica Temporal	On September 1, 1939 Onde: • On: pertence ao padrão de Preposição • September 1, 1939: Reúne os padrões de mês, dia e ano, formando a Estrutura Básica Temporal
Estrutura Pré-temporal + Unidade Temporal + Advérbio	A few weeks later Onde: • A few: pertence ao padrão referente à Estrutura Pré-temporal • Weeks: pertence ao padrão referente à Unidade Temporal • Later: pertence ao padrão referente à Advérbio

Tabela 2 – Exemplos de Regras definidas pelo RISO-TT

Uma etapa muito importante do processo de extração de expressões temporais, que é a ação de percorrer os nós das entidades temporais e convertê-los para o formato padrão, levando em consideração os seguintes casos especiais (HAGÈRE et al., 2010):

- Expressões que possuem elementos em numeração romana. (E.g.,: "*XX*", "*XV*");
- Expressões que envolvem unidade de tempo não representada nos padrões normais. (E.g.,: "*semana*", "*trimestre*" e "*quinzena*");
- Expressões que possuem frações de tempo. (E.g.,: "*dois dias e meio*");
- Expressões que indiquem informalidade para indicar hora. (E.g.,: "*10 pras 6hs*");
- Expressões não numéricas. (E.g.,: "*duas horas da tarde*");
- Expressões que usam advérbios temporais. (E.g.,: "*Ontem*", "*Amanhã*");
- Expressões com marcação com valores relativos. (E.g.,: "*No próximo mês*").

2.3 Contexto Temporal de Conceitos

O processo de determinação do contexto temporal de conceitos em textos consiste na delimitação temporal de objetos e eventos, tais como:

- O período ao qual uma pessoa morou em uma determinada residência, ou foi empregado de uma determinada empresa;
- O período que durou uma guerra entre dois países;
- O tempo preciso no qual um avião aterrissou;

Além disso, informações temporais estão presentes em diversos textos, publicados diariamente (ROTH, 2012) e por este motivo são bastante usados em aplicações de Processamento de Linguagem Natural (NLP), como:

- Inferência Textual;
- Sumarização de documentos;
- *Tracking* de eventos temporais;
- População de bases de conhecimento temporais.

Segundo ROTH (2012), o processo de extração de informações temporais gera eventualidades, que podem ser pontuais (E.g., "*win*", "*reach the summit*", "*die*", "*leave, sneeze*") ou acontecimentos, ou seja, que possuem um tempo de início e fim (E.g., "*Run a mile*", "*play a sonata*", "*drink a glass of beer*"). Esses grupos podem ser mais detalhados:

A) Eventualidades Pontuais:
   - Limites (intrinsecamente instantâneo). E.g., "*Win*", "*reach the summit*", "*die*", "*leave*".
   - Momentos (não intrinsecamente instantâneo). E.g., "*Sneeze*", "*flash*", "*hop*", "*kick*".

B) Acontecimentos
   - Estados (não dinâmicos)
     - Habitual (Estados Não-Autônomos). E.g., "*Use to drink*", "*be a drinker*", "*besilly*".
     - Episódicos (Estados Autônomos): E.g., "*Be drunk*", "*beat summit*", "*be drinking*".
   - Ocorrências (dinâmicos)
     - Processos (não possuem início e fim definidos). E.g., "*Run*", "*drink beer*", "*play the piano*".
     - Eventos (possuem início e fim definidos)
       - Episódios (Não indicam o resultado da ação). E.g., "*Run a mile*", "*play the sonata*".
       - Mudanças (Indica o resultado de uma ação). E.g., "*Run to the summit*", "*drink a glass of beer*".

Pode-se acrescentar aos subgrupos propostos por ROTH (2012) os **Objetos Temporais** que são substantivos que possuem um intervalo de existência atrelado a ele. Por exemplo: Napoleon, Bento XVI, Yugoslavia.

Dentre as aplicações da Extração de Contexto Temporal de Conceitos, está a ordenação de eventos com base no tempo no qual ocorreram, como na ordenação das sentenças abaixo, ilustrado pela Figura 1:

(1) John entered the room at 5:00 pm.

(2) It was pitch black

(3) It had three days since he'd slept.

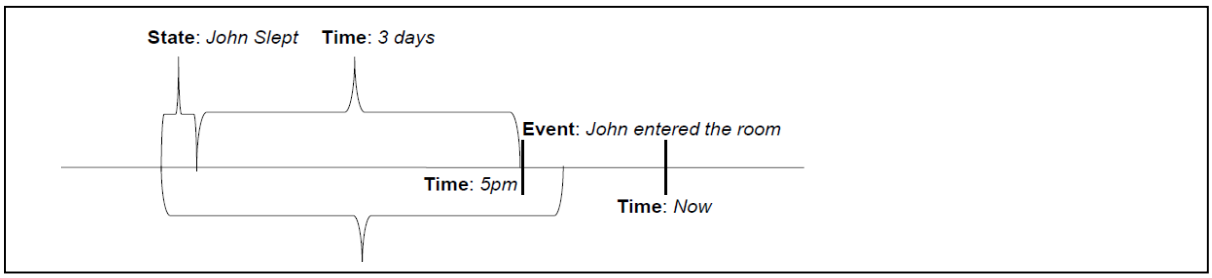


Figura 1 – Ordenação de eventos em discursos

Na Figura 2 é detalhado como as eventualidades estão subdivididas e a relação entre as subdivisões, onde:

- Um **limite** inicia e finaliza um **acontecimento**.
- Um **limite** também culmina um **evento**.
- Um **momento** é uma parte de um **episódio**.
- Um **estado** é o resultado de uma **mudança**.
- Um **estado habitual** é realizado por uma classe de **ocorrências**.
- **Processos** são compostos por vários **eventos**.

(ROTH, 2012)

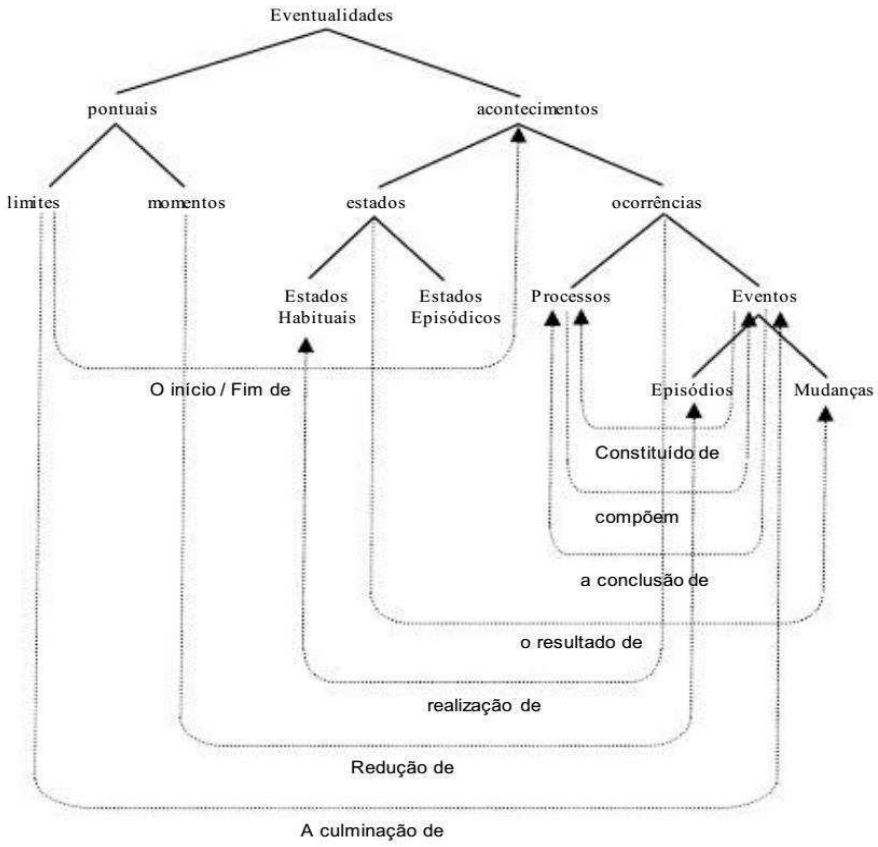


Figura 2 – Subdivisão dos tipos de Eventualidades (Dölling, 2014)

A Temporalização de Conceitos também é útil no retorno de resultados em consultas temporalmente relevantes, bem como a representação de eventualidades e relações entre eles e as informações temporais, respondendo perguntas como "*Who **won** Turing Award in **1966***", "*Who **died** during The Clinton **Administration***", "*On **what day** was **Dec 25th** in **2004***", entre outras (Pustejovsky et. al., 2005).

2.4 Lógica Temporal

A lógica temporal proposta por James F. Allen é uma lógica de 1ª ordem que permite raciocinar sobre a associação entre intervalos de tempo, propriedade e termos que correspondem aos objetos do domínio. Para esta representação é utilizado um pequeno número de predicados, dentre os quais o mais importante é o predicado HOLDS, que identifica que uma determinada propriedade é válida durante um intervalo. Assim HOLDS (*p*, *t*) retorna verdadeiro apenas se a propriedade *p* se mantem durante o intervalo *t* (ALLEN, 1984).

De acordo com ALLEN (1984), há um conjunto de relações básicas mutuamente exclusivas entre intervalos temporais. Cada uma dessas relações é representada por um predicado lógico. Estas relações são ilustradas na Tabela 3.

Relação	Símbolo	Inverso	Exemplo
X before Y	<	>	XXX YYY
X meets Y	m	Mi	XXXXYYYYYY
Y overlaps Y	o	Oi	XXXXX YYYYYYY
X during Y	d	Di	XXX YYYYYYYY
X starts Y	s	Si	XXXX YYYYYYYY
X finishes Y	f	Fi	XXXX YYYYYYYYYY
X equals Y	=	=	XXXX YYYY

Tabela 3 – Relações de temporalidade propostas por (Allen, 1984)

Onde:

- **DURING (t1, t2):** Indica que o intervalo de tempo t1 está totalmente contido dentro do intervalo t2.
- **STARTS (t1, t2):** Indica que o intervalo de tempo t1 compartilha o mesmo início do intervalo t2, mas finaliza antes de t2.
- **FINISHES (t1, t2):** Indica que o intervalo de tempo t1 compartilha o mesmo fim do intervalo t2, mas inicia antes.
- **BEFORE (t1, t2):** Indica que o intervalo de tempo t1 inicia antes de t2 e eles não se sobrepõem.

- **OVERLAP (t1, t2):** Intervalo de tempo t1 inicia antes de t2 e eles se sobrepõem e um determinado momento.
- **MEETS (t1, t2):** Intervalo de tempo t1 é antes do intervalo t2 e não existe nenhum intervalo de tempo entre t1 e t2.
- **EQUAL (t1, t2):** t1 e t2 são exatamente o mesmo intervalo.

Para cada um desses predicados existe um conjunto de axiomas que definem seu comportamento. Dado qualquer intervalo I, um outro intervalo J estará relacionado a ele por uma dessas relações. Existem axiomas que garantem que cada relação é mutuamente exclusiva dos demais e vários axiomas que descrevem possíveis relações entre eles, conforme os exemplos abaixo:

$$\text{BEFORE}(t1, t2) \,\&\, \text{BEFORE}(t2, t3) \Rightarrow \text{BEFORE}(t1, t3) \qquad \text{(T. 1)}$$

$$\text{MEETS}(t1, t2) \,\&\, \text{DURING}(t2, t3) \Rightarrow \text{OVERLAPS}(t1, t3) \vee \text{DURING}(t1, t3) \vee \text{MEETS}(t1, t3) \qquad \text{(T. 2)}$$

O axioma T.1 diz respeito à transitividade do predicado BEFORE, ou seja, se o intervalo de tempo t1 ocorre antes do intervalo t2 e o intervalo t2 ocorre antes de t3, logo, o intervalo t1 ocorre antes de t3.

O axioma T.2. diz que, se o intervalo t1 ocorre antes de t2 e não há nenhum intervalo de tempo entre t1 e t2, e t2 ocorre durante o intervalo t3, então, ou o intervalo t.1 sobrepõe o intervalo t.3, ou o intervalo t.1 ocorre durante o intervalo t.3 ou o intervalo t.1 ocorre antes de t.3 e não há nenhum intervalo de tempo entre t.1 e t.3.

As regras anteriores serão uteis para definir um predicado que sumariza as relações nas quais um intervalo está totalmente contido em outro. Como é ilustrado a seguir:

$$\text{IN}(t1, t2) \Leftrightarrow \text{DURING}(t1, t2) \vee \text{STARTS}(t1, t2) \vee \text{FINISHES}(t1, t2) \qquad \text{(T.3)}$$

Usando este predicado, é introduzida a primeira propriedade crucial do predicado HOLDS: Se uma propriedade *p* se mantem durante um intervalo T, ele se mantém durante todos os subintervalos de T (ALLEN, 1984).

$$\text{HOLDS}(p, T) \Leftrightarrow (\forall\, t \text{ IN } (t,T) \Rightarrow \text{HOLDS}(p, t)) \qquad \text{(H.1)}$$

Este predicado indica que uma determinada propriedade mantida em um intervalo T, está mantida nos diferentes sub-intervalos pertencentes a T (ALLEN, 1984).

Outro predicado proposto por ALLEN (1984) diz respeito a especificação de eventos. Um evento engloba um espaço de tempo fixo, ou seja, nenhum subintervalo deste tempo representa o evento todo. Por exemplo, o evento "*I walked from home to school*" ocorreu (OCCUR) numa manhã em um intervalo de tempo. Desta forma, o predicado OCCUR recebe como parâmetros um evento

e e um intervalo de tempo *T* e retorna verdadeiro apenas se o evento *e* ocorrer neste intervalo *T* e se não existe um subintervalo de *T* sobre o qual o evento tenha ocorrido. Assim, para qualquer evento *e*, e tempos *T* e *t'* existe o seguinte axioma:

$$\text{OCCUR}(e,t) \;\&\; \text{IN}(t',t) \Rightarrow \sim\text{OCCUR}(e,t') \qquad (0.1)$$

Processos são ocorrências mais fracas do que eventos, logo o axioma O.1 não se aplica. Um processo ocorre por algum tempo ao longo de um intervalo de tempo T, mas não no tempo todo. Por exemplo, "*I walked from home to school in the morning*" não significa que o sujeito desta frase passou a manhã inteira indo à escola, mas que caminhou em algum horário da manhã para escola. Ou seja, se um processo está ocorrendo em um intervalo T, deve também estar ocorrendo em pelo menos um subintervalo de T. Para formalizar esta condição, foi criado o predicado OCURRING, para um processo *p* e um tempo *t*:

$$\text{OCCURING}(p,t) \Rightarrow \exists t'.\text{IN}(t',t) \;\&\; \text{OCCURRING}(p,t') \qquad (0.2)$$

O predicado O.2 indica que, se um dado um processo *p* ocorre em um tempo *t*, então existe uma porção *t'* de t na qual o processo *p* também ocorre.

2.5 Topic Maps

Topic Maps foram desenvolvidos no final da década de 1990 como uma maneira de representar estruturas de índices no final de livros, de modo que vários índices de documentos diferentes pudessem ser mesclados. Porém, desenvolvedores de sistemas perceberam que com uma pequena generalização adicional, poderiam criar um meta-modelo com aplicação potencialmente muito mais ampla. O resultado deste trabalho foi publicado em 1999 como ISO/IEC 13250 – Topic Maps (Mapa de Tópicos, 2016).

Topic Maps são estruturas que utilizam uma sintaxe baseada em XML, chamada XML Topic Maps (XTM), e que são utilizadas para representar o conhecimento compostas por 3 elementos: tópicos, associações e ocorrências. Um tópico consiste de um nó, que representa um conceito ou elemento que, por sua vez, relaciona-se com outros tópicos por meio de associações que são relações semânticas entre conceitos (E.g., "*é um*", "*é parte de*"). As características de um tópico são contextualizadas por meio do elemento escopo, que consiste, em linhas gerais, de uma descrição com o objetivo de evitar ambiguidades e tornar o conceito único (ARAÚJO JÚNIOR, 2013).

Topic Maps baseiam-se na representação de **conceitos** (**tópicos**) através de **relações** (**associações**). A partir desse mapeamento de conceitos e associações são construídos caminhos que apontam para **ocorrências** que são informações relevantes para estes conceitos e que determinam quais documentos são indexados pelo respectivo tópico (BAIAO, 2008).

A Figura 3 mostra estes três conceitos de um Topic Map, onde a primeira partição, "*Espaço de Tópicos*", ilustra relações entre tópicos e a segunda partição, "*Espaço de Recursos*" ilustra associações entre recursos e tópicos.

Ainda na Figura 3, os tópicos "*Londres*", "*Reino Unido*" e "*Europa*" estão relacionados entre si. Considerando que as relações são do tipo *pertence à*, Londres *pertence ao* Reino Unido que por sua vez *pertence à* Europa. Cada um desses termos possui informações específicas que os caracterizam. Por exemplo, a cidade de Londres possui informações como data de fundação, quantidade de habitantes, entre outras.

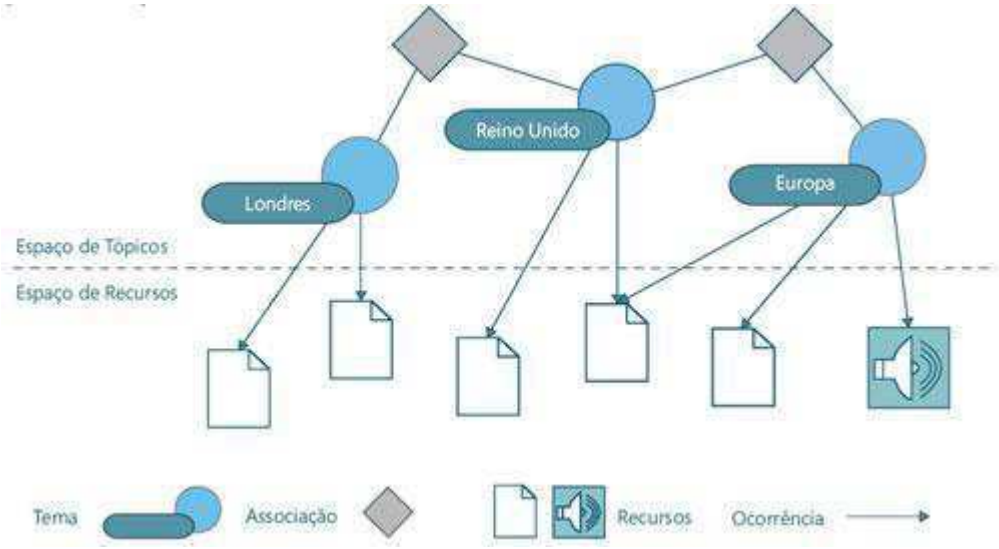


Figura 3 – Ilustração representado os 3 conceitos básicos dos mapas de tópicos (Mapa de Tópicos, 2016).

Um Topic Map pode atuar como uma visão geral de alto nível do conhecimento contido em um conjunto de recursos. Desse modo, especialistas de uma determinada área podem modelar seu conhecimento de forma estruturada, permitindo que não-especialistas aprendam os conceitos básicos e seus relacionamentos e aprofundar através das informações específicas de cada tópico. Topic Maps podem ser utilizados para agrupar documentos semelhantes, uma vez que eles podem capturar informações relacionadas a um determinado tópico de um documento de forma estruturada e manipular tópicos semânticos e suas relações.

Podem ser compostos pelos seguintes elementos:

- **Tópicos.** E.g.,: <*Valência*>, <*Espanha*>;
- **Tipo do Tópico**, indicam a classificação do tópico Exemplos: <*cidade*>, <*país*>;
- **Associação**, responsável por modelar as relações entre os tópicos. Exemplo: <*Valência*> <*está_localizada_em*> <*Espanha*>;

- **Tipo da Associação.** Indica o tipo da relação que é estabelecida entre os tópicos. E.g.,: *<está_localizada_em>*;
- **Papel da Associação.** São padrões de associações. E.g.,: o papel *<cidade><está_localizada_em><país>* que cidades podem possuir a associação "*está localizada em*" em um determinado país.
- **Escopo.** Indica o domínio relacionado ao tópico**.** E.g.,: *<Geography>*
- **Ocorrência.** Informações relevantes a respeito de um tópico. Exemplo: www.turismovalencia.es representa página para assuntos de turismo na cidade de Valência.
- **Tipo da Ocorrência.** Exemplo: WebPage, representa o tipo da informação relacionada ao tópico.
- **Assunto Público.** Representa o local no qual as informações estão disponíveis. Exemplo: http://www.valencia.es/ é o local onde a informação da Home Page de Turismo da cidade de Valência está localizada.
  (TOPIC MAPS, 2015)

No Capítulo 3 serão listados os trabalhos relacionados à abordagem proposta neste trabalho de dissertação.

Capítulo 3 - Metodologia

Neste capítulo será apresentada a metodologia para o desenvolvimento do RISO-GCT e a verificação e validação do protótipo criado. Após introdução do roteiro do estudo experimental, apresentamos a metodologia de trabalho para a realização dos testes de comparação do RISO-GCT com as outras ferramentas. Mostramos a seleção de variáveis e o design do experimento. A aplicação desta metodologia será mostrada no Capítulo 7 sobre Experimentos e Validações.

3.1 Roteiro do Estudo Experimental

Os passos a serem realizados durante o estudo experimental dos resultados obtidos a partir da execução do RISO-GCT são os seguintes:

1) **Tema do Experimento:** Extração de Conceitos presentes em textos e Contextos Temporais correspondentes.
2) **Área de Estudo:** Recuperação de Informação e Indexação de Documentos com base em Informações Temporais.
3) **O problema:** Relacionar corretamente os conceitos e expressões temporais presentes em um documento.
4) **Importância do Problema:** O Topic Map contendo conceitos relacionados às informações temporais podem ser utilizados por indexadores de documentos e motores de busca.
5) **Objetivos**:

   Pergunta I: Os Topic Maps gerados pelo RISO-IC foram enriquecidos corretamente com as informações temporais relacionadas aos conceitos?

   a. **Hipótese Nula:** Não. O Topic Map gerado pelo RISO-IS se manteve sem as informações de contextos temporais.
   b. **Hipótese Alternativa:** Sim. Após a execução do RISO-GCT o Topic Map gerado pelo RISO-IS foi enriquecido com os contextos temporais.

   Pergunta II: Todas as expressões temporais que foram extraídas dos textos e relacionadas aos conceitos foram normalizadas corretamente?

   a. **Hipótese Nula:** Não. A função NORM do RISO-TT não cobriu todos os formatos de expressões temporais presentes nos textos, de maneira que estes não puderam ser normalizados.

b. **Hipótese Alternativa:** Sim. A função NORM do RISO-TT cobriu todos os formatos temporais presentes nos textos, e, consequentemente, todas as datas foram normalizadas corretamente.

3.2 Metodologia de Trabalho

Primeiramente deverá ser definida uma abordagem responsável por realizar a extração de conceitos presentes em textos e possíveis expressões temporais relacionadas. Estas expressões temporais podem ser buscadas no próprio documento ou na base de dados da DBPedia.

O segundo passo será a escolha da linguagem de programação e a base de dados para se realizar a implementação desta abordagem.

O terceiro passo será definir o Corpus a ser utilizado na validação da abordagem proposta.

O quarto passo será pesquisar por ferramentas que se proponham a realizar tarefas semelhantes ao RISO-GCT para que fosse possível comparar as saídas produzidas por estas ferramentas e as saídas produzidas pelo RISO-GCT.

Para cada uma das ferramentas selecionadas para comparação com o RISO-GCT deverá ser realizado o seguinte procedimento:

a) Busca por artigos científicos e informações da especificação do desenvolvimento;

b) Estudo de funcionalidades das ferramentas;

O quinto passo será, a partir do Corpus selecionado, realizar a criação de um gabarito cujas respostas para serem comparadas com a saída produzida pelo RISO-GCT. Em seguida foi a executado o RISO-GCT recebendo como entrada a junção das saídas produzidas pelo RISO-TT e RISO-IS.

No sexto passo será feita a comparação dos resultados obtidos com a execução do RISO-GCT com os resultados presentes no gabarito criado anteriormente para posterior tabulação dos resultados e cálculo da precisão, cobertura e *f-measure*.

O RISO-GCT disponibiliza, ao fim da execução, o resultado da extração dos conceitos presentes no texto e os contextos temporais relacionados a estes conceitos, no qual cada relacionamento criado é disponibilizado em duas versões:

- Conceito + Expressão Temporal Normalizada
- Conceito + Expressão Temporal Não Normalizada

Optou-se por considerar estes formatos devido ao fato de a função NORM do RISO-TT não possuir cobertura para uma parte considerável das datas presentes nos textos. Como o foco principal deste trabalho foi a avaliação da RISO-GCT, verificamos que o mesmo poderia ter sua eficácia penalizada caso fossem consideradas apenas as datas normalizadas.

Com relação à análise da capacidade de relacionar conceitos e expressões temporais presentes em um documento, para se chegar à resposta da pergunta I serão realizados 8 experimentos, cada um com um texto diferente, onde foi investigado se as entidades foram relacionadas com as expressões temporais corretas. Para isto, os resultados obtidos pela execução do RISO-GCT serão comparados com o resultado final da mesma atividade realizada de maneira manual. De maneira similar, para se chegar à resposta da pergunta II a partir dos mesmos experimentos realizados para solução da pergunta I, serão verificados se as expressões temporais extraídas e relacionadas com conceitos presentes no texto foram normalizadas corretamente. Para isto, serão comparados os resultados obtidos pela execução automática com o resultado obtido através da execução manual desta tarefa.

