RECONHECIMENTO DE ENTIDADES NOMEADAS EM NOTÍCIAS DE GOVERNO

Tiago Santos da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Jano Moreira de Souza

Rio de Janeiro
Fevereiro de 2012

RECONHECIMENTO DE ENTIDADES NOMEADAS EM NOTÍCIAS DE GOVERNO

Tiago Santos da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Jano Moreira de Souza, Ph.D.

_____

Prof. Geraldo Bonorino Xexéo, D.Sc.

_____

Prof.ª Vanessa Braganholo Murta, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

FEVEREIRO DE 2012

Silva, Tiago Santos da

Reconhecimento de Entidades Nomeadas em Notícias de Governo/ Tiago Santos da Silva. – Rio de Janeiro: UFRJ/COPPE, 2012.

XIV, 99 p.: il.; 29,7 cm.

Orientador: Jano Moreira de Souza

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2012.

Referências Bibliográficas: p. 93-99.

1. Reconhecimento de Entidades Nomeadas. 2. Análise de Notícias. 3. Integrador de Notícias de Governo. I. Souza, Jano Moreira de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Primeiramente, agradeço à Deus por ter me dado saúde, força e todas as demais condições necessárias para concluir este trabalho.

Agradeço aos meus familiares por estarem sempre apoiando todas as minhas decisões e ajudando nos momentos difíceis.

Agradeço a UFRJ por me conceder a oportunidade de obter uma educação de qualidade e excelência durante toda a minha vida acadêmica.

Agradeço aos meus orientadores Jano Moreira de Souza e Sérgio Assis Rodrigues, que me auxiliaram durante toda essa jornada, dedicando o tempo deles em apresentar críticas e sugestões visando o aperfeiçoamento deste trabalho.

Agradeço a Daiane Evangelista Ferreira, mestranda em Ciência da Informação na Universidade Federal Fluminense (UFF) por dedicar seu tempo auxiliando na anotação da Coleção de Notícias de Governo – UFRJ.

Agradeço a todos os meus amigos que me ajudaram com opiniões, ideias e incentivos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

RECONHECIMENTO DE ENTIDADES NOMEADAS EM NOTÍCIAS DE GOVERNO

Tiago Santos da Silva

Fevereiro/2012

Orientador: Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

Nos dias atuais, o volume de notícias publicadas na web cresce aceleradamente, de maneira que se torna difícil para uma pessoa acompanhar toda a informação disponibilizada. Tendo em vista o grande volume de informações apresentadas ao público, diversos trabalhos na literatura têm realizado pesquisas relacionadas ao tratamento automatizado das notícias. As máquinas de busca disponíveis, atualmente, na web, possibilitam ao usuário buscar informações em grandes bases de notícias, entretanto, tais buscas se baseiam na análise do texto desconsiderando sua semântica implícita. Por exemplo, o termo Brasil pode ser visto como um local, mas também pode ser visto como uma organização. Dentre os diversos trabalhos relacionados à extração de semântica em textos podemos mencionar o Reconhecimento de Entidades Nomeadas (NER, sigla em inglês). A tarefa de Reconhecimento de Entidades Nomeadas consiste na tarefa de reconhecer os elementos como Pessoa, Local, Organização, etc. em um dado contexto. Este trabalho tem como objetivo desenvolver e apresentar os métodos utilizados na concepção de um sistema que extraia as entidades nomeadas em notícias de governo na língua portuguesa. As informações consideradas relevantes nesse trabalho é a identificação de entidades tais como: Pessoa, Local, Organização, Cargos, Programas, Eventos, Siglas e algumas das relações que existem entre essas entidades. Além disso, este trabalho apresenta o Integrador de Notícias de Governo cujo objetivo é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

NAMED ENTITIES RECOGNITION IN GOVERNMENT NEWS

Tiago Santos da Silva

February/2012

Advisor: Jano Moreira da Silva.

Department: System Engineering

Currently, the volume of news published on the web has grown rapidly, so that it becomes difficult for a person to monitor all the information provided. Given the large volume of information presented to the public, several studies in literature have conducted research related to the automated processing of news. Search engines available today in the web allow to the user to search for information in large databases of news, however, such searches are based on analysis of the text regardless of their implicit semantics. For example, the term Brazil can be seen as a location, but can also be seen as an organization. Among several works related to the extraction semantic in texts we can mention the Named Entity Recognition (NER). The task of Named Entity Recognition is the task of recognizing the elements such as Person, Location, Organization, etc. in a given context. This work aims to develop and present the methods used to design a system that extracts named entities on government news in the Portuguese language. The information considered relevant in this work is to identify entities such as Persons, Locations, Organizations, Ranks, Programs, Events, Acronyms and some of the relationships that exist between these entities. Furthermore, this paper presents the Government News Integrator, whose objective is to maintain a centralized database of government news that can be easily accessed by other systems. It acts as a Web Portal where readers rather than humans are programs that consume the metadata and the fonts are web pages.

Sumário

Figuras

Gráficos

Tabelas

Capítulo 1. Introdução

1.1 Motivação

Nos dias atuais, o volume de notícias publicadas na web cresce aceleradamente, de maneira que se torna difícil para uma pessoa acompanhar toda a informação disponibilizada. Mesmo que o escopo das notícias de interesse seja reduzido a um tema específico, a tarefa de se manter atualizado sem o auxílio de uma ferramenta automatizada é desafiadora e não raramente inviável.

Ao longo dos anos, um grande esforço foi, e continua sendo, empregado por pesquisadores de todo o mundo na tentativa de facilitar o acesso às notícias. As máquinas de busca são claros exemplos de sistemas automatizados que auxiliam as pessoas no acesso às informações de seus interesses. Entretanto, as máquinas de buscas tradicionais processam as notícias desconsiderando a semântica do texto e, portanto, apresentam limitações na qualidade dos resultados. Dessa maneira, a habilidade de interpretar automaticamente os elementos semânticos presentes em textos faz-se necessária.

Diversos problemas e soluções relacionados à exploração dos elementos semânticos de um texto foram apresentados na literatura. No entanto, grande parte dos métodos desenvolvidos foram testados, somente, para textos da língua inglesa. Embora muitos algoritmos tenham sido desenvolvidos com a característica de serem multilíngues, poucos destes foram efetivamente testados para a língua portuguesa. Essa situação pode criar a ilusão de que alguns problemas resolvidos efetivamente para uma dada língua também já estão resolvidos para a língua portuguesa.

Por exemplo, no Conference on Computational Natural Language Learning 2002 (CoNLL) foi realizada uma competição entre 16 sistemas de reconhecimento de entidades nomeadas, onde todos os sistemas eram multilíngues e que deveriam classificar entidades em duas bases com idiomas distintos: inglês e alemão. Todos os sistemas compartilhavam o mesmo conjunto de exemplos para treinamento e o mesmo conjunto de testes.

Apesar dos algoritmos utilizados serem os mesmos, houve uma diferença considerável entre o desempenho obtido na classificação de textos em inglês em relação aos textos em

alemão. O F-Score médio de acertos obtidos na avaliação da base inglesa foi de 82,17% enquanto a avaliação da base alemã obteve um F-Score de 65,55%. Logo, os testes realizados sobre um algoritmo com característica multilíngue não necessariamente indicam que este funcione bem em outros idiomas.

Então, neste trabalho, o problema de excesso de informação é trazido para a realidade da língua portuguesa, que é um cenário pouco explorado se comparado à gama de estudos realizados em outras línguas. Logo, a motivação inicial para este trabalho é a carência de estudos relacionados ao tratamento automatizado de notícias sobre a ótica da língua portuguesa.

1.2 Delimitação

A tarefa de extrair informações semânticas de uma notícia é uma tarefa complexa e por isso exige uma área de estudos dedicada somente a ela. Uma notícia pode pertencer a várias categorias tais como esporte, cultura, economia e política. Em meio a essa diversidade de notícias há uma variedade de características particulares a cada uma delas. Então, propor um estudo que englobe todas as categorias existentes dificulta o estudo destas características.

Neste trabalho, apenas as notícias relacionadas ao governo são consideradas como relevantes, limitando o universo de notícias estudadas. Com essa restrição espera-se aproveitar as características intrínsecas a esse domínio na extração de elementos semânticos.

1.3 Objetivo

O objetivo principal deste trabalho é realizar estudos sobre a extração de entidades nomeadas das notícias de governo redigidas na língua portuguesa. O objetivo principal deste trabalho é subdividido em dois objetivos específicos que se incrementam.

1.3.1 Reconhecimento de Entidades Nomeadas

Este trabalho propõe o estudo do Reconhecimento de Entidades Nomeadas na língua portuguesa em notícias de governo. Reconhecimento de Entidades Nomeadas (NER) [SUTTON; MCCALLUM, 2006] é o problema de identificar e classificar nomes próprios em textos, incluindo localizações, tais como Brasil e Rio de Janeiro; pessoas, tais como Dilma e Miriam; e organizações, tais como Ministério da Educação e Ministério da Cultura.

A ideia por trás desta proposta é que as entidades nomeadas, em geral, são elementos fundamentais para a compreensão de uma dada notícia. Esta proposta fornece subsídios para respostas automáticas do tipo: "*sobre quem a notícia se trata?", "sobre que lugar(es) a notícia se trata?",* etc.

Algumas iniciativas relacionadas ao Reconhecimento de Entidades Nomeadas em textos da língua portuguesa têm sido realizadas. A principal delas é o HAREM [MOTA; SANTOS, 2008] que é uma avaliação conjunta na área do reconhecimento de entidades mencionadas (REM, sigla em português), organizada pela Linguateca [SANTOS, 2011].

Entretanto, mesmo os melhores resultados obtidos no HAREM possuem taxas de erros significativas. Por exemplo, segundo dados retirados do artigo de Santos e Cardoso [2006] o melhor F-Score que engloba as categorias Pessoa, Local e Organização foi de 64,37%.

Então, esse trabalho propõe desenvolver um sistema de reconhecimento de entidades nomeadas destinado às notícias de governo escritas em português. A justificativa para esta proposta é que dada a restrição imposta aos textos a serem processados, há uma redução na complexidade das características envolvidas na classificação de uma entidade e, portanto, há a possibilidade de aumento de desempenho em relação aos sistemas desenvolvidos para serem genéricos.

Este trabalho propõe a aplicação do Linear-Chain Conditional Random Field na tarefa de reconhecimento de entidades nomeadas em notícias de governo. O Linear-Chain Conditional Random Field é um caso especial do Conditional Random Field (CRF). O CRF [SUTTON; MCCALLUM, 2006] é uma distribuição condicional $P(Y|X)$ com um modelo gráfico associado. A variável X é um vetor de variáveis aleatórias de entrada e Y é um vetor de variáveis aleatórias de saída. Então $P(Y|X)$ é a probabilidade de dado como entrada o vetor X obter a saída Y. Nesse caso o vetor X é a sequência de palavras em uma notícia e Y uma possível rotulação de entidades para X.

O Linear-Chain Conditional Random Field foi escolhido nesse trabalho por ser um método probabilístico e por isso de fácil adaptação à outras línguas, que nesse trabalho é o português.

1.3.2 Integração das Notícias de Governo

Como discutido no início desta seção, diversas pesquisas relacionadas ao tratamento de notícias foram abordadas na literatura, entretanto, a maioria das pesquisas tinham como

pressuposto uma base de notícias de fácil acesso, que na maioria das vezes foi elaborada com a finalidade de pesquisa.

Já recentemente, trabalhos têm sido realizados utilizando como fonte notícias publicadas por meio de Feeds RSS. Feeds [SILVA, 2010] são listas de atualização de conteúdo de um determinado site, escritos com especificações baseadas em XML. Estes Feeds contêm informações sobre as notícias e links apontando para a página original de cada notícia.

Todavia, no mundo real essa abordagem encontra dois problemas:

Problema 1: Em geral, os Feeds RSS são constantemente atualizados e possuem uma quantidade limitada de notícias, então, se um sistema consumidor destas notícias perde algumas dessas atualizações as notícias serão perdidas, acarretando perda de informação. Já as páginas HTML das notícias tendem a ter um período de persistência muito maior.

Problema 2: Sistemas de tratamento de notícias que utilizam como base Feeds não possuem acesso a sites que não disponibilizam Feeds. Esta situação é bem comum nos sites do domínio .gov.br, já que a maioria dos sites ainda não adotaram o uso de Feeds.

Portanto, para o uso prático das ferramentas de tratamento de notícias é necessário ter um método que permita a extração das notícias. Entende-se por extração o processo de identificar as páginas que possuem notícias, e dada a página, remover todo conteúdo irrelevante, tais como cabeçalho, rodapé, etc.

Este trabalho propõe a concepção do Integrador de Notícias do Governo. O objetivo deste integrador é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

O formato de dados utilizado para permitir a interoperabilidade das notícias é o próprio RSS. Como o RSS é gerado dinamicamente e gerenciado pelo integrador, os problemas 1 e 2 podem ser solucionados, já que um sistema X pode requisitar notícias passadas (Problema 1) e pode requisitar notícias de um site que não possui RSS, já que este pode ser gerado automaticamente (Problema 2).

1.4 Organização do Trabalho

Este trabalho é organizado da seguinte forma: No capítulo 2 é apresentada a revisão literária realizada durante esse trabalho. Essa revisão fornece a base necessária para que o leitor tenha uma visão geral sobre as técnicas utilizadas e o contexto dos problemas propostos.

No capítulo 3 é apresentada a representação semântica de notícias por meio de suas entidades nomeadas, bem como os métodos utilizados para a elaboração de um sistema de reconhecimento de entidades nomeadas para notícias de governo. Além disso, é apresentado a Coleção de Notícias de Governo – UFRJ, que é fruto deste trabalho. A Coleção de Notícias de Governo – UFRJ é uma coleção de notícias retiradas de sites do governo brasileiro, criada exclusivamente para este trabalho. As notícias foram anotadas manualmente com o auxílio de uma mestranda em Ciência da Informação da Universidade Federal Fluminense (UFF).

No capítulo 4 é apresentado o Integrador de Notícias de Governo, os métodos utilizados, as dificuldades encontradas e um estudo de caso do uso do integrador no Portal de Notícias de Governo. Este portal é uma iniciativa que envolve a cooperação da SLTI/MP, SECOM/PR, COPPE/UFRJ e SERPRO no intuito de prover ao público e as secretarias - SECOM e SRI - um ambiente de fácil acesso às notícias.

No capítulo 5 são detalhados os experimentos realizados na concepção do sistema de Reconhecimento de Entidades Nomeadas resultante deste trabalho. Este capítulo apresenta os desempenhos obtidos com a aplicação do Linear-Chain Conditional Random Field sobre a Coleção de Notícias de Governo – UFRJ e sobre a Coleção Dourada – HAREM.

No capítulo 6 são apresentadas as conclusões sobre este trabalho, bem como sugestões para trabalhos futuros.

Capítulo 2. Revisão Literária - Reconhecimento de Entidades Nomeadas

2.1 Conceitos e Desafios

Segundo [BUNESCU, 2007] um componente básico na concepção de um sistema de Extração de Informação (IE) é o reconhecimento de entidades nomeadas. Reconhecimento de Entidades Nomeadas (NER, sigla em inglês) [SUTTON; MCCALLUM, 2006] é o problema de identificar e classificar nomes próprios em textos, incluindo localizações, tais como Brasil e Rio de Janeiro; pessoas, tais com Dilma e Miriam; e organizações, tais como Ministério da Educação e Ministério da Cultura.

A tarefa de reconhecimento de entidades [SUTTON; MCCALLUM, 2006] é, dada uma sequência, segmentá-la em palavras, que são partes de uma entidade, e então classificar cada palavra por tipo (pessoa, organização, local, etc.). Por exemplo, na figura 1 é apresentado um exemplo de identificação de entidades nomeadas. A sentença contém quatro entidades diferentes: Wolff e Del Bosque como pessoas, Argentina como uma localização e Real Madrid como uma organização.

[PESSOA] **Wolff**, atualmente um jornalista na
[LOCAL] **argentina**, jogava com [PESSOA] **Del Bosque** no final dos anos setenta no
[ORGANIZAÇÃO] **Real Madrid**.

Figura 1: Exemplo adaptado de SANG [2002].

O desafio deste problema é que muitas entidades nomeadas raramente aparecem mesmo em grandes conjuntos de treinamento, portanto, o sistema deve identificar tais entidades utilizando apenas o contexto em que esta é empregada. Outro desafio, é que a mesma identidade pode apresentar classificações diferentes dependendo do contexto em que se encontram.

Por exemplo, na sentença 1 a entidade Brasil representa um local, enquanto na sentença 2 ela representa uma organização (Governo Brasileiro). Esses desafios tornam o uso de gazetters obsoletos, já que eles possuem as limitações descritas acima.

> No período entre 2011 e 2014, o PAC 2 investirá R$ 35,1 bilhões para a execução de obras de saneamento básico no País, retomando investimentos em um setor essencial para a saúde e qualidade de vida da população. Todas as regiões do **Brasil** serão beneficiadas, de acordo com o déficit de saneamento dos municípios. [MINISTÉRIO, 2011b]. (Sentença 1)
>
> O projeto apresentado pelo **Brasil** na reunião do Cosiplan incluiu tanto propostas para a integração física das redes já existentes nos países de fronteira quanto para sua integração lógica. [MINISTÉRIO, 2012c]. (Sentença 2)

2.2 Metodologia para Tarefa de Reconhecimento de Entidades Nomeadas

A capacidade de reconhecer as entidades anteriormente desconhecidas é uma parte essencial da tarefa de Reconhecimento de Entidades Nomeadas [NADEAU; SEKINE, 2007]. Inicialmente os algoritmos de NER eram baseados em regras escritas manualmente que indicavam a existência de uma determinada entidade dado um contexto.

Por exemplo, se um substantivo capitalizado é precedido do termo “no” então este substantivo é um Local; se um substantivo capitalizado é precedido do termo “diz” então este substantivo é uma Pessoa.

Entretanto, a escrita manual de regras tem a desvantangem de ser tediosa, pois necessita da escrita de muitas regras, e tais regras precisam cobrir os mais diversos cenários possíveis que dificilmente podem ser descobertos manualmente.

Então, dado a dificuldade de conceber um sistema NER baseado em regras escritas manualmente, diversos estudos foram realizados na tentativa de elaborar métodos de aprendizado automatizado. Os métodos utilizados na elaboração de sistemas de reconhecimento de entidades são divididos em três categorias: aprendizado supervisionado, aprendizado semi-supervisionado e aprendizado não supervisionado.

2.2.1 Aprendizado Supervisionado

O aprendizado supervisionado tem sido a técnica predominante na tarefa de reconhecimento de entidades nomeadas [NADEAU; SEKINE, 2007]. A ideia dos métodos baseados em aprendizado supervisionado consiste em extrair de uma coleção de treinamento as características necessárias para a correta classificação das entidades nomeadas mencionadas em textos desconhecidos.

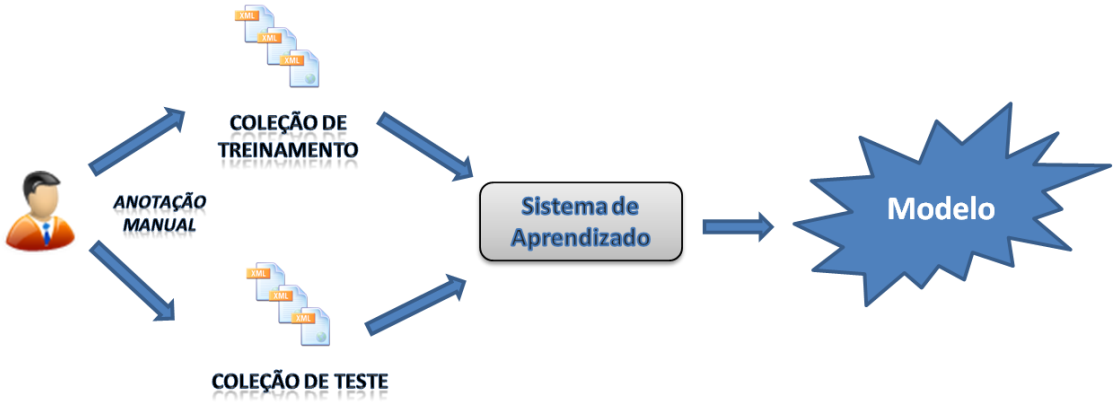

Figura 2: Arquitetura Simplificado do Aprendizado Supervisionado

A figura 2 apresenta uma arquitetura simplificada do aprendizado supervisionado. Inicialmente um ou mais especialistas avaliam um conjunto de textos, para cada texto os especialistas anotam manualmente as suas entidades nomeadas de forma a criar exemplos. Em seguida, o conjunto de textos anotados é dividido em um conjunto de treino e de teste.

O conjunto de treino é utilizado para transferir o conhecimento do especialista para o sistema de aprendizado. O conjunto de teste é utilizado para validar a generalização do aprendizado obtido pelo sistema por meio do conjunto de treino. Para que a validação não seja comprometida é essencial que o conjunto de treino e treinamento seja disjunto evitando que o sistema decore em vez de generalizar.

Alguns dos métodos que empregam o aprendizado supervisionado são: Hidden Markov Models (HMM) [BIKEL et al, 1997], Decision Trees [SEKINE, 1998], Maximum Entropy Models (ME) [BORTHWICK et al, 1998], Support Vector Machines (SVM) [ASAHARA; MATSUMOTO, 2003] e Conditional Random Fields (CRF) [MCCALLUM; LI, 2003].

2.2.1.1 Vantagens

Os métodos de aprendizado supervisionado apresentam a vantagem de possuir uma base rica em exemplos para a correta gereneralização do aprendizado. Como essa base é elaborada por especialista a acurácia da base de exemplos é alta reduzindo o ruído de informação.

Além disso, a elaboração de um conjunto de testes permite a monitorazação automatizada da evolução do aprendizado dos sistemas.

2.2.1.2 Desvantagens

Embora o método de aprendizado supervisionado apresente muitas vantagens há dificultadades em sua concepção. A coleção de exemplos precisa ser completa e correta, ou seja, precisa englobar a maioria dos contextos possíveis e apresentar o menor número de erros possíveis. Caso um contexto não esteja contido nos exemplos o sistema não poderá aprender toda uma classe de sentenças e caso haja algum erro no exemplos o sistema pode generalizar esse erro de modo a amplificar o erro.

Por isso a concepção de uma coleção de exemplos é custosa e tediosa, umas vez que requer uma grande quantidade de tempo e de mão de obra especializada.

2.2.2 Aprendizado Semi Supervisionado

O Aprendizado Semi Supervisionado é o mais utilizado depois do aprendizado supervisionado na tarefa de reconhecimento de entidades nomeadas. Esse tipo de aprendizado consiste em criar manualmente apenas um pequeno conjunto de exemplos e partir deste conjunto de exemplos fornecer a um sistema os requisitos minímos para conseguir aprender a partir de coleções de textos não anotadas.

Figura 3: Arquitetura Simplificado do Aprendizado Semi Supervisionado

Os métodos que empregam o aprendizado Semi Supervisionado necessitam apenas de um pequeno conjunto de exemplos. Esse conjunto reduzido de exemplos são utilizados como ponto de partida para o processo de aprendizado.

A figura 3 apresenta uma arquitetura simplificada do aprendizado semi supervisionado. Nesse tipo de aprendizado um ou mais especialistas elaboram uma coleção

de treino e teste contendo as entidades nomeadas anotadas manualmente da mesma maneira que no aprendizado supervisionado, mas com um tamanho inferior. A partir do conjunto reduzido de treinamento este pode ser automaticamente expandido.

Entidades com nomes que raramente podem ser atribuídos a mais de um tipo de entidade podem ser utilizados para a expansão do conjunto de treinamento. Por exemplo, a expressão *Dilma Rousseff* muito provavelmente é uma entidade do tipo Pessoa, então todas as sentenças na coleção não anotada que possuam esta expressão podem constituir um novo exemplo para o treinamento, dessa maneira expandido o conjunto de exemplos já que novos contextos podem ser adicionados à coleção.

Esse mecanismo de expansão é plausivel porque a qualidade da coleção de treinamento está mais associada à variedade de contextos do que à variedade de nomes. Alguns dos métodos que empregam o aprendizado semi supervisionado podem ser encontrados em: Brin [1998], Collins e Singer [1999], Collins [2002] e Riloff e Jones [1999].

2.2.2.1 Vantagens

Os métodos de aprendizado semi supervisionado tentam balanciar as vantagens dos métodos supervisionados com os métodos não supervisionados. Nos métodos semi supervisionados a necessidade de exemplos elaborados manualmente é reduzida, minimizando os custos de tempo e mão de obra na elaboração da base de treino. E apresenta a vantagem de ter uma base de conhecimento que não está disponível no aprendizado não supervisionado.

2.2.2.2 Desvantagens

A desvantagem deste método reside na complexidade dos algortimos envolvidos, uma vez que estes devem trabalhar com um conjunto mais reduzido de exemplos. Além do algoritmo de treinamento é preciso elaborar algoritmos para a expansão do conjunto de treinamento.

2.2.3 Aprendizado Não Supervisionado

No aprendizado não supervisionado não há um conjunto de exemplos para ser treinado. A típica abordagem adotada nesse tipo de aprendizado é a clusterização. Por exemplo,

podemos criar grupos de entidades nomeadas a partir da similaridade dos contextos em que estas se encontram. Abaixo segue alguns exemplos do emprego dessa técnica:

Alfonseca e Manandhar [2002] estudam o problema de rotular uma palavra de entrada com um tipo apropriado de entidade nomeada. Tipos de entidades são retirados do WordNet (por exemplo, localização > país, animate> pessoa, animate > animal, etc.). A abordagem é para atribuir uma assinatura de tópico para cada termo WordNet listando apenas as palavras que frequentemente co-existem com ele em um grande corpus. Então, dada uma palavra de entrada em um determinado documento, o contexto da palavra (palavras que aparecem em uma janela de tamanho fixo em torno da palavra de entrada) é comparado ao tipo de assinatura e classificado naquele mais semelhante.

Em Evans [2003], o método para a identificação de hipônimos/hiperônimos descritos na obra de Hearst [1992] é aplicado a fim de identificar hiperônimos potenciais de sequências de palavras em maiúsculas que aparecem em um documento. Por exemplo, quando X é uma sequência maiúscula, a consulta "como X", é pesquisada na web e, nos documentos recuperados, o substantivo que precede imediatamente a consulta pode ser escolhido como o hiperônimo de X. Da mesma forma, em Cimiano e Völker [2005], os padrões de Hearst são usados, mas desta vez o recurso consiste em contar o número de ocorrências de passagens como: "cidade como", "organização como", etc.

2.2.3.1 Vantagens

Os métodos de aprendizado não supervisionado possuem a vantagem de não requisitar a criação de uma base de exemplos anotadas manualmente para o treinamento, por isso eles podem ser utilizados sobre uma vasta quantidade de textos no aprendizado.

2.2.3.2 Desvantagens

A desvantagem deste método reside na dificuldade de conceber algoritmos para reconhecimento de entidades nomeadas sem o uso de exemplos. Portanto, o uso dessa alternativa é restrito à algumas classes especiais de problemas de reconhecimento de entidades, que possuem o potencial de formarem clusters em relação a algumas propriedades específicas.

2.3 Métodos Probabilísticos

Diversos métodos de aprendizado supervisionados foram discutidos na literatura com o objetivo de solucionar a tarefa de reconhecimento de entidades nomeadas. Nesse trabalho foram estudados métodos baseados em modelos probabilísticos.