3.3 Seleção de Variáveis

Para este estudo foram utilizadas as seguintes variáveis:

a) Variáveis Independentes:
   - **Documentos sem Marcação (D):**
     Documentos originais a serem processados pelo RISO.
   - **Dados provenientes da DBPedia (DBP):**
     Informações Temporais de diversos conceitos presentes na base de dados da DBPedia.

b) Variáveis Dependentes:
   - **Documentos contendo marcações de classes gramaticais (D-ES):**
     Resultantes da execução do RISO-VTD (BISPO, 2013) e que contém o texto com a indicação da classificação gramatical de cada sintagma presente no texto original processado.
   - **Documentos contendo marcações de temporais (D-TT):**
     Resultantes da execução do RISO-TT (SANTOS, 2013) e que contém o texto com a indicação dos sintagmas que formam expressões temporais.
   - **Documentos contendo a extração de conceitos presentes no texto (D-ENT):**
     Resultante da execução do RISO-VTD (BISPO, 2013) e que contém a indicação dos sintagmas que representam conceitos.
   - **Topic Map gerado pelo RISO-IC (TM):**
     Arquivos que compões o Topic Map e que possuem as informações de conceitos presentes nos documentos e os dados resultantes do enriquecimento semântico destes conceitos.

3.4 Design do Experimento

O design do experimento é formado por Geradores de Contextos Temporais de Conceitos e por Documentos.

3.4.1 Geradores de Contextos Temporais

a) RISO-GCT: Ferramenta desenvolvida para o Projeto RISO.
b) Temporal Fact and Event Extraction from Free Text: Ferramenta desenvolvida pelo Instituto Max-Plack de Ciência da Computação da Universidade de Saarlandes.
c) TM-Gen: Ferramenta desenvolvida pelo Instituto de Investigação Sanitária da Universidade de Saragoça.

3.4.2 Documentos

O Corpus selecionado foi um conjunto de 22 documentos presentes no Wikiwars[12], dos quais 8 documentos foram escolhidos aleatoriamente para a realização dos experimentos.

3.4.3 Variáveis Resposta

Para cada um dos documentos selecionado dentre os demais presentes no Corpus para realização da validação, foi criado um gabarito a partir da realização da extração manual de relações entre expressões temporais (normalizadas e não normalizadas) e conceitos.

Desta forma, os resultados presentes no gabarito são comparados com as saídas geradas a partir da execução do RISO-GCT utilizando como entrada os documentos selecionados do Corpus, para o cálculo das medidas de precisão, cobertura e *f-measure*.

[12] http://timexportal.wikidot.com/wikiwars

Capítulo 4 - Trabalhos Relacionados

O capítulo a seguir apresenta os trabalhos relacionados com a abordagem proposta neste trabalho de dissertação para extração de relações entre conceitos e contextos temporais inclusão em Topic Maps, o *Temporal Fact and Event Extraction from Free Text* (KUZEY, 2011), o GeoST (MATA *et. al.*, 2010), o PorTexTO (MOTA, 2008) e o TM-Gen (GARRIDO et. al., 2013).

Dentre os trabalhos relacionados, nenhum destes possuía código fonte aberto e nem puderam ser disponibilizados para que fossem realizadas as etapas de verificação e validação. Não foram encontradas outras abordagens relacionadas.

4.1 Temporal Fact and Event Extraction from Free Text

Desenvolvido pelo Instituto Max-Planck de Ciência da Computação da Universität des Saarlandes, o framework *Temporal Fact and Event Extraction from Free Text* utiliza um mecanismo de extração de informações que recupera fatos temporais e eventos de texto semiestruturado e livre presentes em artigos da Wikipédia para enriquecer a ontologia temporal T-YAGO, que é uma extensão da ontologia YAGO[13]. Ele utiliza novos predicados para capturar as expressões temporais dos conceitos existentes. T-YAGO contém cerca de 300 mil conceitos extraídos de dados semiestruturados da Wikipedia (KUZEY, 2011).

O framework consiste em duas etapas de extração. Estas etapas serão detalhadas nas seções 4.1.1 e 4.1.2.

4.1.1 Extração Baseada em Padrões por Normalização de Datas Explícitas

Nesta etapa, a ontologia T-YAGO é enriquecida utilizando a própria base de conhecimento para geração de padrões e a partir deles encontrar conceitos e seus respectivos contextos temporais relacionados.

Este processo é subdividido nas seguintes etapas:

- **Seleção de Padrões:** Conceitos temporalizados são extraídos de texto livre através de buscas de padrões de texto obtidos a partir de T-YAGO. A precisão desta técnica depende estritamente da quantidade destes padrões significativos. A frequência dos

[13] YAGO é uma enorme base de conhecimento semântico, derivado de Wikipedia WordNet e GeoNames. Atualmente, YAGO tem conhecimento de mais de 10 milhões de entidades (como as pessoas, organizações, cidades, entre outras.) e contém mais de 120 milhões de fatos sobre essas entidades (MPN, 2014).

padrões é estatisticamente avaliada, de modo a ter padrões frequentes que resultarão em alto *recall* e de alta precisão.

- **Desambiguação de Entidades:** Para extração de informação orientada à ontologia, as locuções verbais do texto precisam ser mapeadas para as entidades corretas da ontologia. Muitas vezes este mapeamento é ambíguo, o que faz a tarefa de encontrar o significado pretendido de uma locução difícil. Por exemplo, a palavra "*Andrew*" pode ser mapeada para Andrew Jackson ou Andrew Johnson.

  É utilizada uma heurística para desambiguar entidades, onde, sabendo-se que um artigo da Wikipedia mapeia para uma única entidade, é possível determinar as palavras que podem mapear para a entidade correta. Por exemplo, se a entidade é *Michael Jackson*, então a palavra a palavra "*Michael"* ou a palavra "*Jackson"* são mapeados para a entidade "*Michael Jackson"*. Entretanto, se ambos as palavras ocorrem em sequência, esta sequência é diretamente mapeada para a entidade.
- **Resolução de Co-referência:** Em linguagem natural existem várias expressões em uma sentença ou documento que referenciam a mesma entidade. Por exemplo, na frase "*John is married to Mary. He loves her so much*." a palavra "*He*" se refere a "*John*". O sistema mapeia os pronomes como "*his", "her", "he", "she", "him"* com o mesmo nome da entidade referente ao artigo corrente.
- **Checagem de Tipo:** Uma ontologia não é uma simples coleção de fatos, mas uma coleção de tipos de entidades e fatos criando uma estrutura taxonômica[14]. Os fatos recém extraídos precisam ser consistentes com o domínio e a gama de relações. Exemplo: Enquanto o fato **f1**: *Nicolas_Sarkozy* `isMayorOf` *Neuilly-sur-Seine* foi corretamente tipado, ou seja, as entidades estão, de fato, no domínio da relação `isMayorOf`, o fato f2: *Nicolas_Sarkozy* `isMayorOf` *French_Economy* não foi corretamente tipado uma vez que *French Economy* não faz parte do domínio `isMayorOf`, de maneira que **f2** precisa ser podado.

Após a extração dos conceitos, estes precisam ser temporalizados. Por exemplo, para a sentença "*Jacques Chirac, born on 29 November 1932, is a French Politician who served as President of France from 1995 to 2007 and served as Prime Minister of France between the years 1974-1976 and from 1986 to 1988*". Para o conceito **f1**: *Jacques_Chirac* `holdsPoliticalPositon` *President_of_France* é preciso extrair o intervalo de tempo para o conceito f1, desta forma, é

[14] Vocabulário controlado de uma determinada área do conhecimento, e, acima de tudo, um instrumento ou elemento de estrutura que permite alocar, recuperar e comunicar informações dentro de um sistema sob uma premissa lógica (ALVARES, 2007). No contexto da Ciência da Computação, estrutura taxonômica é um sistema para classificar e facilitar o acesso à informação. (TERRA et al., 2006)

necessário determinar para f1 a data de referência correta e conectá-lo a **f1** como **f2**: f1`startedOnDate` 17-05-1995, **f3**: `f1endedOnDate` 16-05-2007.

A extração de contextos temporais dos conceitos consiste em 2 fases:

1. **Fase 1:** Detecção e Normalização de Expressões Temporais
2. **Fase 2:** Extração de conceitos e atribuição da informação temporal correta.

   Esta fase, por sua vez, é composta por duas etapas:

   2.1. **Indução Automática por Padrões:** São criados novos padrões através da base de dados temporais já existente (T-YAGO)

   2.2. **Aplicação dos Padrões:** Aplicando os padrões em textos para extrair novos conceitos contextualizados com a informação temporal.

4.1.2 Extração Baseada em Padrões por Normalização de Datas Implícitas

Existem várias informações temporais embutidas em linguagem textual. Para ilustrar, embora a versão mais recente do YAGO saiba o fato de que *Nicolas_Sarkozy* `holdsPoliticialPosition` *Minister_of_Interior*, ele não sabe o fato de *Nicolas_Sarkozy* `holdsPoliticialPosition` *Minister_of_Budget*. No entanto, existe uma frase no artigo da Wikipedia sobre Nicolas Sarkozy que diz "*Nicolas Sarkozy era ministro do Orçamento no governo de Edouard Balladur, durante o último mandato de François Mitterrand*", que mostra que o fato de *Nicolas Sarkozy* `holdsPoliticalPosition` *Minister_of_Budget* pode ser criado com intervalos de tempo válidos uma vez a frase temporal implícita "*no governo de Edouard Balladur*" ou "*durante o último mandato de François Mitterrand*" são detectados e normalizado corretamente.

Uma vez que T-YAGO conhece os fatos abaixo, a ferramenta utiliza das informações presentes em T-YAGO para encontrar datas nestas expressões implícitas, normalmente presentes em locuções adverbais temporais contidas nas frases.

- **f214:** Edouard_Balladur `holdsPoliticialPosition` Prime_Minister_of_France
- **f215:** *f214* `startedOnDate` *29-03-1993*
- **f216:** *f214* `endedOnDate` *10-05-1995*

Detalhamos, nas seções 4.1.2.1 e 4.1.2.2, os processos de Mapeamento de Expressões Temporais e Atribuição destas expressões aos fatos.

4.1.2.1 Mapeamento de Expressões Temporais

O processo de mapeamento de expressões temporais consiste de 3 fases: identificação, interpretação e normalização.

Identificação de Expressões Temporais

Nesta fase a ferramenta *Fact and Event Extraction from Free Text* identifica em artigos da Wikipedia locuções adverbiais compostas pelo padrão <advérbio> <frase nominal>, definidas manualmente. Com isso, a ferramenta irá localizar padrões como "*during French Revolution*", "*during Mitterrand's presidency*", "*after Word War I*", entre outros.

Interpretação

Nesta fase, locuções verbais temporais são mapeadas para um fato temporal existente na base de conhecimento temporal. Ou seja, dada uma base de conhecimento K, o objetivo desta fase é mapear as locuções verbais temporais candidatas encontradas na fase anterior para o fato temporal mais semelhante em K. Por esta razão, foram desenvolvidos dois dicionários, Rich Dicionary System (RDS) e Context Aware System (CAS). Qualquer um desses dicionários *d* pode responder a uma consulta *q* com uma pontuação de confiança onde *q* contém uma locução verbal temporal candidata encontrada na fase de identificação.

Normalização

Dado um conjunto de fatos temporais, a etapa de normalização faz consultas à base T-YAGO, com o objetivo de obter um tempo válido ou um intervalo de tempo válido. Uma vez que obtém a informação temporal, a ferramenta atribui à locução adverbial a informação temporal obtida. Desta forma, a cláusula temporal implícita é mapeada para o intervalo de tempo um ponto na linha do tempo na qual esta cláusula implica.

4.1.2.2 Atribuição das Informações Temporais aos Fatos Temporais

A ferramenta também associa fatos temporais aos fatos presentes na base de conhecimento. Por exemplo, a sentença "*After becoming Governor of California, Schwarzeneger acted in the movie Around the World in 80 days*" contém 3 propriedades interessantes, um advérbio temporal "*after*", um fato contido na base de conhecimento e um fato temporal.

Por exemplo, em uma base de conhecimento temporal que possua os seguintes fatos que implicam na locução verbal "*becoming Governor of California*":

- **f11:** *Arnold_Schwarzeneger* `holdsPoliticianPosition` *Governor_of_California*
- **f12:** f11 `startsOnDate` 17-11-2003
- **f13:** f11 `endsOnDate` 03-01-2011

Além disso, considerando que a base de conhecimento YAGO possua o seguinte fato que implica na frase "*Schwarzeneger get acted in the movie Around the World in 80 days*". Desta forma, este fato não possui nenhuma informação temporal qualificando-o.

- **f31:** Arnold_Schwarzeneger `actedIn` Around_the_World_in_80_days_(2004_film)

A locução "*becoming Governor of California*" pode ser mapeada para o fato f11: **f11:** *Arnold_Schwarzeneger* `holdsPoliticianPosition` *Governor_of_California* pelas técnicas citadas anteriormente. E a locução "*Schwarzeneger acted in the movie Around the World in 80 days*" pode ser mapeada para o fato **f31:** *Arnold_Schwarzeneger* `actedIn` *Around_the_World_in_80_days_(2004_film)*. A ideia é conectar um fato temporal à um fato da base de conhecimento. Desta forma, a informação temporal poderá ser atribuída ao fato conhecido da base de conhecimento.

Advérbios temporais possuem polaridade {-1, 0, 1}, onde -1=before, 0=during e 1=after. A polaridade é usada para determinar a relação correta que vai ser utilizada para ligar um fato temporal à um fato conhecido da base. As relações possíveis são `happenedBefore`, `happenedDuring` e `happenedAfter`. A locução verbal é extraída para um possível par de entidades. Todos os fatos conhecidos da base que possuem entidades presentes na locução verbal são checados utilizando a ferramenta PROSPERA[15] para verificar se a locução mapeia para um fato conhecido da base. Se a ferramenta PROSPERA retornar um fato conhecido da base a partir locução verbal selecionada, então a fato da base e o fato temporal são conectados através da relação determinada pela polaridade do advérbio temporal. Desta forma, a base de conhecimento vai sendo enriquecida pela dimensão temporal. Como exemplo, o fato **f31:** *Arnold_Schwarzeneger* `actedIn` *Arround_theWorld_in_80_Days_(2004_film)* vai ser conectado ao fato **f11:** *Arnold_Schwarzeneger* `holdsPoliticalPosition` *Governor_of_California* através da relação **f39:** f31 `happenedAfter` f11.

4.2 GeoST: Geographic, Thematic and Temporal Information Retrieval from Heterogeneous Web data sources

O GeoST (Geographic, Thematic and Temporal Information fro Heterogeneous Web Data Sources) foi um trabalho desenvolvido pela Unidade Profissional Interdisciplinar em Engenharia e Tecnologias Avançadas (UPIITA) do Instituto Politécnico Nacional (IPN) do México. Foi criada uma abordagem orientada à consulta em ontologias para recuperação de informações temporais e geográficas na web. Consultas espaço-temporais são utilizadas para solução de questões relacionadas às dimensões "*quando*", "*o que*" e "*onde*". GeoST adota uma estratégia baseada em consultas em uma ontologia espaço-temporal executado por um motor de busca (MATA *et. al.*, 2010).

[15] PROSPERA (PRospering knOwledge with Scalability, PrEcision, and RecAll) é um sistema escalável responsável por contruir automaticamente bases de conhecimento a partir de fontes presentes na web. (NAKASHOLE *et al.*, 2011)

Considerando a consulta Q_1= {"*When and where it happened The Mexican Revolution*"}, a informação é coletada a partir da Wikipedia: http://es.wikipedia.org/wiki/Revolución_mexicana. Neste caso, o principal conceito a ser pesquisado na Wikipedia é "*Mexican Revolution*". A Wikipedia possui uma estrutura definida com uma lista de conteúdo, esta lista é utilizada como ponto de partida para encontrar informações relevantes sobre os elementos contidos no texto. Ou seja, cada item da lista aponta para um parágrafo e em cada um desses parágrafos são buscados os termos. Quando algum termo é encontrado, o parágrafo que contém o termo é extraído. Posteriormente, as frases que possuem os termos são extraídas a partir do parágrafo. Informações adicionais que emergem dessas frases são procuradas, isto é, locais e nomes de lugares, eventos e referências de tempo e informação temática. Além disso, quando uma referência para uma determinada data é encontrada (por exemplo, um ano dentro do período 1910-1917), a frase que contém esta data é extraída. Estas frases são processadas sintática e semanticamente e, nesta etapa, são eliminadas as *stop words*, e de acordo com as relações semânticas identificadas na ontologia, os termos relevantes são extraídos e as frases são integradas em um conjunto final de resultados.

Em geral o processo de pesquisa possui três etapas: Primeiramente são buscadas localizações geográficas e nomes de lugares via Web Service. Posteriormente são recuperadas datas, eventos, e referências temporais e, finalmente, são buscadas informações temáticas adicionais.

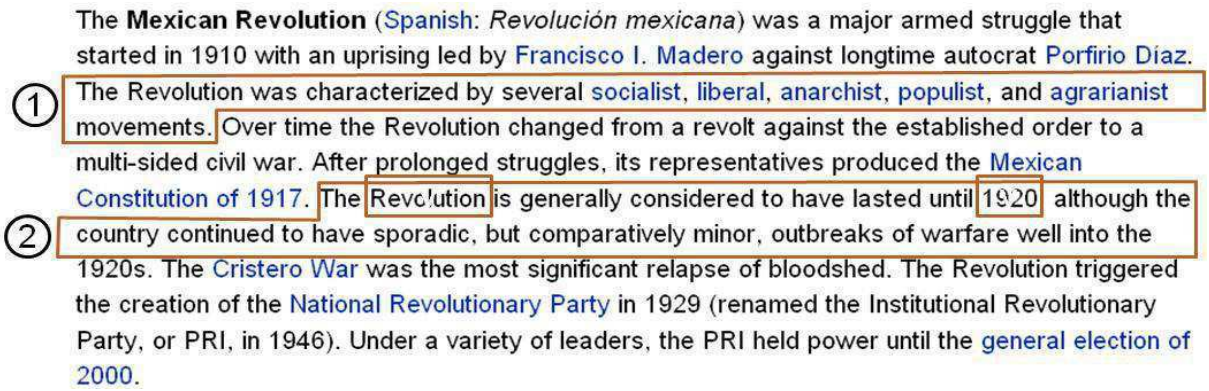

The **Mexican Revolution** (Spanish: *Revolución mexicana*) was a major armed struggle that started in 1910 with an uprising led by Francisco I. Madero against longtime autocrat Porfirio Díaz.
① The Revolution was characterized by several socialist, liberal, anarchist, populist, and agrarianist movements. Over time the Revolution changed from a revolt against the established order to a multi-sided civil war. After prolonged struggles, its representatives produced the Mexican Constitution of 1917. The Revolution is generally considered to have lasted until 1920, although the
② country continued to have sporadic, but comparatively minor, outbreaks of warfare well into the 1920s. The Cristero War was the most significant relapse of bloodshed. The Revolution triggered the creation of the National Revolutionary Party in 1929 (renamed the Institutional Revolutionary Party, or PRI, in 1946). Under a variety of leaders, the PRI held power until the general election of 2000.

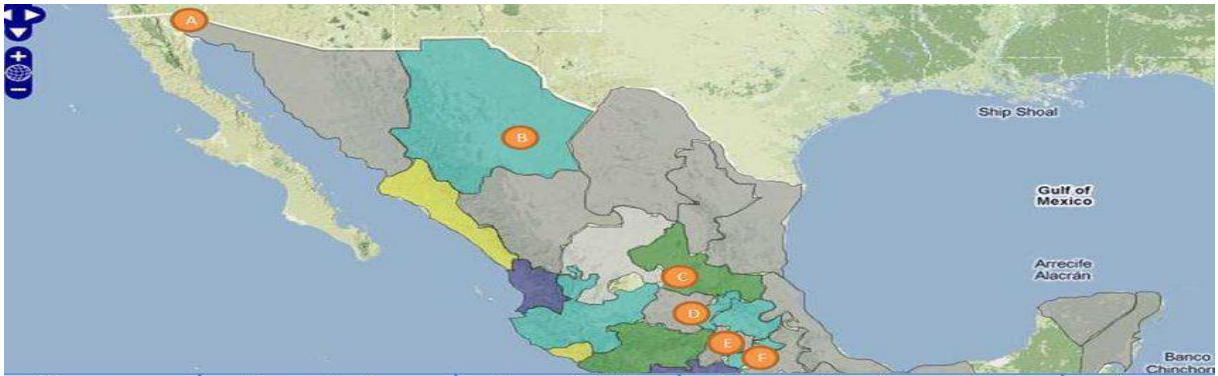
Figura 4 – Fragmento de artigo sobre a Revolução Mexicana na Wikipedia Q_1 (MATA *et. al.*, 2010).

Na Figura 4 é ilustrado o processo de busca utilizando como exemplo a consulta Q_1, onde é mostrado um fragmento do artigo em inglês sobre a Revolução Mexicana extraída da Wikipedia, e como as frases que casam com a consulta Q1 são marcadas. Quando uma data informada na consulta casa com uma data encontrada no texto (e.g., um ano entre o período 1910-1917), a sentença que contém esta data é extraída.

Na primeira etapa do processo são buscados os lexemas que casam com a consulta Q_1 e seu contexto no topo da Wikipedia. Muitos estados e lugares relacionados ao Norte do México (dado como contexto) são recuperados (e.g., estados de San Luis Potosi, Puebla, Cidade do México e outros estados da República do México). Estes estados e lugares são usados como parâmetros de entrada para um *Web Service* conectado à uma base de dados geográfica. O *Web Service* retorna as coordenadas geográficas destes nomes de lugares e estados.

A segunda etapa, procura por datas e expressões temporais (e.g., "*start*", "*finish*", "*before*", after) de acordo com a consulta e com a semântica dos documentos da web. Em particular, o domínio da ontologia identifica se os eventos são Contínuos[16] ou Ocorrências[17]. Por exemplo, para Q1, é verificado que a Revolução Mexicana começou em 1910 e finalizou em 1917, e consequentemente é considerado uma *ocorrência*.

A terceira e última etapa consiste em encontrar informação temática adicional de acordo com a consulta e seu contexto na web. Este processo é baseado em informações presentes na Wikipedia. Por exemplo, a Revolução Mexicana é descrita como um conjunto de movimentos socialistas, liberais, anarquistas, populistas e agrários. Informações adicionais podem ser similarmente derivadas para diferentes locais e estados identificados. A consulta derivada pela aplicação desta busca é inserida em um arquivo XML cujo formato é definido para conter todos os resultados de acordo com os domínios temporais, temáticos e espaciais. Um mapa é finalmente derivado utilizando o servidor geográfico para cada local geográfico relacionado à Revolução Mexicana (Figura 5).



Time Reference	Event starts	Event finish	Geographic places related	identifier
1910			SLP, San Luis Potosí City	C
1910	-	-	Mexicali City	A
1910	-	-	chihuahua City	B
-	1910	1917	queretaro City	D
1917	-	-	Puebla City	F
-	1910	1917	Mexico City	E

Figura 5 – Resultados exibidos para a consulta Q1¬ (MATA et. al., 2010).

[16] Entidades que possuem existência em um tempo contínuo e preservam sua identidade mesmo após mudanças (MATA *et. al.*, 2010)

[17] São eventos e/ou processos que ocorrem e em seguida já não mais existem (e.g., uma epidemia, um deslocamento, entre outros). (MATA *et. al.*, 2010)

A Figura 5 ilustra o mapa que mostra a localização geográfica relacionada à consulta Q_1, e abaixo, os locais relacionados ao período de tempo (e.g., onde um evento começa e termina). A ferramenta utiliza uma API do Google Maps para mostrar os resultados da busca de acordo com o tempo (representado por uma linha que conecta as localizações), isto é, a partir da sequência ordenada das datas relacionadas com as cidades mexicanas envolvidas na revolução (vide Figura 6).

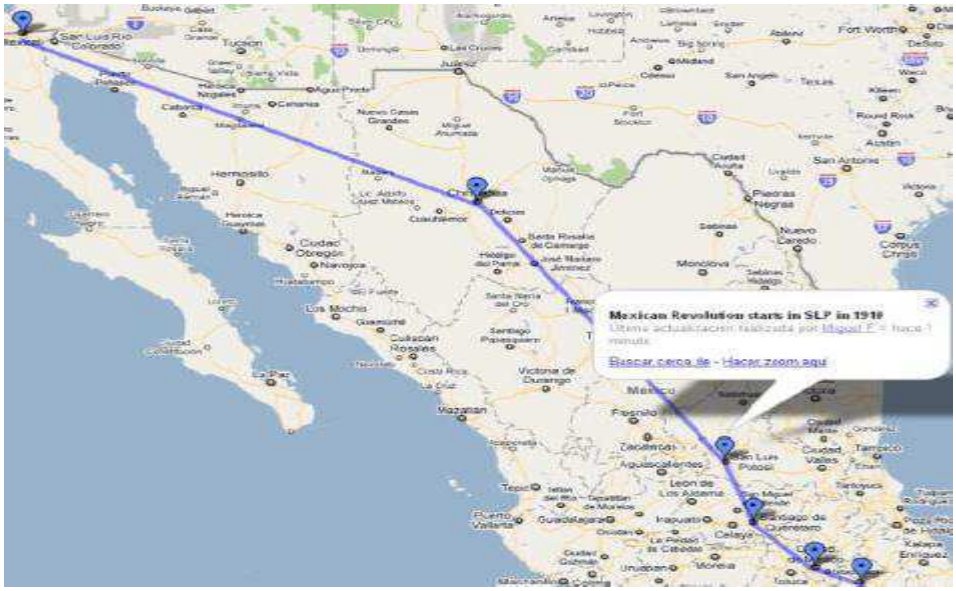


Figura 6 – Resultados ordenados por data retornadas na consulta Q_1 exibidos no Google Maps (MATA *et. al.*, 2010).