Os métodos baseados em modelos probabilísticos foram escolhidos como fonte de estudo nesse trabalho porque eles podem ser facilmente adaptados para o uso em outros idiomas. Uma que vez que o aprendizado é realizado estatisticamente, não há a necessidade de descrever conceitos semânticos próprios de um idioma para efetuar o aprendizado.

A subseção abaixo descreve os dois métodos mais relevantes, baseados em modelos probabilísticos, descritos na literatura e por isso os mais estudados nesse trabalho. O primeiro método, o Hidden Markov Models, é descrito brevemente e em seguida o Linear Conditional Random Field é explicado detalhadamente já que este é derivado do HMM.

Nesse trabalho, o método Linear-Chain Conditional Random Field foi utilizado nos experimentos de reconhecimento de entidades nomeadas em notícias de governo. Esse método foi escolhido por duas razões, a primeira é por ser um modelo probabilístico que não exige conhecimento explícito do idioma a ser estudado, e por isso pôde ser utilizado em textos da língua portuguesa sem muitas complicações.

E a segunda, é porque a metodologia utilizada nesse trabalho é a supervisionada, já que esta permite a transferência de conhecimento por meio da elaboração de uma base de treinamento, além disso, permite a comparação dos resultados obtidos pela base de treinamento elaborada nesse trabalho com a base já existente Coleção Dourada que será detalhada posteriormente.

2.3.1 Hidden Markov Models

O Hidden Markov Models (HMM) [RABINER, 1989] é uma extensão da cadeia de Markov, que adiciona mais um processo estocástico ao modelo, ou seja, o HMM é um modelo estocástico que modela dois processos estocásticos sendo um oculto. Diferentemente da cadeia de Markov as observações do modelo seguem uma função probabilística.

O HMM tem como objetivo simular uma fonte de sinal a partir de um conjunto de observações. Essa fonte de sinal é um sistema X que dado um conjunto de observações $O = \{O_1, O_2, ... O_n\}$ produz os sinais $Q = \{Q_1, Q_2, ... Q_n\}$. O modelo é composto basicamente

pelas observações (tipos distintos de observações), pelos números de estados da cadeia de markov, pela função de transição entre estados, pela função probabilística de emissão de um sinal dado um estado e pela função probabilística do estado inicial.

O HMM tem sido amplamente utilizado em tarefa de processamento de linguagem natural tais como reconhecimento de *part-of-speech (POS)* e o reconhecimento de entidades nomeadas. Em tarefas de processamento de linguagem natural geralmente as observações são as sequências de palavras presentes em um texto e os sinais são as informações semânticas relativas às palavras observadas.

2.3.2 Linear-Chain Conditional Random Field

Um Conditional Random Field (CRF) [SUTTON; MCCALLUM, 2006] é uma distribuição condicional *P(Y|X)* com um modelo gráfico associado. A variável X é um vetor de variáveis aleatórias de entrada e Y é um vetor de variáveis aleatórias de saída. Então *P(Y|X)* é a probabilidade de dado como entrada o vetor X obter a saída Y. CRF tem sido aplicada em diversas áreas tais como, processamento de texto [TASKAR et al., 2002, PENG; MCCALLUM, 2004, SETTLES, 2005, SHA; PEREIRA, 2003], bioinformática [SATO; SAKAKIBARA, 2005, LIU et al., 2005] e visão computacional [He et al., 2004, KUMAR; HEBERT, 2003].

2.3.2.1 Definição

O problema de Reconhecimento de Entidades Nomeadas pode ser formulado como um problema de encontrar uma distribuição condicional. Seja X = [x_1, x_2, x_3 … x_{n-1}, x_n] uma sequência de palavras em um texto, onde x_j é uma palavra localizada na posição j do texto. Seja Y = [y_1, y_2, y_3 … y_{n-1}, y_n] uma sequência contendo um rótulo para cada palavra do vetor X. Então temos que *P(Y|X)* diz a probabilidade do texto representado por X ter a rotulação Y.

Por exemplo, seja o texto *"O Brasil investe R$ 2.000.000.000 no estado do Rio de Janeiro"*, então temos:

X = *{O, Brasil, investe, R$, 2.000.000.000, no, estado, do, Rio, de, Janeiro}*

$\mathbf{Y_1}$ = *{O, LOCAL, O, O, O, O, O, O, LOCAL, LOCAL, LOCAL}*

$\mathbf{Y_2}$ = *{O, ORGANIZACAO, O, O, O, O, O, O, LOCAL, LOCAL, LOCAL}*

$\mathbf{Y_3}$ = *{O, O, O, O, O, O, O, O, LOCAL, LOCAL, LOCAL}*

O vetor X é a representação do texto, os vetores Y_1, Y_2 e Y_3 são possíveis rotulações de X. No texto apresentado o termo Brasil é uma organização já que representa o governo brasileiro e o termo Rio de Janeiro é um local. Então $P(Y_2|X)$ tem que ser maior do que $P(Y_1|X)$ e $P(Y_3|X)$.

Dado X e uma distribuição condicional $P(Y|X)$ temos que a melhor rotulação de X é o vetor Y_k tal que $P(Y_k|X)$ é maior do que qualquer outro Y_j com $P(Y_j|X)$.

A dificuldade da modelagem dada acima é encontrar a distribuição $P(Y|X)$ e dada essa distribuição calcular a melhor rotulação Y dado o vetor de entrada X. Para isso, é utilizado um caso especial do Conditional Random Field que é o Linear-Chain Conditional Random Field. Abaixo segue a definição do Linear-Chain Conditional Random Field dada por Sutton e Mccallum [2006].

Definição: *Sejam X e Y vetores de variáveis aleatórias,* $\Lambda = \{\lambda_k\} \in R^K$ *um vetor de pesos,* $\{ f_k(y, y', x_t) \}_{k=1}^{K}$ *um conjunto de funções características e t uma posição do vetor X. Então o Linear-Chain Conditional Random Field é uma distribuição P(Y|X) que tem a seguinte forma:*

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\},$$

Fórmula 1: Distribuição condicional do Linear-Chain Conditional Random Field [SUTTON; MCCALLUM, 2006]

Onde Z(X) é uma função de normalização dada por:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

Fórmula 2: Função de Normalização [SUTTON; MCCALLUM, 2006]

A definição dada acima é derivada do Hidden Markov Models (HMM) [BIKEL et al, 1997], onde *P(Y|X)* é calculada a partir da distribuição *P(X, Y)* dada pelo HMM. Uma importante característica deste modelo, bem como todos os derivados do Conditional Random Field, é que não há necessidade de se conhecer a distribuição *P(X)*, que pode ser uma distribuição demasiadamente complexa.

Então, pela definição dada acima é preciso definir o vetor de funções características f = (f_1, f_2, f_3, f_4, . . ., f_k) e o vetor de pesos $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$. O vetor Λ aplica pesos ao vetor de funções de características, dessa maneira algumas funções tem maior impacto do que outras na contabilização final da probabilidade estimada.

As funções características têm o formato $f_k(y_t, y_{t-1}, x_t)$, onde y_t é a rotulação da palavra na posição t do vetor de entrada X, y_{t-1} é a rotulação da palavra na posição t-1 do vetor de entrada X e x_t é o vetor formado pelas palavras de X que são relevantes para o cálculo do rótulo y_t.

Segundo Mccallum e Li [2003] uma função característica, por exemplo, pode ter o valor 0 na maioria dos casos, e ter valor 1 se e somente se o rótulo y_{t-1} é do tipo OUTRO, o rótulo y_t é do tipo LOCAL e x_t contém uma palavra presente em uma lista de países. Naturalmente, o exemplo dado é simples, diversas funções que consideram a capitalização das palavras, os radicais, os sufixos, bigramas podem ser utilizados na elaboração das funções características.

2.3.2.2 Estimação dos pesos

Como visto na subseção acima a distribuição *P(Y|X)* depende dos valores dados ao vetor de pesos $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$. Este vetor tem o objetivo de ponderar o impacto de cada função característica. Maiores valores de λ_j implicam em maior importância dada à função característica f_j.

O cálculo do vetor $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$ é realizado por meio de um treino supervisionado. Seja $\theta = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$ e $D = \{x^{(i)}, y^{(i)}\}^N$ o conjunto de treinamento com N exemplos de rotulação, temos que a escolha ótima de Λ é a escolha de θ que maximize o somatório do logaritmos das probabilidades $P(y^{(i)}|x^{(i)})$ conforme explicitado abaixo:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$$

Fórmula 3: Função de treinamento do Linear-Chain Conditional Random Field [SUTTON; MCCALLUM, 2006]

Diversos métodos relacionados à maximização da função l(θ) têm sido publicados na literatura, algumas dessas publicações podem ser encontradas em Bertsekas [1999] e Byrd et al. [1994].

Uma vez que as funções características estejam definidas e que o vetor de pesos $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$ tenha sido calculado, teoricamente a distribuição $P(Y|X)$ pode ser calculada para todo Y e X. Suttom e McCallum [2006] demonstram como calcular eficientemente $Y^* = argmax_y(P(Y|X))$, onde Y^* é a rotulação mais provável de X.

2.4 HAREM – Avaliação de Reconhecimento de Entidades Mencionadas na Língua Portuguesa

HAREM [MOTA; SANTOS, 2008] é uma avaliação conjunta na área do reconhecimento de entidades mencionadas (REM, sigla em português), organizada pela Linguateca [SANTOS, 2011]. Uma avaliação conjunta, em poucas palavras, é uma tarefa que vários sistemas concordam em tentar executar, de forma a comparar o desempenho entre eles, com base em medidas consensuais e com recursos criados por uma comissão idônea.

O HAREM [HAREM, 2006] disponibiliza um conjunto de ferramentas destinadas a auxiliar na avaliação dos sistemas de reconhecimento de entidades nomeadas da língua portuguesa. Umas delas é a Coleção Dourada.

Os melhores resultados obtidos no HAREM foram obtidos pelo sistema Palavras-NER. Este sistema é baseado em Restrições de Gramática, tratando o Reconhecimento de Entidades Nomeadas como uma tarefa integrada da marcação gramatical [BICK, 2007]. O Palavras-NER é constituído por um gazeteer que contém aproximadamente 17.000 registros, um módulo morfológico e um módulo de inferência baseado em contexto.

2.4.1.1 Coleção Dourada

A Coleção Dourada [SANTOS; CARDOSO, 2006] é uma coleção de textos de diversas origens e gêneros, em que as entidades nomeadas foram manualmente identificadas, semanticamente classificadas e morfologicamente etiquetadas.

Esta coleção foi resultado de um esforço conjunto para obter textos de várias proveniências, tais como as oriundas de Portugal, Brasil, Angola, Moçambique, Macau, Índia, Timor Leste e Cabo Verde. As características também são diversas, desde textos retirados da web, textos jornalísticos, transcrições de entrevistas, textos técnicos retirados de relatórios extraídos da Web, textos políticos, entre outros.

O gráfico 1 apresenta as categorias utilizadas na anotação da Coleção Dourada e os melhores resultados de precisão e cobertura obtidos para cada uma das categorias.

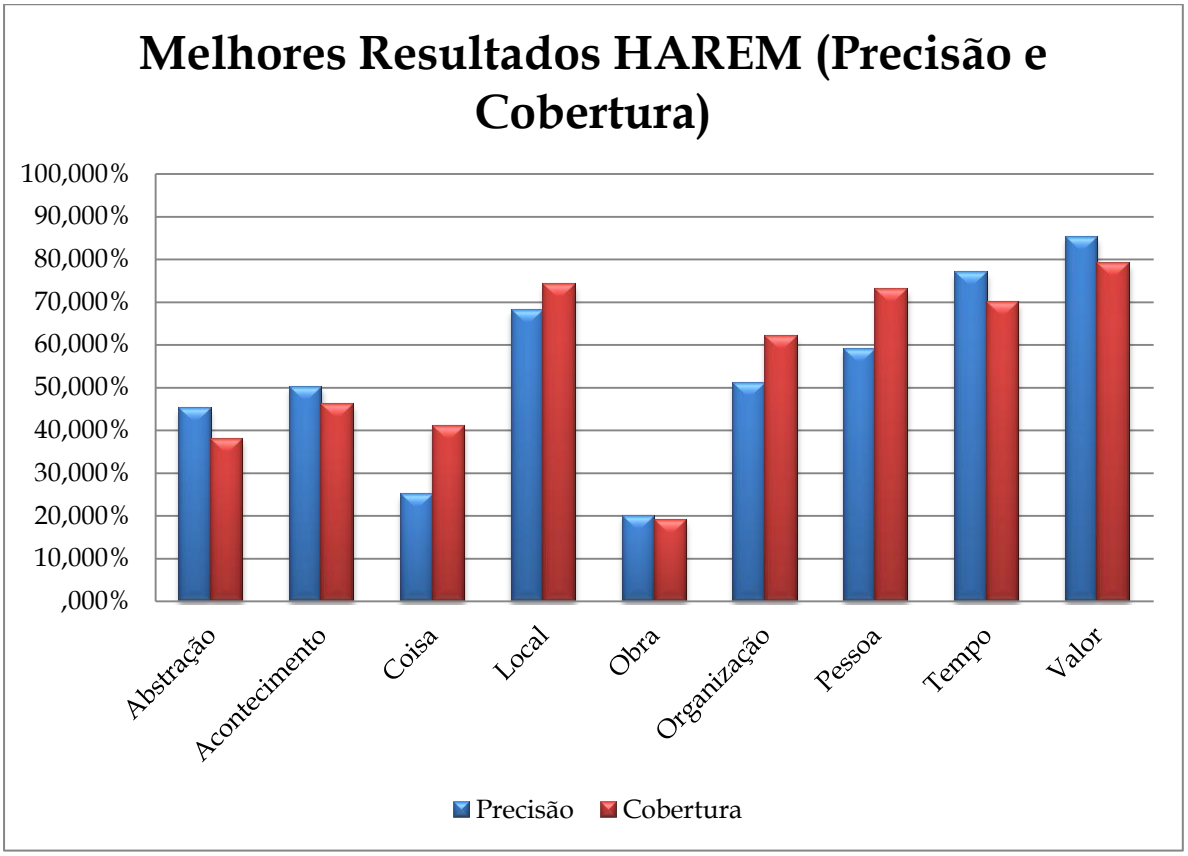

Gráfico 1: Melhores Resultados HAREM (Precisão e Cobertura). Adaptado de SANTOS e CARDOSO [2006]

Capítulo 3. Reconhecimento de Entidades Nomeadas

Este trabalho propõe uma representação semântica da notícia por meio da identificação de suas entidades nomeadas. Reconhecimento de Entidades Nomeadas (NER) [SUTTON; MCCALLUM, 2006] é o problema de identificar e classificar nomes próprios em textos, incluindo localizações, tais como Brasil e Rio de Janeiro; pessoas, tais como Dilma e Miriam; e organizações, tais como Ministério da Educação e Ministério da Cultura.

A ideia por trás desta proposta é que as entidades nomeadas, em geral, são elementos fundamentais para a compreensão de uma dada notícia. Esta proposta fornece subsídios para respostas automáticas do tipo: "*sobre quem a notícia se trata?",* "*sobre que lugar(es) a notícia se trata?",* etc.

Algumas iniciativas relacionadas ao Reconhecimento de Entidades Nomeadas em textos da língua portuguesa têm sido realizadas. A principal delas é o HAREM [MOTA; SANTOS, 2008] que é uma avaliação conjunta na área do reconhecimento de entidades mencionadas (REM, sigla em português), organizada pela Linguateca [SANTOS, 2011].

Entretanto, mesmo os melhores resultados obtidos no HAREM possuem taxas de erros significativas. Por exemplo, segundo dados retirados do artigo de Santos e Cardoso [2006] o melhor F-Score que engloba as categorias Pessoa, Local e Organização foi de 64,37%.

Então, esse trabalho propõe desenvolver um sistema de reconhecimento de entidades nomeadas destinado às notícias de governo escritas em português. A justificativa para esta proposta é que dada a restrição imposta aos textos a serem processados, há uma redução na complexidade das características envolvidas na classificação de uma entidade e, portanto, há a possibilidade de aumento de desempenho em relação aos sistemas desenvolvidos para serem genéricos.

Este trabalho propõe a aplicação do Linear-Chain Conditional Random Field na tarefa de reconhecimento de entidades nomeadas em notícias de governo. O Linear-Chain Conditional Random Field é um caso especial do Conditional Random Field (CRF). O CRF [SUTTON; MCCALLUM, 2006] é uma distribuição condicional $P(Y|X)$ com um modelo gráfico associado. A variável X é um vetor de variáveis aleatórias de entrada e Y é um vetor de

variáveis aleatórias de saída. Então *P(Y|X)* é a probabilidade de dado como entrada o vetor X obter a saída Y. Nesse caso o vetor X é a sequência de palavras em uma notícia e Y uma possível rotulação de entidades para X.

3.1 Reconhecimento de Entidades Nomeadas em Notícias de Governo

3.1.1 Visão Geral

O Reconhecimento de Entidades Nomeadas é uma área muito explorada, sobretudo na língua inglesa. Como pode ser visto em Sang e Meulder [2003], os desempenhos obtidos pelos sistemas de reconhecimento de entidades para a língua inglesa apresentam bons resultados, chegando a alcançar um f-score de até 88,76%, considerando a classificação de quatro categorias: Pessoa, Local, Organização e Diversos.

Este trabalho tem o objetivo de propor a representação semântica de uma notícia por meio de suas entidades. Então, é necessário ter um módulo de reconhecimento de entidades. Uma possível opção para este trabalho seria a de utilizar um sistema de reconhecimento de entidades já existente para a língua portuguesa. No entanto, há duas razões para que essa opção não tenha sido adotada.

Primeiro, porque os sistemas de reconhecimento de entidades nomeadas para a língua portuguesa, estudados neste trabalho, apresentam resultados significativamente inferiores aos obtidos em sistemas desenvolvidos para a língua inglesa.

Por exemplo, os melhores resultados alcançados pelos sistemas de reconhecimento de entidades nomeadas que participaram do HAREM, segundo Santos e Cardoso [2006], foram os F-Scores 65,99% para a categoria Pessoa, 56,26% para a categoria Organização e 70,85% para a categoria Local.

Além disso, tais sistemas foram concebidos para reconhecer as entidades nomeadas presentes nos mais variados tipos de textos da língua portuguesa. Portanto, não se beneficiam com a limitação de ter que classificar apenas notícias relacionadas ao governo.

Segundo, porque dado o contexto de notícias de governo este trabalho propõe a inserção de novos grupos de entidades. Tais grupos não fazem parte dos sistemas estudados neste trabalho, e por isso precisam ser explorados.

3.1.2 Caracterização do Problema

O problema que motiva os estudos dessa seção consiste em investigar a elaboração de um sistema que seja dotado da capacidade de reconhecer entidades nomeadas mencionadas em notícias relacionadas ao governo brasileiro.

Nesse trabalho o problema é dividido em dois cenários diferentes. Em ambos os cenários o universo de textos a serem considerados são as notícias relacionadas ao governo. A diferença entre os cenários é o conjunto de categorias a ser utilizada para a classificação das entidades nomeadas.

No primeiro cenário as categorias consideradas são Pessoa, Local e Organização. Estas categorias são comumente utilizadas na literatura. Logo, este cenário proposto é um subconjunto do problema geral de Reconhecimento de Entidades Nomeadas, já que a única diferença é o universo de textos da língua portuguesa que é reduzido às notícias relacionadas ao governo brasileiro.

No segundo cenário há um incremento no conjunto de categorias a ser considerada. Este trabalho propõe as categorias e subcategorias Programa, Evento, Área Administrativa Cargo e Sigla, como categorias relevantes para o contexto adotado por este trabalho. Esse cenário, diferentemente do primeiro, não é um subconjunto dos demais problemas que foram descritos na seção de trabalhos relacionados, já que essas novas categorias estão sendo estudadas.

3.1.3 Formulação das Hipóteses

Após a caracterização do problema algumas hipóteses foram levantadas acerca das características essenciais para o reconhecimento das entidades nomeadas dado o contexto deste trabalho. Dentre as definições de hipótese encontradas na literatura pode-se destacar a definição abaixo:

> Para RUDIO (1978), hipótese é uma suposição que se faz na tentativa de explicar o que se desconhece. Esta suposição tem por característica o fato de ser provisória, devendo, portanto, ser testada para a verificação de sua validade. Trata-se de antecipar um conhecimento na expectativa de que possa ser comprovado. [AGNER, 2002]

As hipóteses levantadas nesse trabalho podem ser divididas em hipóteses relativas à natureza do problema e em hipóteses relativas aos métodos empregados na solução do problema. A hipótese relativa à natureza do problema é a hipótese principal deste trabalho. As demais hipóteses visam supor quais são as características que devem ser consideradas como relevantes no processo de reconhecimento de entidades. As subseções abaixo descrevem cada uma das hipóteses levantadas neste trabalho.

3.1.3.1 Hipótese 1

A primeira hipótese é relativa à natureza do problema caracterizado. Ela é descrita como:

HIPÓTESE 1: A tarefa de reconhecimento de entidades nomeadas aplicada ao contexto de notícias de governo é uma tarefa mais fácil do que a tarefa de reconhecimento de entidades nomeadas em um contexto geral.

Uma das justificativas de não utilizar um sistema de reconhecimento de entidades nomeadas já existente para a língua portuguesa é baseada nesta hipótese. Naturalmente, a tarefa de reconhecimento das entidades em um contexto limitado não é mais difícil do que a tarefa de reconhecimento das entidades em um contexto geral, já que a solução da segunda pode ser utilizada como solução da primeira tarefa.

Entretanto, comprovar a hipótese 1 não é trivial, embora aparentemente seja intuitiva. No caso da hipótese 1 ser refutada, significa que a tarefa de reconhecimento de entidades no contexto de notícias de governo é tão difícil quanto adotado o contexto geral. Logo, os sistemas de reconhecimento existentes poderiam ser utilizados para solucionar o problema proposto sem perda de desempenho.

A hipótese 1 é sustentada com as premissas de que no contexto de notícias de governo a variação de contextos é menor do que em um contexto geral. O texto é escrito seguindo a linguagem formal e a estruturação do texto segue os modelos de notícia comumente usados. Portanto, a complexidade na extração de padrões é reduzida.

3.1.3.2 Hipótese 2

A segunda hipótese é relativa aos métodos utilizados na solução do problema caracterizado. Ela é descrita como:

HIPÓTESE 2: O reconhecimento de uma entidade X pode ser realizado mediante ao conhecimento da frase em que esta é mencionada, sem necessidade do conhecimento da notícia por completo.

Essa hipótese diz que para o reconhecimento das entidades em uma notícia de governo é necessário apenas obter informações sobre a frase em que ela se encontra. Por exemplo, na frase abaixo é possível identificar que Paulo Mac Donald Ghisi é uma pessoa sem ter o conhecimento do restante da notícia.

> Na oportunidade, o prefeito Paulo Mac Donald Ghisi destacou "a importância do hospital para a população de Foz, para os municípios próximos e, inclusive, para os pacientes de países limítrofes, que vão à cidade para receberem tratamento médico". [MINISTÉRIO, 2011a]

Essa hipótese motiva duas linhas de experimentos. A primeira adota a abordagem de classificar as entidades nomeadas analisando uma frase como um documento separado, ou seja, apenas as características da frase corrente são utilizadas para a classificação de suas entidades nomeadas. A segunda adota a abordagem de classificar as entidades nomeadas analisando uma notícia como um documento separado, ou seja, toda a notícia é utilizada como parâmetro para classificar as entidades nomeadas mencionadas.

A segunda abordagem tem a vantagem de disponibilizar mais informação para o sistema, entretanto, a quantidade demasiada de informação pode tornar o problema de classificação complexo, comprometendo o desempenho do sistema. Tendo em vista essa consideração, é importante formular os experimentos de forma a validar ou refutar esta hipótese.

3.1.3.3 Hipótese 3

A terceira hipótese é relativa aos métodos utilizados na solução do problema caracterizado. Ela é descrita como:

HIPÓTESE 3: Considere o Linear-Chain Conditional Random Fields como método utilizado para o treinamento de entidades nomeadas. Seja P um sistema de reconhecimento de entidades nomeadas para um conjunto de categorias A_1, A_2, A_3 ... A_{n-1}, A_n. Seja S_i um sistema de reconhecimento de entidades nomeadas para uma categoria A_i e S um sistema de reconhecimento de entidades que agrupa os S_1, S_2, S_3 ... S_{n-1}, S_n. O desempenho do Sistema P na classificação de entidades nomeadas é superior ou igual ao desempenho do Sistema S.

A hipótese 3 supõe que um sistema treinado para o reconhecimento de entidades pertencente a um conjunto de categorias tem um desempenho maior do que um conjunto de sistemas treinados cada um para uma única categoria.

Por exemplo, seja S um sistema de reconhecimento de entidades nomeadas para as categorias PESSOA, LOCAL e ORGANIZAÇÃO. Seja P1 um sistema para a categoria PESSOA, P2 um sistema para a categoria LOCAL e P3 um sistema para a categoria ORGANIZAÇÂO. E seja P um sistema que utiliza os resultados de P1, P2 e P3 para realizar a tarefa de reconhecimento de entidades. Segundo a hipótese 3 o sistema S tem um desempenho superior ou igual ao do sistema P.

Essa hipótese tem como justificativa que o sistema P tem mais informações disponíveis do que o sistema S. Pois o sistema P pode utilizar o conhecimento de outras categorias presentes no texto para classificar uma entidade X. O texto abaixo exemplifica como conhecimento de uma entidade de categoria Y pode auxiliar no reconhecimento de uma categoria X.

> O Ministério do Planejamento, por meio do superintendente do **Patrimônio da União no Paraná**, **Dinarte Antonio Vaz** e da gerente-executiva do Instituto Nacional do Seguro Social (INSS), Cleonice Dariva, assinaram, em 13 de maio, em Foz do Iguaçu, a Escritura de Transferência de um terreno com área de 16.300m², do INSS para a União. [MINISTÉRIO, 2011a]

A entidade Dinarte Antonio Vaz pode ter a probabilidade de ser classificada como pessoa se o sistema identificar que Patrimônio da União no Paraná é uma organização e que o padrão Organização + vírgula é predecessor de uma entidade do tipo Pessoa.

Entretanto, o sistema P apresenta um problema que pode comprometer o ganho de informação adicional. Este problema é o de classificar incorretamente uma dada entidade e este se propagar para as outras entidades. Por exemplo, se por algum motivo o sistema classificar Patrimônio da União no Paraná como Lugar é possível que o sistema considere a vírgula como uma lista de lugares. E então classificaria a entidade Dinarte Antonio Vaz como um Lugar.

Tendo em vista esses casos, a hipótese 3 motiva outras duas linhas de experimento. A primeira linha consiste em realizar os experimentos de forma a obter os resultados de um sistema que segue os moldes do sistema P. A segunda linha consiste em realizar os experimentos de forma a obter os resultados de um sistema que segue os moldes do sistema

S. Após os experimentos ambos os resultados serão comparados e analisados a fim de corroborar ou refutar a hipótese 3.

3.2 Coleção de Notícias de Governo – UFRJ

Neste trabalho o método utilizado para o Reconhecimento de Entidades Nomeadas emprega o aprendizado supervisionado. Portanto é preciso ter um conjunto de exemplos para que o sistema possa ser treinado.

Anteriormente foi apresentado a Coleção Dourada, que é uma coleção disponibilizada pelo HAREM contendo uma variedade de textos anotados manualmente. Entretanto, o escopo deste trabalho é a classificação de entidades em notícias de governo, então alguns dos textos da Coleção Dourada são desnecessários para o aprendizado, inclusive podendo dificultar o treinamento.