Na Figura 6 é ilustrado o resultado exibido para a consulta Q1 = "*When and where it happened The Mexican Revolution*" onde são representados no mapa os pontos geográficos onde ocorreram eventos relacionados à Revolução Mexicana no qual, em cada ponto, são atribuídas as datas nos quais os eventos ocorreram. O exemplo em questão na Figura 6 é representada a localização da cidade de San Luis Potosí, local onde se iniciou a Revolução Mexicana em 1910.

4.3 PorTexTO

A ferramenta PorTexto (PORtuguese Temporal Expressions TOol) foi desenvolvida pela Escola Superior de Tecnologia e gestão do Instituto Politécnico de Leiri de Portugal e pelo Departamento de Engenharia Informática da Universidade de Coimbra e que utiliza padrões de expressões temporais criados através de identificação das co-ocorrências e padrões definidos com expressões regulares (MOTA, 2008).

O processamento é feito frase a frase, e estas frases são divididas através pelo Perl's Lingua[18], atomizador da ferramenta Linguateca[19]. A identificação de expressões regulares é feita utilizando expressões regulares com base em co-ocorrências existentes em palavra de referência (meses do ano, dias da semana, estações do ano, festividades, medidas temporais, entre outras).

PorTexTO permite o processamento de documentos tanto em formato de texto simples não-estruturado como em formato estruturado XML. O resultado produzido pelo sistema pode ser um ficheiro no seu formato original, mas com as devidas anotações nas expressões temporais encontradas ou então um ficheiro com todas as expressões temporais encontradas e sua posição relativamente ao texto original.

O sistema tem como entrada o texto original e os padrões de expressões que são previamente criados por outro módulo da Linguateca[20], o Processador de Co-ocorrências[21].
Além da coleção de padrões de expressões, o sistema tem ainda como entrada uma lista de palavras-chave temporais usadas unicamente para excluir do processamento frases que não contenham expressões temporais e assim conseguir diminuir o tempo final de processamento dos documentos. Esta lista funciona como um filtro das frases a serem processadas.

As frases excluídas são todas as que não tem datas, nem nenhuma das palavras-chave temporais. As palavras-chave são definidas de acordo com a lista de expressões temporais que o sistema deverá identificar e classificar, de modo a atingir os objetivos de uma determinada tarefa. Por exemplo, se existirem padrões de expressões temporais com a palavra "*ano*", então "*ano*" deverá existir na lista de palavras-chave temporais.

Nas próximas seções serão detalhados os componentes desta ferramenta.

4.3.1 Módulo Anotador

O módulo Anotador é responsável por identificar expressões temporais, mediante os padrões definidos pelo módulo Processador de co-ocorrências, fazer a sua classificação e, posteriormente, proceder à anotação no texto original.

Os documentos de entrada são trabalhados um de cada vez e o processamento de cada documento é feito numa frase de cada vez. Cada frase será submetida às quatro etapas de processamento do Anotador do PorTexTO. A frase só será dividida nos seus termos caso haja

[18] Esta ferramenta inclui um método configurável para atomização da língua portuguesa. No entanto, também pode ser utilizado para outras línguas (especialmente inglês e francês). (CPAN, 2006)
[19] Centro de Recursos para o processamento computacional da Língua Portuguesa cujo objetivo é facilitar o acesso aos recursos já existentes para o processamento da língua portuguesa. (LINGUATECA, 2015)
[20] http://www.linguateca.pt/
[21] O objetivo principal deste módulo é determinar as expressões temporais mais utilizadas numa determinada colecção segundo uma abordagem estatística e com estas expressões criar os padrões que vão ser posteriormente utilizados no módulo Anotador

necessidade de reconhecer datas com o mês por extenso, ou datas que tenham também o dia da semana. Por exemplo, a frase "*Domingo, 7 de Setembro de 2008*" é dividida nos seguintes termos: "*Domingo*", "*7*", "*de*", "*Setembro*", "*de*", "*2008*". Com estes termos, é verificado se os que estão à esquerda e à direita do mês podem fazer ou não parte de uma data. No caso desses termos poderem constar de uma data então serão incluídos na expressão que vai ser marcada como DATA. No caso contrário, só o mês é que será marcado, mas com o marcador MES. No exemplo apresentado, a frase inicial terá a marcação DATA.

4.3.2 Módulo Processador de Co-ocorrências

Este módulo só é executado quando ainda não existem padrões de expressões temporais ou quando os que existem são insuficientes para a tarefa a desempenhar pelo PorTexTO. O objetivo principal deste módulo é determinar as expressões temporais mais utilizadas numa determinada coleção segundo uma abordagem estatística e com estas expressões criar os padrões que vão ser posteriormente utilizados pelo módulo Anotador. Produz como resultado um arquivo com os padrões definidos através de expressões regulares e a respectiva classificação.

4.4 TM-Gen: A Topic Map Generator from Text Documents

Desenvolvido pelo Instituto de Investigación Sanitaria da Universidad de Zaragoza, TM-GEn (Topic Map GENerator) é um sistema automatizado capaz de extrair informação e conhecimento a partir de um ou mais textos e incluir estas informações em um Topic Map. (GARRIDO et. al., 2013).

Inicialmente é realizado o pré-processamento dos textos para encontrar informações relevantes que serão utilizadas nas fases posteriores. Cada texto é processado separadamente e gera um Topic Map para cada um. Os textos são divididos em sentenças com o propósito de analisá-las separadamente e atribuir a elas uma pontuação de relevância, para elencar as mais importantes no texto. Depois, TM-Gen analisa sintaticamente as sentenças para encontrar os melhores candidatos para ser um tópico, e depois o sistema estabelece as associações entre eles. Em seguida, TM-Gen realiza uma simplificação semântica, caso existam redundâncias, associações incompatíveis ou ambiguidades. Após todos os textos terem sido analisados e terem sido gerados seus respectivos Topic Maps, é realizado um processo de unificação para geração de um único Topic Map para o texto.

As etapas do processamento são detalhadas a seguir:

- **Pré-processamento:** TM-Gen realiza a análise dos textos para obter informações que estão relacionadas às palavras nele contidas. Para isto, é utilizada uma ferramenta de processamento de linguagem natural para extrair as entidades presentes neles.
- Estas informações são guardadas em duas listas. Uma das listas armazenará as entidades identificadas pelo processador de linguagem natural que correspondem a nomes de pessoas, lugares, organizações, entre outros. Cada uma delas possuirá um peso de relevância. A outra lista irá conter as frequências das palavras contidas nos textos.
- **Extração de Palavras-Chave:** Extrai uma lista de palavras-chave para cada texto utilizando TF-IDF (Term Frequency – Inverse Document Frequency)[22]. Posteriormente estas palavras permitem identificar quais partes do texto provê mais conhecimento. Nesta fase são utilizadas as frequências obtidas na fase anterior.
- **Extração de Entidades:** TM-Gen extrai uma lista de entidades em cada texto. Estas entidades são importantes candidatas a comporem no Topic Map.
- **Divisão do Texto em Sentenças:** Divide cada texto em sentenças, onde cada uma será analisada na próxima fase a fim de identificar quais destas sentenças são mais relevantes.
- **Pontuação de Sentenças:** É atribuída uma pontuação à cada sentença para identificar quais delas são mais relevantes. Para esta tarefa, são levadas em consideração a frequência das palavras-chave e entidades contidas em cada sentença.
- **Ordenação das Sentenças:** Uma vez que foi atribuída uma pontuação à cada uma das sentenças, estas são ordenadas da sentença de maior para a de menor relevância (maiores no topo da lista).
- **Análise das Sentença:** Nesta fase as sentenças são analisadas sintática e gramaticalmente para identificar a função de cada palavra presente na sentença e o tipo de cada uma. É importante realizar esta análise uma vez que os melhores candidatos a comporem o Topic Map são as palavras que possuem a função de sujeito.
- **Adição de Tópicos:** TM-Gen adiciona ao Topic Map os candidatos encontrados no passo anterior. Cada candidato, que já foi previamente incluído não é incluído, mas seu peso aumenta.
- **Inclusão de Associações:** Nesta fase são adicionadas ao Topic Map as associações entre os tópicos encontrados em cada sentença. Estas associações são encontradas a

[22] Esta medida leva em conta a capacidade de descrição do termo em relação ao documento. Um termo que é muito frequente em todos os documentos do *corpus* ou parágrafos de um texto não é um bom descritor, e recebe um valor baixo. Um termo que aparece em poucos documentos do *corpus* ou em poucos parágrafos de um texto tende a ser um descritor eficaz, e recebe um valor alto. (RIBEIRO, 2015)

partir dos verbos presentes na sentença. TM-Gen então, realiza esta tarefa procurando pelo sujeito na sentença e associando os verbos relacionados a ele presentes no texto.

- **Simplificação Semântica:** Uma vez que os tópicos e as associações já foram adicionados ao Topic Map, o processo realiza a análise semântica deles e faz a simplificação do Topic Map removendo as redundâncias, associações incompatíveis e ambiguidades.
- **Avaliação:** Cada vez que o sistema inclui um elemento ao Topic Map, um processo de avaliação é realizado. Neste processo, é checado se o número de tópicos adicionados é suficiente. Em caso afirmativo, o processamento é finalizado. Caso contrário o processo continua. Este número é reajustado de acordo com nível de profundidade desejado para o Topic Map.
- **Validação:** O processo checa se o Topic Map foi gerado corretamente e verifica se ocorreram erros nas etapas anteriores.
- **Unificação dos Topic Maps gerados:** Uma vez que os Topic Maps de cada texto foram gerados, é realizada a etapa de unificação.

4.5 Considerações Finais

Não foi encontrado nenhum trabalho relacionado à indexação de documentos, que relaciona termos e informações temporais e inclui estes relacionamentos em Topic Maps, ou seja, nenhum deles produz um *Thesaurus* de conceitos temporalizados relacionados aos documentos indexados.

A Tabela 4 mostra um comparativo entre as ferramentas analisadas:

Ferramenta	Característica								
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Temporal Fact and Event Extraction from Free Text	X	X	X	X	X				X
GeoSGT	X	X	X		X	X			
PorTexTO		X	X						
TM-Gen				X			X		X

Tabela 4 – Tabela Comparativa das ferramentas de Recuperação de Informação

São consideradas as seguintes características:

[1] Retorna Consultas Temporais

[2] Extrai Informações Temporais Implícitas em Textos

[3] Extrai Informações Temporais Explícitas em Textos

[4] Realiza Desambiguação de Termos

[5] Extrai Informações Temporais Normalizadas

[6] Extrai Informações Geográficas

[7] Gera Topic Map com Informações Não-Temporais (Entidades)

[8] Gera Topic Map com Informações Temporais

[9] Relaciona Entidades e Informações Temporais em Documentos

A análise dos trabalhos relacionados concluiu que apenas duas das ferramentas extraem relacionamentos entre conceitos e expressões temporais presentes em documentos ("*Temporal Fact and Event Extraction from Free Text*" e "*TM-Gen*"), dentre as quais, apenas a ferramenta "*TM-Gen*" realiza a inclusão destas informações em Topic Maps. Entretanto, não faz parte do escopo da ferramenta TM-Gen a inclusão de informações temporais relacionadas aos conceitos em Topic Map, uma das funcionalidades implementadas pelo RISO-GCT.

No Capítulo 5 será detalhado o Projeto RISO, responsável pela Recuperação da Informação Semântica de Objetos Textuais. Serão detalhados os componentes do projeto, bem como as evoluções implementadas para validação da abordagem criada neste trabalho de dissertação de mestrado.

Capítulo 5 - Projeto RISO- Recuperação da Informação Semântica de Objetos Textuais

Este capítulo tem como objetivo apresentar o Projeto RISO (Recuperação da Informação Semântica de Objetos Textuais). Serão mostrados todos componentes do projeto, alguns já desenvolvidos outros em desenvolvimento ou a serem desenvolvidos. Um dos componentes é o RISO-GCT, objeto deste trabalho de dissertação.

5.1 Introdução

O projeto RISO (Recuperação da Informação Semântica de Objetos Textuais), está sendo desenvolvido pelo grupo de Sistemas de Informação e Bancos de Dados (SINBAD) do DSC/CEEI/UFCG, e tem como objetivo criar um ambiente de indexação e recuperação semântica de documentos, possibilitando uma recuperação mais acurada, melhorando o fator de precisão dos resultados mediante a diminuição da ambiguidade do sentido dos termos (BISPO, 2013).

O RISO é composto por três componentes:

RISO-VTD (Vocabulários Temáticos de Domínio): Responsável por criar vocabulários específicos para os principais domínios de conhecimento e classificar documentos novos de acordo com estes vetores (BISPO, 2013).

O RISO-VTD é composto por uma fase de treinamento para criação de vocabulários temáticos de domínios e uma fase para, baseado nestes vocabulários, classificar um documento qualquer. Para a primeira fase ele recebe um conjunto de documentos de um domínio de conhecimento específico, realiza a extração de termos formando assim um **vocabulário temático.** Os termos deste vocabulário são ponderados calculando-se a frequência de cada termo, determinando assim a importância de cada termo em relação aos domínios, produzindo assim os **vetores temáticos dos domínios**. Para a classificação de um novo documento o vetor deste documento é comparado com os vetores temáticos dos domínios para detectar o domínio mais próximo do documento.

RISO-IS (Indexação Semântica): Responsável por desambiguar termos (ou sintagmas) existentes nos documentos, transformá-los em conceitos bem definidos, determinar relações semânticas entre os conceitos e criar um Topic Map relacionando conceitos com documentos

(SANTOS, 2013). Esta indexação semântica é realizada em três etapas; A Indexação Conceitual (RISO-IC), a Indexação Temporal (RISO-IT) e a Indexação Espacial (RISO-IE).

- O **RISO-IC (Indexação Conceitual)** é responsável por, a partir de um documento de entrada, realizar a **extração dos termos** presentes no texto. Para cada termo extraído, realiza sua **desambiguação** determinando o conceito representado pelo termo. Os conceitos são submetidos a um processo de **enriquecimento semântico**, onde são acrescidas as relações semânticas linguísticas, obtidas a partir de fontes externas. Ao final, é realizada a **atualização** do Topic Map que poderá servir de índice quais conceitos estão contidos em quais documentos.
- O **RISO-IT** é responsável por identificar, manipular e persistir informações temporais presentes em textos e em fontes externas. Primeiro são encontrados os termos temporais (RISO-TT) para, em seguida, determinar a temporalidade dos conceitos (RISO-GCT).
    - **O RISO-TT (Temporal Tagger)** realiza a leitura de um documento e identifica as expressões temporais presentes nele. Estas expressões são normalizadas, colocando-as em um padrão em que os tempos sejam facilmente por outros componentes do sistema.
    - O **RISO-GCT (Geração de Conceitos Temporais)**, desenvolvido neste trabalho de dissertação de mestrado, é responsável por realizar a leitura de documentos e relacionar cada conceito extraído pelo RISO-VTD com possíveis expressões temporais previamente identificadas pelo RISO-TT. Estas relações são acrescidas ao Topic Map gerado pelo RISO-IC.
- **RISO-IE** é responsável por identificar, manipular e persistir informações geográficas presentes em textos. Analogamente ao RISO-IT são procurados termos espaciais no documento (RISO-TE) para, com isso, determinar características espaciais dos conceitos (RISO-CE).
    - **RISO-TE (Termos Espaciais)** identifica, em documentos textuais, os termos que representam informações geográficas. E.g., "*a cidade de São Paulo*", "*Nova Iorque*", "*Amazonas*", "*Bairro de Bodocongó*", entre outros.
    - **RISO-CE (Conceitos Espaciais)** é responsável por relacionar conceitos presentes em documentos textuais (extraídos pelo RISO-VTD) com as informações geográficas identificadas pelo RISO-TE. Esta operação resulta no enriquecimento do Topic Map com as informações geográficas de cada conceito.

RISO-CS (Consultas Semânticas): Possibilita consultas inteligentes dos usuários onde, a partir de uma entrada inicial dada pelo usuário, a interface de comunicação realiza uma busca no

Thesaurus por todos os conceitos que tenha alguma relação (sintática ou semântica) com o termo definido e mostra opções de desambiguação – conceitual, espacial e temporal. Após a desambiguação são sugeridas expansões da consulta baseadas no enriquecimento semântico do conceito. Com base no conceito expandido o usuário poderá determinar, com precisão, a real necessidade e informação fazendo com que os resultados estejam de acordo com essa necessidade.

Os componentes citados são representados na Figura 7 que ilustra a arquitetura atual do projeto RISO.

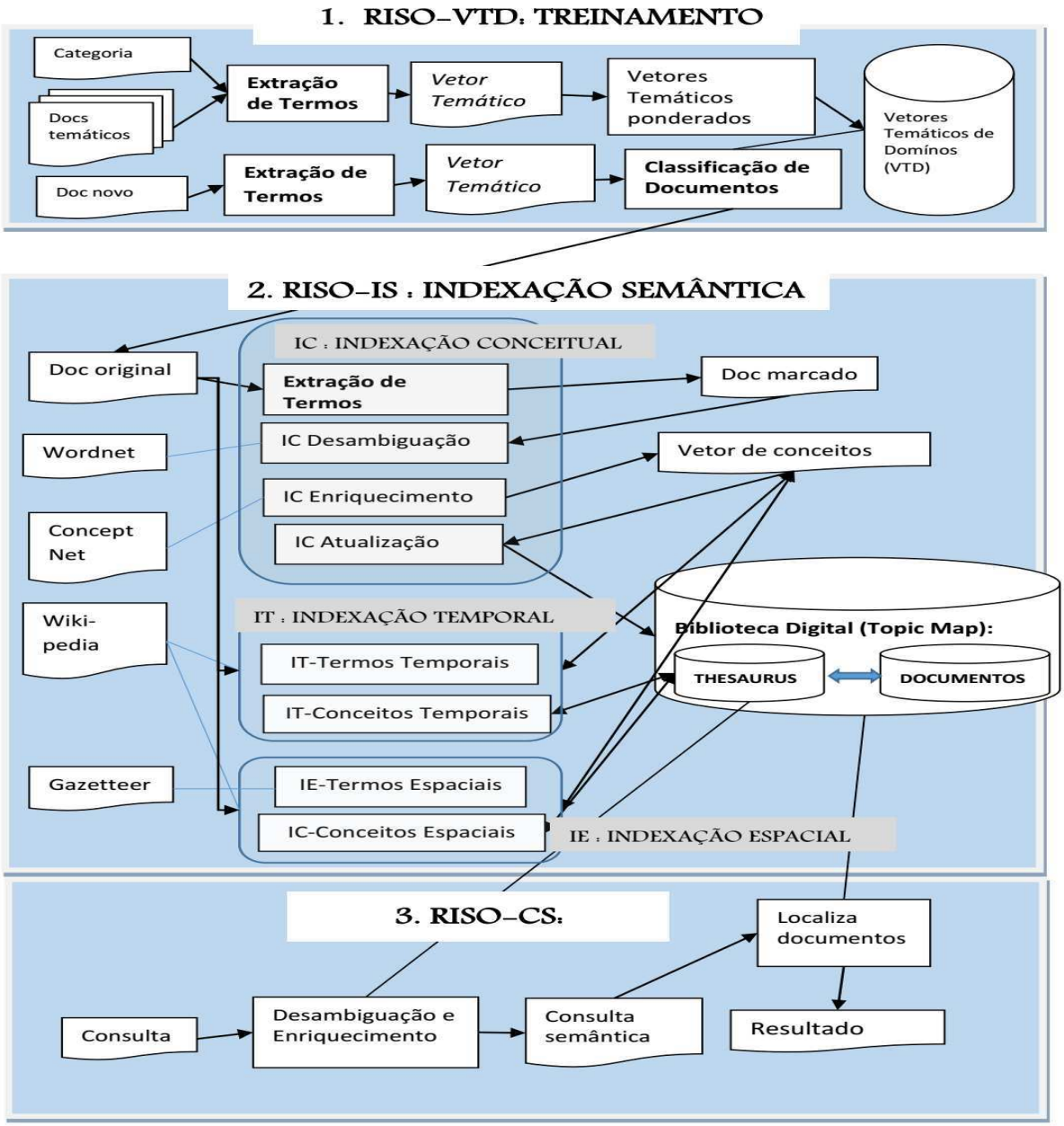


Figura 7 – Componentes atuais do RISO (SCHIEL, 2015)

Nas próximas subseções serão apresentados resumidamente os componentes do RISO em parte já desenvolvidos e que formam a base para esta pesquisa de mestrado.

5.2 RISO-VTD (Sistema de Criação de Vocabulários Temáticos de Domínio para Classificação de Documentos Digitais)

Em um sistema de indexação e recuperação de documentos sem restrições a áreas do conhecimento a possibilidade de ambiguidade nos termos de um documento é muito grande. Por isso uma classificação temática dos documentos a serem indexados pode facilitar a desambiguação dos termos contidos neles. O RISO-VTD (BISPO, 2013), é responsável por criar Vocabulários Temáticos para os principais domínios de conhecimento, por um processo de aprendizado a partir dos termos contidos em documentos típicos. Os Vetores Temáticos podem ser utilizados para classificar novos documentos que irão compor uma biblioteca digital. Com essa classificação, torna-se possível diminuir a ambiguidade dos termos indexados enriquecidos semanticamente, proporcionando uma recuperação de documentos mais precisa. O processo de treinamento do RISO-VTD para a criação dos vocabulários temáticos é ilustrado na Figura 8, que consiste na etiquetação dos termos presentes nos textos com as respectivas classificações gramaticais, para a posterior extração dos termos significativos de cada documento, obtendo assim os termos mais importantes do domínio ao qual cada documento pertence. Este processo é detalhado nas seções 5.2.1 e 5.2.2.

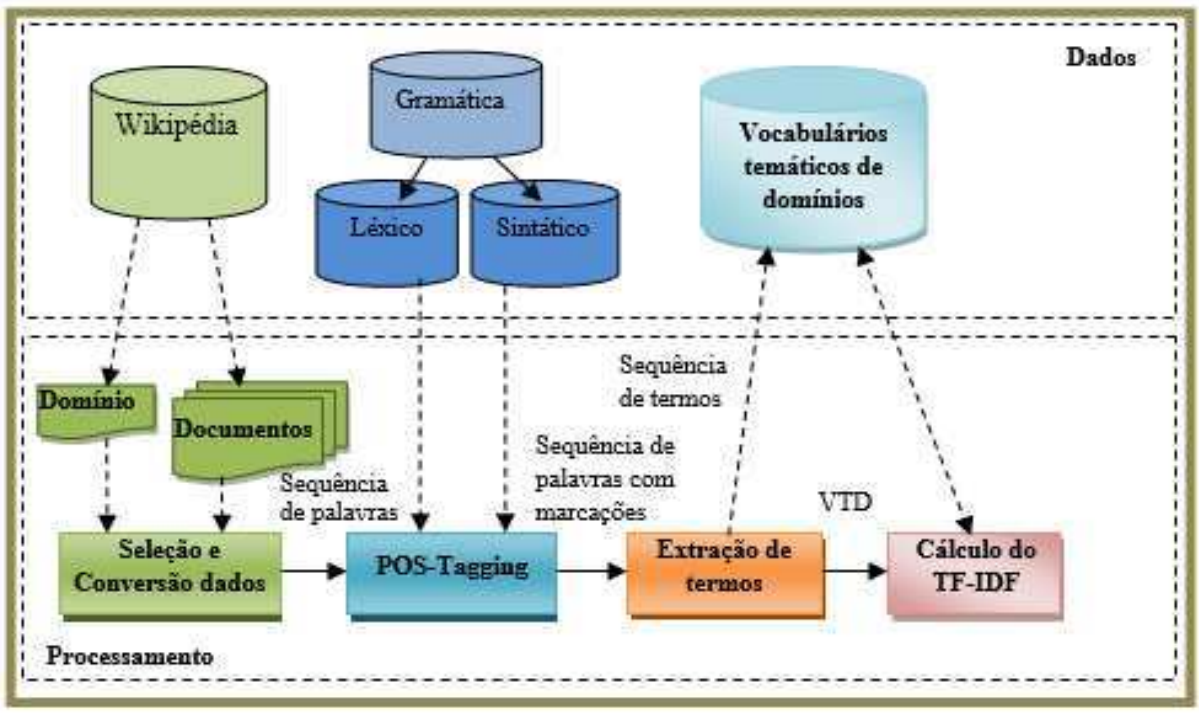


Figura 8 – Arquitetura do processo de treinamento do RISO-VTD (BISPO, 2013)

5.2.1 Processamento dos Documentos utilizando o *MontyLíngua* (*PoS-Tagging*)

Após a seleção de uma coleção de documentos sobre um tema ou domínio, esta coleção é processada utilizando o *PoS-Tagger* do MontyLingua[23]. Esta fase é composta por 5 etapas.

Na etapa de **Separação de Texto**, o texto do documento de entrada é separado em *tokens* com o objetivo de criar estruturas de dados do tipo lista para separar o texto em termos possibilitado, desta maneira, o processo de marcação termo a termo, sem perder a referência do texto.

Na etapa de **Marcação de Texto** é realizada a marcação utilizando a técnica *Part-of-Speech Tagging* (*POS-Tagging*) (BRILL, 1995; BRILL, 1994), que consiste em etiquetar os termos identificados na etapa anterior utilizando *tags* específicas de acordo com uma classe gramatical, indicando se termo é verbo, sujeito, substantivo adjetivo, entre outras. Entretanto, devido à presença de ambiguidade léxica nos termos, determinar a classe gramatical correta do termo em questão torna-se difícil. Logo, para classificar, pode haver a necessidade de utilizar a classificação gramatical do termo vizinho.

Posteriormente, vem a etapa de **Extração de Termos** realizada pelo *RISOExtractor* (Extrator de Termos do RISO) que tem como entrada o texto marcado com *tags* e agrupado em frases nominais, verbais e adjetivas. A etapa de extração contempla as principais alterações feitas com inclusões de regras heurísticas criadas para formação e realização da extração dos termos, com o objetivo de obter uma maior combinação de palavras, obtendo termos mais relevantes. A frase é analisada e, dependendo do seu conteúdo, as frases são criadas com mais de um *token*, sempre iniciando com um substantivo ou adjetivo. O novo termo pode ser composto por *n tokens*. Nesta etapa também ocorre a **determinação dos termos para criação dos vetores temáticos**, que extrai os termos relevantes ao domínio considerado. Esses termos são obtidos de acordo com heurísticas criadas e adicionadas ao *RISOExtractor*, mediante observação da formação dos termos no texto. Os termos são identificados, extraídos e armazenados em uma tabela rotulada com o nome do domínio em evidência.

Após a extração é realizado o **cálculo da frequência do termo,** que determina a importância de cada termo em relação ao domínio em que ele está inserido, denominado *tf* (*term frequency*). Após o processamento de todos documentos é calculada a frequência global dos

[23] http://web.media.mit.edu/~hugo/montylingua/ Consiste em um conjunto de ferramentas individuais de Processamento de Linguagem Natural (PLN), para suprir diversos aspectos de processamento de texto em inglês

termos, que quantifica a importância do termo em relação a todos os domínios, denominada *idf* (*inverse document frequency*).

5.2.2 Armazenamento dos Vetores Temáticos

O armazenamento dos vetores temáticos é realizado em duas etapas: na primeira etapa é realizada a geração do vetor de termos por domínio, sendo cada vetor rotulado com o nome do documento; a segunda etapa é o armazenamento de todos os vetores de termos criados por domínios armazenados numa única base denominada Vetores Temáticos de Domínios (VTD). O VTD tem o mesmo esquema dos vetores criados para cada domínio, cada registro conterá o termo extraído, o domínio ao qual ele pertence, a frequência absoluta (quantidade com que o termo aparece no domínio), a frequência global (número de domínios nos quais aparece o termo), e o peso relativo do termo (*tf-idf*).

5.3 RISO-IS (Indexação Semântica de Documentos)

O RISO-IS (Indexação Semântica de Documentos) é responsável por realizar a indexação semântica de documentos no sentido de obter uma classificação dos termos segundo três categorias: conceitual, temporal e espacial.

É composto por 3 componentes:

- **RISO-IS:** Realiza a indexação conceitual dos documentos;
- **RISO-TT:** Responsável pela determinação da temporalidade dos conceitos.
- **RISO-IE:** É responsável pela determinação de possíveis características espaciais dos conceitos.

Nas seções 5.3.1, 5.3.2 e 5.3.3 serão descritos estes componentes.

5.3.1 RISO-IC (Indexação Conceitual)

O RISO-IC, desenvolvido como RISO-ES por ARAÚJO JÚNIOR (2013), tem como finalidade determinar os conceitos contidos em um texto e promover seu enriquecimento semântico, baseado em informações presentes em fontes externas e heterogêneas ao sistema. Os termos contidos no texto são desambiguados para, em seguida, receberem a serem relacionados com outros conceitos por relações semânticas.

Um conceito constitui o átomo ou a menor unidade de informação capaz de expressar de maneira completa determinado sentido ou ideia, seja ela real ou abstrata. Neste enfoque, a descrição de um conceito pode ser obtida por três tipos de fontes de informação: informações de dicionário (termos que refletem seu papel primeiro diante da língua em consideração),

informações enciclopédicas (informações de conhecimento de mundo) e, por fim, informações de sentido comum (o veredito de um grupo de pessoas sobre instâncias e sua função diária e pragmática). Para a obtenção destas informações são utilizados dados oriundos das bases de dados do WordNet[24], Wikipédia[25] e Verbosity[26], respectivamente.

No item 2 da Figura 7 podem ser vistos os componentes que integram o RISO-IC.

O processo é composto por quatro etapas.

A primeira etapa busca a **identificação e extração de conceitos** presentes em um documento, previamente marcado, utilizando o algoritmo de detecção de conceitos em documentos proposto por BISPO (2013). São identificados os sintagmas nominais e verbais que são as combinações de unidades linguísticas que podem formar orações (nominais ou verbais) e que apresentam características de prováveis conceitos.

A segunda etapa consiste em **determinar os conceitos** baseado em fontes de informação externas ao RISO-IC. São elas: WordNet, Wikipedia, Wiktionary[27], DBPedia, Conceptnet5[28], Verbosity e Reverb[29]. Estas fontes são estruturadas de maneira centralizada e uniforme favorecendo a utilização destas informações no processo de identificação de conceitos presentes nos documentos, bem como o adequado enriquecimento semântico.

Determinados sintagmas são passivos de mais de uma interpretação e os documentos precisam ser indexados ao conceito correto. Desta forma, é necessária a realização do processo de desambiguação, onde são analisadas as possíveis interpretações dos sintagmas identificados, juntamente com os trechos do documento onde são encontrados, para então, determinar conceitualmente, como o documento deverá ser indexado.

A terceira etapa consiste no **enriquecimento semântico** de conceitos. Os conceitos encontrados no texto e que possuem a correspondência exata de acordo com as fontes de informação utilizadas são enriquecidos diretamente sem a necessidade de nenhum processamento complementar. Para os conceitos parcialmente identificados onde apenas partes das palavras que o constitui estão presentes nas fontes de informação, é realizado um procedimento complementar calculando sua proximidade com conceitos presentes nas fontes selecionando o conceito mais próximo, estabelecendo uma relação de hiponímia/hiperonímia.

---

[24] Fornece informações conceituais, relações semânticas entre esses conceitos, tais como: sinônimos, hipônimos, hiperônimos, acrônimos, merônimos e polissemia, além de possíveis interpretações para um determinado vocábulo.

[25] É uma enciclopédia web, multilíngue, construída por meio da "*inteligência coletiva*", sem fins lucrativos e de propósito geral, onde encontram-se milhares de artigos sobre os mais variados assuntos.

[26] Fonte de julgamento popular e validação do sentido comum de expressões.

[27] https://pt.wiktionary.org/wiki/Wikcion%C3%A1rio:P%C3%A1gina_principal

[28] http://conceptnet5.media.mit.edu/

[29] http://reverb.cs.washington.edu/README.html

Na quarta fase ocorre a **atualização** do *thesaurus* com a inclusão do documento na biblioteca digital e atualização do *Topic Map* determinando todas relações entre os conceitos detectados com o novo documento.

5.3.2 RISO-IT (Indexação Temporal)

Para a indexação temporal são procurados os termos temporais contidos no texto (datas, tempos diretos e indiretos) para então determinar a possível temporalidade dos outros conceitos contidos no texto. São detalhadas as funcionalidades do RISO-TT (Temporal Tagger), desenvolvido por SANTOS (2013) e é realizada uma introdução ao RISO-GCT (Geração de Contextos Temporais) desenvolvido por este trabalho de dissertação de mestrado e que será detalhado no Capítulo 6 deste trabalho.

5.3.2.1 RISO-TT (Temporal Tagger)

O RISO-TT (Temporal Tagger) é o extrator de expressões temporais do RISO. Baseado em regras de associações de padrões temporais e inspirado no padrão TIMEX3 (SAQUETE, 2010), ele se difere das demais ferramentas de extração temporal por considerar signos e associações gramaticais mais complexas em seu processo de identificação das expressões temporais (SANTOS, 2013).

Como saídas do processamento de um documento pelo RISO-TT são gerados:

a) Documento marcado (TAG): É criado um documento marcado com as tags *RISOTime* e com o atributo *type* (Ex: *<RISOTime type=Pre-EBT>On September 1, 1939</RISOTime>*). O valor associado ao atributo *type* é o nome da regra da qual a expressão encontrada faz parte (SANTOS, 2013).
b) Vetor temporal (LIST): A partir do documento original é criado um outro documento contendo uma lista das expressões temporais encontradas no documento (Ex: EBT-N -> *from 499 to 493 BC*).
c) Vetor normalizado (NORM): É uma lista das expressões temporais encontradas no documento e seus valores normalizados (Ex: *On September 1, 1939 <--> 1-09-1939*). Os cálculos para o processamento do valor normalizado das expressões temporais do tipo data, hora e minutos são realizados por meio da função *date* da linguagem Python. Além disso, um grupo de expressões temporais complexas foi mapeado para que valores de intervalos fossem levados em consideração (Ex: *from May 10, 2010 to May 10, 2013 < --> 10/05/2010 < X < 10/05/2012*, onde *X* representa o intervalo temporal.

Um padrão, para o RISO-TT, é um conjunto de termos (temporais ou gramaticais) agrupados semanticamente, aos quais se possa atribuir valor a uma expressão temporal. Os padrões são utilizados na formação das regras. São padrões: preposições, advérbios, estações do ano, datas, horas e expressões regulares.

Uma regra é uma sequência ordenada de padrões (temporais e/ou nominais) que caracteriza a formação de expressões temporais. Uma regra considera a posição dos termos que formam uma expressão. Por exemplo, a regra Dia Mês Ano é diferente da regra Mês Dia Ano.

O RISO-TT foi projetado para se tornar uma ferramenta extensível e flexível. Isso significa dizer que, a qualquer momento um novo padrão temporal pode ser inserido às regras definidas no RISO-TT sem a necessidade de desenvolvimento, compilação de código ou adaptação de software de terceiros, bastando apenas que um usuário insira um novo padrão (ou regra, por exemplo) ao dicionário de regras do RISO-TT.

A Figura 9 ilustra os componentes do RISO-TT, que possui as funcionalidades TAG, LIST e NORM. A execução da função TAG resulta em um documento contendo marcações que indicam as expressões temporais presentes no documento original. A função LIST, por sua vez, resulta em um documento contendo a lista das expressões temporais encontradas no texto. Por fim, a função NORM, que resulta em um documento que irá conter a normalização das expressões temporais encontradas no documento original. Para o correto funcionamento das funções do RISO-TT é necessária a definição de padrões e regras que identificam as expressões temporais.

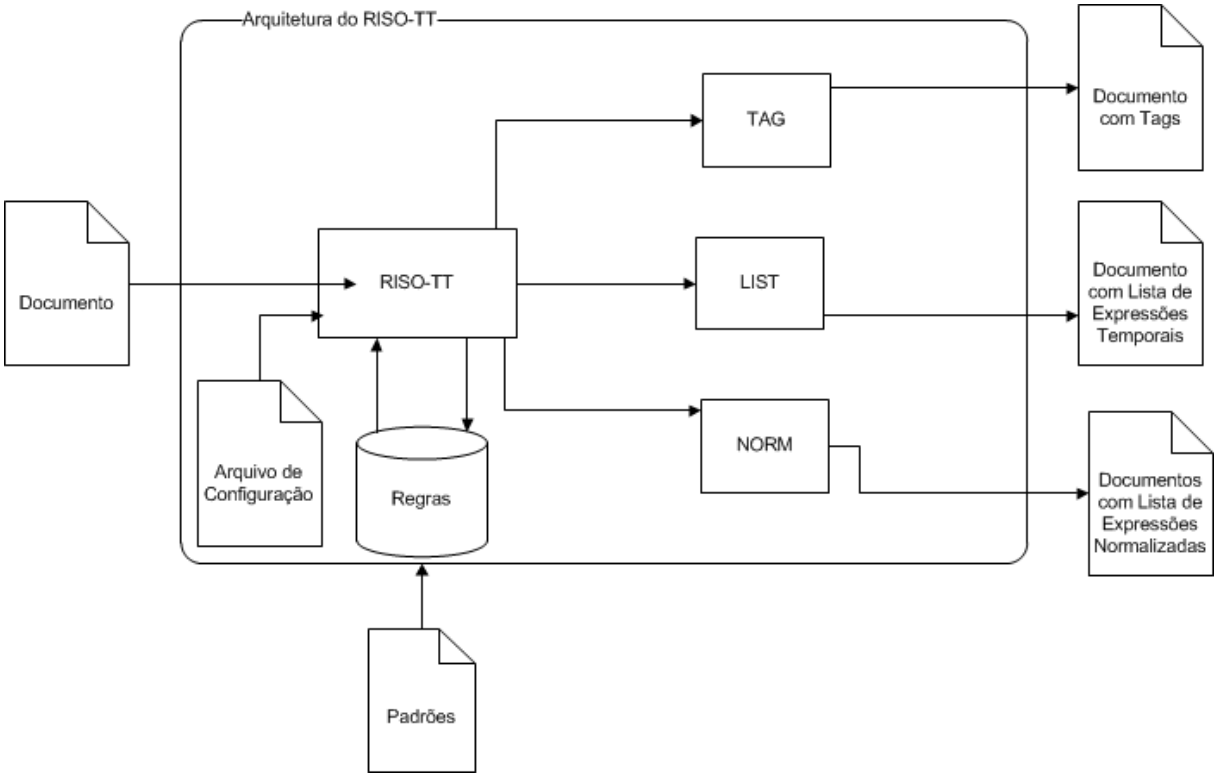


Figura 9 – Arquitetura geral do RISO-TT

O RISO-TT possui um arquivo de padrões, que consiste em um conjunto de termos (temporais ou gramaticais) agrupados semanticamente que possa atribuir valor a uma expressão temporal. São exemplos de padrões: preposições, advérbios, estações do ano, datas, horas e expressões regulares.

Na Tabela 5 são ilustrados alguns exemplos de padrões utilizados pelo RISO-TT.

Sigla	Descrição
I	Intervalos
EPT	Estrutura Pré-temporal
EBT	Estrutura Básica Temporal
Pre	Preposições
EMT	Estrutura Mínima Temporal
UT	Unidade Temporal
Adv	Adverbio
DE	Datas Especiais
H	Padrões de Hora
Dia	Padrões de Dia
Mês	Meses
Ano	Padrões para Ano

Tabela 5 – Exemplos de padrões utilizados pelo RISO-TT para a formação de regras (SANTOS, 2013)

```
1  <?xml version="1.0"?>
2  <simbolo nome="adverbio">
3      <expressao>"early"</expressao>
4      <expressao>"later"</expressao>
5      <expressao>"late"</expressao>
6      <expressao>"earlier"</expressao>
7      <expressao>"past"</expressao>
8      <expressao>"before"</expressao>
9      <expressao>"after"</expressao>
10 </simbolo>
```

Figura 10 – Exemplo de configuração do padrão "advérbio"

A Figura 10 ilustra a definição do padrão "advérbio", composta pelos termos gramaticais "*early*", "*later*", "*earlier*", entre outros.

Uma regra é uma sequência ordenada de padrões (temporais e/ou nominais) que caracteriza a formação de expressões temporais. Uma regra considera a posição dos termos que formam uma expressão. Por exemplo, a regra *Dia Mês Ano* é diferente da regra *Mês Dia Ano*.

Com o uso das regras, o RISO-TT permite que estruturas complexas entre as relações gramaticas e expressões temporais clássicas possam ser reconhecidas como uma única expressão, permitindo o reconhecimento dos mais diversos padrões. Com isso, expressões tais como "*from December 10, 2011 to December 10, 2012*" são classificadas como uma única expressão temporal e não como vários *tokens* temporais.

A Tabela 6 exemplifica as regras que são utilizadas pelo RISO-TT para o reconhecimento de expressões temporais.

Regras	Descrição
EPT-EBT-Adv	estrutura_pre_temporal estrutura_basica_temporal adverbio
EPT-EBT-UT	estrutura_pre_temporal estrutura_basica_temporal unidade_temporal
EPT-UT-Adv	estrutura_pre_temporal unidade_temporal adverbio
Pre-EBT-Adv	preposicoes estrutura_basica_temporal adverbio
Pre-EBT-UT	preposicoes estrutura_basica_temporal unidade_temporal
EPT-EMT-Adv	estrutura_pre_temporal estrutura_minima_temporal adverbio
EPT-EMT-UT	estrutura_pre_temporal estrutura_minima_temporal unidade_temporal
Pre-EA-A	preposicoes estacao_do_ano ano

Tabela 6 – Exemplos de regras utilizadas pelo RISO-TT (SANTOS, 2013)

A Tabela 7 ilustra as expressões que definem ***estrutura_minima_temporal*** e ***estrutura_basica_temporal,*** um dos padrões utilizados na formação das regras descritas na Tabela 6.

Padrão	
estrutura_minima_temporal	<simbolo nome="estrutura_minima_temporal"> <expressao>ano"-"ano</expressao> <expressao>ano "and" ano</expressao> <expressao>ano"-"dia</expressao> <expressao>ano"s"</expressao> <expressao>ano</expressao> </simbolo>
estrutura_basica_temporal	<simbolo nome="estrutura_basica_temporal"> <expressao>mes dia"," ano</expressao> <expressao>mes dia "," ano</expressao> <expressao>mes dia ano</expressao> <expressao>mes dia "of" ano</expressao> <expressao>mes dia "-" mes dia ano</expressao> [...] </simbolo>

Tabela 7 – Alguns dos padrões utilizados pelo RISO-TT para definição de regras.

Na Figura 11 é ilustrado um trecho da definição das regras utilizadas pelo RISO-TT para o reconhecimento das expressões temporais. Onde cada linha possui um tipo de regra seguido dos padrões agrupados para formação da respectiva regra. E.g., a regra "*Pre-EBT*" que é composta

pelos padrões "*preposicoes*" e "*estrutura_basica_temporal*" em sequência identifiria a expressão temporal "*On 7 March*" composta pela preposição "*On*" e pela estrutura básica temporal "*7 March*".

```
<?xml version="1.0"?>
<regras>
    <simbolo nome="expressao_temporal">
        <expressao tipo="I">intervalos</expressao>
        <expressao tipo="Pre-A-Pre-A">preposicoes ano preposicoes ano</expressao>
        <expressao tipo="EPT-EBT-Adv">estrutura_pre_temporal estrutura_basica_temporal adverbio</expressao>
        <expressao tipo="EPT-EBT-UT">estrutura_pre_temporal estrutura_basica_temporal unidade_temporal</expressao>
        <expressao tipo="EPT-UT-Adv">estrutura_pre_temporal unidade_temporal adverbio</expressao>
        <expressao tipo="Pre-EBT-Adv">preposicoes estrutura_basica_temporal adverbio</expressao>
        <expressao tipo="Pre-EBT-UT">preposicoes estrutura_basica_temporal unidade_temporal</expressao>
        <expressao tipo="EPT-EMT-Adv">estrutura_pre_temporal estrutura_minima_temporal adverbio</expressao>
        <expressao tipo="EPT-EMT-UT">estrutura_pre_temporal estrutura_minima_temporal unidade_temporal</expressao>
        <expressao tipo="Pre-EA-A">preposicoes estacao_do_ano ano</expressao>
        <expressao tipo="EPT-EBT">estrutura_pre_temporal estrutura_basica_temporal</expressao>
        <expressao tipo="Pre-EBT">preposicoes estrutura_basica_temporal</expressao>
        <expressao tipo="Pre-EMT">preposicoes estrutura_minima_temporal</expressao>
        <expressao tipo="EPT-UT">estrutura_pre_temporal unidade_temporal</expressao>
        <expressao tipo="EPT-EMT">estrutura_pre_temporal estrutura_minima_temporal</expressao>
        <expressao tipo="Pre-A-Adv">preposicoes ano adverbio</expressao>
        <expressao tipo="EBT-UT">estrutura_basica_temporal unidade_temporal</expressao>
        <expressao tipo="EBT-Adv">estrutura_basica_temporal adverbio</expressao>
        <expressao tipo="D-EMT-Adv">dia unidade_temporal adverbio</expressao>
        <expressao tipo="D-EMT-Adv">adverbio dia unidade_temporal</expressao>
        <expressao tipo="Pre-A-Adv">preposicoes ano adverbio</expressao>
        <expressao tipo="EPT-EN">estrutura_pre_temporal estacao_do_ano</expressao>
        <expressao tipo="Pre-A">preposicoes ano</expressao>
```

Figura 11 – Exemplo de definição de regras utilizadas pelo RISO-TT

Uma estrutura mínima temporal é um padrão temporal que pode ser uma data ou um intervalo onde ambas não estão completas, ou seja não possuem todas as informações necessárias (dia, mês e ano) necessárias para mapear corretamente a data ou intervalo à qual a expressão se refere.

Uma Estrutura Básica Temporal, consiste em um padrão temporal que pode ser uma data ou um intervalo que contém todas as informações necessárias (dia, mês e ano), podendo ser mapeada corretamente a uma data ou intervalo à qual ela se refere.

O RISO-TT possui as seguintes funcionalidades:

a) **TAG:** Um documento sem marcação é passado como parâmetro de entrada, é processado e a saída é um documento marcado com as *tags* do RISOTime e com o atributo *type* (Ex: *<RISOTime type=Pre-EBT>On September 1, 1939</RISOTime>*). O valor atribuído ao atributo *type* é o nome da regra da qual a expressão encontrada faz parte.

b) **LIST:** Um documento sem marcação é passado como parâmetro de entrada, é processado e a saída é um documento com uma lista de expressões temporais encontradas no documento (Ex: EBT-N -> *from 499 to 493 BC*).

c) **NORM:** Um documento sem marcação é passado como parâmetro de entrada, é processado e a saída é um documento com uma lista de expressões temporais encontradas no

documento e seus valores normalizados (Ex: *On September 1, 1939 <--> 1-09-1939*). Os cálculos numéricos para o processamento do valor normalizado das expressões temporais do tipo data, hora e minutos são realizados por meio da função *date*[30] da linguagem Python, que realiza essa normalização de forma transparente. Além disso, um grupo de expressões temporais complexas foi mapeado para que, valores de intervalos, fossem calculados (Ex: *from May 10, 2010 to May 10, 2013 < --> 10/05/2010 > X < 10/05/2012*, onde *X* representa o intervalo temporal.

Na seção 5.3.2.2 são mostradas as evoluções aplicadas no RISO-TT necessárias para a realização deste trabalho de dissertação de mestrado.

5.3.2.2 Evoluções no RISO-TT

Foram necessárias evoluções na função NORM, componente desenvolvido por SANTOS (2013) para que fosse possível a normalização de expressões que não eram normalizados pelo RISO-TT, e consequentemente, enriquecer os resultados gerados pelo RISO-GCT. Estas expressões são caracterizadas pelas regras listadas na Tabela 8:

Regras	Componentes	Exemplo
D-EMT-Adv	qtd. dia, mês ou ano + unidade_temporal + advérbio	*"twelve days later"*
DE	datas_especiais (datas ou períodos especiais)	*"Christmas", "2014 FIFA World Cup"*
EMT	estrutura_minima_temporal	*"2009-12"*
Pre-EMT	preposição estrutura_minima_temporal	*"in 1939"*
Adv-DE	adverbio qtd. unidade_temporal datas_especiais qtd. unidade_temporal adverbio datas_especiais	*"2 days after Christmas, 1950"*
Pre-EMT-EMT	preposição estrutura_minima_temporal- estrutura_minima_temporal	*"from 50s-60s"*
Pre-EBT	preposição estrutura-básica_temporal	*"Nearly 01 October"*
I	intervalos	*"from 01 June, 2015 to 31 June, 2015"*
Pre-EMT-EBT	preposição estrutura_minima_temporal- estrutura_basica_temporal	*"from 01 May to 05 May 2006"*

Tabela 8 – Regras de formação de expressões temporais que não eram normalizadas pelo RISO-TT

[30] Função *Date* da Linguagem Python: http://pythonhelp.wordpress.com/2012/07/10/trabalhando-com-datas-e-horas-em-python-datetime/

Apesar do RISO-TT reconhecer as regras presentes na Tabela 8 na função TAGGER responsável por marcar as expressões temporais presentes no documento (com exceção das datas especiais), não era possível a normalização destas regras a partir da função NORM. Surgiu então a necessidade de se realizar melhorias nesta funcionalidade.

Conforme ilustrado na Tabela 8, neste trabalho o RISO-TT passou a normalizar novas expressões temporais, que são detalhadas a seguir.

Regra D-EMT-Adv

A regra D-EMT-Adv possui os seguintes componentes:

a. Quantidade (E.g., "*one*", "*two*", "*three*")
b. Indicador Temporal (E.g., "*days*", "*months*", "*years*", *"hours*")
c. Advérbio Temporal E.g., ("*before*", "*after*", "*post*")

E.g., "*two months before*"

Identificados estes componentes é realizada a busca por alguma expressão temporal imediatamente anterior à expressão temporal a qual se deseja encontrar a data à qual a mesma se refere, primeiramente na própria frase e, caso não seja encontrada, busca-se na frase anterior, e assim sucessivamente, até que seja encontrada alguma expressão temporal (data específica ou evento). Caso seja encontrada alguma expressão, esta data terá a quantidade de dias e/ou meses e/ou anos adicionados ou subtraídos da data encontrada dependendo do advérbio presente na expressão (E.g., "*after*", "*before*"). O resultado desta operação será uma data que será atribuída à expressão temporal formada pela regra D-EMT-Adv. Caso o indicador temporal possua dimensão menor que dia, mês ou ano (E.g., "*hour*", "*minute*", "*second*", "*miliseconds*"), nenhuma operação é realizada, e a data normalizada encontrada na frase imediatamente anterior será a normalização atribuída à expressão temporal formada pela regraD-EMT-Adv.

Caso não seja encontrada nenhuma expressão temporal, será considerada a data de criação do documento.