A Coleção de Notícias de Governo – UFRJ é uma coleção de notícias retiradas de sites do governo brasileiro, criada exclusivamente para este trabalho. As notícias foram anotadas manualmente com o auxílio de uma mestranda em Ciência da Informação da Universidade Federal Fluminense (UFF). Algumas anotações duvidosas foram discutidas e estudadas até haver um consenso sobre a classificação da entidade.

Esta coleção possui 240 notícias de 26 sites pertencentes ao domínio *.gov.br*. A tabela 6 apresenta a distribuição de notícias por site.

A Coleção de Notícias de Governo – UFRJ foi anotada manualmente de duas formas distintas originando duas bases anotadas. A primeira base consiste na anotação das entidades Pessoa, Local, Organização, Programa e Evento; a segunda base consiste na anotação das entidades Pessoa, Local, Organização, Programa, Evento, Cargo, Sigla e Área Administrativa.

Fonte	*Nº de notícias*
Casa Civil	9
Ministério da Agricultura, Pecuária e Abastecimento	10
Ministério da Ciência, Tecnologia e Inovação	26
Ministério da Cultura	16
Ministério da Educação	5
Ministério da Fazenda	1
Ministério da Integração Nacional	1
Ministério da Justiça	9
Ministério da Pesca e Aquicultura	8
Ministério da Previdência Social	6
Ministério da Saúde	2
Ministério das Cidades	9
Ministério das Comunicações	14
Ministério das Relações Exteriores	1
Ministério de Minas e Energia	2
Ministério do Desenvolvimento Agrário	7
Ministério do Desenvolvimento, Indústria e Comércio Exterior	10
Ministério do Desenvolvimento Social e Combate à Fome	10
Ministério do Meio Ambiente	14
Ministério do Planejamento	20
Ministério do Trabalho e Emprego	13
Ministério do Turismo	10
Ministério dos Esportes	9
Ministério dos Transportes	10
Secretaria dos Portos	8
Secretaria Geral da Presidência	10
TOTAL:	*240*

Tabela 1: Distribuição de notícias por fontes

3.2.1 Primeira Base Anotada (5 Categorias)

A primeira base anotada da Coleção de Notícias de Governo – UFRJ consiste na anotação das 240 notícias coletadas. As notícias foram anotadas considerando as categorias

(figura 4) Pessoa, Local, Organização, Evento e Programa obedecendo às seguintes definições:

PESSOA: São considerados como pessoas, todas as menções de nomes próprios e apelidos que correspondam a um ser humano. As menções realizadas por meios de cargos não são consideradas como uma entidade do tipo Pessoa. Por exemplo, no trecho "o ministro disse" a entidade ministro não é contabilizada como uma entidade do tipo Pessoa.

LOCAL: São considerados como local, todas as menções capitalizadas que podem ser traduzidas como um local geográfico. Por exemplo, dependendo do contexto a entidade "Palácio do Planalto" pode ser considerada como local, já que esta possui uma localização geográfica que inclusive pode ser geo-codificada. Entretanto, o nome de uma localidade quando não usada nesse propósito não é classificada como Local. Por exemplo, na frase "O Brasil investe no Rio de janeiro" a entidade Brasil não se refere a um local e sim a uma organização, que no caso é o Governo Brasileiro.

ORGANIZAÇÃO: São consideradas como organizações todas as menções capitalizadas relacionadas a uma entidade que possui vida própria, tendo uma administração própria e que não são caracterizadas como Pessoas. Por exemplo, ministérios, bancos, hospitais, fundações, etc.

EVENTO: São considerados como eventos todos os acontecimento registrados nas notícias. Todas as menções que podem ser identificadas como um congresso, seminário, conferência, olimpíadas, etc.

PROGRAMA: São considerados como programas, todas as menções capitalizadas de um programa ou plano realizado pelo governo. Por exemplo, PAC, Minha Casa Minha Vida, etc.

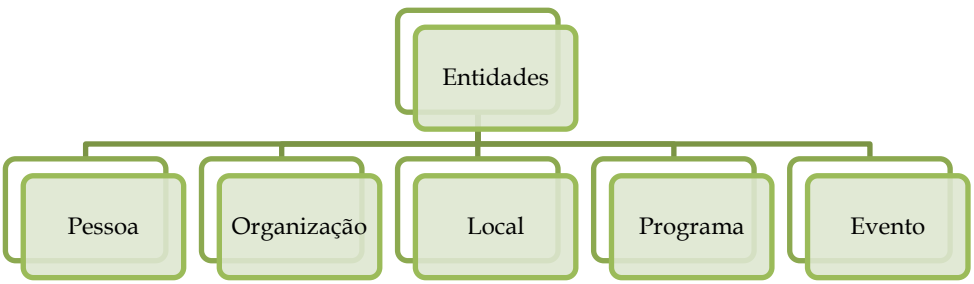

Figura 4: Categorias da Primeira Base Anotada

A primeira base anotada é constituída por 7598 entidades nomeadas, sendo 389 eventos, 1309 locais, 3927 organizações, 1353 pessoas e 620 programas. A tabela abaixo apresenta a distribuição de entidades nomeadas dentro da Coleção de Notícias de Governo – UFRJ.

Nº de Entidades Coleção Notícias de Governo – UFRJ Primeira Base de Notícias Anotadas	
Tipo de Entidade	*Nº de Entidades*
Evento	389
Local	1309
Organização	3927
Pessoa	1353
Programa	620
TOTAL	*7598*

Tabela 2: Nº de entidades da Primeira Base Anotada

A tabela abaixo apresenta uma visão geral do tamanho da Primeira Base Anotada criada a partir da Coleção de Notícias de Governo – UFRJ.

Primeira Base Anotada da Coleção de Notícias de Governo - UFRJ	
Número de notícias	240
Número de sentenças	4110
Número de entidades	7598

Tabela 3: Visão geral do tamanho da Primeira Base Anotada

A sintaxe da anotação realizada na Primeira Base Anotada é semelhante à utilizada na Coleção Dourada. Cada entidade é etiquetada com a tag <EM> ... </EM>, e dentro de cada tag o atributo CATEG pode receber os seguintes valores (em caixa alta): PESSOA, LOCAL, ORGANIZACAO, PROGRAMA e EVENTO. A figura abaixo mostra um exemplo de documento da coleção anotada.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<DOC>
     19 de dezembro de 2011
    <EM CATEG="ORGANIZACAO">Instituto Bacarelli</EM>
    Ministra <EM CATEG="PESSOA">Ana</EM> visita importante projeto cultural apoiado pelo <EM
    CATEG="ORGANIZACAO">MinC</EM> na capital paulista
    Nesta sexta-feira, 16/12, a ministra <EM CATEG="PESSOA">Ana de Hollanda</EM> visitou o <EM
    CATEG="ORGANIZACAO">Instituto Bacarelli</EM>, centro de formação musical e artística localizado em
    <EM CATEG="LOCAL">Heliópolis</EM>, zona sul da cidade de <EM CATEG="LOCAL">São Paulo</EM>.
    Acompanhada pelo Secretário de Políticas Culturais do <EM CATEG="ORGANIZACAO">MinC</EM>, <EM
    CATEG="PESSOA">Sergio Mamberti</EM>, e pelo senador <EM CATEG="PESSOA">Eduardo Suplicy</EM>, a
    ministra conheceu o novo prédio da instituição, construído com recursos da Lei Rouanet, e assistiu a
    apresentações de alunos do instituto, entre eles os integrantes da Sinfônica Heliópolis.

    Talvez o mais conhecido dos projetos do <EM CATEG="ORGANIZACAO">Instituto</EM>, a Sinfônica de
    Heliópolis é formada principalmente por jovens de baixa renda do bairro e é reconhecida
    internacionalmente por sua excelente qualidade musical. Segundo a ministra, "o ensino de música
    oferecido pelo <EM CATEG="ORGANIZACAO">Instituto</EM> é importante para a formação do indivíduos que
    ali estão, mas também é fundamental para a comunidade ". A ministra <EM CATEG="PESSOA">Ana</EM>
    ainda parabenizou os integrantes pelo desempenho do grupo, afirmando que eles são motivo de orgulho
    não apenas para seus familiares, mas para todo o país.

    Em conversa com os diretores do <EM CATEG="ORGANIZACAO">Instituto</EM>, a ministra <EM
    CATEG="PESSOA">Ana</EM> e o secretário <EM CATEG="PESSOA">Mamberti</EM> reforçaram a importância do
    <EM CATEG="PROGRAMA">Plano Nacional de Cultura</EM> como ação estruturante da política cultural do
    país, transformando-a em questão de Estado com metas de longo prazo (leia matéria sobre a assinatura
    da portaria que oficializa as 53 metas do <EM CATEG="PROGRAMA">Plano Nacional de Cultura</EM>), e da
    ratificação do acordo <EM CATEG="ORGANIZACAO">MEC</EM>-<EM CATEG="ORGANIZACAO">MinC</EM>, cujo
    objetivo é desenvolver ações culturais nas escolas da rede pública, "favorecendo a presença de todas
    as áreas culturais nas escolas e possibilitando formação e fruição culturais mais difundidas",
    segundo <EM CATEG="PESSOA">Ana de Hollanda</EM>.

    Para <EM CATEG="PESSOA">Edmilson Venturelli</EM>, diretor de relações institucionais do <EM
    CATEG="ORGANIZACAO">Instituto</EM>, a visita da Ministra da <EM CATEG="ORGANIZACAO">Cultura</EM> é
    uma oportunidade para que o instituto preste contas dos recursos públicos destinados aos seus
    projetos de forma completa, demonstrando ao vivo os resultados de alto nível atingidos pelos alunos
    e a riqueza do trabalho desenvolvido.

</DOC>
```

Figura 5: Exemplo de documento anotado da Primeira Base

3.2.2 Segunda Base Anotada

A Segunda Base Anotada, diferentemente da primeira, permite a anotação de relações entre entidades e possui um número maior de categorias. O objetivo da Segunda Base Anotada é fornecer exemplos de modo a possibilitar a construção de um sistema que consiga extrair informações tais como: Pessoa X tem Cargo Y, Evento X acontece no local Y, etc.

Nessa base 120 notícias das 240 notícias coletadas foram anotadas manualmente, tendo suas entidades e relações identificadas e classificadas. As notícias foram anotadas considerando as categorias Pessoa, Local, Organização, Evento, Programa e Cargo; e as subcategorias Sigla e Área Administrativa. A figura 6 fornece uma visão geral da classificação adotada.

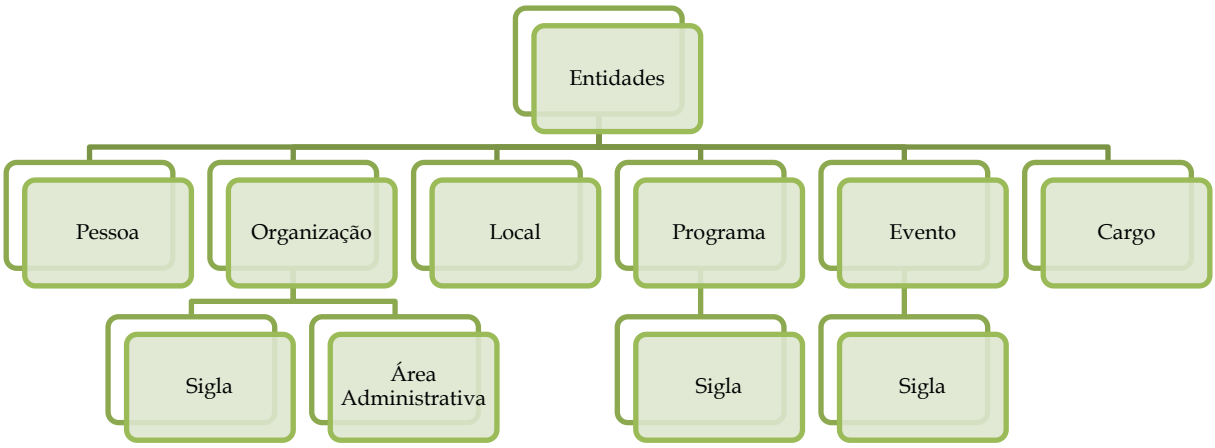

Figura 6: Categorias e Subcategorias da Segunda Base Anotada

A anotação manual das categorias e subcategorias foi realizada segundo os critérios descritos abaixo:

PESSOA: Idem aos critérios utilizados na Primeira Base Anotada.

LOCAL: Idem aos critérios utilizados na Primeira Base Anotada.

ORGANIZAÇÃO: O critério utilizado na classificação desta categoria diverge um pouco do critério adotado na anotação da Primeira Base. Uma entidade é classificada como Organização se ela representa um organismo vivo possuindo uma administração e se o enfoque do texto está na organização e não como complemento do cargo de uma pessoa. Por exemplo, no parágrafo abaixo, o Ministério da Ciência, Tecnologia e Inovação não é complemento, ele é a entidade que realiza uma determinada ação.

> O **Ministério da Ciência, Tecnologia e Inovação** (MCTI) elaborou a publicação Estratégia Nacional de Ciência,Tecnologia e Inovação 2012-2015 e Balanço das Atividades Estruturantes 2011. [MINISTÉRIO, 2012a]

Já no parágrafo abaixo a entidade Ciência, Tecnologia e Inovação é complemento do cargo de ministro, indicando a que organização o cargo se refere. Nesse casso a entidade não é classificada como organização, já que esta passa a ser incorporada à entidade cargo.

> O ministro da **Ciência, Tecnologia e Inovação**, Marco Antonio Raupp, tornou-se membro do Conselho Curador da Empresa Brasil de Comunicação (EBC). Raupp assumiu vaga de conselheiro no lugar de seu antecessor na pasta, o atual ministro da Educação, Aloizio Mercadante, nesta quarta-feira. [MINISTÉRIO, 2012b]

EVENTO: Idem aos critérios utilizados na Primeira Base Anotada.

PROGRAMA: Idem aos critérios utilizados na Primeira Base Anotada.

CARGO: São consideradas como cargos, todas as menções a cargos que pertençam a uma relação Pessoa => Cargo ou Cargo => Organização, ou seja, um cargo só é classificado como tal, se este se encontra próximo a Pessoa que ocupa o cargo ou a Organização a que o cargo se refere. Além disso, se um cargo está textualmente diretamente ligado a uma organização ambos são considerados como um único cargo. Palavras que indicam cargos sem a ocorrência da Pessoa que o ocupa ou a organização a que este pertence são desconsiderados.

Por exemplo, abaixo segue duas citações que exemplificam o critério adotado para a classificação. Na primeira citação o cargo diretor-presidente está próximo à pessoa que o ocupa, portanto diretor-presidente é classificado como cargo e a relação pessoa => cargo também é anotada.

> No encontro, o <*CARGO*>*diretor-presidente*</*CARGO*> da empresa, <*PESSOA*>*Nelson Breve*</*PESSOA*>, convidou o conselho a participar do planejamento estratégico da instituição, em fase inicial. [MINISTÉRIO, 2012b]

Na segunda citação o cargo de ministro é seguido da organização a que este pertence, então ambas as entidades são rotuladas como uma única entidade do tipo Cargo. O motivo para adotar essa abordagem é a de possibilitar a distinção entre cargos do tipo ministro. Nesse caso, após anotar a relação, fica explícito que Marco Antonio Raupp não é apenas ministro, mas também ministro da Ciência e Tecnologia.

> O <*CARGO*>*ministro da Ciência, Tecnologia e Inovação*</*CARGO*>, <*PESSOA*>*Marco Antonio Raupp*</*PESSOA*>, tornou-se membro do Conselho Curador da Empresa Brasil de Comunicação (EBC). Raupp assumiu vaga de conselheiro no lugar de seu antecessor na pasta, o atual ministro da Educação, Aloizio Mercadante, nesta quarta-feira. [MINISTÉRIO, 2012b]

SIGLA: A subcategoria Sigla é uma subdivisão das categorias: Organização, Programa e Evento. Nesse trabalho são considerados como Siglas apenas as menções explícitas, que seguem a ordem Organização => Sigla, Evento => Sigla, Programa => Sigla. Siglas desacompanhadas de seus nomes por extenso não são consideradas como Sigla, tendo apenas a sua categoria anotada. O objetivo deste tipo de anotação é conseguir fornecer exemplos para elaborar um sistema que possa associar uma sigla a um nome por extenso.

ÁREA ADMINISTRATIVA: A subcategoria Área Administrativa refere-se a uma organização que represente um país, estado, município, etc. Em geral, essa subcategoria é um nome de local que é utilizado como uma Organização. Por exemplo, na frase "O Brasil

investe no Rio de Janeiro" a entidade Brasil é classificada como uma organização que tem subcategoria Área Administrativa.

A Segunda Base Anotada é constituída por 3550 entidades nomeadas, sendo 177 eventos, 593 locais, 1487 organizações, 607 pessoas e 309 programas. As tabelas 4 e 5 apresentam a distribuição de entidades nomeadas dentro da Segunda Base Anotada.

Nº de Entidades Coleção Notícias de Governo – UFRJ Segunda Base de Notícias Anotadas (Categorias)	
Tipo de Entidade	*Nº de Entidades*
Evento	177
Local	593
Organização	1487
Pessoa	607
Programa	309
Cargo	377
TOTAL	*3550*

Tabela 4: Nº de entidades da Segunda Base Anotada (Categorias)

Nº de Entidades Coleção Notícias de Governo – UFRJ Segunda Base de Notícias Anotadas (Sub-Categorias)	
Tipo de Entidade	*Nº de Entidades*
Sigla	249
Área Administrativa	198
TOTAL	*447*

Tabela 5: Nº de entidades da Segunda Base Anotada (Sub-Categorias)

A tabela abaixo apresenta uma visão geral do tamanho da Segunda Base Anotada criada a partir da Coleção de Notícias de Governo – UFRJ. As 120 notícias foram escolhidas aleatoriamente.

Segunda Base Anotada da Coleção de Notícias de Governo - UFRJ	
Número de notícias	120
Número de sentenças	-
Número de entidades	3550

Tabela 6: Visão geral do tamanho da Primeira Base Anotada

A sintaxe da anotação realizada na Segunda Base Anotada é semelhante à utilizada na Coleção Dourada. Cada entidade é etiquetada com a tag <EM> ... </EM>, e dentro de cada tag o atributo CATEG pode receber os seguintes valores (em caixa alta): PESSOA, LOCAL, ORGANIZACAO, PROGRAMA, EVENTO e CARGO; algumas entidades também podem receber o atributo SUBCATEG que pode assumir os valores SIGLA e ÁREA ADMINISTRATIVA. A figura abaixo mostra um exemplo de documento da coleção anotada.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<DOC>
16/01/2012 às 17h45 - <EM CATEG="ORGANIZACAO" ID="1">Secretaria Executiva</EM> assume
coordenação das ações da <EM CATEG="EVENTO" ID="2">Copa 2014</EM> no <EM
CATEG="ORGANIZACAO" ID="3">Ministério do Esporte</EM>

O <EM CATEG="CARGO" ID="4">ministro do Esporte</EM>, <EM CATEG="PESSOA" ID="5"
REL="4">Aldo Rebelo</EM>, anunciou nesta segunda-feira (16.01), durante coletiva de
imprensa em <EM CATEG="LOCAL" ID="6">Brasília</EM>, a transferência da coordenação
dos assuntos relacionados à preparação do país para a <EM CATEG="EVENTO" ID="7">Copa
do Mundo FIFA 2014</EM>. Os trabalhos passam a ser coordenados pela <EM
CATEG="ORGANIZACAO" ID="8">Secretaria Executiva</EM>, no lugar da <EM
CATEG="ORGANIZACAO" ID="9">Secretaria Nacional de Futebol e Defesa dos Direitos do
Torcedor</EM>. Tomei a atitude de trazer para a <EM CATEG="ORGANIZACAO"
ID="10">Secretaria Executiva</EM> a condução direta da questão da <EM CATEG="EVENTO"
ID="11">Copa do Mundo</EM>, que ficará a cargo do <EM CATEG="CARGO"
ID="12">secretário</EM> <EM CATEG="PESSOA" ID="13" REL="12">Luis Fernandes</EM>, um
homem experiente e também apaixonado por futebol, disse o ministro.

<EM CATEG="PESSOA" ID="14">Aldo Rebelo</EM> explicou que a medida foi tomada depois
de <EM CATEG="PESSOA" ID="15" REL="16">Alcino Reis</EM>, que era o <EM CATEG="CARGO"
ID="16">secretário de Futebol</EM>, pedir afastamento do <EM CATEG="ORGANIZACAO"
ID="17">Ministério do Esporte</EM>. A <EM CATEG="ORGANIZACAO" ID="18">Secretaria
Executiva</EM> auxilia o ministro na supervisão e coordenação das atividades das
secretarias nacionais, integradas à estrutura do ministério, e na definição das
diretrizes e políticas no âmbito da Política Nacional do Esporte.

Confira a reportagem em áudio

</DOC>
```

Figura 7: Exemplo de documento anotado da Primeira Base Anotada

3.3 Biblioteca de Anotação Semântica de Notícias de Governo

Neste trabalho foi concebida uma biblioteca de anotação semântica de notícias que, dada uma notícia, anota suas entidades nomeadas em um formato XML. Esta biblioteca utiliza a implementação do Linear-Chain Conditional Random Field disponibilizada pela universidade de Stanford.

O primeiro passo no desenvolvimento desta biblioteca foi implementar um módulo de tratamento das notícias. O objetivo deste módulo é criar uma correspondência entre o formato em que as notícias são apresentadas no mundo real e o formato utilizado pelo Linear-Chain Conditional Random Field.

Nesse sentido as notícias são transformadas em dois vetores de variáveis aleatórias, chamados neste trabalho de X e Y, onde o vetor X é a segmentação das notícias em tokens e o vetor Y é a entidade de cada um dos token.

O próximo passo deste trabalho foi descobrir um modelo de distribuição condicional *P(Y|X)* que se aplique às notícias de governo na língua portuguesa. Em outras palavras, foi descobrir um modelo que defina qual a probabilidade de uma rotulação de entidades Y dada uma notícia representada por um vetor X.

Como visto na seção 2.3.2, o Linear-Chain Conditional Random Field oferece um modelo para essa distribuição condicional. Entretanto este modelo depende de um vetor de pesos $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_k)$, onde cada λ_i é um parâmetro que precisa ser calculado. Para isso é utilizada a implementação disponibilizada pela universidade de Stanford. O software disponibilizado gera o modelo a partir de um conjunto de exemplos $\{X^{(i)}, Y^{(i)}\}^N$.

As figuras 8 e 9 apresentam a arquitetura utilizada na concepção dos modelos de reconhecimento de entidades nomeadas para a língua portuguesa no contexto de notícias de governo. Estes modelos representam a distribuição condicional obtida após os procedimentos de treino realizados.

Os exemplos utilizados neste trabalho foram retirados da Primeira Base Anotada e Segunda Base Anotada da Coleção de Notícias de Governo – UFRJ. Ambas originaram dois modelos distintos, que podem ser utilizados a critério do usuário da biblioteca de Anotação Semântica de Notícias.

O modelo 1 originado do treinamento realizado sobre a Primeira Base Anotada, contempla as categorias Pessoa, Local, Organização, Programa e Evento. O melhor modelo desenvolvido nesse trabalho considerando essa base obteve os F-Scores de 55% para a categoria Evento, 74% para a categoria Local, 79% para a categoria Organização, 92% para a categoria Pessoa e 72% para a categoria Programa.

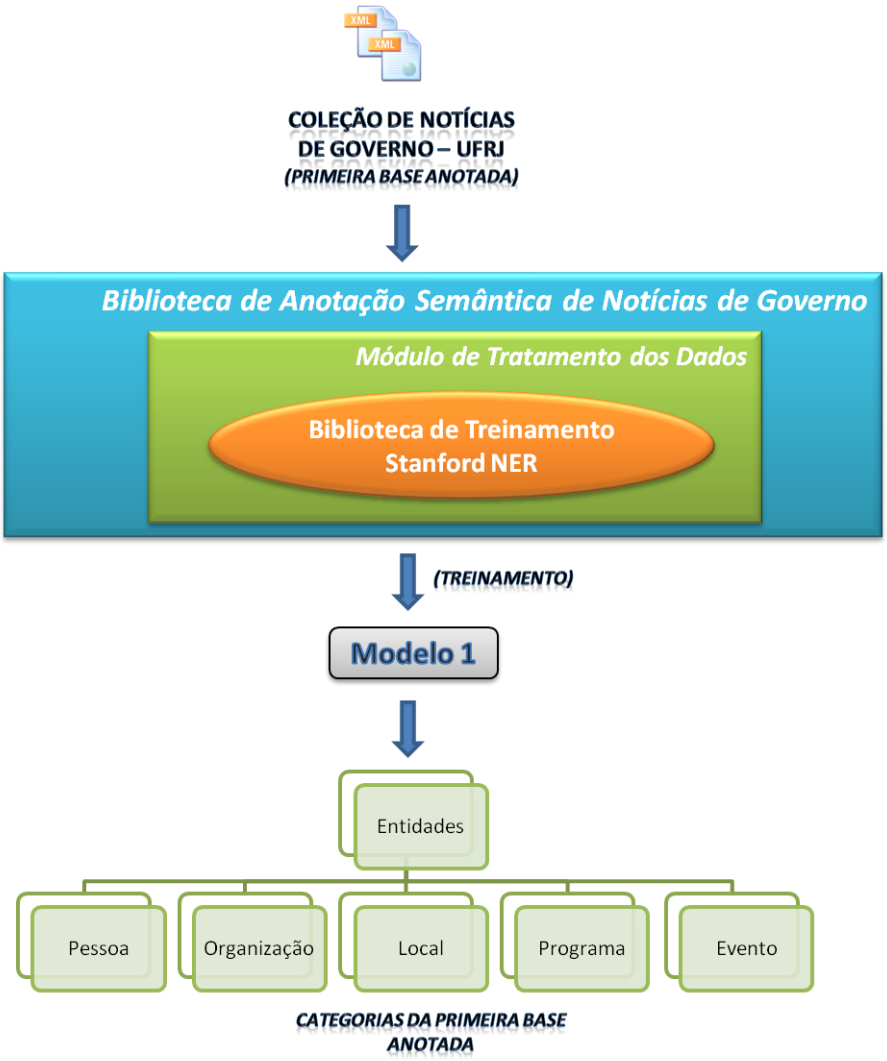

Figura 8: Arquitetura utilizada na concepção do modelo 1 utilizados na biblioteca de Anotação Semântica

O modelo desenvolvido apresentou destaque na classificação da categoria Pessoa, apresentando um F-Score muito próximo a resultados obtidos por humanos. A categoria Organização tinha potencial para apresentar um F-Score mais alto, entretanto, a presença de vírgulas em nomes de Organizações dificultou a tarefa de reconhecimento, já que este modelo compreendeu a vírgula como um separador de entidades, o que é verdade na maioria dos casos.

A tabela 7 apresenta os resultados obtidos pelo melhor modelo criado neste trabalho a partir dos treinos realizados com a Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ. Como resultados gerais foram calculados os F-Score geral e o F-Score contabilizando apenas as categorias Pessoa, Local e Organização. O primeiro apresentou um F-Score de 79% enquanto o segundo apresentou um F-Score de 80%. A variação entre os

dois F-Scores não foi grande devido ao número reduzido de entidades de eventos e programas em comparação com as outras categorias.

Sistema de Reconhecimento de Entidades – Primeira Base Anotada			
Tipo de Entidade	*Precisão*	*Cobertura*	*FScore*
Evento	0.66	0.47	0.55
Local	0.80	0,68	**0.74**
Organização	0.76	0.82	**0.79**
Pessoa	0.89	0.94	**0.92**
Programa	0.83	0.64	0.72
TOTAL (P, L, O)	*0,79*	*0,82*	***0,80***
TOTAL (GERAL)	*0,79*	*0,79*	*0,79*

Tabela 7: Resultados do sistema de reconhecimento de entidades – Primeira Base Anotada

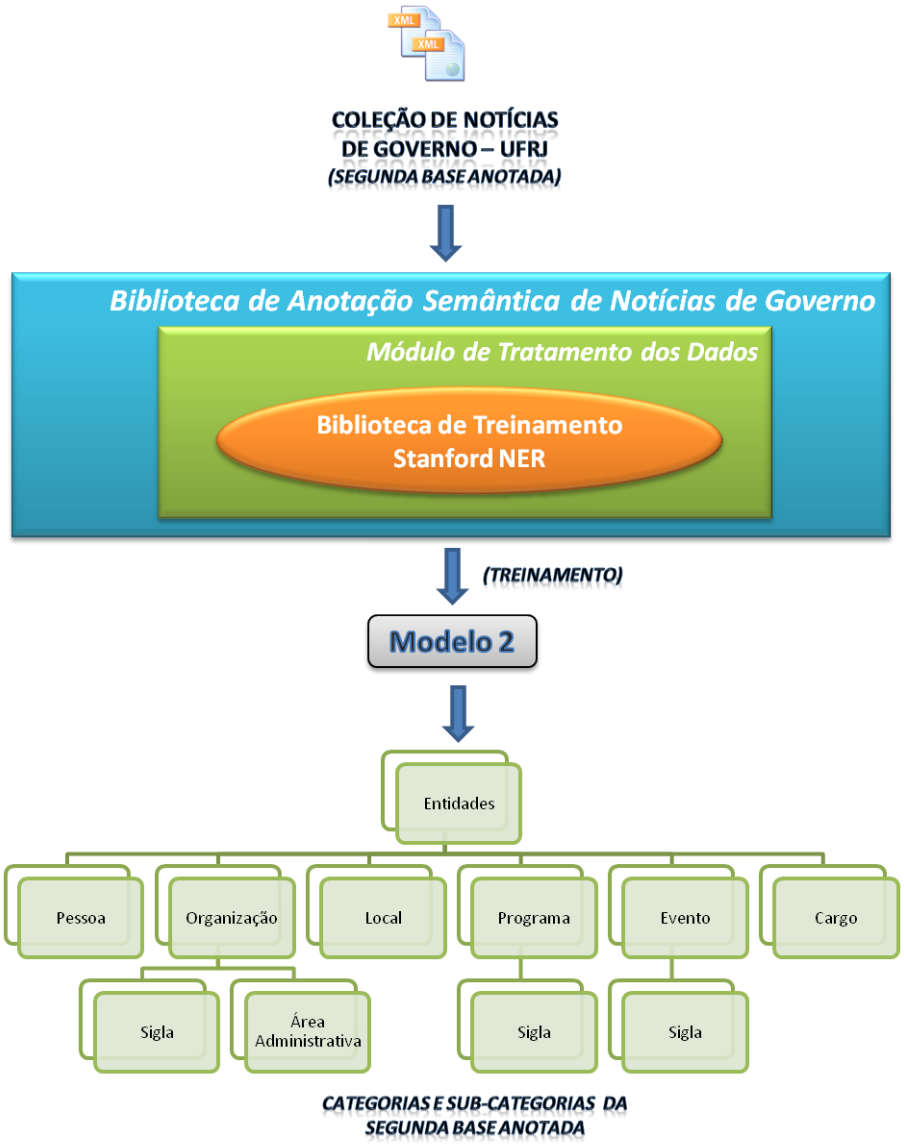

Figura 9: Arquitetura utilizada na concepção do modelo 2 utilizados na biblioteca de Anotação Semântica

O modelo 2 originado do treinamento realizado sobre a Segunda Base Anotada contempla as categorias e subcategorias Pessoa, Local, Organização, Programa, Evento, Cargo, Sigla e Área Administrativa. O modelo desenvolvido nesse trabalho considerando essa base obteve os F-Scores de 71% para a subcategoria Área Administrativa, 76% para a categoria Cargo, 41% para a categoria Evento, 70% para a categoria Local, 66% para a subcategoria Organização, 89% para a categoria Pessoa, 53% para a categoria Programa e 72% para a subcategoria Sigla.

O modelo desenvolvido obteve um bom desempenho nas categorias e subcategorias Área Administrativa, Cargo, Local, Organização, Pessoa e Sigla. Já nas categorias Evento e Programa o sistema apresentou dificuldade na classificação.

A tabela 8 apresenta os resultados obtidos pelo modelo criado neste trabalho a partir dos treinos realizados com a Segunda Base Anotada da Coleção de Notícias de Governo – UFRJ. Como resultados gerais foram calculados os F-Score geral e o F-Score contabilizando apenas as categorias Pessoa, Local e Organização. O primeiro apresentou um F-Score de 72% enquanto o segundo apresentou um F-Score de 67%.

Sistema de Reconhecimento de Entidades – Segunda Base Anotada			
Tipo de Entidade	*Precisão*	*Cobertura*	*FScore*
Área Administrativa	0.70	0.71	0.71
Cargo	0.81	0.72	0.76
Evento	0.71	0.28	0.41
Local	0.77	0.63	**0.70**
Organização	0.66	0.67	**0.66**
Pessoa	0.88	0.90	**0.89**
Programa	0.74	0.42	0.53
Sigla	0.69	0.75	0.72
TOTAL (P, L, O)	*0,75*	*0,72*	***0,73***
TOTAL (GERAL)	*0,75*	*0,67*	*0.71*

Tabela 8: Resultados do sistema de reconhecimento de entidades – Segunda Base Anotada

Capítulo 4. Integrador de Notícias de Governo

Diversas pesquisas relacionadas ao tratamento de notícias foram abordadas na literatura, entretanto, a maioria das pesquisas tinham como pressuposto uma base de notícias de fácil acesso, que na maioria das vezes foi elaborada com a finalidade de pesquisa.

Já recentemente, trabalhos têm sido realizados utilizando como fonte notícias publicadas por meio de Feeds RSS. Feeds [SILVA, 2010] são listas de atualização de conteúdo de um determinado site, escritos com especificações baseadas em XML. Estes Feeds contêm informações sobre as notícias e links apontando para a página original de cada notícia. Todavia, no mundo real essa abordagem encontra dois problemas:

Problema 1: Em geral, os Feeds RSS são constantemente atualizados e possuem uma quantidade limitada de notícias, então, se um sistema consumidor destas notícias perde algumas dessas atualizações as notícias são perdidas, acarretando perda de informação. Já as páginas HTML das notícias tendem a ter um período de persistência muito maior.

Problema 2: Sistemas de tratamento de notícias que utilizam como base Feeds não possuem acesso a sites que não disponibilizam Feeds. Esta situação é bem comum nos sites do domínio .gov.br, já que a maioria dos sites ainda não adotaram o uso de Feeds.

Portanto, para o uso prático das ferramentas de tratamento de notícias é necessário ter um método que permita a extração das notícias. Entende-se por extração o processo de identificar as páginas que possuem notícias e, dada a página, remover todo conteúdo irrelevante, tais como cabeçalho, rodapé, etc.

Este trabalho propõe a concepção do Integrador de Notícias do Governo. O objetivo deste integrador é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores, em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

O formato de dados utilizado para permitir a interoperabilidade das notícias é o próprio RSS. Como o RSS é gerado dinamicamente e gerenciado pelo integrador, os problemas 1 e 2 podem ser solucionados, já que um sistema X pode requisitar notícias passadas (Problema 1)

e pode requisitar notícias de um site que não possui RSS, já que este pode ser gerado automaticamente (Problema 2).

4.1 Integrador de Notícias de Governo

O objetivo do Integrador de Notícias do Governo é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

Para isso o sistema deve localizar de forma automatizada as páginas de notícias publicadas em um determinado grupo de sites do governo, extrair essas notícias, estruturá-las em metadados e disponibilizar mecanismos de recuperação para que sistemas possam consultá-los.

Com o intuito de simplificar a comunicação entre os sistemas foi utilizado o protocolo HTTP para a troca de mensagens e o formato RSS para a representação das notícias. O formato RSS foi escolhido porque ele já tem sido largamente utilizado para o compartilhamento de notícias.

O uso do formato RSS apresenta algumas vantagens, como por exemplo, comunicação direta entre os agregadores de Feeds, disponibilização de links com Feeds atualizados automaticamente para sites do governo que ainda não possuem seus próprios Feeds e compatibilidade com os diversos sistemas e técnicas apresentados na literatura científica que usam como base Feeds, tais como os publicados por Silva [2010], Pera e Ng [2008], Thelwall e Prabowo [2007] e Shaikh et al. [2010].

A comunicação entre um sistema qualquer e o Integrador de Notícias de Governo é realizada por uma requisição GET ou POST. A resposta é um arquivo no formato RSS contendo as notícias que satisfazem os parâmetros escolhidos na requisição. O resultado apresenta o título, a descrição, o link original, a data de publicação e o órgão publicador das notícias. Os parâmetros para efetuar a requisição são:

Fonte: Este parâmetro filtra as notícias por órgão. Uma lista de ids das fontes deve ser repassada.

Busca: Este parâmetro filtra as notícias como um campo de busca. O formato adotado é o utilizado pelo framework Lucene [LUCENE, 2010]. Este framework foi utilizado para a implementação desta funcionalidade.

Step: Este parâmetro define o número de notícias a serem recuperadas.

Page: Este parâmetro é utilizado para paginação das notícias.

4.1.1 Módulos do Integrador de Notícias de Governo

Este sistema é composto por dois módulos principais, o módulo responsável pela coleta e monitoramentos das páginas web do governo e o módulo responsável pela seleção, processamento e conversão da página web em uma representação estruturada.

O sistema monitora periodicamente as páginas publicadas nos sítios do governo e realiza a avaliação de padrões a fim de classificar uma dada página como notícia ou não, em caso de ser classificada como notícia a página é submetida a um processamento com o objetivo de extrair de forma automatizada os metadados da notícia descoberta.

4.1.1.1 Módulo Coletor de Notícias

Atualmente a quantidade de informação disponível na web vem crescendo rapidamente. Por isso uma varredura completa pelas páginas web a fim de atualizar uma determinada base não é uma tarefa trivial e dependendo da taxa de atualização necessária da base, pode se tornar inviável. O problema se agrava ainda mais quando o objetivo é coletar notícias, já que a publicação de novas notícias é realizada rapidamente. Em um intervalo de duas horas uma base contendo notícias previamente cadastradas da internet pode se tornar desatualizada.

Então, nesse trabalho houve a necessidade de desenvolver um módulo capaz de gerenciar a base de notícias atualizando-a constantemente. Desta forma, foi desenvolvido o Módulo Coletor que é um Web Crawler desenvolvido especialmente para navegar pelos sites do governo a procura de páginas com potencial de terem como conteúdo notícias.

Para reduzir o espaço de busca foram utilizadas heurísticas baseadas no comportamento humano quando este procura por notícias em sítios da web. As heurísticas adotadas consideram a localidade das páginas web nos seus respectivos sítios e a existência de termos demarcadores que indicam a possível existência de notícias na página. Estas heurísticas são baseadas em três premissas que são descritas abaixo:

PREMISSAS 1: *As notícias recentes de um sítio web estão nas vizinhanças da página principal.*

PREMISSAS 2: *As páginas contendo as notícias propriamente ditas ou uma lista delas possuem textos que o identificam como notícias.*

PREMISSAS 3: *O sistema se atualizará periodicamente.*

Sítios da web que seguem as boas práticas na elaboração da navegação de suas páginas estão enquadrados na premissa número 1. Com essa restrição imposta pela premissa número 1 uma redução considerável do espaço de busca é realizada.

A premissa número 2 baseia-se no fato de que humanos precisam identificar dentro dos sites os espaços dedicados às notícias e para que isso seja possível o próprio site deve fornecer evidências disso, por exemplo, links com texto como "notícias", "destaque", etc. A premissa número 2 é utilizada com o objetivo de pontuar determinada página [ADAR et al., 2009] como mais provável ou não a ser uma notícia. A fim de evitar falsos negativos essa hipótese é utilizada para indicar a direção da busca e não para a redução do espaço de busca a não ser que o tempo não seja suficiente para visitar todo o espaço devido ao timeout de busca.

A premissa número 3 é utilizada apenas para fortalecer a premissa número 1, pois considerando que o Coletor esteja periodicamente rodando apenas as notícias recentes são o foco da busca, já que as demais notícias já devem ter sido encontradas em rodadas anteriores.

O procedimento adotado pelo módulo coletor deste trabalho consiste em simular N filas, sendo N o número de sites monitorados. Cada site é modelado como um grafo direcionado G = {V, E}, onde V são páginas webs e E é o conjunto de transições entre páginas web.

Seja R = $\{v_{r1}, v_{r2}, ..., v_{rm}\}$ um conjunto de páginas raízes de busca, seja K1 e K2 parâmetros numéricos, seja $V' = \{c_1, c_2, ..., c_t\}$ o conjunto de todas as páginas de um site e $H(v_i, vj)$ uma função binária que diz se um link de rótulo x presente em v_i que aponte para v_j está relacionado à alguma notícia. Então, o grafo G pode ser construído da seguinte forma:

PASSO 1: Tornar V = R;

PASSO 2: Adicionar a V todas as páginas que estejam a distância K1 ou menor de V e adicionar a E todas as arestas envolvidas nesses caminhos;

PASSO 3: Adicionar a V todas as páginas que estejam a distância K2 ou menor de V cujo caminho é formato por arestas que tenham $H(v_i, c_j) = 1$, com exceção da última aresta que não possui restrição. Além disso, adicionar as arestas envolvidas em E.

No caso de K1 = 0, o grafo é formado pelas páginas raízes somadas das páginas conectadas as páginas raízes que tenha $H(v_i, c_j) = 1$. Por exemplo, se K1 = 1, K2 = 5 e a função H for 1 se o nome do link atende a expressão regular "(notícia)|[0-9]+" e 0 caso contrário, o resultado de um possível grafo seria o formato pela página principal, a primeira página que contém a lista de notícias de um site e todas as páginas de paginação de notícias que estejam no máximo a 5 páginas de distância. No caso de $K1 = \infty$ o grafo engloba todas as páginas contidas no site que podem ser alcançadas a partir das páginas raízes.

O parâmetro K1 determina até que profundidade as páginas serão coletadas normalmente sem aplicação de filtros, já o parâmetro K2 determina até que profundidade as páginas serão coletadas tendo a função $H(v_i, c_j)$ como guia. Resumidamente, a ideia é que o WebCrawler navegue inicialmente na tentativa de criar um conjunto de páginas que tenha uma boa probabilidade de conter um link apontando para a página principal de notícias deste site. Então, a ideia seguinte é que o WebCrawler passe a navegar apenas pelas páginas de interesse: páginas que tenham as notícias ou páginas que listam as notícias.

A navegação adotada pelo WebCrawler dentro de um site terá uma eficiência proporcional à qualidade da função $H(v_i, c_j)$.

Neste trabalho a função $H(v_i, c_j)$ é elaborada por meio de uma expressão regular que diz se um rótulo de um link é relevante ou não, por exemplo, se no rótulo contém os termos "notícias", "veja mais", etc. Além disso, também são adicionadas possíveis referências a paginações tais como "1", "2", "próximo", "anterior", etc.

Nos experimentos realizados nesse trabalho os parâmetros K1 = 1 e K2 = 3 obtiveram os melhores resultados considerando custo-benefício. Isso porque, em geral, as páginas relacionadas às notícias estão muito próximas à página principal do site, então não é necessária uma pesquisa profunda para encontrar a listagem de notícias.

4.1.1.2 Módulo de Processamento de Notícias

A maior parte das notícias publicadas nos sites do governo estão armazenadas de forma desestruturada e mesmo alguns sites que possuem RSS disponibilizam apenas títulos ou resumos e não o texto da notícia na íntegra.

As notícias disponíveis na web são constantemente visitadas por milhões de pessoas todos os dias. Cada pessoa que tem acesso à web e busca por novas notícias empenha de forma natural um conjunto de ações necessárias para obter acesso a elas. Por exemplo, um indivíduo ao acessar a página principal de um site, busca de forma intuitiva a área de notícias e após identificar essa área, procura por uma notícia de interesse, e então o indivíduo naturalmente consegue discernir o que é o título da notícia, o corpo da notícia e até mesmo a sua data de publicação.

De alguma forma o conhecimento necessário para discernir as partes integrantes da notícia encontra-se de forma tácita no indivíduo, então para que seja possível automatizar esse processo foi realizado um estudo a fim de documentar o processo intuitivo utilizado pelas pessoas na análise de uma notícia. A partir de um conjunto de premissas extraídas do estudo acima, foi desenvolvido o Módulo de Processamento.

O Módulo de Processamento é responsável por ler uma página HTML e extrair a notícia desta e então fragmentá-la em metadados a fim de estruturá-la. Para realizar a atualização da base de dados o Módulo de Processamento e o Módulo Coletor são integrados. O Módulo Coletor transmite para o Módulo de Processamento as páginas encontradas, em seguida este módulo analisa cada página e armazena os metadados na base de dados caso a página seja considerada uma notícia.

O fluxo da figura 10, adotado para o Módulo de Processamento, é dividido em quatro etapas. A primeira consiste em realizar um pré-processamento da página, a segunda consiste em filtrar as páginas relevantes para o sistema, a terceira consiste no processamento da notícia e a quarta e última consiste em atualizar a base de dados com as informações extraídas.

A etapa de pré-processamento consiste em fazer uma raspagem do código HTML e uma preparação do texto para a sua posterior mineração. Os métodos implementados para a raspagem visam remover do código HTML áreas como menus, botões, anúncios e etc. Os métodos adotados para a raspagem foram adaptações de técnicas existentes na literatura como as encontradas em Gupta et al. [2005], Yi et al. [2003] e Ramaswamy et al. [2003]. Para a preparação do texto foi utilizada a biblioteca HtmlUnit [HTMLUNIT, 2011]. Uma das funcionalidades desta biblioteca é a de disponibilizar a partir de uma página web o conteúdo visível pelo usuário em um browser. O resultado da etapa de pré-processamento é o texto

visível pelo usuário da página e a parte do código HTML resultante da eliminação dos ruídos de informação.

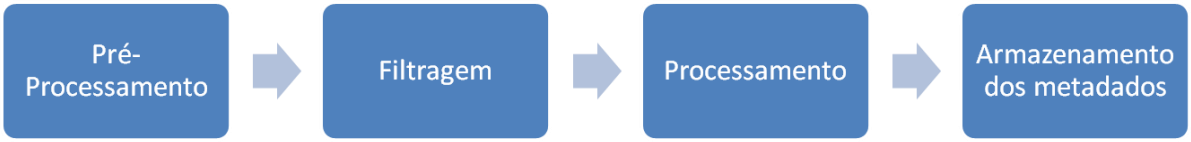

Figura 10: Fluxo do Módulo de Processamento

O Módulo Coletor consegue filtrar apenas partes das páginas que não são notícias a fim de diminuir o fluxo de páginas não relevantes que vão para o Módulo de Processamento. As demais páginas não relevantes são filtradas pelo Módulo de Processamento já que este dispõe de mecanismos de mineração de página mais robustos do que o Módulo Coletor.

Na etapa de filtragem das páginas é realizada uma análise sobre a estrutura do texto retornado pelo pré-processamento. Esse texto é submetido a uma série de verificações que checam se o texto atende às propriedades mínimas encontradas em textos de notícias, o qual algumas foram inspiradas em Norvag e Oyri [2005]. Tais propriedades refletem a estrutura posicional do texto e as evidências léxicas de termos relacionados à área de notícias. Por exemplo, a densidade dos parágrafos, o número de parágrafos, presença de datas recentes, posição absoluta do texto na página e etc. As propriedades padrões para a checagem das páginas foram ajustadas após vários testes realizados sobre uma base de notícias preparada para este trabalho.

A etapa de filtragem retém um número considerável de páginas não relevantes para o sistema. Entretanto, algumas páginas não relevantes só poderão ser descartadas na etapa de processamento quando uma mineração mais apurada será realizada.

A etapa de processamento é responsável por selecionar a notícia dentro da página HTML e extrair os metadados relevantes para a representação da notícia em um formato RSS. Os metadados armazenados da notícia são o título, o corpo do texto, a data de publicação, a data de visitação da página pelo coletor, o órgão fonte e o link de publicação.

Para extrair a notícia e a separar em suas partes integrantes algumas heurísticas baseadas na estrutura posicional do texto são utilizadas. Segue abaixo as principais premissas que servem como base para as heurísticas utilizadas:

1. *Uma notícia é minimamente composta por um título, descrição e data de publicação.*

2. O título de uma notícia está localizado próximo a descrição e sempre acima dela.

3. A data de publicação de uma notícia está localizada próximo ao título ou após a descrição.

4. A data de publicação de uma notícia é uma data dentro de um intervalo considerado aceitável para o escopo da busca.

5. As propriedades HTML da descrição da notícia são as mesmas para todas as palavras da notícia ou pelo menos para a maior parte delas. O mesmo é válido para o título e a data de publicação.

6. Um título e a descrição possuem tamanhos característicos, tendo um tamanho e densidade de palavras mínimas e máximas aceitáveis.

7. Sejam três parágrafos A, B e C dispostos em sequência. Se os parágrafos A e C pertencem à descrição da notícia então o parágrafo B também faz parte da descrição ou é uma legenda de uma imagem ou tabela.

8. Elementos como descrição, título e data de publicação tendem a ter propriedades semelhantes na estrutura de tags do código HTML bem como características de fontes semelhantes.

9. O uso de nomeações de variáveis, de classes e ids nos códigos fontes das páginas HTML podem ser indícios de título, descrições e datas de publicação. Por exemplo, " A educação no Brasil ... "

10. A descrição da notícia é a área com maior densidade de palavras e linhas da página.

Apesar de algumas das premissas serem consideradas como óbvias para seres humanos elas não são para o computador e por isso precisam ser descritas e implementadas.

Baseado nas premissas acima foram implementadas três vetores de funções características $t = (t_1, t_2, t_3, t_4, \ldots, t_k)$, $d = (d_1, d_2, d_3, d_4, \ldots, d_k)$ e $p = (p_1, p_2, p_3, p_4, \ldots, p_k)$. Essas funções características representam as características dos metadados, onde t são funções características relacionadas ao título da notícia, d função características relacionadas à descrição da notícia e p funções características relacionadas à data de publicação.

Seja D um vetor de possíveis descrições da notícia de uma página web P. Então temos que a pontuação de uma possível descrição D_i é dada por

$$pontuaca\_descrição(D_i) = \sum_{j=1}^{k} \lambda_{dj} d_j(P, D_i)$$

,onde D^* = $argmax_{di}$(pontuação_descricao(D_i)). D^* é a descrição mais provável da notícia presente na página P.

Seja T um vetor de possíveis títulos da notícia de uma página web P, Então temos que a pontuação de um possível título T_i é dado por:

$$pontucao\_titulo(T_i) = \sum_{j=1}^{k} \lambda_{tj} d_j(P, T_i, D^*)$$

,onde T^* = $argmax_{ti}$(pontuação_titulo(T_i)). T^* é o título mais provável da notícia presente na página P dado a descrição D^*.

Seja DP um vetor de possíveis datas de publicação da notícia de uma página web P, então temos que a pontuação de uma possível data DP_i é dada por:

$$pontucao\_data\_publicacao(DP_i) = \sum_{j=1}^{k} \lambda_{pj} p_j(P, DP_i, D^*, T^*)$$

,onde DP^* = $argmax_{dpi}$(pontuação_data_publicacao(DP_i)). DP^* é a data de publicação mais provável da notícia presente na página P dado a descrição D^* e o título T^*.

Os vetores $\lambda_d = (\lambda_{d1}, \lambda_{d2}, \lambda_{d3}, \lambda_{d4}, \ldots, \lambda_{dk})$, $\lambda_t = (\lambda_{t1}, \lambda_{t2}, \lambda_{t3}, \lambda_{t4}, \ldots, \lambda_{tk})$ e $\lambda_p = (\lambda_{p1}, \lambda_{p2}, \lambda_{p3}, \lambda_{p4}, \ldots, \lambda_{pk})$. São vetores de pesos relacionados às funções características. O objetivo desses parâmetros é ponderar as características de acordo com a sua importância relativa. Nesse trabalho os parâmetros λ foram escolhidos empiricamente por meio de vários experimentos.

4.2 Estudo de Caso – Portal de Notícias de Governo

Uma versão preliminar do Integrador de Notícias de Governo, implementado nesse trabalho, apresentado nesse capítulo foi utilizada em um sistema Web do Ministério do Planejamento chamado Portal de Notícias do Governo. Nesse cenário o objetivo do Integrador de Notícias de Governo é fornecer ao Portal uma base de notícias constantemente atualizada com as notícias publicados nos mais diversos sites de interesse.

O Portal de Notícias do Governo é uma iniciativa que envolve a cooperação da SLTI/MP, SECOM/PR, COPPE/UFRJ e SERPRO. O seu objetivo é atender às necessidades

da SECOM e prover ao público um ambiente intuitivo e robusto para a pesquisa de notícias publicadas nos sítios do governo. Este Portal foi desenvolvido em J2EE (Java 2 Enterprise Edition) fazendo uso do Padrão MDA (Model Driven Architecture) através do Framework do Ministério da Defesa e Ministério do Planejamento chamado MDArte [PINEL et. al., 2011].

Este Portal é dividido em duas áreas, a área de pesquisas destinada a qualquer usuário da web e a área de gestão de notícias destinada à SECOM e à SRI. As próximas subseções abordam a motivação inicial do projeto Portal de Notícias de Governo, a área de pesquisas do Portal, a área de gerência de notícias, e o uso de Feeds dinâmicos no Portal.

4.2.1 Motivação

A Secom [2010] é responsável pela comunicação do Governo Federal, coordenando um sistema que interliga as assessorias dos ministérios, das empresas públicas e das demais entidades do Poder Executivo Federal. Ela atua para que as ações de comunicação obedeçam a critérios de sobriedade e transparência, eficiência e racionalidade na aplicação dos recursos, além de supervisionar a adequação das mensagens aos públicos. Também observa o respeito à diversidade étnica nacional e regionalização no material de divulgação, avaliando os resultados.

Portanto, é de responsabilidade da SECOM analisar as notícias disponibilizadas ao público pelo Governo Federal, através das assessorias de imprensa dos vários órgãos que o constituem. Por exemplo, a análise permite localizar notícias que possam apresentar evidências de discriminação racial, religiosa, etc., o que pode afetar a imagem do governo.

Atualmente, a localização das notícias para a análise é realizada por analistas de forma manual. Um analista com posse de uma lista de sites acessa um por um procurando por notícias recentes. Esse processo, por ser manual, gera um custo de tempo elevado. Esse elevado custo de tempo, por sua vez, impõe um limite ao número de sites que podem ser analisados pela SECOM, reduzindo a eficácia da análise geral.

Com o objetivo de reduzir os custos de tempo e ampliar a capacidade de análise de notícias o Ministério do Planejamento (MP) propôs um sistema informatizado que permita a gerência das notícias de interesse da SECOM.

Entretanto, essas notícias encontram-se descentralizadas e desestruturadas dificultando a sua recuperação automática. Por exemplo, no ano de 2009 estudos realizados pelo Projeto

Censo Web .br [CGI.br e NIC.br, 2010] identificaram um total de 11.856 sítios sob o domínio .gov.br, sendo visitadas um total de 6.331.256 páginas no formato HTML (figura 11). Todas as notícias publicadas por entidades do governo estão espalhadas por esta vasta quantidade de sites e páginas da web.