Exemplo:

"*Darryl bought his current home on January 5, 2013 and married two months later.*"

Nesta frase ocorre a expressão temporal "*two months later*" que deve ser associada a '*married*'. Para encontrar a data à qual a expressão de refere, é necessário buscar alguma outra expressão que ocorreu anteriormente no texto, que neste exemplo é a expressão "*January 5, 2013*". Consequentemente, para obter a data correta que a expressão temporal "*two months later*" se refere, é necessário fazer a adição de 2 meses à data anterior, resultando-se em "*March 5, 2013*".

O cálculo realizado acima vai depender da unidade temporal presente na expressão que vai indicar uma quantidade "*x*" de dias, meses ou anos. Adicionalmente vai depender da categoria

à qual o advérbio está classificado, uma vez que existem advérbios que indicam anterioridade ou posterioridade, conforme descrito na Tabela 9.

Categoria	Advérbios
Advérbios que indicam anterioridade	before, early, earlier
Advérbios que indicam posterioridade	after, late, later

Tabela 9 – Tabela indicativa dos advérbios temporais

Regra DE

Datas Especiais são entidades ou eventos que possuem uma data ou mais datas atreladas. Para o mapeamento de expressões temporais formados pela regra DE, é realizada uma consulta na base de dados criada pelo RISO-RID (vide seção 6.2.1.2) que irá buscar informações temporais provenientes da ***DBPedia*** referente à esta data especial (mais detalhes na seção 6.2.1.2).

Regras EMT e Pre-EMT

Para o mapeamento destas regras, é verificada a estrutura da expressão temporal.

a. Caso tenhamos uma expressão que possua mês (MM) e ano (YYYY) e não tenha dia (DD), esta expressão será mapeada para uma data igual a [?]-MM-YYYY
b. Caso tenhamos uma expressão que possua apenas o ano (YYYY), esta expressão será mapeada para uma data igual a [?]-[?]-YYYY.
c. Caso tenhamos uma expressão que possua apenas o mês (MM), é buscado o ano [YYYY] em parágrafos anteriores. O ano considerado nesta normalização será o ano encontrado na frase imediatamente anterior, caso seja encontrado. Caso não seja encontrado esta expressão será mapeada para uma data igual a [?]-[?]-YYYY.
d. Caso tenhamos uma expressão temporal que possua apenas o dia (DD), são buscados mês (MM) e ano (YYYY) em parágrafos anteriores. O ano e/ou mês considerado nesta normalização será o ano e/ou mês encontrado na frase imediatamente anterior, caso seja encontrado. Caso não sejam encontrados esta expressão será mapeada para uma data igual a DD-[?]-[?].

Regra Pre-EMT-EBT

Nesta regra existe uma preposição seguida de uma estrutura mínima temporal (dia + mês) e uma estrutura básica temporal (dia + mês + ano).

Neste caso, é assumido que ambas as datas se possuem o mesmo valor referente ao ano. Desta forma, o exemplo "*from 01 May to 05 May 2006*" da Tabela 7 ao ser normalizado, resulta em "*from 01 May 2006 to 05 May 2006*".

5.3.2.3 RISO-GCT (Geração de Contextos Temporais)

O RISO-GCT foi desenvolvido neste trabalho de dissertação de mestrado, e é responsável pela geração de relacionamentos entre conceitos presentes em um documento e as expressões temporais. É composto por 3 módulos: o RISO-UDM, o RISO-RID e o RISO-RCT.

O RISO-UDM (Unifica Documentos Marcados) é responsável por gerar um arquivo contendo a unificação das saídas geradas pelo *PoS-Tagging* do RISO-VTD e o RISO-TT, que contém, respectivamente, as classificações gramaticais dos sintagmas e as marcações referentes às expressões temporais presentes no texto.

O RISO-RID (Recuperação de Informações provenientes da DBPedia) é responsável por recuperar as informações temporais presentes na DBpedia para cada uma das entidades extraídas pelo RISO-VTD. Como resultado da execução deste módulo é criada uma base de dados local que será utilizada tanto pela função NORM do RISO-TT, quanto pelo RISO-RCT.

O RISO-RCT (Relaciona Conceitos e Tempos) é responsável por realizar a leitura do arquivo com as saídas unificadas do RISO-TT e RISO-VTD e gerar relacionamentos entre conceitos e seus respectivos contextos temporais. Estes relacionamentos serão incluídos no Topic Map previamente gerado pelo RISO-IC.

Mais detalhes sobre o funcionamento do RISO-GCT são apresentados no Capítulo 6 deste documento.

5.3.3 RISO-IE (Indexação Espacial)

O RISO-IE é o módulo responsável por realizar a Indexação Espacial de documentos, ainda está em desenvolvimento e possuirá 2 componentes.

Extrator de Termos Espaciais: Será responsável por extrair a localização geográfica dos conceitos presentes no texto. Este módulo terá como base o algoritmo utilizado pelo GeoSEn, proposto por CAMPELO (2008) para identificar e extrair de um texto o máximo de dados que possuam potencial de serem convertidos em informações geográficas.

Gerador de Contextos Espaciais: Será responsável por utilizar os termos espaciais determinados na fase anterior para determinar contextos espaciais de outros conceitos presentes no texto.

Após a aplicação dos três componentes do RISO-IS (Indexação Conceitual, Indexação Temporal e Indexação Espacial), será possível determinar, para os termos contidos em um documento seus três contextos: Conceito, Tempo e Espaço. Se o formato completo de um conceito for da forma: < TERMO:(DOMÍNIO, ESPAÇO, TEMPO) > o termo "*Waterloo*" poderia ficar como <Waterloo: (Batalha; 18/06/1815; loc (Waterloo:Cidade - Bélgica)) >.

5.4 RISO-CS (Consulta Semântica)

Trata-se de um *front-end* ainda em desenvolvimento cuja finalidade é possibilitar consultas inteligentes a uma biblioteca digital indexada pelo RISO.

A partir de uma entrada inicial dada pelo usuário, a interface de comunicação realizará uma busca no *Thesaurus* por todos os conceitos que tenha alguma relação (sintática ou semântica) com o termo definido e trará opções de desambiguação – conceitual, espacial e temporal - e enriquecimento semântico como sugestão. Com base no contexto determinado dos termos da consulta, o usuário poderá determinar, com precisão, a real necessidade e informação fazendo com que os resultados estejam de acordo com essa necessidade. A arquitetura do RISO-CS é descrita no item 3 da Figura 7.

5.5 Conclusão

O projeto RISO tem como objetivo converter a indexação e recuperação da informação contida em documentos textuais de seu atual foco em aspectos sintáticos dos termos para levar em consideração seu significado semântico. Para cada termo ou sintagma encontrado é determinado seu contexto conceitual, temporal e espacial. Um termo contextualizado poderá ser inserido corretamente um uma rede semântica linguística. Esta mesma delimitação contextual é determinada na formulação da consulta à base de documentos. O conceito definido na consulta poderá, por sua vez, ser expandido por outros conceitos linguisticamente relacionados a ele. Essa abordagem tem como objetivo melhorar significativamente tanto a cobertura (pela expansão linguística) como a precisão (pela desambiguação) da recuperação de documentos.

As informações temporais presentes nos textos são identificadas e normalizadas através do módulo RISO-TT. No entanto, existem alguns padrões de expressões temporais que não haviam sido mapeados pela função do RISO-TT responsável por normalizar as expressões temporais presentes, e consequentemente não são normalizados. Desta forma, foram necessárias evoluções neste sistema, de modo a incluir a funcionalidade de normalização de alguns dos padrões não mapeados.

Adicionalmente, como o módulo RISO-IC não inclui informações temporais na sua etapa de indexação de documentos, foi desenvolvido durante este trabalho de dissertação um novo componente do RISO, o RISO-GCT, responsável por recuperar e incluir nos Thesaurus gerado pelo RISO-IC informações temporais relacionadas aos conceitos presentes no texto, conceitos estes identificados previamente pelo RISO-VTD.

No Capítulo 6 será apresentado em detalhes o RISO-GCT, componente de determinação do contexto temporal dos conceitos identificados em textos será apresentado em detalhes.

Capítulo 6 - RISO-GCT – Sistema de Geração de Contextos Temporais

Neste capítulo é apresentada a arquitetura, estrutura e as características do módulo RISO Geração de Contextos Temporais (RISO-GCT).

6.1 Introdução

Com o objetivo de possibilitar a extração de conceitos temporalizados em textos, foi criado uma extensão do RISO, denominado RISO-GCT (Geração de Contextos Temporais) responsável por determinar para cada conceito de um texto uma tupla $<C_n, T_n>$ e incluí-los em um Topic Map, onde:

- C_n: Consiste em um conceito extraído do texto analisado que pode ser uma propriedade (aqui chamado de entidade), evento ou processo (ALLEN, 1968).
- T_n: É uma expressão temporal que consiste em um ou mais intervalos ou instantes na linha do tempo que representa o tempo no qual o conceito C_n foi válido ou ocorreu.

Para conseguir este objetivo, para cada frase presente no texto, é necessário obtermos os conceitos presentes nesta frase, juntamente com as possíveis expressões temporais também contidas na frase. Expressões temporais são sintagmas que expressam a ideia de tempo.

Estas expressões temporais podem descrever um valor **explícito** ou **implícito** (SANTOS, 2013).

Entende-se como Expressões Temporais Explícitas aquelas que para as quais não é necessário analisar informações adicionais para obter o valor normalizado[31] da data à qual a expressão se refere. Exemplos: "*February 24th, 2009*", "*10-12-2015*", "*01/01/1994*", "*February 24th, 2009 to March 12th, 2010*". Note que estas expressões temporais estão completas, ou seja, todas possuem valores referentes a dia, mês e ano.

Expressões Temporais Implícitas são aquelas em que, para obtermos o tempo ao qual a expressão se refere, é necessário obter informações presentes no próprio texto como expressões temporais presentes em frases/orações anteriores, como nos exemplos "*May, 05*", "*two days after*" ou "*at the begining of June*" quando é necessário que seja verificada a informação que falta em frases anteriores do texto ou informações temporais relacionadas à algum conceito presente na

[31] Valor da data em um formato específico. Exemplo: DD-MM-YYYY onde DD se refere ao dia, MM se refere ao mês e YYYY se refere ao ano.

oração. Quando não for encontrado esta informação pode ser útil a data de criação do documento. Também são Expressões Temporais Implícitas aquelas que para as quais é necessário recuperar informações presentes em fontes externas como a *DBPedia* como no exemplo: "*two days before Christmas 1994*", onde é necessário recuperar na *DBPedia* a data referente à expressão "*Christmas 1994*".

Obtidas estas informações, cada frase é dividida em orações, a partir das classes gramaticais de cada palavra contida na frase. Posteriormente é feita a associação entre os termos e as expressões temporais presentes na própria oração ou em orações correlacionadas na mesma frase (explícitas ou implícitas). Por exemplo, na frase "*The next John Mayer concert in Brazil will be two days after Christmas 2015*", "*John Mayer*" é um conceito ao qual será associado à data a qual a expressão temporal "*two days after Christmas 2015*" se refere, ou seja, "*December 27, 2015*".

Caso alguma oração possua um termo que esteja relacionado à alguma expressão temporal presente em outra oração, estes também podem ser relacionados. Por exemplo, na frase "*After drink a few beers, Marcelo left the bench and scored the league title the goal, that fateful day May 1, 2015*." existem 3 orações: "*After drinking a few beers"*, "*Marcelo left the bench*" e "*scored the league title the goal, that fateful day May 1, 2015*". A entidade "*Marcelo*" está relacionada à data "*May 1, 2015*" pois, apesar da entidade "*Marcelo*" e a data não ocorrerem na mesma oração, elas ocorrem em orações vizinhas e a oração que contem a data não possui nenhuma outra entidade.

Também pode haver casos em que a frase de um termo não contenha nenhuma expressão temporal. Neste caso, assim como na fase de normalização de expressões temporais implícitas, é necessário buscar em fonte externa (DBPedia) as datas relacionadas ao termo. E.g.,: data de nascimento e falecimento (para pessoas), data de fundação e data de fechamento (para empresas, instituições, entre outras.), data de inauguração e data de desaparecimento (para monumentos, entre outras.), data de início e fim (para eventos).

O RISO-GCT se difere das demais ferramentas pelo fato de determinar o contexto temporal de entidades e eventos presentes no texto analisado e incluí-las no Topic Map.

6.2 RISO-GCT

As seções 6.2.1.1, 6.2.1.2 e 6.2.1.3 detalham, respectivamente, o funcionamento dos componentes do RISO-GCT (Geração de Contextos Temporais), principais contribuições e limitações da ferramenta.

6.2.1 Arquitetura

O RISO-GCT (Geração de Contextos Temporais) é composto por 3 componentes:

- RISO-UDM (Unifica Documentos Marcados);

- RISO-RID (Recuperação de Informações provenientes da DBPedia);
- RISO-RCT (Relaciona Conceitos e Tempos).

A Figura 12, ilustra os componentes do RISO-GCT, que serão detalhadas nas seções 6.2.1.1, 6.2.1.2 e 6.2.1.3.

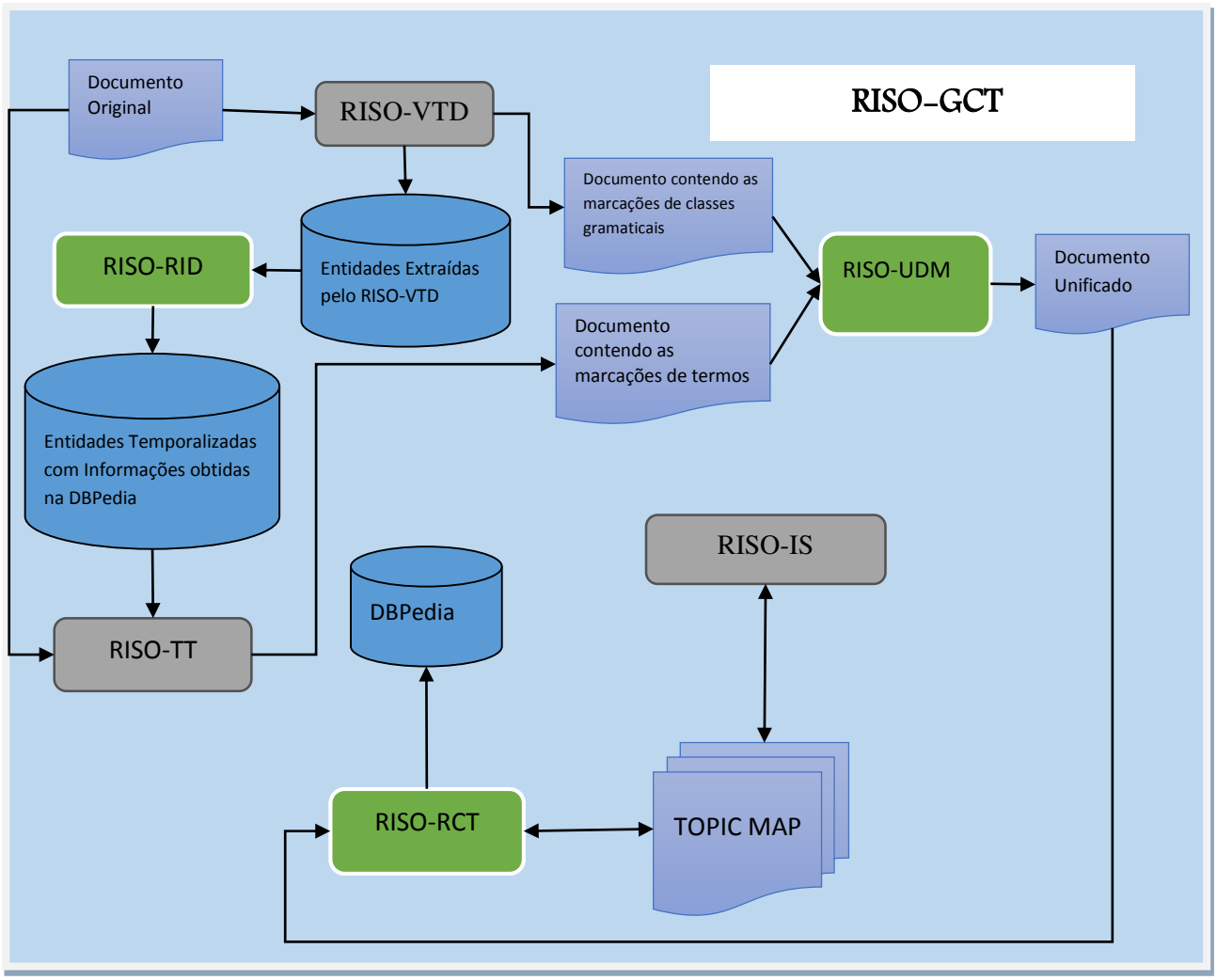


Figura 12 – Estrutura Geral do RISO-GCT

Foi escolhida a linguagem de programação *Java*[32], devido a familiaridade com a linguagem e a portabilidade que esta linguagem oferece. A base de dados escolhida foi o Postgres[33], pela facilidade em armazenar formatos de dados estruturados, e pelo fato de os demais módulos do Projeto RISO já utilizarem este banco de dados.

O primeiro componente, chamado de RISO-UDM (Relaciona Documentos Marcados) é responsável por gerar um arquivo que irá conter o texto marcado com as informações temporais

[32] https://www.java.com/pt_BR/about/
[33] http://www.postgresql.org/.

resultantes da execução do RISO-TT e as informações resultantes da execução do *POS-Tagging* do RISO-VTD que possuem classificações gramaticais das palavras presentes no texto.

O segundo componente, o RISO-RID (Recuperação de Informações provenientes da DBPedia), recebe como entrada as entidades extraídas pelo RISO-VTD e, a partir disto, realiza a busca das datas relacionadas à cada uma destas entidades na DBPedia.

Para o correto funcionamento de primeiro componente é mandatório que, antes da execução do RISO-UDM, tenham sido executados o RISO-VTD e o RISO-TT.

O terceiro componente, RISO-RCT (Relaciona Conceitos e Tempos), recebe como entrada o arquivo contendo as informações unificadas geradas pelo RISO-UDM. Posteriormente, o RISO-RCT cria relacionamentos entre informações temporais e as entidades presentes no texto, utilizando para isto, informações presentes no próprio texto que está sendo analisado ou em fonte externa (DBPedia).

Nas seções 6.2.1.1, 6.2.1.2 e 6.2.1.3, estes componentes serão mais detalhados.

6.2.1.1 RISO-UDM (Unifica Documentos Marcados)

O RISO-UDM (Unifica Documentos Marcados) gera um arquivo contendo marcações específicas para os termos temporais e outros termos candidatos a serem temporalizados. Para a geração deste arquivo, é necessária a execução dos módulos RISO-TT e do componente de *POS-*

```
The settlements at Tilsit gave Napoleon time to organize his empire. One
of his major objectives became enforcing the Continental System against
the British. He decided to focus his attention on the Kingdom of Portugal,
which consistently violated his trade prohibitions. After defeat in the
War of the Oranges in 1801, Portugal adopted a double-sided policy. At
first, John VI agreed to close his ports to British trade. The situation
changed dramatically after the Franco-Spanish defeat at Trafalgar; John
grew bolder and officially resumed diplomatic and trade relations with
Britain.
```

Figura 13 – Texto original recebido como entrada pelo RISO-TT e RISO-VTD
Tagging do RISO-VTD.

A Figura 13 mostra um exemplo de um texto original a ser processado. A Figura 14 por sua vez, mostra o resultado do processamento do texto original pelo RISO-TT, que insere as marcações das expressões temporais identificadas no texto. Na Figura 15 é ilustrada a saída gerada pelo RISO-VTD contendo as marcações das classes gramaticais dos sintagmas presentes no texto original. Estes arquivos serão unificados pelo RISO-UDM (Figura 16), com o intuito de gerar um arquivo consolidado contendo as marcações dos sistemas RISO-TT e RISO-VTD que será lido pelo RISO-RCT.

```
The settlements at <RISOTime type=DE>Tilsit</RISOTime> gave <RISOTime
type=DE>Napoleon</RISOTime> time to organize his empire. One of his major
objectives became enforcing the Continental System against the British.
He decided to focus his attention on the Kingdom of <RISOTime
type=DE>Portugal</RISOTime>, which consistently violated his trade
prohibitions. After defeat in the War of the <RISOTime
type=DE>Oranges</RISOTime> <RISOTime type=Pre-EMT>in 1801</RISOTime>,
<RISOTime type=DE>Portugal</RISOTime> adopted a double-sided policy. At
first, John VI agreed to close his ports to British trade. The situation
changed dramatically after the Franco-Spanish defeat at Trafalgar; John
grew bolder and officially resumed diplomatic and trade relations with
Britain.
```

Figura 14 – Saída gerada pelo RISO-TT

```
The/DT settlements/NNS at/IN Tilsit/NNP gave/VBD DATA_RISO/NNP time/NN
to/TO organize/VB his/PRP$ empire/NN ./. One/CD of/IN his/PRP$ major/JJ
objectives/NNS became/VBD enforcing/VBG the/DT Continental/NNP System/NNP
against/IN the/DT British/JJ ./. He/PRP decided/VBD to/TO focus/VB
his/PRP$ attention/NN on/IN the/DT Kingdom/NNP of/IN Portugal/NNP :/:
which/WDT consistently/RB violated/VBD his/PRP$ trade/NN prohibitions/NNS
./. After/IN defeat/NN in/IN the/DT War/NNP of/IN the/DT Oranges/NNPS
DATA_RISO/NNP :/: Portugal/NNP adopted/VBD double/NN sided/VBD policy/NN
./. At/IN first/JJ :/: John/NNP VI/NNP agreed/VBD to/TO close/VB his/PRP$
ports/NNS to/TO British/JJ trade/NN ./. The/DT situation/NN changed/VBN
dramatically/RB after/IN the/DT Franco/NNP Spanish/JJ defeat/NN at/IN
Trafalgar/NNP ;/: John/NNP grew/VBD bolder/JJR and/CC officially/RB
resumed/VBD diplomatic/JJ and/CC trade/NN relations/NNS with/IN
Britain/NNP ./.
```

Figura 15 – Saída gerada pelo *POS-Tagging*

A saída resultante deste pré-processamento é um texto anotado com as TAGs que indicam as classes gramaticais de cada *token*, juntamente com as TAGs que indicam as expressões temporais encontradas no texto, conforme é ilustrado na Figura 16.

Ao final, apenas as entidades cujas datas correspondentes foram localizadas na DBPedia são marcadas como sendo expressões do tipo DE.

```
The/DT settlements/NNS at/IN <RISOTime_type=DE>Tilsit</RISOTime>
gave/VBD <RISOTime_type=DE>Napoleon</RISOTime> time/NN to/TO organize/VB
his/PRP$ empire/NN ./. One/CD of/IN his/PRP$ major/JJ objectives/NNS
became/VBD enforcing/VBG the/DT Continental/NNP System/NNP against/IN
the/DT British/JJ ./. He/PRP decided/VBD to/TO focus/VB his/PRP$
attention/NN on/IN the/DT Kingdom/NNP of/IN
<RISOTime_type=DE>Portugal</RISOTime> :/: which/WDT consistently/RB
violated/VBD his/PRP$ trade/NN prohibitions/NNS ./. After/IN defeat/NN
in/IN the/DT War/NNP of/IN the/DT <RISOTime_type=DE>Oranges</RISOTime>
<RISOTime_type=Pre-EMT>in_1801</RISOTime> :/:
<RISOTime_type=DE>Portugal</RISOTime> adopted/VBD double/NN sided/VBD
policy/NN ./. At/IN first/JJ :/: John/NNP VI/NNP agreed/VBD to/TO close/VB
his/PRP$ ports/NNS to/TO British/JJ trade/NN ./. The/DT situation/NN
changed/VBN dramatically/RB after/IN the/DT Franco/NNP Spanish/JJ
defeat/NN at/IN Trafalgar/NNP ;/: John/NNP grew/VBD bolder/JJR and/CC
officially/RB resumed/VBD diplomatic/JJ and/CC trade/NN relations/NNS
with/IN Britain/NNP ./.
```

Figura 16 – Saída Unificada gerada pelo RISO-UDM

6.2.1.2 RISO-RID (Recuperação de Informações provenientes da DBPedia)

Nesta etapa, a partir dos termos extraídos do documento pelo RISO-VTD, serão recuperadas informações temporais relacionadas a elas na *DBPedia*. Ou seja, para cada termo extraído pelo RISO-VTD são feitas consultas na base de dados da *DBPedia* utilizando SPARQL[34], uma linguagem estruturada para expressar consultas em diversas fontes de dados armazenados como RDF (Resource Description Framework)[35], um modelo padrão para intercâmbio de dados na Web.

Para obter as informações temporais são utilizados atributos adequados dos conceitos presentes na *DBPedia* conforme ilustrado na Tabela 10.

Entidade	Informações Temporais obtidas na DBPedia	Predicado
Instituição	Data de Fundação	Established Date
Cidade	Data de Fundação	FoundingDate CurrentCatDate (Status de cidade desde)
Empresa	Data de Fundação	FoundingDate
Estado	Data de Fundação	Admittance
Evento	Período	Period Date (Start, Finish)
Feriado	Data	Date
Local	Data de Inauguração	Opening Date Openned Created Opening Start Date Completion Date
País		Established Date Founding Date
Pessoa	Data de Nascimento/Falecimento	Birth Date Death Date
Pessoa Cargo	Data de Início/Fim na qual exerceu algum cargo	Active Years End Date Active Years Start Date
	Ano de Início/Fim na qual exerceu algum cargo	Active Years Start Year Active Years End Year
	Período de Reinado	Reign

Tabela 10 – Datas a serem recuperadas de acordo com a entidade/evento

Uma das limitações desta etapa se deve ao fato de não ser feita a etapa de desambiguação. Desta forma, as datas associadas aos sintagmas são verificadas de acordo com a ordem descrita na Tabela 10. Ou seja, são realizas consultas seguindo esta sequência, e caso seja encontrada

[34] http://pt.dbpedia.org/sparql
[35] http://www.w3.org/RDF/

alguma informação temporal durante a consulta de um tipo específico de entidade (E.g., "*instituição*", "*cidade*", "*empresa*"), considera-se como sendo a informação temporal atribuída ao sintagma e são desconsideradas as consultas subsequentes.