NÚMERO DE PÁGINAS HTML E SITES DA WEB - .GOV.BR		
NÚMERO DE SITES WEB	NÚMERO TOTAL DE PÁGINAS HTML DA WEB	NÚMERO MÉDIO DE PÁGINAS HTML POR SITE
11.856	6.331.256	534,01

Figura 11: Número de Páginas HTML e Sítios da Web no domínio .GOV.BR. Adaptado de CGI.br e NIC.br [2010]

Uma possível maneira de estruturar as notícias dos sítios do governo é o uso de Feeds/RSS [HAMMERSLEY, 2005] contendo as notícias mais recentes. Entretanto, apesar de seu uso ter sido amplamente adotado na web, ainda não é utilizado pela maioria dos sítios do governo. Além disso, a grande maioria dos sítios do governo que apresentam RSS disponibiliza apenas o título, ou o título e um resumo da notícia e não o texto na íntegra.

Então, para solucionar a deficiência de estruturação e descentralização das notícias nos sites do governo foi utilizado o Integrador de Notícias de Governo. O Integrador de Notícias de Governo é responsável por manter atualizada a base de notícias utilizando o ambiente informatizado chamado Portal de Notícias de Governo.

4.2.2 Área de Pesquisas de Notícias

A área de pesquisas do Portal de Notícias do Governo (figura 12) é uma interface especializada para a realização de consultas sobre as notícias publicadas pelos diversos sítios do governo. O portal é disponibilizado ao público e permite aos usuários realizarem consultas genéricas ou consultas especializadas.

Para a realização de buscas é utilizado o framework Lucene [LUCENE, 2010]. O Lucene é uma biblioteca desenvolvida totalmente em JAVA e disponibiliza diversas opções de máquinas de busca. O Lucene oferece uma implementação com alto desempenho na realização de consultas sobre volumosas bases de dados.

O Portal de Notícias por meio do Lucene oferece ao usuário a possibilidade de realizar consultas utilizando operadores AND, OR, NOT, + e – tais como as utilizadas nas diversas máquinas de busca disponibilizadas na web. Exemplos de consultas são:

- *PAC AND Salvador*
- *"PAC 2" OR "Minha Casa, Minha Vida"*

Além de realizar uma consulta genérica o usuário pode realizar consultas restringindo os sítios a serem consultados. O usuário pode selecionar os sítios de interesse ou as categorias de interesse em que deseja realizar a busca. Por exemplo, o usuário pode realizar uma consulta buscando por todas as notícias publicadas pelos ministérios em que o programa "Minha Casa, Minha Vida" é mencionado.

Atualmente o portal abrange um total de 138 sites divididos em 10 categorias. As categorias são:

Categorias	*Nº de Sites*
Ministérios	23
Secretarias	8
Conselhos	2
Agências	5
Bancos	4
Associações Municipais	12
Governos Estaduais e Distrito Federal	9
Prefeituras e Capitais	7
Fundações	4
Outros	64
TOTAL	138

Tabela 9: Distribuição de Sites por categorias

Figura 12: Área pública do Portal de Notícias do Governo

4.2.3 Área de Gestão de Notícias

O Portal de Notícias possui um ambiente de acesso restrito. O desenvolvimento desse ambiente é uma iniciativa de informatização do ambiente de trabalho da SECOM e da SRI, cuja finalidade é o monitoramento interno das notícias de interesse de ambos os órgãos.

A área restrita (figura 13) contém todas as funcionalidades disponibilizadas na área pública, entretanto, informações adicionais são vinculadas às notícias para a gerência das informações contidas na base de notícias. Com o ambiente de monitoramento informatizado os analistas podem avaliar as notícias e manter um controle interno destas.

A informatização do monitoramento das notícias possui como principais vantagens os seguintes itens:

Fácil acesso às notícias – Os analistas responsáveis pela análise das notícias possuem a tarefa de localizar as notícias nos sítios do governo periodicamente. Essa busca demanda tempo já que diversas páginas web precisam ser acessadas, e muitos destes acessos são realizados desnecessariamente, tendo em vista que nem todos os sítios são atualizados diariamente. A área restrita do Portal de Notícias do Governo disponibiliza ao analista a lista de notícias publicadas recentemente de forma centralizada, reduzindo o tempo gasto na procura pelas notícias do dia.

Vincular informações às notícias – Os analistas ao analisarem as notícias eventualmente precisam realizar anotações sobre as notícias em análise. Uma vez que as notícias publicadas são automaticamente armazenadas na base de dados, há a possibilidade de vincular informações a essas notícias. Por exemplo, informações como: potencial de replicação da notícia, avaliação numérica dada à notícia, etc.

Geração de relatórios – Com um sistema informatizado é possível gerar relatórios sobre as notícias publicadas e as informações vinculadas às notícias. Por exemplo, criar um relatório com todas as notícias da semana atual e que foram avaliadas pelos analistas como relevantes para serem publicadas no Portal Brasil.

Figura 13: Área restrita do Portal de Notícias do Governo

4.2.4 Feeds RSS e o Portal de Notícias

As notícias publicadas no Portal de Notícias do Governo podem ser integradas com outros sistemas através da disponibilização das notícias em formato RSS. RSS [BLEKAS et. al., 2006] é uma família de formatos de arquivos XML que resumem informações de web sites. Ele é utilizado em sites que contêm uma elevada periodicidade de atualização, como por exemplo, sites de notícias web.

No formato RSS as informações relativas às notícias são armazenadas em campos separados (figura 14). Dessa forma as notícias que estão desestruturadas nas páginas HTML

encontram-se estruturadas em formato de XML e em um padrão que possibilita a integração entre sistemas.

O Portal de Notícias gera um link RSS para cada pesquisa realizada pelo usuário. Esse RSS é gerado dinamicamente a cada requisição, contendo todas as notícias que atendem à consulta realizada. O usuário pode vincular o RSS gerado a agregadores de Feeds RSS. Tais agregadores podem ser utilizados em computadores pessoais, smartphones, tablets, etc. Dessa maneira é possível o usuário se manter atualizado sobre uma consulta constantemente. A figura 15 mostra um exemplo da integração do RSS gerado pelo Portal de Notícias com o Google Reader.

RSS tag	**Descrição**
<rss>	Inicializar a informação RSS
<item>	Informação resumida
<ttl>	Tempo de vida (em minutos). Indica a quantidade de tempo durante o qual a informação é considerada válida
<title>	O título da informação
<description>	Uma pequena descrição da informação
<link>	Link para a informação na página web

Figura 14: Principais tags do RSS - adaptado de BLEKAS et al [2006]

Figura 15: Integração do RSS gerado pelo Portal de Notícias e o Google Reader

4.3 Ambiente de Análise de Notícias de Governo

Este capítulo apresenta um protótipo em andamento idealizado neste trabalho que é o Ambiente de Análise de Notícias de Governo, este sistema está sendo desenvolvido a partir do Sistema de Reconhecimento de Entidades Nomeadas no capítulo 3 e do Integrador de Notícias de Governo descrito nas subseções acima.

O objetivo deste protótipo é prover funcionalidades de extração de informação de um dado conjunto de notícias e disponibilizar essa informação visualmente, a fim de proporcionar ao analisador um panorama das notícias referentes a um contexto específico. Este protótipo ilustra o potencial do uso do Sistema de Reconhecimento de Entidades Nomeadas, quando este é combinado com o Integrador de Notícias de Governo.

Não é objetivo deste trabalho finalizar o protótipo, mas sim demonstrar algumas potencialidades que podem ser alcançadas quando o Sistema de Reconhecimento de Entidades Nomeadas e o Integrador de Notícias de Governo são usados em conjunto.

O protótipo Ambiente de Análise de Notícias de Governo é composto por dois módulos principais de visualização. O primeiro é o de geocodificação das localidades de um dado conjunto de notícias, o segundo é a representação de um dado conjunto de notícias por meio

de Nuvens de Tags. Ambos os módulos fornecem ao usuário uma visualização geral das informações contidas nas notícias.

4.3.1 Geocodificação das Notícias de Governo

Um importante exemplo de informação que pode ser obtida de uma notícia é o local a qual esta se refere. Perguntas como *onde?, de onde?* são perguntas comuns quando se tratam de notícias. Por exemplo, quais são os lugares mais mencionados quando o Programa de Aceleração do Crescimento (PAC) é mencionado? Quais localidades tiveram destaques em número de notícias em um dado período de tempo? Quais localidades tiveram o maior número de registro de dengue?

A análise visual dessas informações pode auxiliar os analistas na verificação de padrões de um determinado acontecimento, já que uma análise visual é mais fácil de ser compreendida do que uma análise analítica quando o número de notícias é grande.

Com a finalidade de fornecer essa visualização foi acoplado ao protótipo a API do Google Maps [TUTORIAL, 2010] e a API de Geocodificação do Google [GEOGOOGLE, 2008]. O funcionamento desta visualização se dá a partir de um conjunto de notícias de interesse do usuário. Destas notícias são extraídas as entidades que indicam uma localidade, então essas entidades são geocodificadas e plotadas no mapa.

Como o número de notícias pode ser demasiadamente grande, a plotagem de todos os lugares podem se tornar inviável, então é utilizada uma API que agrupa os marcadores no mapa de acordo com suas proximidades. A figura 16 exemplifica a geocodificação das notícias.

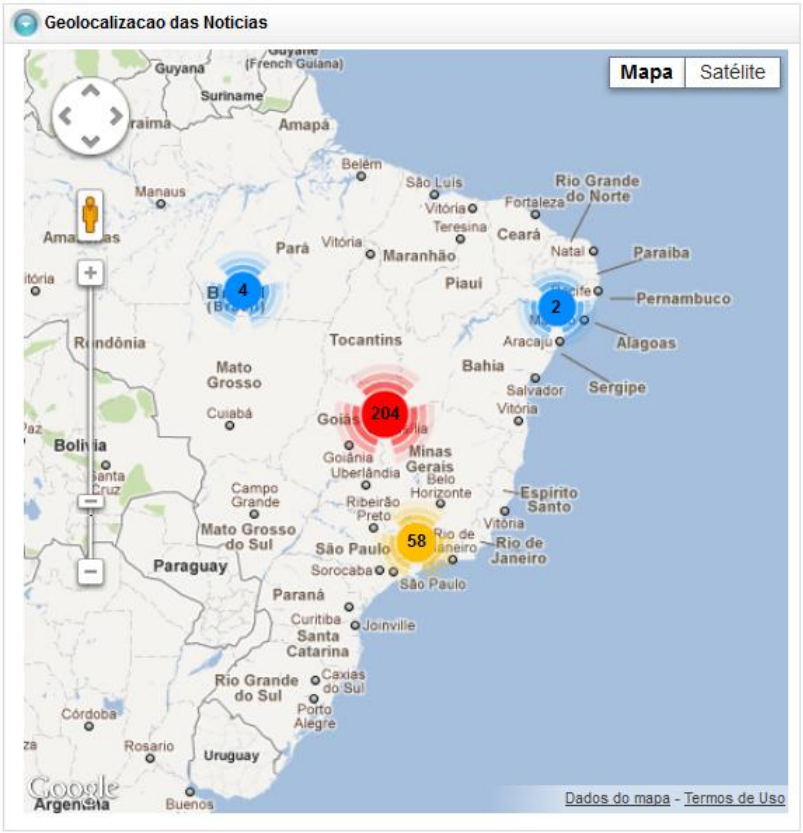

Figura 16: Mapa utilizado na visualização das localidades

4.3.2 Representação do Resultado por Tags de Entidades

Outra informação importante quando se trata de notícias é descobrir qual são os temas mais abordados durante um dado período de tempo. Neste trabalho é sugerida uma visualização de resultados de uma pesquisa de notícias usando Nuvens de Tags. As Nuvens de Tags sugeridas neste trabalho são formadas por entidades, onde cada categoria de entidade corresponde a uma Nuvem de Tags.

Segundo Morgado [2010] "Nuvens de Tags são representações visuais de um conjunto de palavras, tipicamente um conjunto de tags selecionadas por algum método racional, na qual atributos do texto como tamanho, estilo, ou cor são usados para representar características dos termos associados".

A visualização das Nuvens de Tags apresenta quais as entidades mais mencionadas, dado um contexto de interesse. Por exemplo, um filtro de notícias por dia teria como resultado Nuvens de Tags que correspondem as Pessoas, as Organizações e os Eventos mais mencionados durante o dia.

A ideia da construção das Nuvens de Tags propostas neste trabalho é simples considerando que já existe um sistema de reconhecimento de entidades. Basta o sistema classificar as notícias de interesse e contabilizar a frequência das entidades de cada categoria, e para cada categoria criar uma Nuvem de Tags com as entidades de maior frequência.

Neste trabalho foi utilizado a API Term Cloud [TERMCLOUD] disponibilizada pelo Google. A figura abaixo ilustra o uso desta API sobre a categoria Pessoa.

Figura 17: Exemplo de Nuvem de Tags da categoria Pessoa

Capítulo 5. Experimentos e Resultados

Este capítulo visa detalhar os principais experimentos realizados durante este trabalho. Os métodos utilizados, os processos de treinamento, bem como os resultados obtidos, que são descritos de forma a embasar as conclusões obtidas nesse trabalho.

Este capítulo é organizado da seguinte forma: a seção 5.1 apresenta os métodos utilizados, de maneira a proporcionar uma visão geral dos experimentos.

A seção 5.2 descreve os detalhes da biblioteca de reconhecimento de entidades nomeadas disponibilizada pela universidade de Stanford. O objetivo desta seção é descrever a biblioteca de Stanford, a fim de proporcionar aos leitores as informações necessárias para reproduzir os experimentos, uma vez que não há uma documentação formalizada da biblioteca.

A seção 5.3 descreve os experimentos realizados para o reconhecimento de entidades nomeadas em textos da língua portuguesa em um contexto global, ou seja, o reconhecimento de entidades nomeadas independente do contexto. O objetivo dos experimentos realizados nesta seção é a construção de um baseline para os demais experimentos.

As seções 5.4 e 5.5 descrevem os experimentos realizados no reconhecimento de entidades nomeadas em textos da língua portuguesa no contexto de notícias de governo, ou seja, o conjunto de textos relevantes se restringe às notícias relacionadas ao governo brasileiro. O sistema resultante da seção 5.3 é utilizado como baseline dos experimentos dessa seção.

5.1 Métodos do Experimento

Esta subseção visa apresentar um panorama das etapas utilizadas para a realização dos experimentos. A figura 18 sintetiza as etapas envolvidas na execução dos experimentos.

A primeira etapa é a Coleta de Dados que consiste na pesquisa de quais dados são relevantes para a pesquisa. Para a coleta de dados foram utilizadas duas fontes, a base de textos anotada, disponibilizada pela Linguateca, chamada de Coleção Dourada; e a Primeira e Segunda Base Anotada da Coleção de Notícias de Governo – UFRJ detalhada no capítulo 3.

A segunda etapa é a Preparação da Base de Textos que consiste em preparar os dados coletados na etapa anterior para as etapas seguintes do experimento. Nessa etapa a Coleção Dourada e a Coleção de Notícias de Governo são tratadas de forma a atender as especificidades destes experimentos. O resultado desta etapa é um conjunto de arquivos de treino e teste utilizados em cada um dos experimentos.

Da Coleção Dourada são extraídos os exemplos destinados aos experimentos de reconhecimento de entidades sem restrição dos textos, cujo objetivo é definir um baseline para os experimentos seguintes. Da Coleção de Notícias de Governo – UFRJ são extraídos os exemplos destinados aos experimentos de reconhecimento de entidades em notícias relacionadas ao governo.

A terceira etapa é a preparação para a execução da primeira série de experimentos. O objetivo dessa série de experimentos é de definir um baseline para os experimentos realizados sobre as notícias de governo. A quarta etapa é análise dos resultados obtidos nos experimentos preparados na etapa anterior.

As próximas etapas visam realizar a execução dos experimentos relativos ao reconhecimento de entidades nomeadas em notícias de governo conforme proposto neste trabalho.

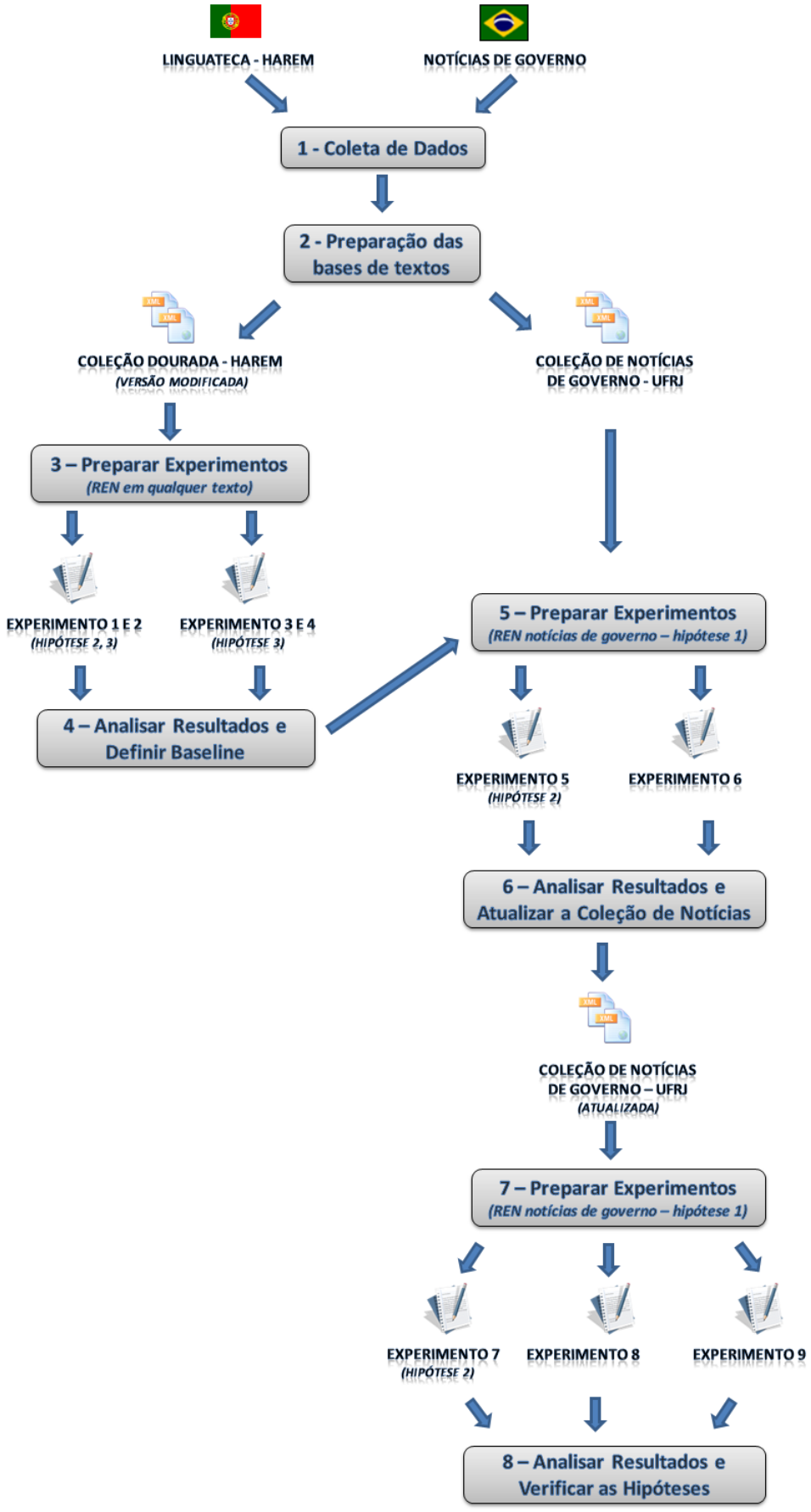

Figura 18: Etapas realizadas na execução dos experimentos

5.2 Biblioteca de Reconhecimento de Entidades Nomeadas de Stanford: Stanford NER

O Stanford NER [THE STANFORD, 2012] é utilizado no aprendizado supervisionado de extração de entidades nomeadas e também para a extração das entidades baseada no modelo gerado pelo treinamento. A figura 19 mostra os passos envolvidas no uso do Software NER para o reconhecimento de entidades nomeadas. Estes passos são divididos em duas etapas distintas, a primeira é treinar o sistema com a finalidade de aprender a rotular novas sentenças e a segunda é a de rotular novas sentenças.

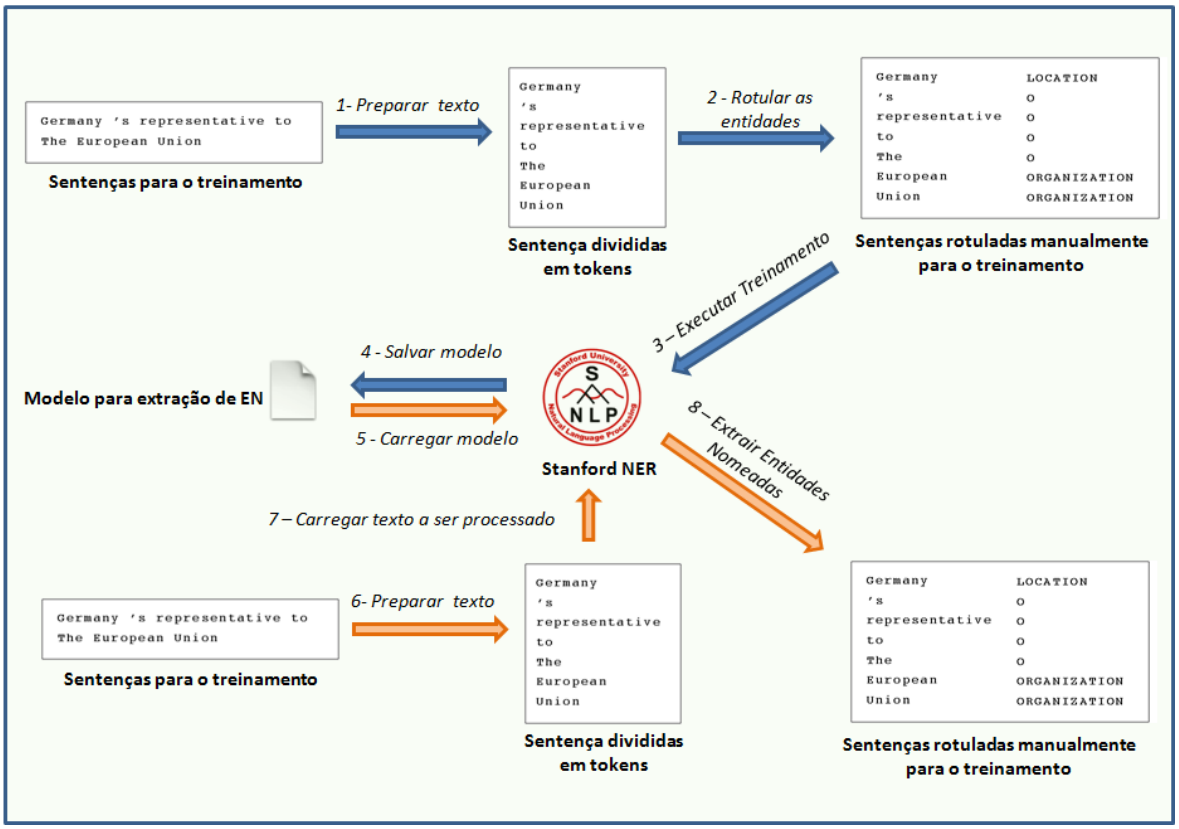

Figura 19: Processo de execução do Software NER

As subseções abaixo explicam detalhadamente cada passo do uso do Stanford NER no reconhecimento de entidades nomeadas.

5.2.1 Etapa de Treinamento

5.2.1.1 Preparação do texto

A primeira etapa para utilizar o Stanford NER consiste em treinar o sistema de maneira que ele possa aprender a rotular novas sentenças corretamente. O primeiro passo desta etapa é selecionar um conjunto de sentenças que será utilizado para criar os exemplos destinados ao treinamento.

As sentenças selecionadas precisam ser diversificadas de forma a abranger o maior número de contextos diferentes possíveis. Essa diversificação é importante porque o algoritmo utilizado pelo Stanford NER se baseia no contexto em torno das palavras para identificar a entidade nomeada. Então o importante não é diversificar os nomes das entidades, e sim os contextos em que estas entidades estão inseridas.

O algoritmo utilizado pelo Stanford NER não utiliza, explicitamente, os elementos morfológicos das sentenças no processo de aprendizado. Esta característica permite a este software ser considerado multilíngue, ou seja, ele pode ser utilizado em diversas línguas. Portanto, as sentenças utilizadas para o treinamento podem ser de variados idiomas, como por exemplo, o inglês, o português e o alemão.

As sentenças selecionadas precisam ser preparadas para o treinamento. O Linear-Chain Conditional Random Field implementado pelo Stanford NER atua sobre uma sequência linear de palavras, por isso é preciso transformar as sentenças que estão em linguagem natural em uma cadeia linear de tokens.

O Stanford NER disponibiliza um programa para a realização deste pré-processamento.

A figura 20 mostra um simples exemplo da transformação do texto em um sequência linear.

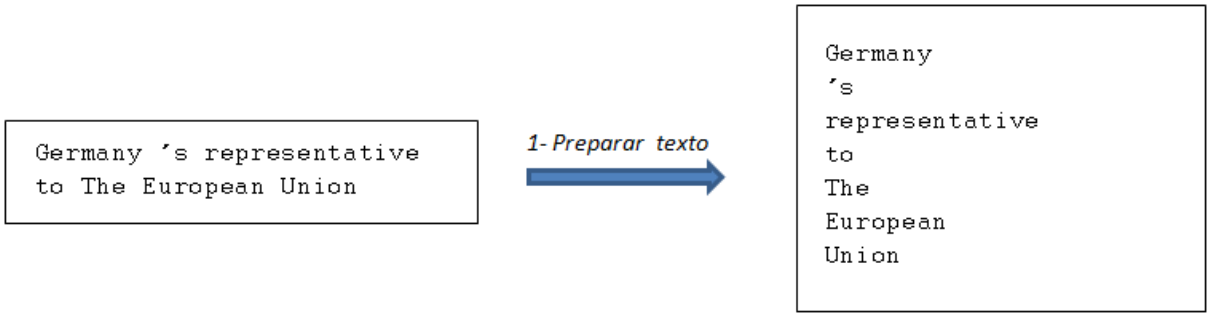

Figura 20: Exemplo de pré-processamento das sentenças

5.2.1.2 Anotação das entidades

O segundo passo da etapa de treinamento consiste na rotulação do arquivo resultante do pré-processamento das sentenças selecionadas.

Seja P uma lista das palavras contidas na cadeia linear de tokens construída conforme a subseção 5.2.1.1, e R é o conjunto de rótulos necessários adicionado de um rótulo default O. Então, temos que o objetivo do passo de rotulação é atribuir a cada P_j um rótulo R_i.

Por exemplo, na figura 21, temos que P = {Germany, 's, representative, to, The, European, Union} e R = {Location, Organization, O}, onde O é o rótulo default.