Ao fim desta etapa é criada uma base de dados contendo as informações extraídas para cada uma das entidades presentes no texto de acordo com a Tabela 11.

data	nome_entidade
15-08-1769 < X < 05-05-1821	Napoleon
[?]-[?]-1552	Tilsit
[?]-5-2010 < X < [?]-05-2010	Tilsit
25-04-1974	Portugal
[?]-11-2012 < X < [?]-11-2012	Oranges

Tabela 11 – Exemplo de informações extraídas da DBPedia e inseridos na base de dados do RISO

Na base de dados do RISO serão incluídas todas as datas recuperadas pelo processamento do RISO-RCT. Como no exemplo ilustrado na Figura 17, para a entidade "Napoleon" a informação incluída na base de dados é referente a junção de duas informações presentes na DBPedia, a data de nascimento e a data de morte do imperador Napoleão Bonaparte.

Figura 17 – Informações recuperadas na DBPedia para o sintagma "*Napoleon*"

De maneira simular na Figura 18, para a entidade "Tilsit" a informação incluída na base de dados é referente às duas informações temporais encontradas na DBPedia (data do último senso e data na qual a a cidade de "*Tilsit*" conquistou o status de cidade).

Figura 18 – Informações recuperadas na DBPedia para o sintagma "Tilsit"

6.2.1.3 RISO-RCT (Relacionamento de Conceitos e Tempos)

Este componente é responsável por realizar a associação entre as entidades e as expressões temporais presentes no texto. Estes relacionamentos são incluídos no Topic Map do documento criado pelo RISO-IS. As etapas necessárias para realização desta atividade são descritas nas seções 6.2.1.3.1, 6.2.1.3.2 e 6.2.1.3.3.

A formalização do algoritmo implementado pelo RISO-RCT é mostrada no Apêndice A deste trabalho. No Apêndice B, por sua vez, são mostradas as funções utilizadas pelo algoritmo implementado no RISO-RCT.

6.2.1.3.1 Divisão da Frase em Orações

Orações são sequencias de palavras que, estruturadas em torno de um verbo, formam uma mensagem. Se não houver um verbo, a estrutura linguística não será uma oração[36]. As frases são divididas de maneira que cada trecho da frase contenha um verbo, acompanhado ou não de sujeito e/ou predicado.

O processo de divisão das frases em orações inicia-se buscando por conjunções, previamente identificadas pelo *POS-Tagging* do RISO-VTD e sinais de pontuação (vírgula e ponto-e-vírgula). Estes sinais e estas conjunções delimitam o conjunto de blocos nos quais a frase será dividida. Por exemplo, a frase "*Jessica ran for 1.5 km, ate a snack, and then went home*" será dividida em 3 blocos: "*Jessica ran for 1.5 km*", "*ate a snack*", "*then went home*".

Para cada um dos blocos obtidos é verificado se eles possuem ou não verbos, também previamente identificados pelo *POS-Tagging*, conforme ilustrado na Figura 19:

```
Jessica/NNP ran/VBD for/IN 1/CD ./. km/NN :/: ate/VBD snack/NN :/: and/CC
then/RB went/VBD home/NN.
```

Figura 19 – Frase marcada pelo *POS-Tagging* do RISO-VTD

A frase descrita na Figura 19 contém 3 verbos ("*ran*", "*ate*", "*went*"), ambos no pretérito.

A Tabela 12 ilustra a TAGs utilizadas pelo *POS-Tagging* para identificar os verbos.

Tag	Descrição	Exemplos
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing.present, non-3d	take
VBZ	verb, 3rdperson sing. present	takes

Tabela 12 – Tags do *POS-Tagging* que identificam verbos

[36] http://www.gramatica.net.br/oracao/

Caso um bloco não possua um verbo, é verificado o próximo bloco da frase, até que se encontre algum que contenha verbo, de maneira que cada bloco sem verbo lido é unificado ao bloco lido anteriormente, formando assim um único bloco. O processo continua até que todos os blocos tenham sido verificados. Com isso, ao final desta fase terão sido formados *n* blocos, de maneira que em cada bloco terá um conjunto de palavras, dos quais uma destas palavras será um verbo. Os blocos extraídos ao final deste processo formam as orações da frase.

- **Oração 1:** "*Jessica run for 1.5 km*"
- **Oração 2:** "*ate a snack*"
- **Oração 3:** "*and then went home*"

As orações extraídas das frases são submetidas posteriormente à etapa de Verificação de Correlação entre Orações.

6.2.1.3.2 Verificação de Correlação entre Orações

A partir das orações obtidas na etapa anterior, são verificadas, uma a uma, quais delas possuem nomes próprios (identificados pelo *POS-Tagging* conforme descrito na Tabela 13). Caso não possuam, é verificada a oração seguinte, até que se encontre alguma que contenha nomes próprios, de maneira que todas as orações processadas até se encontrar uma oração que contenha um nome próprio são consideradas correlacionadas.

Tag	Descrição	Exemplos
NNP	proper noun, singular	John
NNPS	proper noum, plural	Vikings

Tabela 13 – Tags do POS Tagging que identificam nomes próprios

Caso seja encontrada uma oração que possua um nome próprio, são buscadas novas orações correlacionadas. Para isto, são buscadas novas orações que contenham substantivo próprio. Se esta busca tiver lido todas as orações até o fim da frase sem encontrar nenhuma oração que contenha substantivo próprio, todas as orações lidas até aquele momento, são correlacionadas ao grupo de orações correlacionadas imediatamente anterior. O processo continua até que todas as orações da frase tenham sido verificadas.

A seguir é ilustrado um exemplo da aplicação desta abordagem.

"Although she was tired, Jessica ran for 1.5 km swam for minutes and then returned home."

Neste caso são extraídas 4 orações:

- **Oração 1:** "*Although she was tired*"
- **Oração 2:** "*Jessica ran for 1.5 km*"
- **Oração 3:** "*ate a snack*"
- **Oração 4:** "*and then went home*"

Os seguintes passos são realizados:

1. A primeira oração é lida e não é encontrado nenhum nome próprio. Esta oração é armazenada para ser correlacionada à alguma outra oração posteriormente.
2. A segunda oração é lida. Nesta é encontrada o nome próprio "*Jessica*". Com isto, a Oração 1 e a Oração 2 são correlacionadas.
3. A terceira oração é lida e não é encontrado nenhum nome próprio. Esta oração é armazenada para ser correlacionada à alguma outra oração posteriormente.
4. A quarta oração é lida e não é encontrado nenhum nome próprio. Esta oração é armazenada para ser correlacionada à alguma outra oração posteriormente.
5. Não há mais orações a serem lidas e ainda existem duas orações pendentes de serem correlacionadas: As orações 3 e 4. Estas orações são relacionadas ao grupo de orações correlacionadas imediatamente anterior às estas orações. Neste caso, as orações 3 e 4 são correlacionadas às orações 1 e 2, que já haviam sido correlacionadas anteriormente.

Ou seja, todas as orações desta frase estão correlacionadas.

Existem cenários especiais, como, por exemplo, orações iniciadas por advérbios, conforme é ilustrado na Figura 20.

"After being captured by the Fourth Crusade in 1204 and then recaptured by the forces of Nicaea, under the command of Michael VIII Palaeologus in 1261, Constantinople and the Byzantine Empire were taken by the Ottoman Empire May 29, 1453."

Figura 20 – Exemplo de frase que contém orações iniciadas por advérbios.

No exemplo ilustrado na Figura 20, a entidade "*Fourth Crusade*" é associada à expressão temporal "*in 1204*"pois estão presentes na mesma oração. Assim como as entidades "*forces of Nicaea*" e "*Michael VIII Palaeologus*"que são associadas à expressão "*in 1261*" por também pertencerem à mesma oração. No entanto, mesmo estando em orações diferentes, as entidades "*Constantinople*", "*Byzantine Empire*" e "*Ottoman Empire*" além de estarem associadas à expressão temporal "*May 29, 1453*", estão associadas às expressões temporais "*in 1204*" e "*in 1261*", uma vez que as orações são correlacionadas. Esta relação de correlação se estabelece pelo advérbio "*after*" que inicia a oração vizinha.

6.2.1.3.3 Associação de Entidades e Contextos Temporais

Nesta última etapa, são verificados quais os contextos temporais dos conceitos presentes na frase. Desta forma, as entidades são associadas às expressões temporais presentes na oração

na qual elas ocorrem, ou a expressões temporais presentes em oração correlacionada à presente na frase. Caso a frase não possua nenhuma expressão temporal mas possua alguma entidade, serão obtidas as informações presentes na base de dados criada pelo RISO-RID com as informações obtidas da DBPedia referente à esta entidade.

Ao final da execução do fluxo descrito acima, o Topic Map referente ao documento que está sendo analisado é enriquecido com as informações dos conceitos encontrados juntamente com os contextos temporais a ele relacionados (ambos associados ao documento). Na Figura 21 que representa a relação do tópico "*Waterloo*" relacionado ao documento "*Napoleon.txt*" no trecho destacado é mostrado a inclusão do instante temporal "*18-06-1815*" representado pelo intervalo de tempo "*18-06-1815 < X < 18-06-1815*".

```
<rdf:RDF
[…]
<rdf:Description
rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Napoleon.txt">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>
<rdf:Description
rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/mansfield/n/a_town_in_north_
central_ohio">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <rel:PartOf
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/ohio/n/a_midwestern_state
_in_north_central_united_states_in_the_great_lakes_region"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/grape">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <rel:AtLocation
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/field"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Waterloo
/r/HasDate">
    <rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/18-
06-1815_to_18-06-1815 /r/HasDate"/>
  </rdf:Description>
```

Figura 21 – Exemplo de Topic Map enriquecido com as informações extraídas pelo RISO-GCT

6.3 Considerações Finais

O RISO é um projeto que se propõe a ser uma nova opção dentre as opções já existentes na área da Recuperação da Informação, utilizando para isto, o enriquecimento semântico dos termos tanto na consulta quanto na fase de indexação dos documentos, resultando em consultas mais acuradas.

Com a finalidade de melhorar a acurácia dos resultados da consulta semântica do RISO (RISO-CS) foi necessária a inclusão do enriquecimento dos conceitos com base em informações temporais relacionadas a estes conceitos. Para esta finalidade, foi criado o RISO-GCT (Geração de

Contextos Temporais), responsável por relacionar as informações temporais relacionadas aos conceitos e incluir estas informações em Topic Maps.

O RISO-GCT utiliza como entrada o documento original a ser analisado, o mesmo documento contendo as marcações de classificações gramaticais realizadas pelo RISO-VTD, o documento contendo as informações obtidas pelo RISO-TT e as informações presentes na base de dados da DBPedia. O processamento do documento resulta na extração dos relacionamentos entre conceitos e as expressões temporais presentes no documento ou em na base de dados da DBPedia e na inclusão destas informações obtidas nos Topic Maps utilizados para indexação dos documentos.

Com isto, espera-se uma recuperação mais precisa de documentos, possibilitando, por exemplo, a recuperação sem ruídos de documentos referentes à acontecimentos de uma determinada época.

Capítulo 7 – Experimentos e Validações

Neste capítulo são apresentados a análise experimental realizada com o RISO-GCT bem como a verificação e a validação dos resultados alcançados.

7.1 Verificação

O processo de verificação foi realizado para que fosse possível responder as perguntas definidas para a análise experimental, onde são verificadas a eficácia da estratégia utilizada pelo RISO-GCT para relacionar conceitos e expressões temporais e, adicionalmente, são verificadas se as melhorias realizadas na função NORM do RISO-TT foram suficientes para normalizar todas as expressões temporais presentes no texto ou na base de dados da DBPedia.

Para responder a estas questões, inicialmente foi selecionado o *corpus* formado por 8 textos aleatórios presentes no Wikiwars. Realizada a etapa de pré-processamento dos documentos, onde são unificadas as saídas geradas pelo RISO-TT, RISO-VTD e pelo RISO-UDM (vide seção 6.2.1.1 deste documento), foi executado o processo principal do RISO-GCT, responsável por relacionar os conceitos e as expressões temporais presentes no texto.

Ao final do processo de verificação descrito acima, além do Topic Map no formato RDF onde os conceitos são relacionados às expressões temporais, foi gerada uma planilha no formato "*.csv*" contendo os mesmos relacionamentos conceito/tempo incluídos no Topic Map. Adicionalmente são incluídos nesta planilha os relacionamentos conceito/tempo na forma não-normalizada.

Em paralelo, foi realizada manualmente a tarefa de encontrar relacionamentos entre os conceitos e as expressões temporais relacionadas. Estas informações foram incluídas em um gabarito para que fosse possível comparar com os resultados obtidos pela execução automática por meio RISO-GCT, e assim, ser possível avaliar a sua eficácia. O gabarito criado possui 2382 relacionamentos extraídos manualmente.

As informações presentes na planilha são comparadas com os valores do gabarito criado previamente, e a partir disto, é calculada a eficácia do processo realizado pelo RISO-GCT. Este processo de comparação é realizado de maneira automática por um programa desenvolvido especialmente para este fim.

Formalizando o processo de verificação, um documento *D* possui um conjunto de conceitos *c* e expressões temporais *t* relacionadas a estes conceitos. Uma das saídas geradas pelo RISO-GCT, irá resultar em *n* triplas compostas por $<c_i, t_i, tn_i>$ onde c_i é o i-ésimo conceito presente no texto, t_i é a expressão temporal relacionada a este conceito e *tn* é a expressão temporal após o

processo de normalização realizado pela função NORM do RISO-TT. Outro documento *G*, consiste no gabarito que por sua vez, irá conter *n* triplas <cg_i, tg_i, tng_i> que consiste na tripla descrita anteriormente, mas representando as informações produzidas manualmente.

Para que uma tripla <c_i, t_i, tn_i> gerada pelo RISO-GCT seja considerada correta, é necessário que exista no gabarito *G* uma tripla <cg_j, tg_j, tng_j> tal que $c_i = cg_{ij}$, $t_i = tg_j$ e $tn_j = tng_j$. Caso estas condições sejam validadas, então <c_i, t_i, tn_i> $\subseteq$ C, onde C é o conjunto de triplas geradas corretamente.

7.2 Validação

Para validar o desenvolvimento do RISO-GCT, as evoluções realizadas no RISO-TT e analisar sua eficácia, foi realizado um experimento comparativo entre os resultados obtidos pelo RISO-GCT e a os resultados presentes no gabarito gerado. O resultado desta comparação serviu como base para comparação do RISO-GCT com as ferramentas com funcionalidades semelhantes.

Dentre as ferramentas estudadas, foram selecionadas *TM-Gen* e *Temporal Fact and Event Extraction from Free Text* que também realizam a extração de relacionamentos entre conceitos e expressões temporais. Entretanto, não foi possível obter o código fonte nem o executável destas ferramentas. O *TM-Gen* e o *Temporal Fact and Event Extraction from Free Text* não foram disponibilizados pois são projetos não abertos e financiados por empresas privadas.

Foi optado por utilizar o mesmo Corpus utilizado na validação do RISO-TT, uma vez que o mesmo é composto por documentos extraídos da Wikipédia e foi validado pela *Association for Computational Linguistics* e publicada na *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (MAZUR, 2010).

Os resultados obtidos são discutidos na seção 7.3 deste trabalho.

7.3 Análise dos Resultados

Com base no gabarito criado previamente, foi calculado a eficácia da abordagem implementada pelo RISO-GCT.

Para este estudo, foi avaliado o desempenho do RISO-GCT considerando os seguintes cenários:

- Acertos obtidos considerando todos os relacionamentos presentes no gabarito, incluindo os conceitos cujas datas precisam ser recuperadas da DBPedia.
- Acertos obtidos considerando todos os relacionamentos presentes no gabarito, exceto os conceitos cujas datas precisam ser recuperadas da DBPedia.

Para ambos os cenários acima, foram considerados os seguintes sub-cenários:

- Acertos obtidos validando as datas normalizadas, resultantes da execução da função NORM do RISO-TT.
- Acertos obtidos **não** validando as datas normalizadas, resultantes da execução da função NORM do RISO-TT.

A Tabela 14 mostra os resultados nos dois cenários, um onde são considerados todos os relacionamentos conceitos/expressões temporais gerados pelo RISO-GCT, e outro desconsiderando apenas os relacionamentos gerados através de consultas à base de dados criada com informações obtidas da DBPedia.

	Considerando Datas Presentes na DBPedia				Desconsiderando Datas presentes na DBPedia			
Texto	**Qtd. Total Gabarito**	**Qtd. Total RISO-GCT**	**Qtd. Acertos (NORM)**	**Qtd. Acertos (Ñ NORM)**	**Qtd. Total Gabarito**	**Qtd. Total RISO-GCT**	**Qtd. Acertos (NORM)**	**Qtd. Acertos (Ñ NORM)**
Napoleon	**388**	426	295	308	**278**	273	230	243
WW1	**254**	243	161	181	**179**	151	122	139
IraqWar	**297**	285	197	204	**234**	178	154	161
AmCiv War	**258**	256	197	241	**224**	255	165	208
Vietnan War	**175**	136	114	131	**168**	132	110	127
AmRev War	**393**	381	285	367	**312**	374	208	290
French Rev	**158**	177	131	140	**138**	177	111	120
Korean War	**241**	259	155	216	**212**	242	128	189

Tabela 14 – Acertos obtidos pelo RISO-GCT

Com base nas informações obtidas após a execução do RISO-GCT, foram calculadas a Precisão, Cobertura e *F-Measure* das amostras obtidas. Os resultados são exibidos na Tabela 15 e Tabela 16 que representam, respectivamente, o cenário onde são considerados todos os relacionamentos gerados pelo RISO-GCT e o cenário onde são desconsideradas às consultas a base de dados composta de informações provenientes da DBPedia.

Pode-se verificar, em ambos os cenários, que os resultados obtidos desconsiderando o processo de normalização do RISO-TT obtiveram melhores resultados em comparação com as informações contidas no gabarito, com valores de *precisão, cobertura e f-measure* melhores. Isto se deve ao fato de ainda serem necessárias evoluções significativas no RISO-TT que possibilitem o reconhecimento de novos padrões de expressões temporais.

O mesmo pode-se afirmar, a respeito dos resultados obtidos desconsiderando o processo de obtenção de informações temporais advindas da DBPedia. Isto se deve ao fato de que o processo de recuperação de informações da DBPedia realizado pelo componente RISO-RID do RIS-GCT não possuir no momento consultas abrangentes o suficiente para recuperação de informações temporais de alguns conceitos.

	Datas Normalizadas			Datas Não Normalizadas		
	Precisão	**Cobertura**	**F-Measure**	**Precisão**	**Cobertura**	**F-Measure**
Napoleon	0,69248826	0,760309278	0,724815725	0,72300469	0,793814433	0,756756757
WW1	0,66255144	0,633858268	0,647887324	0,74485597	0,712598425	0,728370221
IraqWar	0,69122807	0,663299663	0,676975945	0,71578947	0,686868687	0,701030928
AmCivWar	0,76953125	0,763565891	0,766536965	0,94140625	0,934108527	0,937743191
VietnanWar	0,83823529	0,651428571	0,733118971	0,96323529	0,748571429	0,84244373
AmRevWar	0,7480315	0,72519084	0,736434109	0,96325459	0,933842239	0,948320413
FrenchRev	0,74011299	0,829113924	0,782089552	0,79096045	0,886075949	0,835820896
KoreanWar	0,5984556	0,643153527	0,62	0,83397683	0,89626556	0,864

Tabela 15 – Tabela de Valores de Precisão, Cobertura e F-Measure referente aos resultados obtidos pelo RISO-GCT

	Datas Normalizadas			Datas Não Normalizadas		
	Precisão	**Cobertura**	**F-Measure**	**Precisão**	**Cobertura**	**F-Measure**
Napoleon	0,84249084	0,827338129	0,834845735	0,89010989	0,874100719	0,882032668
WW1	0,80794702	0,681564246	0,739393939	0,9205298	0,776536313	0,842424242
IraqWar	0,86516854	0,658119658	0,747572816	0,90449438	0,688034188	0,781553398
AmCivWar	0,64705882	0,736607143	0,688935282	0,81568627	0,928571429	0,868475992
VietnanWar	0,83333333	0,654761905	0,733333333	0,96212121	0,755952381	0,846666667
AmRevWar	0,55614973	0,666666667	0,606413994	0,77540107	0,929487179	0,84548105
FrenchRev	0,62711864	0,804347826	0,704761905	0,6779661	0,869565217	0,761904762
KoreanWar	0,52892562	0,603773585	0,563876652	0,78099174	0,891509434	0,832599119

Tabela 16 – Tabela de Valores de Precisão, Cobertura e F-Measure referente aos resultados obtidos pelo RISO-GCT, exceto datas obtidas na DBPedia.

Os resultados gerados pelo RISO-GCT, bem como o gabarito utilizado para validação desta pesquisa encontram-se disponíveis em: http://bit.ly/1NpVWFi.

7.3.1 Discussão

Os experimentos realizados tiveram como objetivo avaliar a eficácia da estratégia utilizada pelo RISO-GCT em relacionar conceitos presentes em documentos e expressões temporais também obtidas no próprio documento ou em fonte externa (DBPedia). Também foi avaliado se os Topic Maps gerados pelo RISO-IS, foram enriquecidos corretamente com os contextos temporais extraídos do texto. Para isto, foi selecionado um subconjunto de documentos presentes no corpus *Wikiwars*.

O gabarito gerado para validação dos resultados, possui os relacionamentos encontrados entre conceitos e expressões temporais divididos em 2 categorias. Na primeira, as datas são inseridas normalizadas, da mesma forma em que são inseridas nos Topic Maps. Na segunda, as datas são inseridas da forma original na qual elas estavam presentes no documento, ou seja, sem que fossem normalizadas. Esta divisão foi necessária para avaliar a eficácia da função de normalização de expressões temporais do RISO-TT.

Respondendo às perguntas, com relação à Pergunta I, foram comparados os itens presentes no Topic Map gerados pelo RISO-IC, antes e depois de cada execução RISO-GCT. Pode-se afirmar que o Topic Map foi enriquecido com os contextos temporais. Desta forma, os documentos poderão ser indexados de acordo com dados temporais.

Ainda sobre a Pergunta I, analisando-se as medidas de precisão, ou seja, a quantidade de acertos em relação à quantidade de respostas geradas pelo RISO-GCT, nota-se que dentre as demais medidas esta foi a que teve o pior desempenho. Isto ocorreu devido ao número elevado de falsos-positivos gerados pelo componente RISO-RCT do RISO-GCT, ou seja, foram extraídas informações temporais que não estavam diretamente relacionadas aos conceitos presentes no texto, apesar de estarem na mesma frase.

Em alguns textos analisados, a extração de falsos-positivos resultou, inclusive, em um número maior de relacionamentos conceitos/contexto temporal extraídos pelo RISO-GCT em comparação com o próprio gabarito. Casos como estes foram verificados em alguns cenários em frases que possuam mais de uma expressão temporal distribuídas em diferentes orações na mesma frase, onde expressões temporais foram erroneamente associadas a conceitos. Os valores de precisão obtidos tiveram valor médio de 0,713 no pior caso (desconsiderando informações temporais obtidas por meio de consultas à DBPedia e considerando a normalização das datas na validação dos resultados) e 0,84 no melhor caso (desconsiderando normalização das datas e as desconsiderando informações temporais obtidas através de consultas à DBPedia).

Sobre as medidas de cobertura, ou seja, a quantidade de acertos com relação à quantidade de respostas presentes no gabarito, os resultados do RISO-GCT obtiveram cobertura média de,

aproximadamente, 0,84 no melhor caso (desconsiderando a normalização das datas e as datas obtidas por meio de consultas à DBPedia) e de 0,70 no pior caso (considerando a normalização das datas na comparação dos resultados) com valor máximo de 0,934. Não houveram variações significativas nas medidas de cobertura entre os cenários que consideravam as datas obtidas da DBPedia e os que não consideraram estas informações. Ao desconsiderar a normalização das expressões temporais presentes no texto, os resultados melhoraram em torno de 10%.

Analisando a medida F-Measure[37], ou seja, a média ponderada harmônica entre as medidas de precisão e cobertura, foram obtidos os valores médios de 0,83 no melhor caso (desconsiderando as informações provenientes da DBPedia e desconsiderando a normalização das expressões temporais na comparação dos resultados) e 0,702 no pior caso (cenário que desconsiderava as informações provenientes de DBPedia, mas considerava a normalização das datas na comparação dos resultados).

Com base nas medidas de precisão, cobertura e *f-measure*, obtivemos um desempenho melhor que a ferramenta TM-Gen, considerando as medidas de cobertura, precisão e *f-measure* contidas na Figura 22 em comparação com os valores médios das mesmas medidas obtidas pelo RISO-GCT (tanto para o pior quanto para o melhor caso), embora não seja possível afirmar com certeza, pois as ferramentas não foram executadas com a mesma entrada.