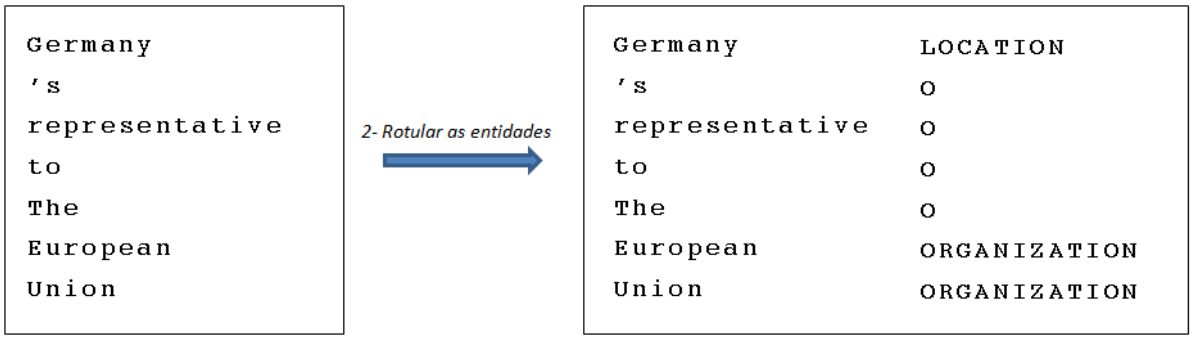

Figura 21: Exemplo de rotulação

A sintaxe utilizada pelo Stanford NER exige que a atribuição dos rótulos siga o seguinte padrão: <palavra> + <tab> + <nome do rótulo>.

O resultado desta subseção é um arquivo de treinamento contendo os exemplos de rotulação necessários para que o sistema Stanford NER possa generalizar.

5.2.1.3 Execução do treinamento

O terceiro passo da etapa de treinamento é executar o treinamento propriamente dito. Para a execução do treinamento é preciso configurar o arquivo de configuração e ter o arquivo resultante da subseção 5.2.1.2.

O arquivo de configuração armazena as informações necessárias para o treinamento, como, por exemplo, a origem do caminho do arquivo de treinamento e o destino do modelo gerado a partir da execução do treinamento. A figura 22 mostra um exemplo do arquivo de configuração. Os campos mais importantes são o "trainFile" e "serializeTo".

- **trainFile –** Caminho do arquivo contendo a rotulação das sentenças dispostas linearmente.
- **serializeTo –** Destino do modelo gerado após o treinamento.

Após realizar as configurações necessárias é preciso executar o seguinte comando: para a realização do treinamento "*java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop <nome do arquivo de configuração>*". O resultado é o modelo gerado especificado no campo serializeTo.

```
#location of the training file
trainFile = base.txt
#location where you would like to save (serialize to) your
#classifier; adding .gz at the end automatically gzips the file,
#making it faster and smaller
serializeTo = modelo.ser.gz

#structure of your training file; this tells the classifier
#that the word is in column 0 and the correct answer is in
#column 1
map = word=0,answer=1

#these are the features we'd like to train with
#some are discussed below, the rest can be
#understood by looking at NERFeatureFactory
useClassFeature=true
useWord=true
useNGrams=true
#no ngrams will be included that do not contain either the
#beginning or end of the word
noMidNGrams=true
useDisjunctive=true
maxNGramLeng=6
usePrev=true
useNext=true
useSequences=true
usePrevSequences=true
maxLeft=1
#the next 4 deal with word shape features
useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=chris2useLC
```

Figura 22: Arquivo de configuração para o treinamento

5.2.2 Etapa de reconhecimento das entidades nomeadas

Na etapa de treinamento o objetivo é gerar um modelo que atenda ao contexto proposto. Este modelo contém as informações aprendidas para o reconhecimento das entidades nomeadas. Então após realizar o treinamento o sistema está hábil para rotular uma nova sentença.

O uso do Standford NER para a rotulação de sentenças pode ser realizado de duas maneiras: a primeira é através de um executável disponibilizado por Stanford, a segunda é utilizar a biblioteca disponibilizada.

5.2.2.1 Binário

O reconhecimento de entidades nomeadas pode ser realizado via prompt utilizando o jar disponibilizado pelo Grupo de Stanford juntamente com o modelo contendo as informações para o reconhecimento das entidades.

Para realizar a anotação dos tokens os seguintes passos precisam ser realizados:

1. Executar o comando de pré-processamento abaixo:
    a. java -cp stanford-ner.jar edu.stanford.nlp.process.PTBTokenizer <*entrada*> > <saida>
    b. Onde <entrada> é o nome do arquivo texto que deseja ser rotulado e <saída> é o arquivo texto de entrada transformado em uma cadeia linear de tokens.
2. Executar o comando de reconhecimento das entidades nomeadas abaixo:
    a. java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -<*modelo*> -testFile <*resultado*>
    b. Onde <modelo> é o nome do modelo gerado na etapa de treinamento e <resultado> é o arquivo de saída com a cadeia linear de tokens devidamente rotulada. O resultado segue o mesmo padrão do apresentado na figura 19.

5.2.2.2 API

A biblioteca contida no pacote Stanford NER pode ser adicionada normalmente a qualquer projeto Java. Desta maneira o software Stanford NER pode ser facilmente acoplado a outros sistemas baseados na linguagem Java.

A classe principal no pacote jar do Stanford NER é chamada de CRFClassifier. Os métodos importantes para o reconhecimento das entidades são:

- **loadClassifier(String filename) –** Método responsável por carregar o modelo gerado a partir do treinamento.
- **apply(String texto) –** Método responsável por reconhecer as entidades nomeadas presentes na String texto. A saída desta função é um texto XML, onde as entidades nomeadas encontradas são rotuladas com o nome dos rótulos apropriados. A figura 23 mostra um exemplo de saída.

```
O ministro da <ORGANIZACAO>Ciência, Tecnologia e Inovação (MCTI)</ORGANIZACAO>, <PESSOA>Aloizio
Mercadante</PESSOA>, foi o primeiro entrevistado do ano do programa Bom Dia, Ministro e destacou a
políticas que vêm sendo implementadas pelo governo brasileiro na área, que incluem o <PROGRAMA>Programa
Ciência sem Fronteiras</PROGRAMA>. A entrevista foi  transmitida ao vivo pela TV NBR.
```

Figura 23: Exemplo de Saída da função apply

A figura 24 mostra um exemplo de código escrito em Java que utiliza a biblioteca Stanford NER para o reconhecimento de entidades nomeadas. O código abaixo lê um arquivo texto comum e imprime o arquivo com o texto devidamente rotulado.

```
1  public class MainClassification {
2
3       static private String model = "ner-model-pot-geral.ser.gz";
4       static private String file = "exemplo.txt";
5
6        public static void main(String[] args) throws Exception {
7
8            Properties p = new Properties();
9            CRFClassifier c = new CRFClassifier(p);
10           c.loadClassifier(model);
11
12           BufferedReader br = new BufferedReader(new FileReader(file));
13
14           while(br.ready())
15           {
16               System.out.println(c.apply(linha));
17           }
18
19           br.close();
20       }
21 }
22
```

Figura 24: Exemplo de código para reconhecimento de entidades nomeadas utilizando a biblioteca de Stanford

5.3 Execução dos Experimentos: Baseline

Essa subseção descreve os experimentos realizados para o reconhecimento de entidades nomeadas em textos da língua portuguesa em um contexto global, ou seja, o reconhecimento de entidades nomeadas independente do contexto.

A importância destes experimentos baseia-se na necessidade de delimitar um baseline para os experimentos descritos na próxima sessão. Pois, esse trabalho levanta a hipótese de que os resultados obtidos na literatura podem ser melhorados, já que o escopo deste trabalho é restrito às notícias de governo.

Então, nesse trabalho é desenvolvido um sistema de reconhecimento de entidades nomeadas sem a restrição de utilizar apenas as notícias de governo. O objetivo deste sistema é ser utilizado para comparar os resultados obtidos com os sistemas de reconhecimento de entidades restritos às notícias de governo.

A base de textos anotados utilizado nessa série de experimentos é a Coleção Dourada descrita na seção 2.4.1.1, já que esta coleção contempla diversas origens de textos da língua portuguesa.

Essa série de experimentos é constituída por quatro experimentos individuais. Cada experimento tem como objetivo obter informações sobre a validade das hipóteses levantadas neste trabalho.

5.3.1 Experimento 1

Nesse experimento três categorias foram avaliadas simultaneamente, as categorias são: Pessoa, Local e Organização. Os arquivos de treinamento foram obtidos a partir da Coleção Dourada. Foram gerados 2 arquivos, um arquivo de treinamento e outro arquivo de teste.

Nesse experimento as hipóteses 2 e 3 são utilizadas na elaboração dos arquivos de treinamento. Conforme a hipótese 2 cada frase presente nas notícias é utilizada como um exemplo individual sem interação com outras frases, ou seja, cada sequência a ser rotulada é definida por uma frase.

Experimentos 1 e 2 – Arquivo de Treino	
Número de tokens	72477
Número de sequências	2733
Número médio de tokens por sequência	26,5

Experimentos 1 e 2 - Arquivo de Teste	
Número de tokens	19874
Número de sequências	842
Número médio de tokens por sequência	23,6

Tabela 10: Dados sobre os arquivos de treino e teste dos experimentos 1 e 2

A tabela 10 apresenta os dados relativos aos arquivos de treino e teste utilizados neste experimento. Estes dados apresentam o tamanho dos arquivos utilizados nesse experimento e a dimensão das sequências de tokens a serem rotuladas conforme abordado na seção 5.2.1.2.

As tabelas 11 e 12 apresentam a distribuição de entidades nos arquivos de treino e de teste. O arquivo de treino possui 80% das entidades enquanto o arquivo de teste possui os 20% restantes das entidades.

Experimento 1 – Arquivo de Treino	
Tipo de Entidade	*Nº de Entidades*
Local	990
Organização	717
Pessoa	696
TOTAL	*2403*

Tabela 11: Distribuição das entidades por categoria no arquivo de treino do experimento 1

Experimento 1 – Arquivo de Teste	
Tipo de Entidade	*Nº de Entidades*
Local	189
Organização	176
Pessoa	203
TOTAL	*568*

Tabela 12: Distribuição das entidades por categoria no arquivo de teste do experimento 1

O gráfico 2 apresenta os resultados obtidos a partir do treinamento realizado neste experimento. Os valores de cada coluna se referem aos resultados obtidos ao utilizar o sistema na classificação das entidades do arquivo de teste.

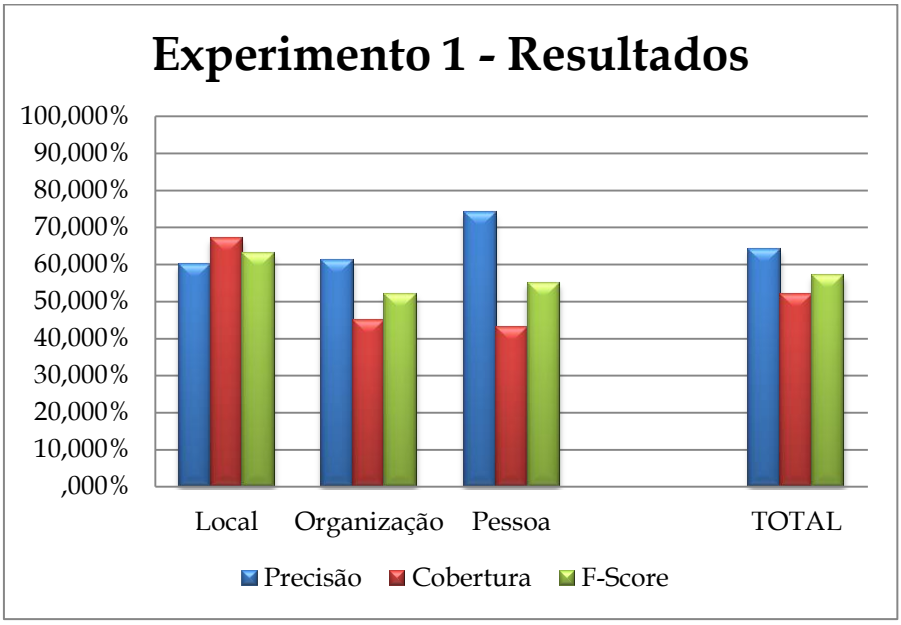

Gráfico 2: Resultados do experimento 1

Como resultado do experimento 1 o sistema resultante apresentou um F-score geral de 57%, sendo 63% para a categoria Local, 52% para a categoria Organização e 54% para a

categoria Pessoa. Os resultados obtidos demonstram valores baixos de cobertura para as categorias de Organização e Pessoa, mas apresenta uma razoável F-Score para a categoria Local.

5.3.2 Experimento 2

Nesse experimento onze categorias foram avaliadas simultaneamente; as categorias são: Abstração, Acontecimento, Coisa, Local, Objeto, Obra, Outro, Organização, Pessoa, Tempo e Valor. Os arquivos de treinamento foram obtidos a partir da Coleção Dourada.

Nesse experimento as hipóteses 2 e 3 são utilizadas na elaboração dos arquivos de treinamento. Conforme a hipótese 2 cada frase presente nas notícias é utilizada como um exemplo individual sem interação com outras frases, ou seja, cada sequência a ser rotulada é definida por uma frase. Os dados relativos aos arquivos de treino e teste utilizados neste experimento podem ser visualizados na tabela 10.

As tabelas 13 e 14 apresentam a distribuição de entidades nos arquivos de treino e de teste. O arquivo de treino possui 80% das entidades enquanto o arquivo de teste possui os 20% restantes das entidades.

Experimento 2 - Treino	
Tipo de Entidade	*Nº de tokens*
Abstração	277
Acontecimento	90
Coisa	100
Local	990
Objeto	1
Obra	143
Outro	20
Organização	717
Pessoa	696
Tempo	320
Valor	341
TOTAL	*3695*

Tabela 13: Distribuição das entidades por categoria no arquivo de treino do experimento 2

Experimento 2 - Teste	
Tipo de Entidade	*Nº de tokens*
Abstração	88
Acontecimento	19
Coisa	28
Local	189
Objeto	0
Obra	39
Outro	18
Organização	177
Pessoa	204
Tempo	76
Valor	83
TOTAL	*921*

Tabela 14: Distribuição das entidades nomeadas por categoria no arquivo de teste do experimento 2

O gráfico 3 apresenta os resultados obtidos a partir do treinamento realizado neste experimento. Os valores de cada coluna se referem aos resultados obtidos ao utilizar o sistema na classificação das entidades do arquivo de teste.

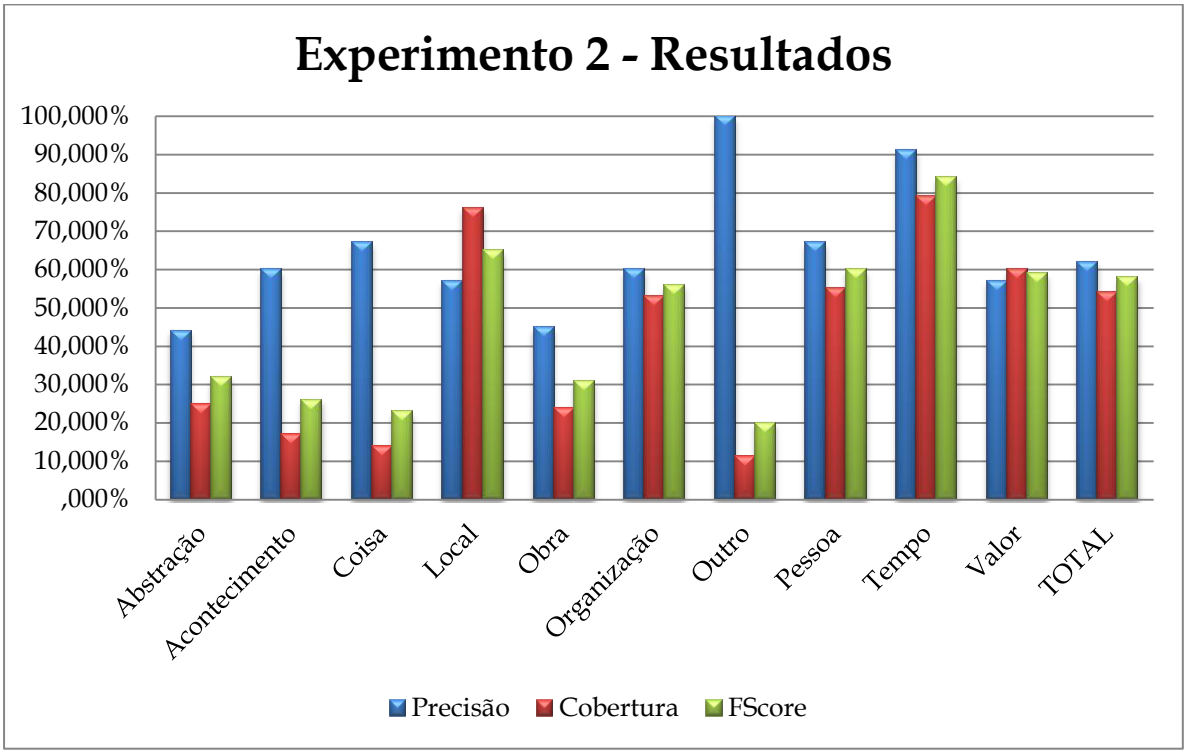

Gráfico 3: Resultados do experimento 2

Como resultado do experimento 2 o sistema resultante apresentou um F-Score geral de 57%, sendo este o mesmo resultado obtido no experimento 1. As categorias Abstração, Acontecimento, Coisa, Obra e Outro apresentaram valores baixos assim como os obtido no Harem. Os demais apresentaram F-Scores razoáveis se comparados aos F-Score publicados no Harem. Além disso, considerando apenas as categorias Pessoa, Local e Organização o F-Score obtido foi de 61%.

Um importante ponto a ser observado é que houve uma pequena melhora no desempenho das categorias Pessoa, Local e Organização em relação aos resultados obtidos no experimento 1. Essa observação está de acordo com a hipótese 3, já que ao incluir novas categorias a taxa de acerto do sistema foi ampliada.

5.3.3 Experimento 3

Este experimento é semelhante ao experimento 1. Entretanto, nesse experimento a hipótese 2 não é utilizada na elaboração dos arquivos de treino e teste. Então, os arquivos de treino e de teste têm como uma sequência a notícia por completo, ou seja, a classificação de uma entidade em uma dada notícia é realizada por meio da avaliação completa desta.

A tabela 15 apresenta os dados relativos aos arquivos de treino e teste utilizados neste experimento. Estes dados apresentam o tamanho dos arquivos utilizados nesse experimento e a dimensão das sequências de tokens a serem rotuladas conforme abordado na seção 5.2.1.1.

Experimentos 3 e 4 - Treino	
Número de tokens	72477
Número de sequências	103
Número médio de tokens por sequência	703,7

Experimentos 3 e 4 - Teste	
Número de tokens	19874
Número de sequências	26
Número médio de tokens por sequência	764,4

Tabela 15: Dados sobre os arquivos de treino e teste dos experimentos 3 e 4

As tabelas 16 e 17 apresentam a distribuição de entidades nos arquivos de treino e de teste. O arquivo de treino possui 80% das entidades enquanto o arquivo de teste possui os 20% restantes das entidades.

Experimento 3 - Treino	
Tipo de Entidade	*Nº de Entidades*
Local	983
Organização	713
Pessoa	694
TOTAL	*2390*

Tabela 16: Distribuição das entidades por categoria no arquivo de treino do experimento 3

Experimento 3 - Teste	
Tipo de Entidade	*Nº de Entidades*
Local	189
Organização	175
Pessoa	196
TOTAL	*562*

Tabela 17: Distribuição das entidades por categoria no arquivo de teste do experimento 3

O gráfico 4 apresenta os resultados obtidos a partir do treinamento realizado neste experimento. Os valores de cada coluna se referem aos resultados obtidos ao utilizar o sistema na classificação das entidades do arquivo de teste.

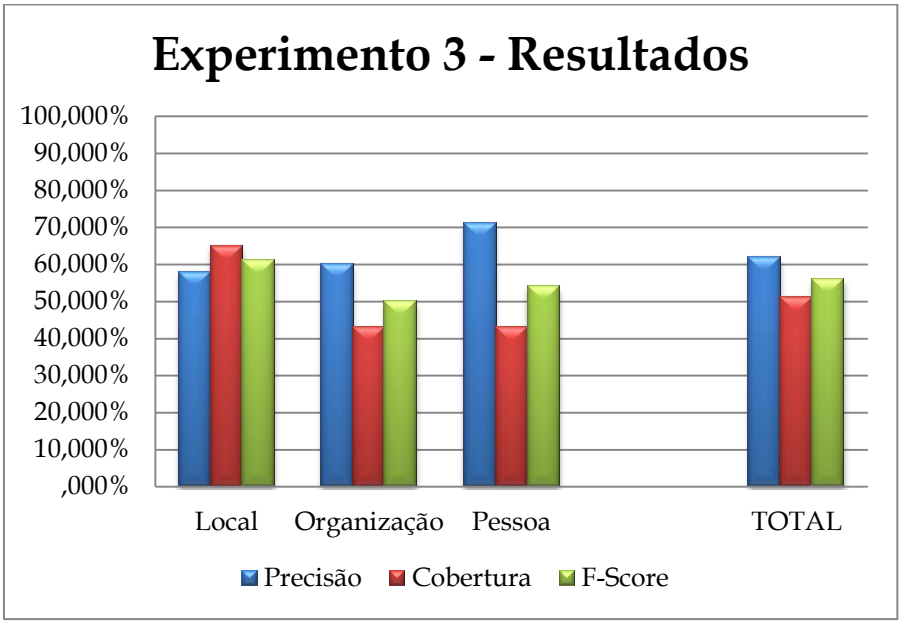

Gráfico 4: Resultados do experimento 3

Como resultado do experimento 3 o sistema resultante apresentou um F-Score geral de 55%. Este é levemente inferior ao obtido no experimento 1. Além disso, todas as categorias apresentaram um desempenho levemente inferior aos obtidos no experimento 1.

5.3.4 Experimento 4

Este experimento é semelhante ao experimento 2. Entretanto, nesse experimento a adaptação da hipótese 2 não é utilizada na elaboração dos arquivos de treino e teste. Os dados relativos aos arquivos de treino e teste utilizados neste experimento podem ser visualizados na tabela 15.

As tabelas 18 e 19 apresentam a distribuição de entidades nos arquivos de treino e de teste. O arquivo de treino possui 80% das entidades enquanto o arquivo de teste possui os 20% restantes das entidades.

Experimento 4 - Treino	
Tipo de Entidade	*Nº de tokens*
Abstração	276
Acontecimento	89
Coisa	100
Local	983
Objeto	1
Obra	143
Outro	20
Organização	713
Pessoa	694
Tempo	319
Valor	340
TOTAL	*3678*

Tabela 18: Distribuição das entidades por categoria no arquivo de treino do experimento 4

Experimento 4 - Teste	
Tipo de Entidade	*Nº de tokens*
Abstração	88
Acontecimento	19
Coisa	28
Local	189
Objeto	0
Obra	38
Outro	18
Organização	176
Pessoa	197
Tempo	76
Valor	83
TOTAL	*912*

Tabela 19: Distribuição das entidades por categoria no arquivo de teste do experimento 4

O gráfico 5 apresenta os resultados obtidos a partir do treinamento realizado neste experimento. Os valores de cada coluna se referem aos resultados obtidos ao utilizar o sistema na classificação das entidades do arquivo de teste.

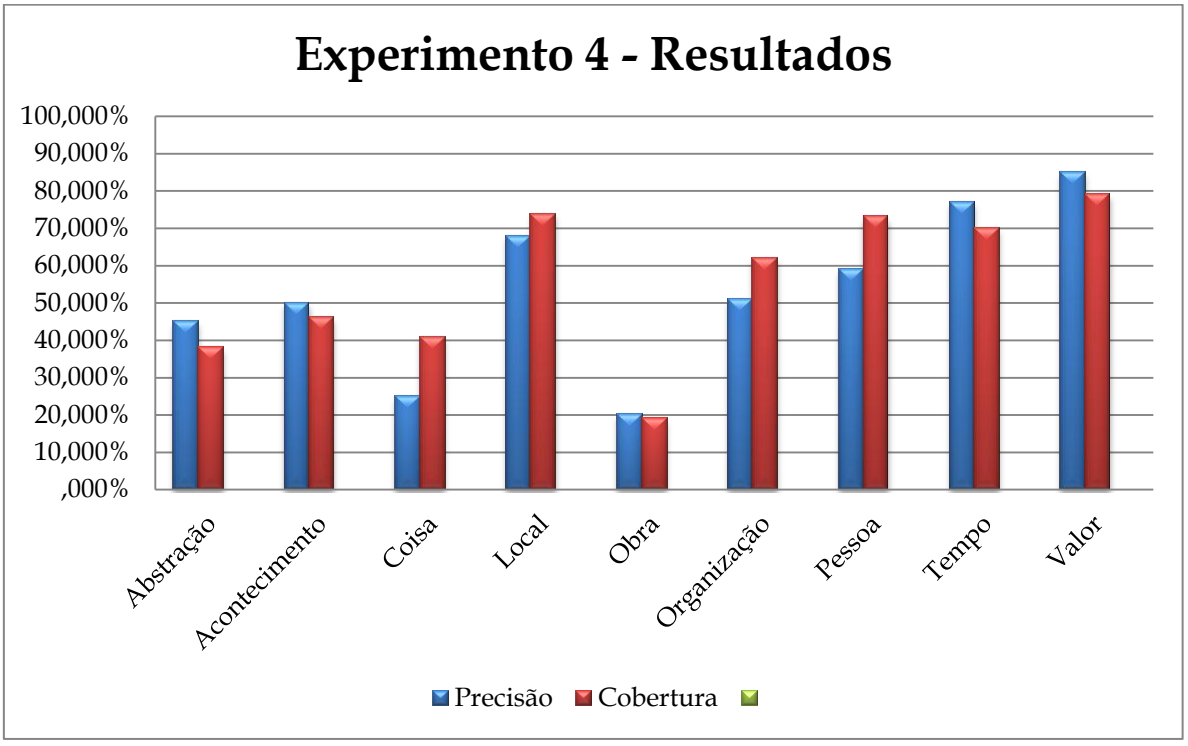

Gráfico 5: Resultados do experimento 4

Como resultado do experimento 4 o sistema resultante apresentou um F-Score geral de 58%, apresentando uma leve melhora em relação ao experimentos anteriores. Além disso, considerando apenas as categorias Pessoa, Local e Organização o F-Score obtido foi de 60%.

5.3.5 Conclusão

A série de experimentos dessa subseção visa fornecer dados suficientes para desenvolver um sistema que seja utilizado como baseline dos experimentos seguintes. Os quatro experimentos realizados retratam quatro características diferentes no método empregado para treino. Nos experimentos 1 e 2 a base de textos Coleção Dourada foi dividida em frases para o treinamento de três e onze categorias respectivamente. Nos experimentos 3 e 4 a base de textos Coleção Dourada foi dividida em textos completos para o treinamento de três e onze categorias respectivamente.

A análise realizada sobre esses experimentos tem como ênfase a avaliação das categorias Pessoa, Local e Organização. O F-Score geral obtido considerando essas três categorias foram: 57% para o experimento 1; 61% para o experimento 2; 56% para o experimento 3 e 60% para o experimento 4. A maior variação entre os resultados foi a obtida comparando-se o experimento 2 com o experimento 3, em que a variação de desempenho foi de aproximadamente 9%.

Os experimentos que utilizavam as frases como elementos de treino e os experimentos que utilizavam os textos completos não tiveram variações significativas em seus resultados. Essa observação está de acordo com a hipótese 2, uma vez que utilizar apenas a frase para classificar suas entidades demonstrou ter o mesmo resultado do que utilizar todo o texto em que a frase está inserida.

O modelo resultante do experimento 2 foi o modelo adotado nesse trabalho para a elaboração do baseline dos experimentos seguintes. Este experimento foi escolhido por apresentar o melhor F-Score geral considerando as categorias Pessoa, Local e Organização.