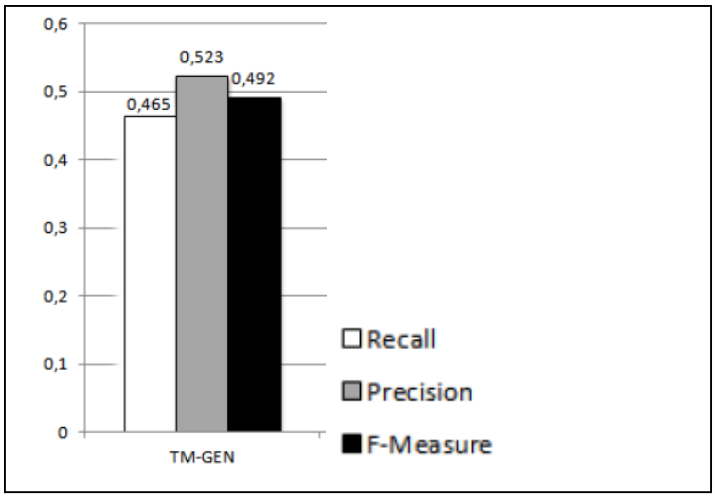


Figura 22 – Desempenho obtido pela ferramenta TM-Gen (GARRIDO, 2013)

O RISO-GCT obteve um desempenho aparentemente superior à ferramenta *Temporal Fact and Event Extraction from Free Text*, uma vez que a mesma considerou em seu estudo apenas relações do tipo `holdsPoliticialPosition` ou seja, o período no qual um político XPTO desempenhou determinada posição política. Ainda assim, obteve precisão de 67%. Muito abaixo do desempenho do RISO-GCT que abrangia relacionamentos dos mais diversos tipos.

[37] Calculado a partir do (2*precisão*cobertura)/(precisão+cobertura)

Relation	Number of Seeds	Input Articles	Extracted Facts	Precision
holdsPoliticalPos	50	100	221	67%

Tabela 17 – Desempenho obtido pela ferramenta Temporal Fact and Event Extraction from Free Text (KUZEY, 2003)

Já com relação à Pergunta II, o RISO-GCT apresenta um resultado satisfatório tanto nos cenários que consideram todos os relacionamentos extraídos pelo RISO-GCT, quanto nos cenários onde são desconsiderados os relacionamentos extraídos, cujas datas foram obtidas a partir da base de dados criada com informações da DBPedia.

Dentre os 4 sub-cenários analisados, o RISO-GCT se saiu melhor nos 2 sub-cenários que consideravam os resultados contendo apenas os relacionamentos extraídos a partir de conceitos presentes no texto, não considerando os resultados contendo expressões temporais obtidas na base de dados populada com dados da DBPedia. Isto ocorreu, devido ao fato de o RISO-GCT não possuir consultas que abranjam todas as ontologias (country, things, location, people, entre outras) aos quais alguns dos conceitos pertenciam. De maneira que seriam preciso melhorias no RISO-GCT para incluir novas consultas SPARQL para abranger ainda mais ontologias da DBPedia.

Dentre os 2 sub-cenários que melhores avaliados, o sub-cenário no qual eram consideradas os resultados sem a realização da normalização das datas se saiu melhor que os resultados que consideravam o processo de normalização das expressões temporais pela função NORM do RISO-GCT. A causa para este comportamento, foi devido ao fato de a função NORM do RISO-TT ainda não estar abrangendo a normalização uma porção considerável de expressões temporais.

7.3.2 Ameaças a Validade

Como fator que pode ameaçar a validade dos resultados obtidos, pode-se elencar a possível interferência do participante do experimento na fase de extração manual dos relacionamentos de conceitos/expressões temporais, devido ao fato do próprio ter desenvolvido a técnica de extração automática de relacionamentos e também ter desenvolvido a ferramenta RISO-GCT que implementa esta técnica.

O fato de ter sido utilizado um subconjunto de um Corpora finito (*Wikiwars*), ameaça a validade externa do experimento, uma vez que não é possível realizar afirmações sobre o desempenho da abordagem implementada pelo RISO-GCT sem que haja novos experimentos.

Também pode-se citar, por exemplo, a qualidade da extração manual realizada nos documentos. Apesar de durante a fase de avaliação dos resultados obtidos pelo RISO-GCT com os

resultados presentes na extração manual terem sido realizadas algumas correções no gabarito, é provável que alguns erros tenham passado despercebidos.

Não pôde ser realizada uma comparação mais efetiva do RISO-GCT com ferramentas similares, uma vez que, apesar do contato realizado com os desenvolvedores das ferramentas, não foi disponibilizado código fonte, nem os textos utilizados nos respectivos experimentos de validação destas ferramentas.

Capítulo 8 - Conclusões

Neste capítulo são mostradas as conclusões acerca deste trabalho, bem como as sugestões de trabalhos futuros.

Para resolver o problema da determinação do contexto temporal de conceitos contidos em documentos textuais foi desenvolvida uma estratégia que consiste em, a partir dos conceitos/entidades presentes no texto, tentar determinar as informações temporais relacionadas a eles.

Este processo é feito com base nas extrações de entidades/conceitos e expressões temporais realizadas respectivamente, pelo RISO-VTD proposto por BISPO (2013) e pelo RISO-TT proposto SANTOS (2013), além marcação com TAGs indicando as classes gramaticais das palavras contidas no texto, também pelo RISO-VTD. Com os dados resultantes destas extrações, é realizada a extração das frases contidas no documento com todos seus termos marcados.

Cada frase é dividida em orações, onde cada oração é um trecho da frase (ou frase inteira) que se estrutura com base em um verbo, podendo uma frase pode conter uma ou mais orações. A partir das orações extraídas, são verificadas quais das orações da frase são relacionadas, ou seja, se referem às mesmas entidades/conceitos. Para isto, são verificadas as orações que não possuem entidades/conceitos (sujeito) e estas são relacionadas às orações da mesma frase que continham estas informações, seguindo critérios de proximidade.

Posteriormente, entidade e conceitos são associados às informações temporais contidas nas orações relacionadas, ou em casos onde as frases não possuem informações temporais explícitas, são recuperadas as informações relacionadas às entidades/conceitos a partir de consultas realizada à base de dados da DBPedia.

O trabalho foi validado utilizando-se um sub-conjunto de 9 documentos presentes no *corpus* Wikiwars composto por 22 documentos. Neste *corpus* encontram-se documentos que possuem narrativas de guerras que ocorreram ao longo da história da humanidade.

Com base nos resultados obtidos pelo RISO-GCT, presentes na Tabela 14 e na Tabela 15 deste trabalho, foram realizadas comparações com os resultados obtidos pelas ferramentas TM-Gen e *Temporal Fact and Event Extraction from Free Text*, utilizando as informações disponibilizadas pelos autores nos respectivos trabalhos, uma vez que não foi possível obter acesso ao código fonte e ao binário executável destas ferramentas durante a fase de validação, bem como ao *corpus* utilizado. Nestas comparações há fortes indícios de que o RISO-GCT se saiu melhor que estas ferramentas e os objetivos propostos para estes trabalhos foram atingidos.

8.1 Contribuições

Considerando as informações apresentadas nas seções 5.3.2.2 do Capítulo 5 e das seções 6.2.1.1, 6.2.1.2 e 6.2.1.3 do Capítulo 6, as principais contribuições do RISO-GCT são:

a) Determinação do valor correto de expressões temporais presentes no texto que necessitam de análise de informações presentes em diferentes partes do texto ou em fontes externas ao texto (*DBPedia*).
b) Normalização de padrões que até então não eram normalizados pelo RISO-TT.
c) Independência de software de terceiros (que não sejam os do próprio RISO).
d) Extração de informações temporais contidas em documentos, criação dos relacionamentos entre estas informações temporais e os objetos a eles relacionados e atualização de Topic Maps contendo estas informações.

8.2 Limitações/Outliers

Não fizeram parte do escopo desta etapa do projeto a implementação das seguintes características:

a) **Melhorias de Tempo de Processamento:** A etapa responsável por realizar a busca das informações temporais relacionadas às entidades demanda um alto tempo para conclusão. Uma vez que, para cada termo extraído pelo RISO-VTD é necessária a execução de cerca de 10 consultas à base de dados da *DBPedia*, caso a entidade em questão não esteja presente na base de dados do RISO juntamente com as informações temporais relacionadas.
b) **Normalização:** Ainda existem um grande número de expressões temporais que não foram normalizados, uma vez que esta tarefa envolve processamento de linguagem natural e as expressões temporais podem apresentar diversas formas no texto.

   Neste projeto limitamos a granularidade temporal das expressões temporais processadas pelo RISO em dia, mês e ano. Sendo necessárias evoluções para que o RISO-GCT passe a funcionar para outras granularidades como segundos, nano segundos e séculos, por exemplo. Para a expressão temporal "*One century after the Waterloo Battle*", o RISO-GCT não está preparado para calcular o século a qual esta expressão se refere. Ao contrário da expressão temporal "*Two years after the Waterloo Battle*" que seria processada pelo RISO-GCT e resultaria data 18-06-1817, ou seja, 2 anos depois da data na qual ocorreu a Batalha de Waterloo, em 18-06-1815.
c) **Integração dos Componentes do RISO:** Para o correto funcionamento do RISO-GCT é mandatório que sejam executados o RISO-VTD e o RISO-TT para que o RISO-GCT

utilize as saídas geradas por eles. No entanto, atualmente estes sistemas precisam ser executados manualmente. Para agilizar o processo, seria necessária a criação de um script que automatizasse a execução do RISO-VTD, RISO-TT e RISO-GCT.

d) **Etapa de Desambiguação dos Conceitos:** Nesta etapa do projeto não houve a etapa de desambiguação dos termos. De maneira que não há diferenciação entre, por exemplo, a Cidade de São Paulo, o Estado de São Paulo ou o santo São Paulo. De maneira que, nesta versão do projeto, ambos são considerados como sendo o mesmo conceito. Apesar do RISO-IC realizar este tratamento, o mesmo não ocorre no RISO-GCT.

8.3 Trabalhos Futuros

Como trabalhos futuros, podem ser considerados:

- Melhorias no processo de reconhecimento da data na qual expressões do tipo "*D-EMT-Adv*" se referem. Atualmente as expressões temporais deste tipo são processadas pelo RISO-GCT buscando-se por datas citadas anteriormente no texto. Conforme o exemplo: "*Darryl bought his home on January 5, 2013 and married two months later*" onde a expressão temporal "*two months later*" se refere a dois meses depois de 05/01/2013. No entanto, existem casos ainda não mapeados pelo RISO-GCT onde, estas expressões se referem a datas que ocorrem posteriormente à ocorrência delas, como no exemplo "*Darryl bought his home two months after the date of his marriage, wich was January 5, 2013*". O RISO-GCT precisará ser alterado para que possa mapear corretamente as expressões "*D-EMT-Adv*" da maneira correta em ambos os cenários citados.
- Melhorias no algoritmo de busca de informações temporais de conceitos presentes na base de dados da DBPedia. Uma porção considerável das datas relacionadas a determinados conceitos ainda não estão sendo recuperadas nas consultas SPARQL realizadas pelo RISO-GCT.
- Melhorias na funcionalidade de reconhecimento de conceitos presentes em documentos, uma vez que muitos dos conceitos extraídos pelo RISO-IC não são necessariamente palavras presentes no texto, e sim, partes de outras palavras compostas, sendo necessário a realização de tratamento de falsos-positivos.
- Melhorias na função NORM do RISO-TT uma vez que, mesmo após a inclusão de funcionalidades para normalização de novos formatos de expressões temporais,

ainda existe uma parte considerável de expressões temporais que não estão sendo normalizados, a exemplo da expressão temporal "*Summer of 1959*".

- Melhorias no processo de extração de relacionamentos entre conceitos e expressões temporais presentes em frases com mais de uma expressão temporal, onde nem sempre ambas as expressões estão relacionadas a todos conceitos presentes na frase. Neste estudo, foram incluídos tratamentos para diversas tipos de ocorrências destes cenários, mas ainda assim, houve casos em que foram gerados falsos-positivos.
- Para entidades temporais poderá ser feito uma distinção entre fato, evento e processo, de acordo com Allen. Para isto poderão ser utilizados os atributos da Wikipedia/DBPedia.
- Atualmente, a granularidade das fontes indexadas pelos tópicos no RISO considera documentos, o que poderia ser refinado para considerar parágrafos ou sentenças.

Referências Bibliográficas

Allen, J. F. (1984). Towards a General Theory of Action and Time. *Artificial Intelligence Vol. 23, pp. 123-154*: Elsevier Science Publishers Ltd. Essex, UK.

Almeida, G. M. B.; Vale, O. A. (1996). Do texto ao termo: Interação entre terminologia, morfologia e linguística de corpus na extração semi-automática de termos.

Alvares, L. (2007). Taxonomia. Faculdade de Ciência da Informação. Universidade de Brasília, 2007.

Anáfora (linguística). Disponível em: https://pt.wikipedia.org/wiki/An%C3%A1fora_(lingu%C3%ADstica). Acesso em 06 out. 2015.

ANNIE: A Nearly-New Information Extraction System. Disponível em: https://gate.ac.uk/sale/tao/splitch6.html#chap:annie. Acesso em 06 out. 2015.

Araújo Júnior, J. G. (2013). Documentos Textuais Baseada Em Fontes Heterogêneas de Informação. Dissertação Mestrado – Programa de Pós-Graduação em Ciência da Computação – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande.

Barros, L. A. (2004). Curso Básico de terminologia. Acadêmica (São Paulo, Brasil), Edusp.

Baião, G.; Lima, G. A. B. de O. (2008) A utilização de Mapas de Tópicos na Compatibilização de Conteúdos Hipertextuais Semanticamente Estruturados. Dissertação Mestrado em Ciência da Informação – Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

Bispo, M. C. T. (2013) Criação de Vetores Temáticos de Domínios para a Desambiguação Polissêmica de Termos. Dissertação Mestrado – Programa de Pós-Graduação em Ciência da Computação – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande.

Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In Proceedings of the twelfth national conference on artificial intelligence. Seattle, Wa. American Association for Artificial Intelligence.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21(4), 543–565.

Campelo, C. (2008). GeoSEn: um Motor de Busca com Enfoque Geográfico. Dissertação Mestrado – Programa de Pós-Graduação em Ciência da Computação – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande.

Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V. (2002). GATE: an Architecture for Development of Robust HLT Applications. Department of Computer Science. University of Sheffield.

DBPEDIA. Disponível em: http://pt.dbpedia.org/. Acesso em 06. Dec. 2015.

Dicio. Significado de Termo. Disponível em: http://www.dicio.com.br/termo/. Acesso em: 09 jul. 2015.

Dicionário Informal (2009). Corpus. Disponível em: http://www.dicionarioinformal.com.br/corpus/ Acesso em 15 jun. 2015.

Dölling, J. (2014). Aspectual Coertion and Eventualy Structure. Institut für Linguistik. Universität Leipzig

Garrido, A.; Buey, M. G.; Escudero, S.; Ilari, S.; Mena, E.; Silveira, S. B. (2013). TM-Gen: A Topic Map Generator from Text Documents. IIS Department, University of Zaragoza. Zaragoza, Spain. Computer Department, University of Lisbon. Lisbon, Portugal.

Hagège, C.; Baptista, J.; Mamede, N. (2010). Caracterização e processamento de expressões temporais em português. In: **Linguamatica**, v. 2, n. 1, p. 63-77, abr. 2010.

Kageura K., Umino B. (1996). Methods of automatic term recognition - a review. Terminology.

Kuzey, E. (2011). Extraction of Temporal Facts and Events from Wikipedia. Universität des Saarlandes. Max-Planck-Institut für Informatik. Master's Thesis in Computer Science.

Laguna, M. S. C. (2014). Extração automática de termos simples baseada em aprendizado de máquina. Tese (Doutorado – Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

LINGUA-PT-PLNBase. Disponível em: http://search.cpan.org/dist/Lingua-PT-PLNbase/lib/Lingua/PT/PLNbase.pm. Acesso em: 01 out. 2015.

LINGUATECA. Disponível em http://www.linguateca.pt/. Acesso em: 01 out. 2015.

Mani, I.; Wilson G. (2000). Processing of News. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000), pages 69–76.

Marsic, G. (2011). Temporal Processing of News: Annotation Of Temporal Expressions, Verbal Events And Temporal Relations. 2011. (PhD Thesis). University of Wolverhampton, Wolverhampton, 2011.

Mata, F.; Claramunt, C. (2010). GeoST: Geographic, Thematic and Temporal Information Retrieval from Heterogeneous Web data sources. Unidade Profissional Interdisciplinar em Engenharia e Tecnologias Avançadas (UPIITA). Instituto Politécnico Nacional (IPN).

Mazur, P.; DALE, R. WikiWars: A New Corpus for Research on Temporal Expressions.In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**, Association for Computational Linguistics, p. 913–922, MIT, Massachusetts, USA, 9-11 October 2010.

Mapa de Tópicos - Uma introdução dos Mapas de Tópicos. Disponível em: https://msdn.microsoft.com/pt-br/library/aa480048.aspx. Acesso em: 03 de mar. 2016.

MontyLingua (2005). A FREE, Commonsense-Enriched Natural Language Understander for English. Disponível em: http://web.media.mit.edu/~hugo/montylingua/. Acesso em 08 out. 2015.

Mota, C.; Santos, D. (2009). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, 2008, Capítulo 8, p. 159–170

MPN (2014). Disponível em; http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/. Acesso em 09 ago. 2015

Nakashole, N.; Theobald, M.; Weikum, G. (2011). Scalable Knowledge Harvesting with High Precision and High Recall. . Universität des Saarlandes. Max-Planck-Institut für Informatik.

Negri, M.; Marseglia, L. (2004). Recognition and Normalization of Time Expressions: ITC-ist at TERN 2004. ITC-int, Centro per la Riceca Scientifica e Tecnologia.

Neves, P. I.; Corrêa, D. A.; Cavalcanti, M. C. (2013). Uma análise sobre abordagens e ferramentas para Extração de Informação. Seção de Engenharia e Computação – Instituto Militar de Engenharia (IME). Departamento de Informática – Universidade Federal Rural do Rio de Janeiro (UFRRJ). Laboratório Nacional de Computação Científica (LNCC).

Nunes, F.; Hounsell, M. S.; Junior, R. S. U. R. (2012). Desenvolvimento de um Simulador de CLP como um Compilador. Laboratory for Research on Visual Applications (LARVA), Depto. De Ciência da Computação (DCC), Universidade do Estado de Santa Catarina (UDESC).

Pazienza, M.T.; Pennacchiotti, M., Zanzotto, F.M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis S (ed) Knowledge Mining Series: Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, Berlin.

Ribeiro, L. B. (2015). Análise de Sentimentos em comentários sobre aplicativos para dispositivos móveis. Estudo de Impacto do pré-processamento. Departamento de Ciência da Computação, Universidade de Brasília. Brasília, DF.

Roth, D.; Ji, H.; Cassidy, T., Do, Q. (2012). Temporal Information Extraction and Shallow Temporal Reasoning. Computer Science Department, University of Illinois at Urbana-Champaign. Computer Science Department and Linguistic Department, Queeens College and the Graduate Center City University of New York.

Sales, R. (2007). A questão da Linguagem usada dentro das Organizações: Um Levantamento Biográfico. Disponível em: http://revista.acbsc.org.br/racb/article/view/486/624. Acesso em 16 ago. 2015.

Saquete, E. (2010). A System for Recognizing and Normalizing TIMEX3. In: Proceedings of the 5th International Workshop on Semantic Evaluation.

Santos, A. A. (2013). RISO-TT – Extração de Expressões Temporais em Textos. Dissertação Mestrado – Programa de Pós-Graduação em Ciência da Computação – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande.

Schiel, U. (2012). Projeto RISO-M. SINBAD - Recuperação Semântica de Objetos Multimídia. Grupo de Sistemas de Informação e Banco de Dados. Departamento de Sistemas e Computação, Universidade Federal de Campina Grande.

Schiel, U. (2015). RISO - Recuperação Semântica de Objetos Multimídia - Arquitetura. Grupo de Sistemas de Informação e Banco de Dados. Departamento de Sistemas e Computação, Universidade Federal de Campina Grande.

Silva, T. de M. S. (2003). Extração de Informação para Busca Semântica na Web Baseada em Ontologias. Dissertação Mestrado – Programa de Pós-Graduação em Engenharia Elétrica. Universidade Federal de Santa Catarina.

Teline, M. F.; Almeira, G. M. B.; Aluísio, S. M. (2004). Extração Manual e Automática de Terminologia: Comparando Abordagens e Critérios. Núcleo Interinstitucional de Linguística Computacional, ICMC-USP, São Carlos-SP, Brasil. Departamento de Letras, UFSCAR, São Carlos-SP, Brasil.

Teline, M. F. (2004). Avaliação de Métodos de Extração Automática de Terminologia para textos em Português. Dissertação Mestrado – Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Terra, J. C. C. et al. (2006). Taxonomia: Elemento fundamental para gestão do conhecimento. Biblioteca Terra Fórum Consultores. Disponível em <http://portais.integra.com.br/sites/terraforum/Biblioteca/libdoc00000102v003taxonomia_%20fundamental_GC.pdf>. Acesso em: 12 jan. 2006.

Topic Maps (2015). Disponível em http://ocw.uc3m.es/ingenieria-informatica/information-engineering/lecture-notes-1/05-Topic_Maps.pdf. Acesso em: 05 abr. 2016

Vivaldi, J.; Rodríguez, H. (2007). Evaluation of terms and term extraction system. A practical approach. Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication.

Wikipedia, a Enciclopédia Livre. Disponível em: https://pt.wikipedia.org/wiki/Wikip%C3%A9dia. Acesso em 24. Nov. 2015

Yangarber, R.; Grishman, R. (2000) *Extraction Pattern Discovery through Corpus Analysis*. TR- 00-143, The Proteus Project, New York University. In: Proceedings of the Workshop Information Extraction meets Corpus Linguistics, Second International Conference on Language Resources and evaluation (LREC 2000), Athens, Greece, 2000.

Zambenedetti, C. Extração de Informação sobre Bases de Dados Textuais. Dissertação Mestrado – Programa de Pós-Graduação em Ciência da Computação – Instituto de Informática, Universidade Federal do Rio Grande do Sul.

Zavaglia, C.; Oliveira, L. H. M.; Nunes, M. G. V.; Teline, M. F.; Aluisio, S. M (2005). Avaliação de Métodos de Extração Automática de Termos para a Construção de Ontologias. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP.