5.4 Execução dos Experimentos: Primeira Base Anotada

Esta subseção descreve os experimentos realizados no reconhecimento de entidades nomeadas em textos da língua portuguesa no contexto de notícias de governo, ou seja, o

conjunto de textos relevantes se restringe às notícias relacionadas ao governo brasileiro. O sistema resultante da seção 5.3 é utilizado como baseline dos experimentos dessa seção.

Dado a dificuldade de anotar manualmente as entidades de um grande número de notícias, esta série de experimentos é dividida em três etapas. A primeira etapa consiste em utilizar a primeira versão da Coleção de Notícias de Governo – UFRJ no treinamento dos sistemas de reconhecimento de entidades nomeadas.

A segunda etapa consiste em utilizar os resultados obtidos na etapa anterior para classificar uma coleção de notícias não anotadas. Então, a classificação realizada automaticamente é revisada manualmente e corrigida. As novas notícias anotadas são adicionadas à primeira versão da Coleção de Notícias de Governo, dando origem à versão completa da Coleção de Notícias de Governo.

A terceira etapa consiste em utilizar a Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ, revisada na segunda etapa, no treinamento dos sistemas de reconhecimento de entidades nomeadas. O resultado esperado é que com a inclusão de novos exemplos os desempenhos obtidos sejam aperfeiçoados.

A etapa 1 e a etapa 3 são constituídas por dois experimentos cada, onde os experimentos 5 e 6 pertencem à etapa 1 e os experimentos 7 e 8 pertencem à etapa 3. Semelhantemente à série de experimentos realizada na seção 5.3, os experimentos 5 e 7 utilizam as frases presentes nas notícias como exemplos unitários, já os experimentos 6 e 8 utilizam as notícias como exemplos unitários.

Em todos os quatro experimentos as categorias adotadas para classificação são: Pessoa, Local, Organização, Evento e Programa. As categorias Pessoa, Local e Organização possuem como baseline o sistema escolhido na subseção 5.3. Para cada conjunto de testes utilizados nesses experimentos o seu baseline foi calculado e utilizado para comparação. As categorias Evento e Programa por serem categorias novas não possuem um baseline para ser utilizado como comparação.

5.4.1 Experimento 5

Este experimento tem como objetivo estudar o treinamento de um sistema de reconhecimento de entidades considerando o uso do experimento 2. Este experimento tem como restrição o reconhecimento de entidades nomeadas em textos da língua portuguesa cuja origem seja notícias relacionadas ao governo brasileiro.

A base de textos utilizada nesse experimento é a primeira versão da Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ. A partir da base de textos anotada e da hipótese 2, os arquivos de treino e de testes (tabela 20) foram gerados.

Experimento 5 – Treino	
Número de tokens	37236
Número de sequências	1191
Número médio de tokens por sequência	31,3

Experimento 5 - Teste	
Número de tokens	29060
Número de sequências	918
Número médio de tokens por sequência	31,6

Tabela 20: Dados sobre os arquivos de treino e teste utilizados no experimento 5

As tabelas 21 e 22 apresentam a distribuição de entidades nos arquivos de treino e de teste deste experimento.

Experimento 5 e 6- Treino	
Tipo de Entidade	*Nº de Entidades*
Evento	162
Local	377
Organização	1139
Pessoa	423
Programa	126
TOTAL	*2227*

Tabela 21: Distribuição das entidades presentes no arquivo de treino dos experimentos 5 e 6

Experimento 5 e 6 - Teste	
Tipo de Entidade	*Nº de Entidades*
Evento	83
Local	262
Organização	892
Pessoa	264
Programa	146
TOTAL	*1647*

Tabela 22: Distribuição das entidades presentes no arquivo de teste dos experimentos 5 e 6

Após realizar o treinamento com os arquivos de treino, o sistema de reconhecimento resultante classificou o arquivo de teste. Como resultado o sistema apresentou a precisão de

60%, 77%, 71%, 85% e 75% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e a cobertura de 43%, 59%, 77%, 90% e 45% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e o F-Score de 50%, 67%, 74%, 87% e 56% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa.

A partir dos resultados individuais foram calculados valores gerais considerando as três categorias principais (Pessoa, Local e Organização) e as cinco categorias do experimento. Considerando apenas as três principais categorias o sistema apresentou os valores de 75%, 76% e 75% para precisão, cobertura e F-Score respectivamente. Considerando todas as categorias utilizadas nesse experimento o sistema apresentou os valores de 74%, 72% e 73% para precisão, cobertura e F-Score.

O sistema escolhido, conforme descrito na seção 5.3, foi utilizado para classificar as entidades do arquivo de treino utilizado nesse experimento. Dessa maneira foi estabelecido o baseline deste experimento. Os F-Scores calculados foram de 48%, 51% e 72% para as categorias Local, Organização e Pessoa respectivamente. O F-Score geral foi 54%.

Ao comparar os resultados obtidos nesse experimento com o baseline é constatado um ganho de desempenho nas três categorias estudadas. O ganho de desempenho verificado é de aproximadamente 39% sobre o resultado obtido como baseline. Esse resultado está de acordo com a hipótese 1.

O gráfico 6 sintetiza os resultados obtidos nesse experimento. Em roxo estão as taxas de acertos obtidas com o sistema treinado com a Coleção Dourada (baseline deste trabalho) e em verde as taxas de acerto obtidas com o sistema treinado nesse experimento.

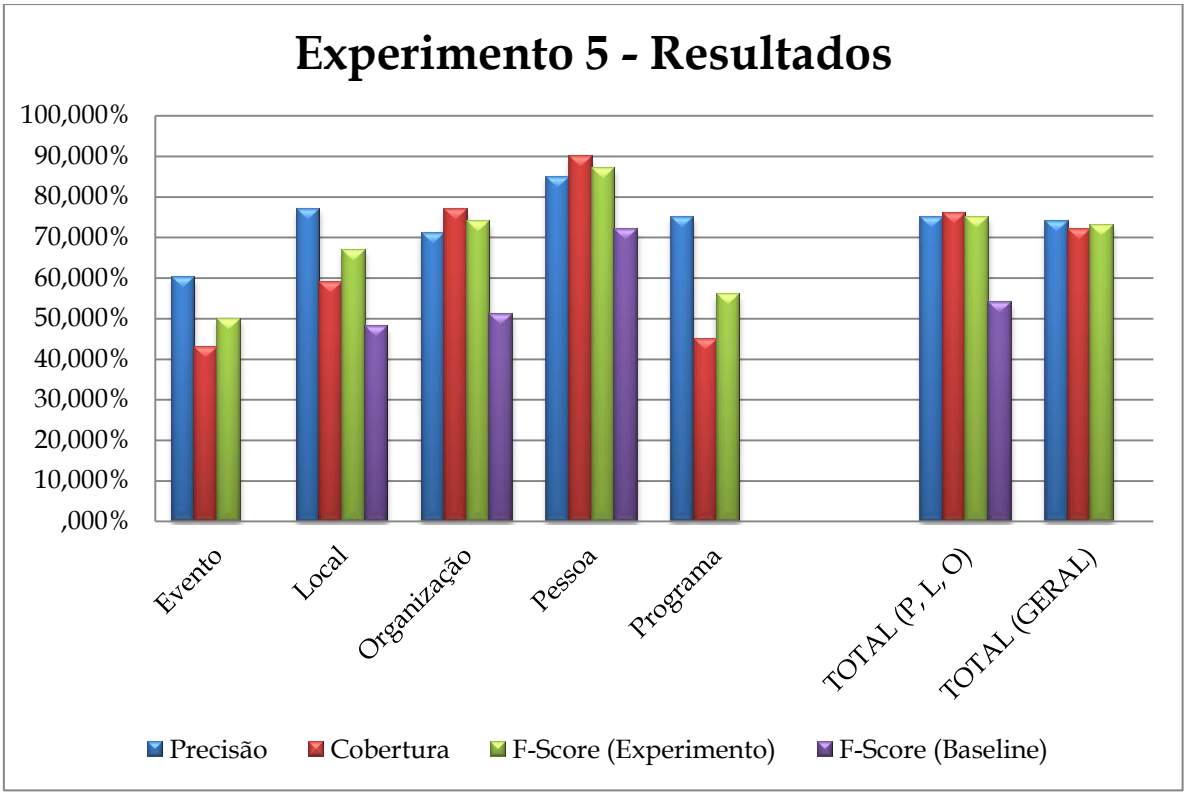

Gráfico 6: Resultados obtidos no experimento 5

5.4.2 Experimento 6

Este experimento estuda o treinamento de um sistema de reconhecimento de entidades com as mesmas restrições impostas ao experimento 5. Entretanto, nesse experimento a hipótese 2 não é utilizada, a fim de possibilitar a comparação dos resultados obtidos e com isso a avaliação da hipótese levantada. Dessa maneira os exemplos unitários utilizados nesse experimento são as notícias e não as frases.

A base de textos utilizada nesse experimento é a primeira versão da Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ. A partir da base de textos anotada, os arquivos de treino e de testes (tabela 23) foram gerados.

Experimento 6 – Treino	
Número de tokens	37236
Número de sequências	72
Número médio de tokens por sequência	517,17

Experimento 6 - Teste	
Número de tokens	29060
Número de sequências	48
Número médio de tokens por sequência	605,42

Tabela 23: Dados sobre os arquivos de treino e teste utilizados no experimento 6

As tabelas 21 e 22 apresentam a distribuição de entidades nos arquivos de treino e de teste deste experimento. A distribuição das entidades é a mesma apresentada no experimento 5.

Após realizar o treinamento com os arquivos de treino, o sistema de reconhecimento resultante classificou o arquivo de teste. Como resultado o sistema apresentou a precisão de 60%, 74%, 71%, 86% e 76% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e a cobertura de 43%, 60%, 77%, 90% e 45% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e o F-Score de 50%, 67%, 74%, 88% e 57% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa.

A partir dos resultados individuais foram calculados valores gerais considerando as três categorias principais (Pessoa, Local e Organização) e as cinco categorias do experimento. Considerando apenas as três principais categorias o sistema apresentou os valores de 75%, 76% e 75% para precisão, cobertura e F-Score respectivamente. Considerando todas as categorias utilizadas nesse experimento o sistema apresentou os valores de 74%, 72% e 73% para precisão, cobertura e F-Score.

O sistema escolhido, conforme descrito na seção 5.3, foi utilizado para classificar as entidades do arquivo de treino utilizado nesse experimento. Dessa maneira foi estabelecido o baseline deste experimento. Os F-Scores calculados foram 48%, 51% e 72% para as categorias Local, Organização e Pessoa respectivamente. O F-Score geral foi 54%.

Ao comparar os resultados obtidos nesse experimento com o baseline é constatado um ganho de desempenho nas três categorias estudadas assim como as obtidas no experimento 5. O ganho de desempenho verificado é de aproximadamente 39% sobre o resultado obtido como baseline. Esse resultado está de acordo com a hipótese 1.

Os resultados obtidos no experimento 6 são aproximadamente os mesmos obtidos no experimento 5 apresentando pequenas variações em torno de +-0,01. Essa pequena variação está de acordo com a hipótese 2, já que o uso de toda a notícia para a classificação das entidades não apresentou uma vantagem significativa em relação à abordagem que utiliza apena a frase como fonte de informação.

O gráfico 7 sintetiza os resultados obtidos nesse experimento. Em roxo estão as taxas de acertos obtidas com o sistema treinado com a Coleção Dourada (baseline deste trabalho) e em verde as taxas de acerto obtidas com o sistema treinado nesse experimento.

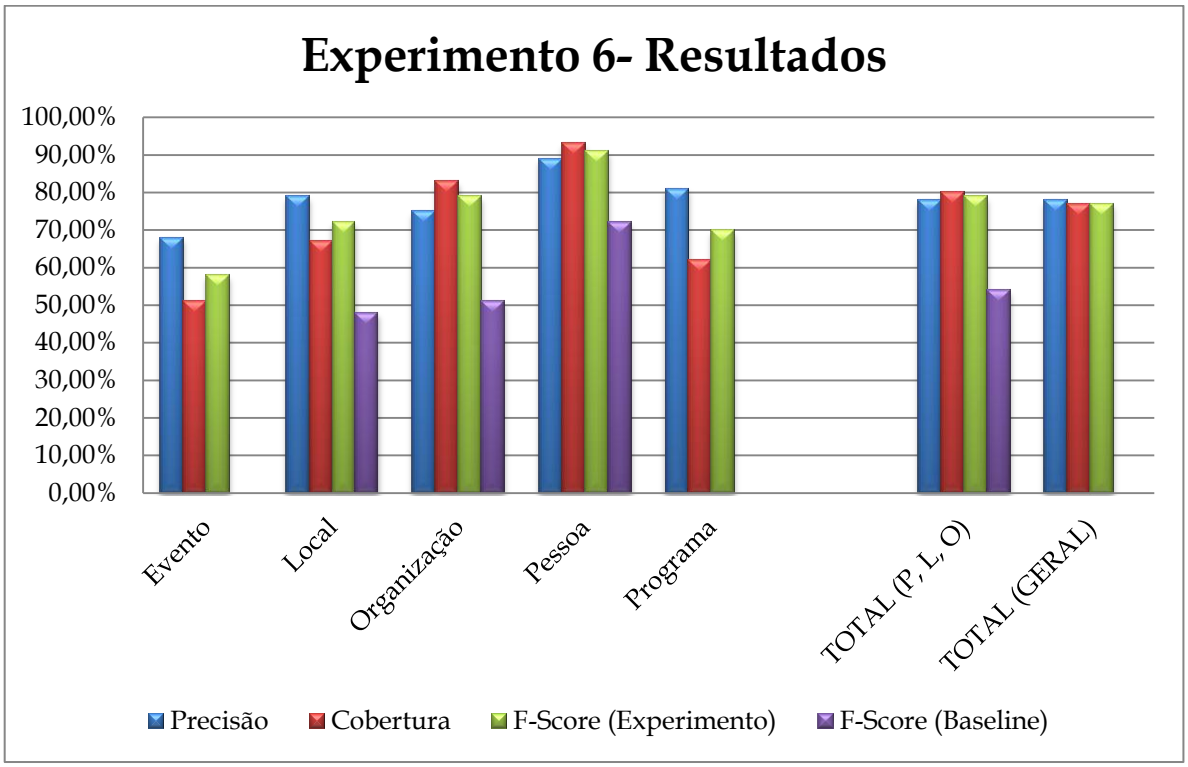

Gráfico 7: Resultados obtidos no experimento 6

5.4.3 Experimento 7

Este experimento tem as mesmas características do experimento 5. A diferença reside na inclusão de novas notícias na Coleção de Notícias de Governo – UFRJ, ou seja, esse experimento é o mesmo realizado no experimento 5 mas com uma carga maior de exemplos.

A base de textos utilizada nesse experimento é a versão completa da Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ que é detalhada na subseção 3.2. A partir da base de textos anotada e da hipótese 2, o arquivo de treino é gerado (tabela 24). O arquivo de teste utilizado é exatamente o mesmo utilizado no experimento 5.

Experimento 7 – Treino	
Número de tokens	100242
Número de sequências	3192
Número médio de tokens por sequência	31,4

Experimento 7 - Teste	
Número de tokens	29060
Número de sequências	918
Número médio de tokens por sequência	31,6

Tabela 24: Dados sobre os arquivos de treino e teste utilizados no experimento 5

As tabelas 25 e 26 apresentam a distribuição de entidades nos arquivos de treino e de teste deste experimento.

Experimento 7 e 8 - Treino	
Tipo de Entidade	*Nº de Entidades*
Evento	306
Local	1047
Organização	3035
Pessoa	1089
Programa	474
TOTAL	*5951*

Tabela 25: Distribuição das entidades presentes no arquivo de treino dos experimentos 7 e 8

Experimento 7 e 8 - Teste	
Tipo de Entidade	*Nº de Entidades*
Evento	83
Local	262
Organização	892
Pessoa	264
Programa	146
TOTAL	*1647*

Tabela 26: Distribuição das entidades presentes no arquivo de teste dos experimentos 7 e 8

Após realizar o treinamento com os arquivos de treino, o sistema de reconhecimento resultante classificou o arquivo de teste. Como resultado o sistema apresentou a precisão de 68%, 79%, 75%, 89% e 81% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e a cobertura de 51%, 67%, 83%, 93% e 62% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e o F-Score de 58%, 72%, 79%,

91% e 70% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa.

A partir dos resultados individuais foram calculados valores gerais considerando as três categorias principais (Pessoa, Local e Organização) e as cinco categorias do experimento. Considerando apenas as três principais categorias o sistema apresentou os valores de 78%, 80% e 79% para precisão, cobertura e F-Score respectivamente. Considerando todas as categorias utilizadas nesse experimento o sistema apresentou os valores de 78%, 77% e 77% para precisão, cobertura e F-Score.

O sistema escolhido, conforme descrito na seção 5.3, foi utilizado para classificar as entidades do arquivo de treino utilizado nesse experimento. Dessa maneira foi estabelecido o baseline deste experimento. Os F-Score calculados foram 48%, 51% e 72% para as categorias Local, Organização e Pessoa respectivamente. O F-Score geral foi 54%.

Ao comparar os resultados obtidos nesse experimento com o baseline é constatado um ganho de desempenho nas três categorias estudadas. O ganho de desempenho verificado é de aproximadamente 46% sobre o resultado obtido como baseline. Esse resultado está de acordo com a hipótese 1.

Após incluir novos exemplos para treinamento foi constatado um ganho de desempenho em relação ao experimento 5. O ganho de desempenho obtido nesse experimento é aproximadamente de 5%. A categoria Programa apresentou o maior ganho de desempenho que foi aproximadamente de 25%.

O gráfico 8 sintetiza os resultados obtidos nesse experimento. Em roxo estão as taxas de acertos obtidas com o sistema treinado com a Coleção Dourada (baseline deste trabalho) e em verde as taxas de acerto obtidas com o sistema treinado nesse experimento.

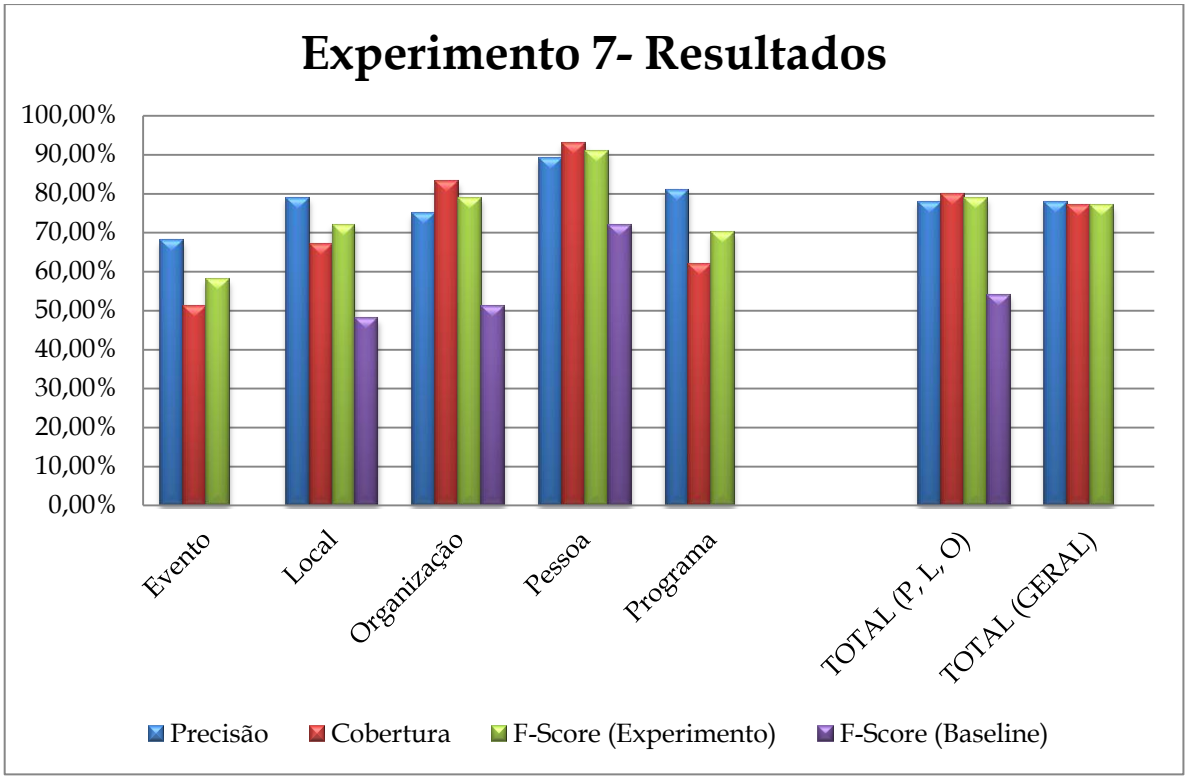

Gráfico 8: Resultados obtidos no experimento 7

5.4.4 Experimento 8

Este experimento tem as mesmas características do experimento 6. A diferença reside na inclusão de novas notícias na Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ, ou seja, esse experimento é o mesmo realizado no experimento 6, mas com uma carga maior de exemplos.

A base de textos utilizada nesse experimento é a Primeira Base Anotada da Coleção de Notícias de Governo – UFRJ que é detalhada na subseção 3.2. A partir da base de textos anotada, o arquivo de treino é gerado (tabela 27). O arquivo de teste utilizado é exatamente o mesmo utilizado no experimento 6.

Experimento 8 – Treino	
Número de tokens	100242
Número de sequências	192
Número médio de tokens por sequência	522,09

Experimento 8 - Teste	
Número de tokens	29060
Número de sequências	48
Número médio de tokens por sequência	605,42

Tabela 27: Dados sobre os arquivos de treino e teste utilizados no experimento 8

As tabelas 25 e 26 apresentam a distribuição de entidades nos arquivos de treino e de teste deste experimento. A distribuição das entidades é a mesma apresentada no experimento 7.

Após realizar o treinamento com os arquivos de treino, o sistema de reconhecimento resultante classificou o arquivo de teste. Como resultado o sistema apresentou a precisão de 66%, 80%, 76%, 89% e 83% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e a cobertura de 47%, 68%, 82%, 94% e 64% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa; e o F-Score de 55%, 74%, 79%, 92% e 72% respectivamente para as categorias Evento, Local, Organização, Pessoa e Programa.

A partir dos resultados individuais foram calculados valores gerais considerando as três categorias principais (Pessoa, Local e Organização) e as cinco categorias do experimento. Considerando apenas as três principais categorias o sistema apresentou os valores 79%, 82% e 80% para precisão, cobertura e F-Score respectivamente. Considerando todas as categorias utilizadas nesse experimento o sistema apresentou os valores de 79%, 79% e 79% para precisão, cobertura e F-Score.

O sistema escolhido, conforme descrito na seção 5.3, foi utilizado para classificar as entidades do arquivo de treino utilizado nesse experimento. Dessa maneira foi estabelecido o baseline deste experimento. Os F-Score calculados foram 48%, 51% e 72% para as categorias Local, Organização e Pessoa respectivamente. O F-Score geral foi 54%.

Ao comparar os resultados obtidos nesse experimento com o baseline é constatado um ganho de desempenho nas três categorias estudadas. O ganho de desempenho verificado é de aproximadamente 48% sobre o resultado obtido com o baseline. Esse resultado está de acordo com a hipótese 1.

Os resultados obtidos no experimento 7 são aproximadamente os mesmos obtidos nesse experimento, apontando um pequeno acréscimo de desempenho. Essa pequena variação está de acordo com a hipótese 2, já que o uso de todo a notícia para a classificação das entidades não apresentou uma vantagem significativa em relação à abordagem que utiliza apena a frase como fonte de informação.

O gráfico 9 sintetiza os resultados obtidos nesse experimento. Em negrito estão os resultados que são comparados com o baseline. Em roxo estão as taxas de acertos obtidas com o sistema treinado com a Coleção Dourada (baseline deste trabalho) e em verde as taxas de acerto obtidas com o sistema treinado nesse experimento.

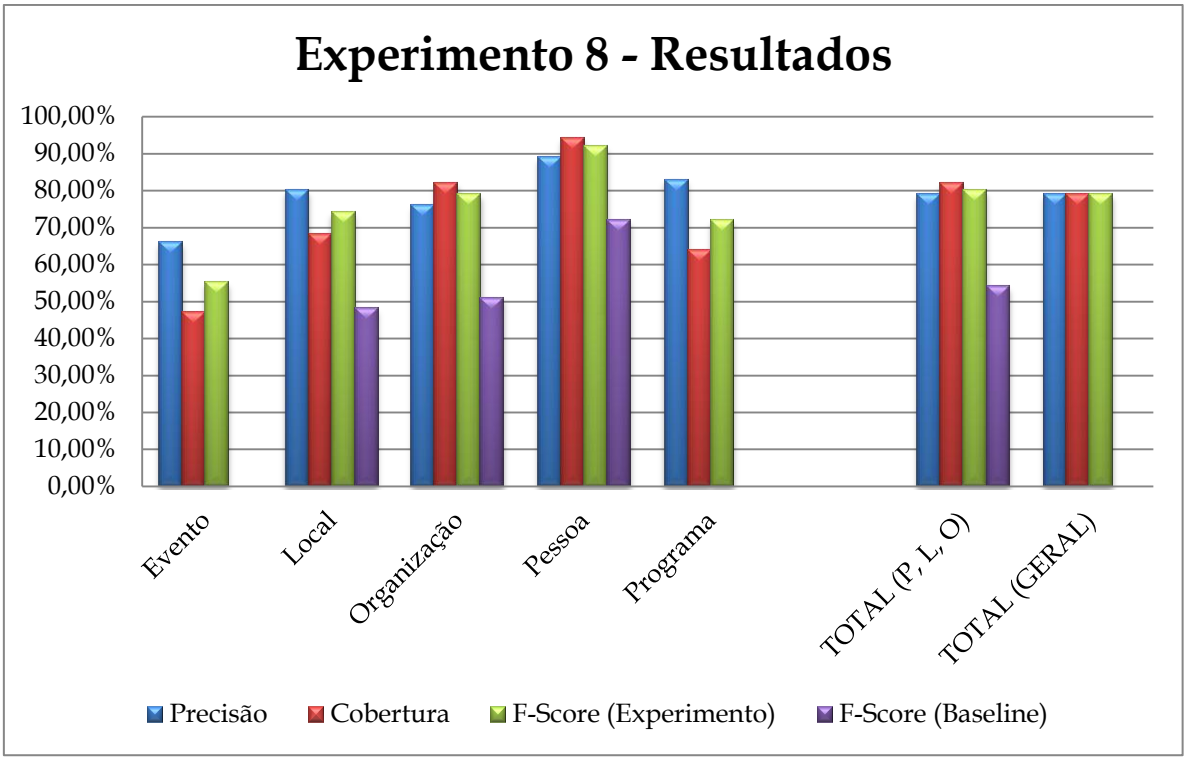

Gráfico 9: Resultados obtidos no experimento 8

5.5 Execução do Experimento: Segunda Base Anotada

Esta subseção descreve os experimentos realizados no reconhecimento de entidades nomeadas em textos da língua portuguesa no contexto de notícias de governo, ou seja, o conjunto de textos relevantes se restringe às notícias relacionadas ao governo brasileiro.

O experimento descrito nessa subseção utilizou os exemplos fornecidos pela Segunda Base Anotada da Coleção de Notícias de Governo – UFRJ no processo de treinamento do sistema de reconhecimento de entidades proposto.

Nesse experimento os tipos de entidades adotadas para classificação foram: Pessoa, Local, Organização, Evento, Programa, Cargo, Sigla e Área Administrativa.

5.5.1 Experimento 9

A base de textos utilizada nesse experimento é a Segunda Base Anotada da Coleção de Notícias de Governo – UFRJ detalhada na subseção 3.2. A tabela 28 descreve os arquivos de teste e treinamento.