Apêndice A – Algoritmo para Geração de Contextos Temporais

O trecho do algoritmo ilustrado a seguir é responsável por realizar a divisão de cada frase de um documento em orações. A tarefa consiste em identificar conjuntos de palavras que se estruturam em torno de um ou mais verbos.

```
1.    ∀f_i ∈ F(t_n)
2.        // inicializa conjunto vazio
3.        Inicializa(conjAux)
4.        ∀ bloco ∈ Divide(f_i, DELIM) = {e_i, e_i+1, e_i+2, … , e_n}
5.        // divide a frase a partir dos delimitadores presentes no conjunto DELIM
6.            Se ContémVerbo(bloco_i) AND ContémEntidade(bloco_i)
7.            // Caso o bloco possua um ou mais verbos e um ou mais entidades
8.                Se size(ORAÇÕES) > 0
9.                AND NOT ContémEntidade(GetItemConjunto(ORAÇÕES, size(ORAÇÕES)-1)))
10.               // Se alguma outra oração que contenha entidade foi extraída anteriormente, ou...
11.               OR
12.               Se StartWithVerb(bloco_i) OR StartWithAdverb(bloco_i)
13.               // Se o bloco inicia com um verbo ou um advérbio
14.                   oração = GetItemConjunto(ORAÇÕES, size(ORAÇÕES)-1)
15.                   oração.concat(bloco_i)
16.                   // recupera a oração anterior e concatena com o bloco que está sendo lido
17.               Senão
18.               // caso contrário
19.                   oração = bloco_i
20.                   AddItemConjunto(oração, ORACÕES)
21.                   // adiciona o bloco que está sendo lido ao conjunto de orações extraídas.
22.           Se ContémVerbo(bloco_i) AND NOT ContémEntidade(bloco)
23.           // Se o bloco contém verbo, mas não contém uma entidade...
24.               Se size(ORAÇÕES) > 0
25.               // Se alguma outra oração foi extraída anteriormente, ou...
26.               OR
27.               Se StartWithVerb(bloco_i) OR StartWithAdverb(bloco_i)
28.               // Se o bloco da frase em questão inicia com verbo ou advérbio
29.                   oração = GetItemConjunto(ORAÇÕES, size(ORAÇÕES)-1)
30.                   oração = oração.concat(bloco_i)
31.                   // recupera a oração extraída anteriormente e concatena com o bloco que
32.                   está sendo lido.
33.               Senão
34.               // Caso contrário
35.                   AddItemConjunto((bloco_i, ORACÕES)
36.                   // Adiciona o bloco ao conjunto de orações.
37.           Se NOT ContémVerbo(bloco_i) AND ContémEntidade(bloco)
38.           // Caso o bloco lido não possua um verbo, mas possua uma entidade
39.               Se size(ORAÇÕES) > 0
40.               // Se alguma outra oração tiver sido extraída anteriormente, ou...
41.               OR
42.               Se StartWithVerb(bloco_i) OR StartWithAdverb(bloco_i)
43.               // Se o bloco que está sendo lido inicia com um verbo ou um advérbio
44.                   oração = GetItemConjunto(ORAÇÕES, size(ORAÇÕES)-1)
45.                   oração = oração.concat(bloco_i)
46.                   // Recupera a oração extraída anteriormente e concatena com o bloco que
47.                   está sendo lido.
48.               Senão
49.               // Caso contrário
50.                   AddItemConjunto((bloco_i, ORACÕES)
51.                   // Adiciona o bloco que está sendo lido ao conjunto de orações extraídas
52.           Senão
53.           // Caso contrário
54.               oração = oração.concat(bloco_i)
55.               // Adiciona o bloco que está sendo lido ao conjunto de orações extraídas.
```

```
56.                 Se i = Tamanho(Divide(f,DELIM))
57.                 // Caso a frase inteira já tenha sido lida
58.                         AddItemConjunto(bloco_i, ORACÕES)
59.                         // Adiciona o bloco que está sendo lido à lista de orações.
```

O trecho do algoritmo ilustrado a seguir é responsável por identificar as relações entre os conceitos e as expressões temporais presentes nas orações.

```
1.  ∀f_i ∈ F(t_n)
    [...]
60.         // Verifica a quantidade de expressões temporais presentes no texto (implícitas ou explícitas)
61.         Se GetQtdDates(f_i) = 0
62.         // Caso não possua, nenhuma expressão temporal (explícita ou implícita), nenhuma ação será
63.         tomada. Segue para a próxima frase
64.                 proximaFrase(f_i)
65.         Se GetQtdDates(f_i) = 1
66.         // Caso a frase possua apenas uma expressão temporal, esta será relacionada à todas as entidades
67.         presentes nesta frase.
68.                 ∀s_k ∈ ENT(o_z)
69.                 ∀t_j ∈ ET(o_z) ^ ∀t_j ∈ TT(o_z)
70.                         AssociaEntidadeData(s_k, VerificaTemporalidadeTermo (t_j))
71.         Se GetQtdDates(f_i) > 1
72.         // Caso a frase possua mais de uma expressão temporal, as expressões temporais e as entidades
73.         presentes na mesma oração são associadas,
74.                 ∀o_z ∈ ORAÇÕES
75.                         Se StartWith(f_i, ADVTEMP_POS) = "false" e StartWith(f_i, ADVTEMP_ANT) =
76.                         "false"
77.                         // Caso a frase não se inicie com nenhum advérbio temporal...
78.                                 Se GetQtdDates(o_z) > 0
79.                                 // Caso contrário, relaciona as expressões temporais e as entidades
80.                                 presentes na mesma oração.
81.                                         ∀s_k ∈ Terms(o_z)
82.                                         ∀t_j ∈ Temp(o_z)
83.                                                 tf = VerificaTemporalidadeTermo (t_j)
84.                                                 AssociaEntidadeData(s_k, tf)
85.                         Senão
86.                                 Se StartWith(f_i, ADVTEMP_POST) = "true"
87.                                 // Se frase se inicia com um advérbio temporal que indica
88.                                 posterioridade...
89.                                         Se GetQtdDates(o_z) > 0
90.                                         // Caso a oração que está sendo lida no momento possua
91.                                         uma ou mais expressões temporais
92.                                                 ∀s_k ∈ ENT(o_z)
93.                                                 ∀t_j ∈ ET(o_z) ^ ∀t_j ∈ TT(o_z)
94.                                                 // Associa todas as expressões temporais
95.                                                 presentes no conjunto conjAux às entidades
96.                                                 presentes na oração que está sendo lida no
97.                                                 momento.
98.                                                         tf = VerificaTemporalidadeTermo (t_j)
99.                                                         AssociaEntidadeData(s_k,tf)
100.                                        Se StartWith(o_z, ";") = "true"
101.                                        // Se a oração inicia com o caractere ponto-e-vírgula
102.                                                ∀s_k ∈ ENT(o_z)
103.                                                ∀ t_j ∈ conjAux
104.                                                        // Associa as expressões temporais
105.                                                        presentes no conjunto conjAux às
106.                                                        Entidades presentes na oração.
107.                                                        tf = VerificaTemporalidadeTermo (t_j)
108.                                                        AssociaEntidadeData(s_k,After(tf))
109.                                                clean(conjAux)
                                                    // Limpa conjunto auxiliar
                                            Se z <> n então
```

```
                                        // se não chegou ao fim da frase, adiciona as expressões
                                        temporais presentes na oração para serem associadas a
                                        termos presentes em outras orações.
                                                ∀s_k ∈ ENT(o_z)
                                                ∀t_j ∈ ET(o_z) ^ ∀t_j ∈ TT(o_z)
                                                // Associa todas as expressões temporais
                                                presentes na oração que está sendo lida às
                                                Entidades presentes na mesma oração.
                                                        tf = VerificaTemporalidadeTermo (t_j)
                                                        AddItemConjunto(tf,conjAux)
                                        Senão
                                        // se chegou ao fim da frase
                                                ∀s_k ∈ ENT(o_z)
                                                ∀ t_j ∈ conjAux
                                                        // Associa todas as expressões temporais
                                                        presentes no conjunto conjAux às
                                                        entidades presentes na oração que está
                                                        sendo lida no momento.
                                                        tf = VerificaTemporalidadeTermo (t_j)

                                                        AssociaEntidadeData(s_k,After(tf))
                                Se StartWith(f_i, ADVTEMP_ANT) = "true"
                                // Se frase inicia com um advérbio temporal que indica
                                anterioridade...
                                        Se GetQtdDates(o_z) > 0
                                        // Caso a oração que está sendo lida no momento possua
                                        uma ou mais expressões temporais
                                                ∀s_k ∈ TNT(o_z)
                                                ∀t_j ∈ TTE(o_z) ^ ∀t_j ∈ TTNE(o_z)
                                                        // Associa todas as expressões temporais
                                                        presentes no conjunto conjAux às
                                                        entidades presentes na oração que está
                                                        sendo lida no momento.

                                                        tf = VerificaTemporalidadeTermo (t_j)
                                                        AssociaEntidadeData(s_k, tf)
                                        Se StartWith(o_z, ";") = "true"
                                        // Caso a oração inicie com o caractere ponto-e-vírgula
                                                ∀ t_j ∈ conjAux
                                                        tf = VerificaTemporalidadeTermo (t_j)
                                                        AssociaEntidadeData(s_k,Before(tf))
                                                clean(conjAux)
                                                // Limpa conjunto auxiliar
                                        Se z <> n então
                                        // Caso não tenha chegado ao fim da frase
                                                tf = VerificaTemporalidadeTermo (t_j)
                                                AddItemConjunto(tf,conjAux)
                                        Senão
                                                ∀ s_j ∈ conjAux
                                                        tf = VerificaTemporalidadeTermo (t_j)
                                                        AssociaEntidadeData(s_k,Before(tf))

        Return ItensTopicMap(D)
```

Apêndice B – Funções utilizadas pelo Algoritmo para Geração de Contextos Temporais

Abaixo são listadas as funções que são utilizadas no algoritmo formalizado no Apêndice A.

- $F(t) = \{f_1, f_2, f_3, \ldots, f_n\}$
  É a função que retorna o conjunto de frases *f* presentes em um texto *t*.
- DELIM: É o conjunto de caracteres que utilizados na divisão das frases em blocos necessários à extração de orações (vide função Orações(*f*)). Este conjunto é composto pelos caracteres vírgula e conjunções.
  E.g., DELIM = {",", "*and*", "*or*", ...}
- ADVTEMP_POS: É o conjunto de advérbios temporais identificados pelo RISO-TT que indicam posterioridade
  E.g., {"*after*", "*late*"}
- ADVTEMP_ANT: É o conjunto de advérbios temporais identificados pelo RISO-TT que indicam anterioridade.
- Divide(*f*,*DELIM*) = $\{b_1, b_2, \ldots, b_n\}$: Função que realiza a divisão da frase *f* em blocos a partir de uma lista de caracteres delimitadores *DELIM*.
- ContémVerbo(*b*): Função que verifica se um determinado trecho de um texto contém um verbo extraído pelo RISOExtractor.
- ContémEntidade(*b*): Função que verifica se um determinado trecho de um texto contém uma entidade extraída pelo RISOExtractor.
- StartWithVerb(*b*): Verifica se um determinado trecho de um texto se inicia com verbo.
- StartWithAdverb(*b*): Verifica se um determinado trecho de um texto se inicia com algum advérbio.
- GetQtdDates(*b*): Retorna a quantidade de expressões temporais presentes em uma frase *f*.
- Tamanho(C): Função que retorna a o tamanho da lista C.
- TT(o) é a função que retorna os termos temporais explícitos presentes na oração *o*.
- ET(o) é a função que retorna os termos temporais implícitos presentes na oração *o*. Estes termos podem ser:
  - Eventos: acontecimentos sociais, artísticos, desportivos, datas comemorativas, entre outros.
  - Objeto: constitui a existência de alguma coisa. Pode ser uma pessoa, uma instituição, um monumento, um país, um estado, uma cidade, entre outros.

As informações temporais relacionadas aos termos temporais implícitos são obtidas através de consultas à base de dados da *DBPedia*.

Entidades/Eventos Temporais possuem como atributo uma lista de datas associadas à eles.

s→listaExprTemporais = $\{t_1, t_2, ... t_n\}$

Onde *t* são as expressões temporais associadas implícitas associadas ao evento/entidade temporal *s*.

- ENT (o) é a função que retorna os termos não-temporais presentes na oração *o*.
  Os termos são classificados conforme a regra abaixo:
  - Se RisoEXTClassific(t) = "*NN*" OR "*NNP*" OR (RisoEXTClassific(t) = "*NNP*" AND WikiTemp (t) = ø) → t ∈ ENT.
  - Se (RisoEXTClassific(t) = "*NNP*" AND WikiTemp (t) ≠ ø) → t ∈ ET.
  - Se RisoTTDate(t) = TRUE → t ∈ TT.

  Onde:
  - NN: É a classificação dada pelo *POS-Tagging* do RISO-VTD para substantivos comuns.
  - NNP: É a classificação dada pelo *POS-Tagging* do RISO-VTD para substantivos próprios
- T(*t*): Retorna a temporalidade de um termo temporal não explícito presente em TT.
- RisoEXTClassific(t)
  Função que retorna a classificação gramatical dada à um termo pelo *POS-Tagging* do RISO-VTD.
- RisoTTDate(t)
  Função que retorna TRUE se o termo *t* é uma data marcada pelo RISO-TT.
- WikiTemp (t)
  Retorna a temporalidade de um termo de acordo com as informações presentes na Wikipedia.
- VerificaTemporalidadeTermo (t)
  Função que recebe um termo *t* pertencente aos conjuntos ET ou TT e retorna a data à qual ao termo se refere.
  Caso seja um termo temporal (TT), é retornada a própria data.

Caso seja um evento temporal (ET), são recuperadas informações relacionadas a este termo na base de dados da DBPedia.

- NORM(t) = Retorna a expressão temporal normalizada.
- maisPróximo(s) = Retorna a data mais próxima da entidade *s*. Entende-se como distância a quantidade de palavras existentes entre a entidade e uma data presente na oração.
- GetTemp(s) = Retorna o tempo relacionado ao sintagma *s*.
- ItensTopicMap (D) = $\{e_1, e_2, e_3, ..., e_n\}$ é o conjunto de itens que serão adicionados a um Topic Map de um documento D.
  Onde cada e_i é um par <s,t> sendo *s* um termo (em TT) e *t* um tempo (em ETT)

  e_i = <s_i,GetTemp(s_i)=t> ϵ ItensTopicMap (D) se:

  Dado uma frase *f* de um documento *D* e uma oração *o* em *D,* valem as seguintes regras de criação de Indices <s,t> para um documento D

  - R1: ET(o) = ø ^WikiTemp(s) = ø → ∀ *s* em Terms(*o*) vale <s> em ENT(*o*)
  - R2: ET(o) = ø ^ WikiTemp (s) =t' → ∀ *s* em Terms(*o*) <s, GetTemp(*s*)=t`> em ITT(o)
  - R3: ET(o) = t' → <s,t> em ITT(*o*)
  - R4: ET(o) = $T = t'_1, t'_2, \dots, t'_n$ → t=maisProximo(*p*,*T*) vale <*s*,*t*> em TT(*o*)
- AssociaEntidadeData(e,t) =
  Função que adiciona o termo temporal normalizado *t* à lista de expressões temporais que estão relacionadas a um termo temporal implícito *s*. O termo temporal implícito *e* é acrescido ao dicionário de termos temporais que é o conjunto de todas as entidades e suas expressões temporais relacionadas.

AssociaEntidadeData(s,t) = AddItemConjunto(NORM(*t*), s→listaExprTemporais) | <s,t> ϵ ItensTopicMap(D)

- StartsWith(f, C):
  Função que retorna "*true*" caso a frase *f* se inicie com a palavra *C* e "*false*" caso contrário.
- After(*t*): Função que retorna a expressão temporal que se refere a uma data posterior à data informada. E.g., After(10-may-1988) = X > 10-may-1988
- Before(*t*): Função que retorna a expressão temporal que se refere a uma data anterior à data informada. E.g., Before(10-may-1988) = X < 10-may-1988
- clean(*C*): Função que retira de um conjunto C todos os itens presentes neste conjunto.
- Inicializa(*C*): Função que inicializa a lista C vazia.
- proximaFrase(f_i): Vai para a frase imediatamente posterior à frase f_i.

Apêndice C – Exemplo de Documento Processado pelo RISO-GCT

A seguir é exibido um exemplo de documento processado pelo RISO-GCT, mais especificamente um fragmento do Documento "03_AmCivWar.sgm" do *corpus* WikiWars.

Lincoln's victory in the presidential election of 1860 triggered South Carolina's declaration of secession from the Union

By February 1861, six more Southern states made similar declarations

On February 7, the seven states adopted a provisional constitution for the Confederate States of America and established their temporary capital at Montgomery, Alabama

On March 4, 1861, Abraham Lincoln was sworn in as President

Under orders from Confederate President Jefferson Davis, troops controlled by the Confederate government under P.G.T. Beauregard bombarded the fort with artillery on April 12, forcing the fort's capitulation

For months before that, several Northern governors had discreetly readied their state militias; they began to move forces the next day

Anaconda Plan and blockade, 1861

In May 1861, Lincoln enacted the Union blockade of all Southern ports, ending regular international shipments to the Confederacy

By late 1861, the blockade stopped most local port-to-port traffic

The Union victory at the Second Battle of Fort Fisher in January 1865 closed the last useful Southern port and virtually ended blockade running

Western Theater 1861-1863

Upon the strong urging of President Lincoln to begin offensive operations, McClellan attacked Virginia in the spring of 1862 by way of the peninsula between the York River and James River, southeast of Richmond

A seguir é ilustrado o resultado do pré-processamento do documento pelo RISO-GCT responsável por incluir as marcações referentes às classificações gramaticais dos sintagmas e as expressões temporais contidas no documento.

Lincoln/NNP victory/NN in/IN the/DT presidential/JJ election/NN <RISOTime_type=Pre-EMT>of_1860</RISOTime> triggered/VBD <RISOTime_type=DE>South_Carolina</RISOTime> declaration/NN of/IN secession/NN from/IN the/DT Union/NNP ./.

<RISOTime_type=Pre-EBT>By_February_1861</RISOTime> :/: six/CD more/JJR Southern/NNP states/NNS made/VBN similar/JJ declarations/NNS ./.

<RISOTime_type=Pre-EBT>On_February_7</RISOTime> :/: the/DT seven/CD states/NNS adopted/VBN provisional/JJ constitution/NN for/IN the/DT Confederate/NNP States/NNPS of/IN America/NNP and/CC established/VBN their/PRP$ temporary/JJ capital/NN at/IN Montgomery/NNP :/: Alabama/NNP ./.

A/DT pre/NN war/NNP <RISOTime_type=EBT>February</RISOTime> Peace/NNP Conference/NNP <RISOTime_type=Pre-EMT>of_1861</RISOTime> met/VBD in/IN Washington/NNP in/IN failed/VBN attempt/NN at/IN resolving/VBG the/DT crisis/NN ./.

<RISOTime_type=Pre-EBT>On_March_4,_1861</RISOTime> :/: Abraham/NNP Lincoln/NNP was/VBD sworn/VBN in/IN as/IN President/NNP ./.

Under/IN orders/NNS from/IN Confederate/NNP President/NNP Jefferson/NNP Davis/NNP :/: troops/NNS controlled/VBN by/IN the/DT Confederate/NNP government/NN under/IN P./NNP G./NNP T./NNP Beauregard/NNP bombarded/VBD the/DT fort/NN with/IN <RISOTime_type=DE>artillery</RISOTime> <RISOTime_type=Pre-EBT>on_April_12</RISOTime> :/: forcing/VBG the/DT fort/NN capitulation/NN ./.

For/IN <RISOTime_type=CE>months_before_that</RISOTime> :/: several/JJ Northern/NNP governors/NNS had/VBD discreetly/RB readied/VBN their/PRP$ state/NN militias/NNS ;/: they/PRP began/VBD to/TO move/VB forces/NNS <RISOTime_type=EPT-UT>the_next_day</RISOTime> ./.

Anaconda/NNP Plan/NNP and/CC blockade/NN :/: <RISOTime_type=EMT>1861</RISOTime> Winfield/NNP Scott/NNP :/: the/DT commanding/VBG general/JJ of/IN the/DT <RISOTime_type=DE>U.S.</RISOTime> Army/NNP :/: devised/VBN the/DT Anaconda/NNP Plan/NNP to/TO win/VB the/DT war/NN with/IN as/IN little/JJ bloodshed/NN as/IN possible/JJ ./.

<RISOTime_type=Pre-EBT>In_May_1861</RISOTime> :/: Lincoln/NNP enacted/VBD the/DT Union/NNP blockade/NN of/IN all/DT Southern/NNP ports/NNS :/: ending/VBG regular/JJ international/JJ shipments/NNS to/TO the/DT Confederacy/NNP ./.

The/DT Union/NNP victory/NN at/IN the/DT Second/NNP Battle/NNP of/IN Fort/NNP Fisher/NNP <RISOTime_type=Pre-EBT>in_January_1865</RISOTime> closed/VBD the/DT last/JJ useful/JJ Southern/NNP port/NN and/CC virtually/RB ended/VBD blockade/NN running/VBG ./.

<RISOTime_type=DE>Easter</RISOTime>n/NNP Theater/NNP <RISOTime_type=EMT>1861-1863</RISOTime> ./.

Upon/IN the/DT strong/JJ urging/NN of/IN President/NNP Lincoln/NNP to/TO begin/VB offensive/JJ operations/NNS :/: McClellan/NNP attacked/VBD <RISOTime_type=DE>Virginia</RISOTime> in/IN <RISOTime_type=EPT-EMT>the_spring_of_1862</RISOTime> by/IN way/NN of/IN the/DT peninsula/NN between/IN the/DT York/NNP River/NNP and/CC James/NNP River/NNP :/: southeast/RB of/IN Richmond/NNP ./.

A seguir é ilustrado o RDF que irá compor o Topic Map, responsável por indexar o documento, onde as informações inseridas pelo RISO-GCT aparecem destacadas em **negrito**.

Devido à grande quantidade de informações inseridas no arquivo, parte do RDF foi suprimido.

```
<rdf:RDF
[...]
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/south_carolina">
  <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/state"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
 </rdf:Description>
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/South
Carolina@CONTEXT">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rel:hasContext
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/american_state/n/one_of_the_50_state
s_of_the_united_states"/>
  <rel:hasContext
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/deep_south/n/south_carolina_and_geo
rgia_and_alabama_and_mississippi_and_louisiana"/>
  <rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/state"/>
  <rel:hasContext
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/carolina/n/the_area_of_the_states_of_n
orth_carolina_and_south_carolina"/>
 </rdf:Description>
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/South_Carolina
/r/HasDate">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-
1860_to__-__-1860 /r/HasDate"/>
 </rdf:Description>
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Lincoln@CONTEXT">
  <rel:hasContext
rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/president_of_unite_state/n/the_person
_who_holds_the_office_of_head_of_state_of_the_united_states_government"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
```

<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/state_capital/n/the_capital_city_of_a_political_subdivision_of_a_country"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/domestic_sheep/n/any_of_various_breeds_raised_for_wool_or_edible_meat_or_skin"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/lawyer/n/a_professional_person_authorized_to_practice_law"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/lincoln_car"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/nebraska/n/a_midwestern_state_on_the_great_plains"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Lincoln/r/HasDate">
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en//r/HasDate"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-1860_to__-__-1860 /r/HasDate"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-05-1861_to__-05-1861 /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/presidential election@CONTEXT">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/change_society"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/hold_every_four_year"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/hold_to_elect_president"/>
</rdf:Description>

<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/presidential_election /r/HasDate">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-1860_to__-__-1860 /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/secession@CONTEXT">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/artistic_movement/n/a_group_of_artists_who_agree_on_general_principles"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/separation/n/the_act_of_dividing_or_disconnecting"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/secession"/>
<rel:hasContext rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/school/n/a_body_of_creative_artists_or_writers_or_thinkers_linked_by_a_similar_style_or_by_similar_teachers"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/secession /r/HasDate">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-1860_to__-__-1860 /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/union">
<rel:CapableOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/demand_high_wage"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rel:AtLocation rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/workplace"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Union /r/HasDate">
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-1862_to__-__-1862 /r/HasDate"/>

<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-1860_to__-__-1860 /r/HasDate"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/ /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Southern ports@CONTEXT">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Southern_states /r/HasDate">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-02-1861_to__-02-1861 /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/made /r/HasDate">
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/5-09-__to_5-09-__ /r/HasDate"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-02-1861_to__-02-1861 /r/HasDate"/>
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/ /r/HasDate"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/similar_declarations /r/HasDate">
<rdfs:subClassOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-02-1861_to__-02-1861 /r/HasDate"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Western Theater@CONTEXT">
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>

```
 </rdf:Description>
 <rdf:Description  rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Western_Theater
/r/HasDate">
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
   <rdfs:subClassOf          rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/__-__-
1861_to__-__-1863 /r/HasDate"/>
 </rdf:Description>
[...]
</rdf:RDF>
```

Os arquivos completos gerados pelo RISO-GCT encontram-se disponíveis em: http://bit.ly/1NpVWFi.

Apêndice D – Trecho das Saídas Geradas pelo RISO-GCT

A tabela deste apêndice ilustra um trecho dos resultados obtidos após a execução do RISO-GCT pelo RISO-GCT recebendo como entrada o documento "*Napoleon.txt*". Estes resultados foram comparados com o gabarito gerado para validação deste trabalho. A primeira coluna o conceito extraído do documento a segunda e terceira colunas representam a lista de datas relacionadas aos conceitos na forma normalizada pelo RISO-TT e as datas na forma não-normalizada, respectivamente.

Entidade	ET normalizada	ET não-normalizada
admitted	|[?]-05-[?]|[?]-[?]-1784	|In May|in 1784|
age	|[?]-[?]-1821	|in 1821|
Ajaccio	|15-08-1769	|on 15 August 1769|
Allied victory	|	|six years|
Allies	| |[?]-04-1814	|The next year|April 1814|
Amiens	|[?]-[?]-1802	|in 1802|
annihilated	|[?]-[?]-1807	|in 1807|
artillery	|[?]-[?]-1789| |[?]-09-1785	|in 1789|one year|in September 1785|
artillery officer	|[?]-[?]-1789|	|in 1789|one year|
attack	|[?]-[?]-1809	|in 1809|
attention	|[?]-[?]-1807	|in 1807|
attorney	|[?]-[?]-1777	|present|in 1777|
Auerstedt	|[?]-[?]-1807	|in 1807|
Austria	|[?]-[?]-1796|[?]-[?]-1809| |[?]-[?]-1810|	|in 1796|in 1809|the fall of 1809|in 1810|In early 1813|later in the year|
Austrians	|[?]-[?]-1796|[?]-[?]-1809| |	|in 1796|in 1809|In early 1813|later in the year|
Autun	|[?]-01-1779	|In January 1779|
Auxonne	|[?]-[?]-1789	|in 1789|
banished	|[?]-[?]-1789|[?]-[?]-1793	|in 1789|in 1793|
Battle of Austerlitz	|02-12-1805 < X < 02-12-1805	|02-12-1805 < X < 02-12-1805|
Battle of Friedland	|[?]-[?]-1807	|in 1807|
Battle of Leipzig	|[?]-10-1813	|In October 1813|
Battle of Trafalgar	|[?]-10-1805	|In October 1805|
Battle of Wagram	|07-06 < X < 07-06|06-07-1809 < X < 06-07-1809|06-07-[?] < X < 06-07-[?]	|07-06 < X < 07-06|06-07-1809 < X < 06-07-1809|06-07-[?] < X < 06-07-[?]|
Battle of Waterloo	|[?]-06-[?]	|in June|
battles	|[?]-[?]-1815	|in 1815|
battles of Jena	|[?]-[?]-1807	|in 1807|
Beauharnais	|[?]-[?]-1796	|in 1796|
become	|	|one year|
blockade	|[?]-[?]-1812	|in 1812|
Saddam Hussein	|13-12-2003 < X < 13-12-2003	|on December 13, 2003|
dashed	|[?]-05-[?] < X < [?]-05-[?]	|in May|
Task Force	|22-07-[?] < X < 22-07-[?]	|On July 22|

invasion phase	|15-04-[?] < X < 15-04-[?]|19-03-2016 < X < 30-04-2016	|on April 15|March 19-April 30|
dramatic visit	|1-05-2003 < X < 1-05-2003	|On May 1, 2003|
next few months	|	|the next few months|
appointment	|	|until May 11, 2003|
Qusay	|22-07-[?] < X < 22-07-[?]	|On July 22|
al Qaeda	|[?]-[?]-2004 < X < [?]-[?]-2004	|of 2004|
affiliated al Qaeda group	|[?]-[?]-2004 < X < [?]-[?]-2004	|of 2004|
former regime	|[?]-[?]-2003 < X < [?]-[?]-2003	|of 2003|
military	|[?]-11-2004 < X < [?]-11-2004	|in November 2004|
Operation Red Dawn	|13-12-2003 < X < 13-12-2003	|on December 13, 2003|
day	|22-07-[?] < X < 22-07-[?]	|On July 22|

O gabarito utilizado para validação desta pesquisa encontra-se disponível em: http://bit.ly/1NpVWFi.