Experimento 9 – Teste	
Número de tokens	26902
Número de sequências	48
Número médio de tokens por sequência	560,4

Experimento 9 – Treino	
Número de tokens	37817
Número de sequências	77
Número médio de tokens por sequência	491,1

Tabela 28: Dados sobre os arquivos de treino e teste utilizados no experimento 9

As tabelas 29 e 30 apresentam a distribuição de entidades nos arquivos de treino e de teste deste experimento.

Experimento 9 – Treino	
Tipo de Entidade	*Nº de tokens*
Área Administrativa	137
Cargo	263
Evento	101
Local	405
Organização	740
Pessoa	411
Programa	141
Sigla	158
TOTAL	*2356*

Tabela 29: Distribuição das entidades presentes no arquivo de treino do experimento 9

Experimento 9 – Teste	
Tipo de Entidade	*Nº de tokens*
Área Administrativa	87
Cargo	160
Evento	78
Local	260
Organização	475
Pessoa	268
Programa	140
Sigla	105
TOTAL	*1573*

Tabela 30: Distribuição das entidades presentes no arquivo de teste do experimento 9

Após realizar o treinamento com os arquivos de treino, foi realizado o teste de validação. Os resultados obtidos no processo de testes são apresentados no gráfico 10. Em roxo estão as taxas de acertos obtidas com o sistema treinado com a Coleção Dourada (baseline deste trabalho) e em verde as taxas de acerto obtidas com o sistema treinado nesse experimento.

A partir dos resultados individuais foram calculados valores gerais considerando as três categorias principais (Pessoa, Local e Organização) e as oito categorias do experimento. Considerando apenas as três principais categorias o sistema apresentou os valores de 75%, 76% e 75% para precisão, cobertura e F-Score respectivamente. Considerando todas as categorias utilizadas nesse experimento o sistema apresentou os valores de 74%, 72% e 73% para precisão, cobertura e F-Score.

As categorias Pessoa e Cargo apresentam os melhores resultados enquanto as categorias Evento e Sigla apresentaram os piores resultados. A categoria Evento obteve uma boa precisão, mas apresentou uma baixa taxa de cobertura.

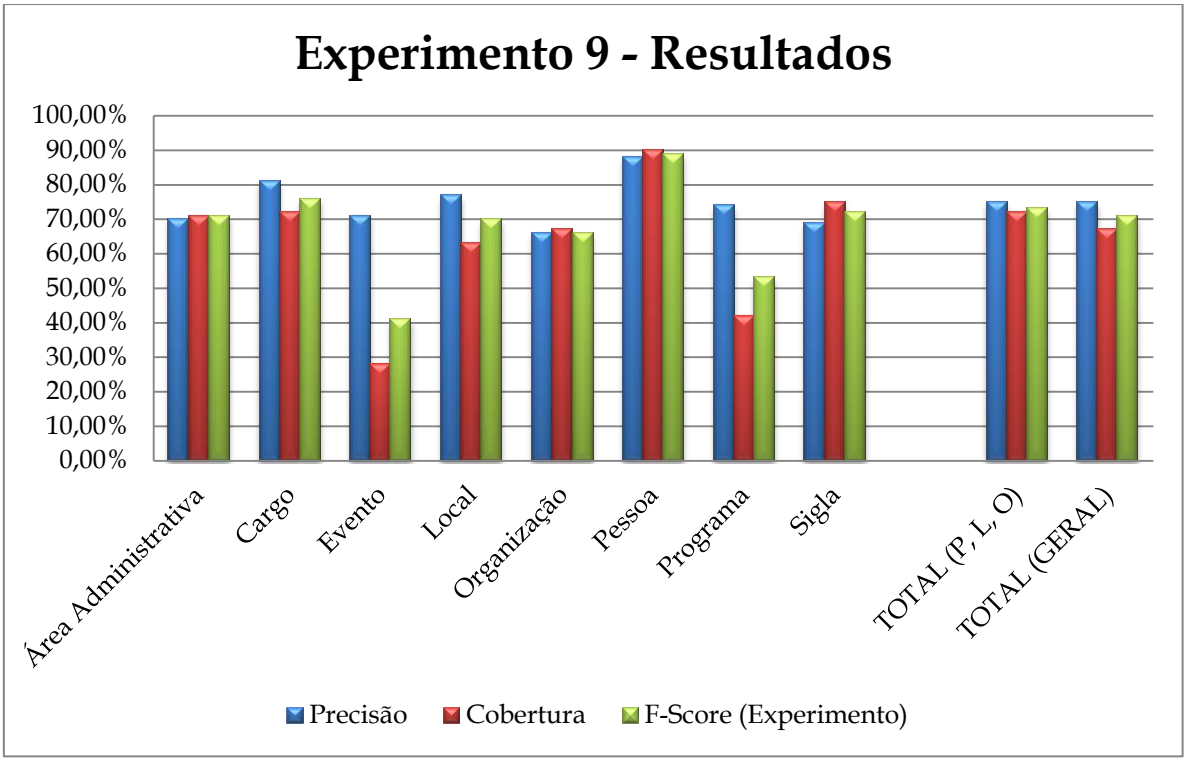

Gráfico 10: Resultados obtidos no experimento 9

5.6 Conclusão

Nesse trabalho foram conduzidos nove experimentos, sendo que os quatro primeiros experimentos tiveram como objetivo analisar o reconhecimento de entidades nomeadas na língua portuguesa usando uma base de exemplos já estudada na literatura, que é a Coleção Dourada. O resultado destes quatro experimentos foi a concepção de um sistema de reconhecimento de entidades nomeadas sem restringir os textos a apenas notícias de governo.

Os demais experimentos seguintes tiveram como objetivo analisar a capacidade de aprendizado de reconhecimento de entidades no contexto de notícias de governo. Para realizar essa análise foram comparados sistemas treinados com a Coleção Dourada e com a Coleção de Notícias - UFRJ.

Após a realização dos experimentos foram constatados ganho de desempenho nos sistemas de reconhecimento que foram desenvolvidos especialmente para o contexto de notícias de governo. O ganho de desempenho foi observado em todas as classes comparadas, que foram: Pessoa, Local e Organização. O gráfico 11 apresenta a comparação entre os resultados obtidos com o sistema treinado a partir de textos da língua portuguesa em geral e

com o sistema treinado a partir de textos restritos às notícias de governo redigidas na língua portuguesa.

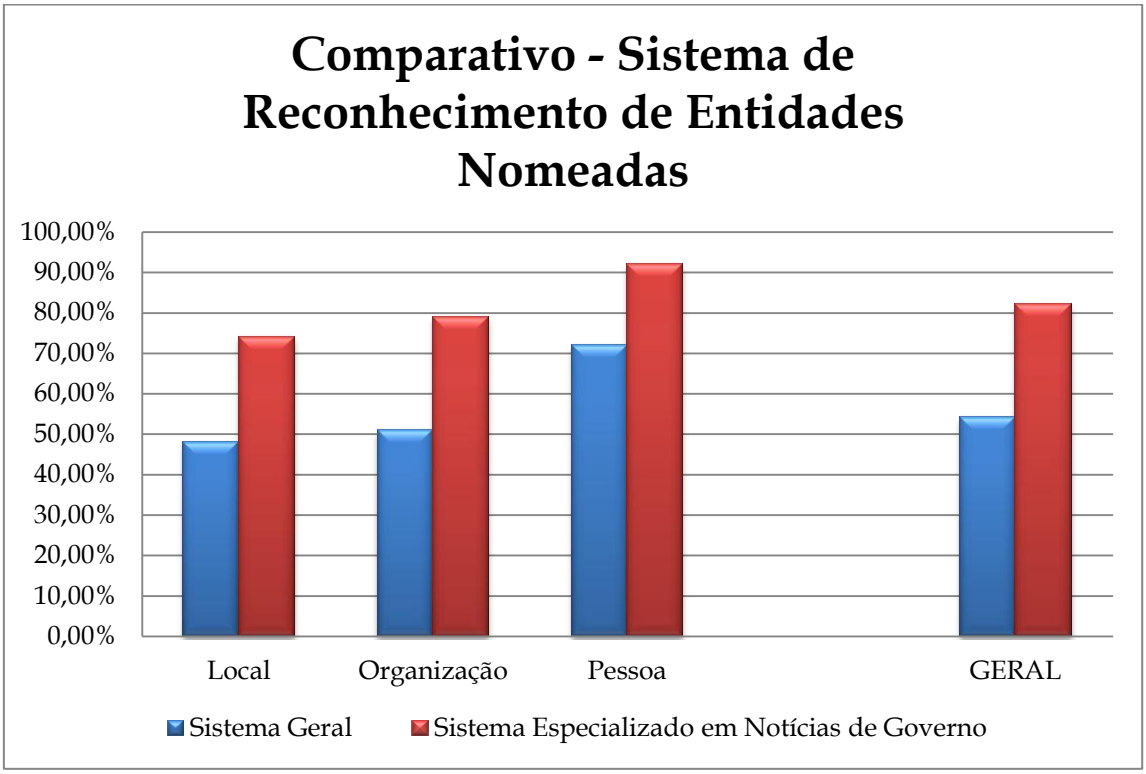

Gráfico 11: Comparativo entre os sistemas de reconhecimento de entidades

Esse comparativo fornece evidências de que ao limitar o escopo dos textos utilizados há uma redução na complexidade das características intrínsecas ao problema, facilitando o processo de aprendizado.

Além do comparativo realizado nesses experimentos, pode-se concluir que o uso de frase ou textos completos, como exemplos, não inferiu em alterações significativas no treinamento dos sistemas de reconhecimento de entidades nomeadas.

Capítulo 6. Conclusão e Trabalhos Futuros

Nos dias atuais, o volume de notícias publicadas na web cresce aceleradamente, de maneira que se torna difícil para uma pessoa acompanhar toda a informação disponibilizada. Tendo em vista o grande volume de informações apresentadas ao público, diversos trabalhos na literatura têm realizado pesquisas relacionadas ao tratamento automatizado das notícias.

Neste trabalho foram realizados estudos que visavam tratar a notícia, bem como o acesso a esta. Para isso foram realizadas duas propostas. A primeira propõe o reconhecimento de entidades nomeadas em notícias de governo, onde a tarefa de Reconhecimento de Entidades Nomeadas é caracterizada como um problema de identificar e classificar nomes próprios em textos, incluindo localizações, tais como Brasil e Rio de Janeiro; pessoas, tais como Dilma e Miriam; e organizações, tais como Ministério da Educação e Ministério da Cultura.

A tarefa de Reconhecimento de Entidades proposta nesse trabalho teve seu escopo reduzido à notícias de governo redigidas na língua portuguesa. Com a redução do escopo e o uso do Linear-Chain Conditional Random Field foi observado um notável ganho de desempenho em relação aos resultados obtidos por um sistema de reconhecimento de entidades genérico. Essa observação fornece evidências de que ao limitar o escopo dos textos utilizados há uma redução na complexidade das características intrínsecas ao problema, facilitando o processo de aprendizado.

Além disso, como fruto deste trabalho foi concebido a Coleção de Notícias de Governo – UFRJ, que demonstrou ser uma importante base para o treinamento de entidades nomeadas devido aos bons resultados obtidos após sua utilização. A Coleção de Notícias de Governo – UFRJ é uma coleção de notícias retiradas de sites do governo brasileiro, criada exclusivamente para este trabalho. As notícias foram anotadas manualmente com o auxílio de uma mestranda em Ciência da Informação da Universidade Federal Fluminense (UFF).

A segunda proposta propõe a concepção do Integrador de Notícias de Governo, cujo objetivo é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

Nesse trabalho a arquitetura do Integrador de Notícias de Governo foi modelada e implementada. O uso do Integrador foi validado em um estudo de caso, de modo que este foi utilizado em um projeto do Ministério de Planejamento chamado Portal de Notícias de Governo com a finalidade de fornecer ao Portal uma base de notícias constantemente atualizada com as notícias publicados nos mais diversos sites de interesse.

O Portal de Notícias do Governo é uma iniciativa que envolve a cooperação da SLTI/MP, SECOM/PR, COPPE/UFRJ e SERPRO. O seu objetivo é atender às necessidades da SECOM, da SRI e prover ao público um ambiente intuitivo e robusto para a pesquisa de notícias publicadas nos sites do governo.

Como trabalhos futuros, este trabalho sugere a expansão da Coleção de Notícias de Governo – UFRJ com novas categorias, novos tipos de relacionamentos entre entidades e novos exemplos, a fim de prover uma base cada vez mais robusta. Além disso, a tarefa de reconhecimento de Relações entre Entidades Nomeadas é um problema não solucionado no meio científico e que precisa ser mais explorado.

Outra importante sugestão neste trabalho é a inclusão de ontologias na representação semântica das notícias. Por exemplo, por meio do reconhecimento de entidades e de algumas de suas relações é possível popular automaticamente uma ontologia através da leitura das notícias conforme elas vão sendo publicadas. Então, posteriormente, essas ontologias podem ser utilizadas para a realização de inferências sobre as notícias.

Referências Bibliográficas

[ADAR et al., 2009] ADAR, Eytan; TEEVAN, Jaime; DUMAIS, Susan T.; ELSAS, Jonathan L. The web changes everything: understanding the dynamics of web content. 2009. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*, Ricardo Baeza-Yates, Paolo Boldi, Berthier Ribeiro-Neto, and B. Barla Cambazoglu (Eds.). ACM, New York, NY, USA, p. 282-291.

[AGNER, 2002] AGNER, Luiz C. *Otimização do diálogo usuários-organizações na World Wide Web*: estudo de caso e avaliação ergonômica de usabilidade de interfaces humano-computador. Rio de Janeiro, 2002. (Dissertação de Mestrado). Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Design. PUC-Rio, 2002. Cáp. 6.

[ALFONSECA; MANANDHAR, 2002] ALFONSECA, Enrique; MANANDHAR, S. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. 2002. In: *Proc. International Conference on General WordNet*.

[ASAHARA; MATSUMOTO, 2003] ASAHARA, Masayuki; MATSUMOTO, Y. Japanese Named Entity Extraction with Redundant Morphological Analysis. 2003. In: *Proc. Human Language Technology conference – North American chapter of the Association for Computational Linguistics*.

[BERTSEKAS, 1999] BERTSEKAS, Dimitri P. Nonlinear Programming. *Athena Scientific*, 2nd edition, 1999.

[BICK, 2007] BICK, Eckhard. "Functional Aspects on Portuguese NER". In: Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 145-155.

[BLEKAS et al., 2006] BLEKAS, Alexander; GAROFALAKIS, John; STEFANIS, Vasilios. Use of RSS feeds for content adaptation in mobile web browsing. 2006. In: *Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A)*: Building the mobile web: rediscovering accessibility? (W4A '06). ACM, New York, NY, USA, p. 79-85.

[BIKEL et al., 1997] BIKEL, Daniel M.; MILLER, S.; SCHWARTZ, R.; WEISCHEDEL, R. Nymble: a High-Performance Learning Name-finder. 1997. In: *Proc. Conference on Applied Natural Language Processing*.

[BORTHWICK et al., 1998] BORTHWICK, Andrew; STERLING, J.; AGICHTEIN, E.; GRISHMAN, R. NYU: Description of the MENE Named Entity System as used in MUC-7. 1998. In: *Proc. Seventh Message Understanding Conference*.

[BRIN, 1998] BRIN, Sergey. Extracting Patterns and Relations from the World Wide Web. 1998. In: *Proc. Conference of Extending Database Technology*. Workshop on the Web and Databases.

[BUNESCU, 2007] BUNESCU, Razvan Constantin. *Learning for Information Extraction*: From Named Entity Recognition and Disambiguation to Relation Extraction. Ph.D. Thesis, University of Texas at Austin. 2007.

[BYRD et al., 1994] BYRD, Richard H.; NOCEDAL, Jorge; SCHNABEL, Robert B. Representations of quasi-Newton matrices and their use in limited memory methods. *Math. Program.*, 63 (2):129–156, 1994.

[CGI.br e NIC.br, 2010] CGI.br e NIC.br. *Dimensões e características da Web brasileira*: um estudo do .gov.br. 2010. Disponível em: <http://www.cgi.br/publicacoes/pesquisas/govbr/cgibr-nicbr-censoweb-govbr-2010.pdf>. Acesso em: 05 fev. 2011.

[CIMIANO; VÖLKER, 2005] CIMIANO, Philipp; VÖLKER, J. Towards Large-Scale, Open-Domain and Ontology-Based Named Entity Classification. 2005. In: *Proc. Conference on Recent Advances in Natural Language Processing*.

[COLLINS, 2002] COLLINS, Michael. Ranking Algorithms for Named–Entity Extraction: Boosting and the Voted Perceptron. 2002. In: *Proc. Association for Computational Linguistics*.

[COLLINS; SINGER, 1999] COLLINS, Michael; SINGER, Y. Unsupervised Models for Named Entity Classification. 1999. In: *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[EVANS, 2003] EVANS, Richard. A Framework for Named Entity Recognition in the Open Domain. 2003. In: *Proc. Recent Advances in Natural Language Processing*.

[GEOGOOGLE, 2008] GEOGOOGLE – Google Geocoder Java API. 2008. Disponível em: <http://geo-google.sourceforge.net/>. Acesso em: 15 fev. 2012.

[GUPTA et al., 2005] GUPTA, Suhit; KAISER, Gail E.; GRIMM, Peter; CHIANG, Michael F.; STARREN, Justin. Automating Content Extraction of HTML Documents. 2005. *World Wide Web*, v. 8, n. 2, p. 179-224, June 2005.

[HAMMERSLEY, 2005] HAMMERSLEY, Ben. *Developing Feeds with Rss and Atom*. First ed. O'Reilly. 2005.

[HAREM, 2006] HAREM (Avaliação de Reconhecimento de Entidades Mencionadas). 2006. Disponível em: <http://www.linguateca.pt/primeiroHAREM/harem_introducao.html>. Acesso em: 10 fev. 2012.

[HE et al., 2004] HE, Xuming; ZEMEL, Richard S.; CARREIRA-PERPIÑIÁN, Miguel Á. Multiscale conditional random fields for image labelling. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[HEARST, 1992] HEARST, Marti. Automatic Acquisition of Hyponyms from Large Text Corpora. 1992. In: *Proc. International Conference on Computational Linguistics*.

[HTMLUNIT, 2011] HTMLUNIT. *Welcome to HtmlUnit*. 2011. Disponível em: <http://htmlunit.sourceforge. net/>. Acesso em: 05 fev. 2011.

[KUMAR; HEBERT, 2003] KUMAR, Sanjiv; HEBERT, Martial. Discriminative fields for modeling spatial dependencies in natural images. In: THRUN, Sebastian; SAUL, Lawrence; SCHÖLKOPF, Bernhard (Editores), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2003.

[LIU et al., 2005] LIU, Yan; CARBONELL, Jaime; WEIGELE, Peter; GOPALAKRISHNAN, Vanathi. Segmentation conditional random fields (SCRFs): A new approach for protein fold recognition. In: *ACM International conference on Research in Computational Molecular Biology (RECOMB05)*, 2005.

[LUCENE, 2010] LUCENE. *Apache Lucene* - Overview: um estudo do .gov.br. 2010. Disponível em: <http://lucene.apache.org/>. Acesso em: 05 fev. 2011.

[MCCALLUM; LI, 2003] MCCALLUM, Andrew; LI, W. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. 2003. In: *Proc. Conference on Computational Natural Language Learning*.

[MINISTÉRIO, 2012a] MINISTÉRIO da Ciência e Tecnologia e Inovação (MCTI). *Livro traz estratégia para próximos anos e atividades de 2011*. 2012. Disponível em: <http://www.mct.gov.br/index.php/content/view/335668.html>. Acesso em: 12 fev. 2012.

[MINISTÉRIO, 2012b] MINISTÉRIO da Ciência e Tecnologia e Inovação (MCTI). *Ministro Raupp toma posse no Conselho Curador da EBC*. 2012. Disponível em: <http://www.mct.gov.br/index.php/content/view/336014.html>. Acesso em: 12 fev. 2012.

[MINISTÉRIO, 2012c] MINISTÉRIO das Comunicações. *Anel vai interconectar as redes de fibra óptica dos países sul-americanos, reduzindo custos de conexão*. 2012. Disponível em: <http://www.mc.gov.br/notícias-do-site/24049-030212-minicom-apresenta-proposta-tecnica-para-construcao-de-anel-optico-na-america-do-sul>. Acesso em: 12 fev. 2012.

[MINISTÉRIO, 2011a] MINISTÉRIO do Planejamento. *MP regulariza imóvel do Centro de Atendimento Médico em Foz. 2011*. Disponível em: <http://www.mp.gov.br/notícia.asp?p=not&cod=7310&cat=69&sec=9>. Acesso em: 29 jan. 2012.

[MINISTÉRIO, 2011b] MINISTÉRIO do Planejamento. *Presidenta anuncia obras de saneamento do PAC 2 em mais de 1.100 municípios brasileiros*. 2011. Disponível em: <http://www.mp.gov.br/notícia.asp?p=not&cod=7939&cat=475&sec=61>. Acesso em: 12 fev. 2012.

[MORGADO, 2010] MORGADO, Fernando Fernandes. Representação de documentos através de nuvens de termos. Rio de Janeiro, 2010. (Dissertação de Mestrado). Universidade Federal do Estado do Rio de Janeiro (UFRJ). COPPE - Programa de Engenharia de Sistemas e Computação. 2010.

[MOTA; SANTOS, 2008] MOTA, Cristina; SANTOS, Diana (Editoras). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*: O Segundo HAREM.

Linguateca, 2008. Disponível em: <http://www.linguateca.pt/LivroSegundoHAREM/>. Acesso em: 10 jan. 2012. (ISBN: 978-989-20-1656-6).

[NADEAU; SEKINE, 2007] NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3--26, 2007.

[NORVAG; OYRI, 2005] NORVAG, Kjetil; OYRI, Randi. News Item Extraction for Text Mining inWeb Newspapers. 2005. In: *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI '05)*. IEEE Computer Society, Washington, DC, USA, p. 195-204.

[PENG; MCCALLUM, 2004] PENG, Fuchun; MCCALLUM, Andrew. Accurate information extraction from research papers using conditional random fields. 2004. In: *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*, 2004.

[PERA; NG, 2008] PERA, Maria Soledad; NG, Yiu-Kai. Utilizing phrase-similarity measures for detecting and clustering informative RSS news articles. 2008. *Integr. Comput.-Aided Eng*., v. 15, n. 4, p. 331-350, December 2008.

[PINEL et al., 2011] PINEL, Roque Elias; CARMO, Filipe Braida do; MONTEIRO, Rodrigo Salvador; ZIMBRÃO, Geraldo. Improving tests infrastructure through a model-based approach. SIGSOFT. 2011. *Softw. Eng. Notes*, v. 36, n. 1, p. 1-5, January 2011.

[RABINER, 1989] RABINER, Lawrence R. A tutorial on hidden markov models and selected applications in speech recognition. 1989. In: *Proceedings of the IEEE*, v. 77, n. 2, p. 257-286, February 1989.

[RAMASWAMY et al., 2003] RAMASWAMY, Lakshmish; IYENGAR, Arun; LIU, Ling; DOUGLIS, Fred. Techniques for efficient fragment detection in web pages. 2003. In: *Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03)*. ACM, New York, NY, USA, p. 516-519.

[RILOFF; JONES, 1999] RILOFF, Ellen; JONES, R. Learning Dictionaries for Information Extraction using Multi-level Bootstrapping. 1999. In: *Proc. National Conference on Artificial Intelligence*.

[SANG, 2002] SANG, Erik F. Tjong Kim. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: *Proceedings of CoNLL-2002*, Taipei, Taiwan, 2002, pp. 155-158.

[SANG; MEULDER, 2003] SANG, Erik F. Tjong Kim; MEULDER, Fien De. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 142-147.

[SANTOS, 2011] SANTOS, Diana. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language 3.2* (2011), pp. 113-128. ISSN: 18909639. Volume editado por J.B.Johannessen, Language variation infrastructure.

[SANTOS; CARDOSO, 2006] SANTOS, Diana; CARDOSO, Nuno. A golden resource for named entity recognition in Portuguese. 2006. In: VIEIRA, Renata; QUARESMA, Paulo; NUNES, Maria da Graça Volpes; MAMEDE, Nuno J.; OLIVEIRA, Claudia; DIAS, Maria Carmelita (Editores). *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Language, PROPOR'2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 69-79. Itatiaia, Rio de Janeiro, Brasil, 13-17, Maio 2006. Springer.

[SATO; SAKAKIBARA, 2005] SATO, Kengo; SAKAKIBARA, Yasubumi. RNA secondary structural alignment with conditional random fields. Bioinformatics, 21:ii237–242, 2005.

[SECOM, 2010] SECOM. Secretaria de Comunicação Social da Presidência da República. *Secretaria*. 2010. Disponível em: <http://www.secom.gov.br/sobre-a-secom/a-secretaria>. Acesso em: 05 fev. 2011.

[SEKINE, 1998] SEKINE, Satoshi. 1998. Nyu: Description of the Japanese NE System Used For Met-2. 1998. In: *Proc.Message Understanding Conference*.

[SETTLES, 2005] SETTLES, Burr. Abner: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[SHA; PEREIRA, 2003] SHA, Fei; PEREIRA, Fernando. Shallow parsing with conditional random fields. In: *Proceedings of HLT-NAACL*, pages 213–220, 2003.

[SHAIKH et al., 2010] SHAIKH, Mostafa Al Masum; PRENDINGER, Helmut; ISHIZUKA, Mitsuru. Emotion Sensitive News Agent (ESNA): A system for user centric emotion sensing from the news. 2010. *Web Intelli. and Agent Sys*., v. 8, n. 4, p. 377-396, December 2010.

[SILVA, 2010] SILVA, Marcelino Campos Oliveira. *Dados autonômicos*. Rio de Janeiro, 2010. (Dissertação de Mestrado). Universidade Federal do Estado do Rio de Janeiro (UFRJ). COPPE - Programa de Engenharia de Sistemas e Computação. 2010.

[SUTTON; MCCALLUM, 2006] SUTTON, Charles; MCCALLUM, Andrew. An introduction to conditional random fields for relational learning. 2006. In: GETOOR, Lise; TASKAR, Ben (Editores). *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[TASKAR et al., 2002] TASKAR, Ben; ABBEEL, Pieter; KOLLER, Daphne. Discriminative probabilistic models for relational data. 2002. In: *Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02)*, 2002.

[TERMCLOUD] TERMCLOUD - Sample. Disponível em: <http://visapi-gadgets.googlecode.com/svn/trunk/termcloud/doc.html>. Acesso em: 15 fev. 2012.

[THELWALL; PRABOWO, 2007] THELWALL, Mike; PRABOWO, Rudy. Identifying and characterizing public science-related fears from RSS feeds: Research Articles. *Journal of the American Society for Information Science and Technology*, v. 58, n. 3, p. 379-390, February 2007.

[THE STANFORD, 2012] THE STANFORD NLP (Natural Language Processing) Group. 2012. Disponível em: <http://www-nlp.stanford.edu/software/CRF-NER.shtml>. Acesso em: 26 jan. 2012.

[TUTORIAL, 2010] TUTORIAL da Google Maps Javascript API V3. 2010. Disponível em: <http://code.google.com/intl/pt-BR/apis/maps/documentation/javascript/tutorial.html>. Acesso em: 15 fev. 2012.

[YI et al., 2003] YI, Lan; LIU, Bing; LI, Xiaoli. Eliminating noisy information in Web pages for data mining. 2003. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03)*. ACM, New York, NY, USA, p. 296-305.