PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CONSTRUÇÃO DE UM CORPUS ANOTADO PARA CLASSIFICAÇÃO DE ENTIDADES NOMEADAS UTILIZANDO A WIKIPEDIA E A DBPEDIA

CRISTOFER WEBER

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dra. Renata Vieira

Porto Alegre
2015

Dados Internacionais de Catalogação na Publicação (CIP)

W376c Weber, Cristofer

Construção de um corpus anotado para classificação de entidades nomeadas utilizando a Wikipedia e a Dbpedia / Cristofer Weber. – Porto Alegre, 2015.
84 p.

Dissertação (Mestrado) – Faculdade de Informática, PUCRS.
Orientador: Prof. Dr. Renata Vieira.

1. Informática. 2. Processamento da Linguagem Natural.
3. Linguítica Computacional. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Construção de um Corpus anotado para classificação de Entidades Nomeadas Utilizando a Wikipedia e a Dbpedia" apresentada por Cristofer Weber como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, aprovada em 16/03/2015 pela Comissão Examinadora:

Profa. Dra. Renata Vieira– PPGCC/PUCRS
Orientadora

Prof. Dr. Rafael Heitor Bordini– PPGCC/PUCRS

Profa. Dra. Viviane Moreira– UFRGS

Homologada em 19/11/2015, conforme Ata No. 021 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320–3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Venho de uma família que nunca mediu esforços para educar seus filhos. Devo muito aos meus pais, Luiz Inácio Weber e Iene Arend, pelos estímulos ao longo da minha vida, pela compreensão quando dos meus tropeços, e pelo orgulho que brilha em seus olhos à cada conquista. Dedico também à minha esposa, Sinara K. F. Weber, que embarcou comigo nesta jornada. Sem ela eu não chegaria até aqui; com ela eu vou mais longe. Sua presença, suas palavras nos momentos mais difíceis, sua compreensão sempre que me ausentei, foram fundamentais para o fechamento deste ciclo. Amo vocês, e espero corresponder à tudo que fizeram por mim.

"There is a road that I must travel May it be paved or unseen May I be hindered by a thousand stones Still onward I'd crawl down on my knees."
(Disillusion - Back to Times of Splendor)

AGRADECIMENTOS

Agradeço aos meus irmãos, Tatiana, Jônatas, e Larissa, pelo estímulo. Vocês estiveram um passo à frente, mostrando o caminho. Agora vamos juntos :-)

Agradeço à minha orientadora, Professora Doutora Renata Vieira, por me desafiar a fazer mais. Você esteve presente quando precisei, e confiou em mim quando eu quis arriscar. Foi o que precisei para crescer. Muito obrigado!!!

CONSTRUÇÃO DE UM CORPUS ANOTADO PARA CLASSIFICAÇÃO DE ENTIDADES NOMEADAS UTILIZANDO A WIKIPEDIA E A DBPEDIA

RESUMO

Algumas tarefas de processamento de linguagem natural podem ser aprendidas por algoritmos a partir de *corpus* de exemplo, mas a obtenção destes exemplos pode ser um gargalo. Neste trabalho nós investigamos como a Wikipedia e a DBpedia, dois recursos de linguagem disponíveis de forma gratuita, podem ser utilizados como *corpus* para a classificação de entidades nomeadas, uma tarefa fundamental de extração de informações e um passo necessário para outras tarefas como extração de relações e resolução de co-referências.

Palavras Chave: Processamento de Linguagem Natural, Construção de *Corpus*, Recursos de Linguagem, Extração de Informações, Classificação de Entidades Nomeadas.

BUILDING A CORPUS FOR NAMED ENTITY RECOGNITION USING PORTUGUESE WIKIPEDIA AND DBPEDIA

ABSTRACT

Some natural language processing tasks can be learned from example corpora, but having enough examples for the task at hands can be a bottleneck. In this work we address how Wikipedia and DBpedia, two freely available language resources, can be used to support Named Entity Recognition, a fundamental task in Information Extraction and a necessary step of other tasks such as Co-reference Resolution and Relation Extraction.

Keywords: Natural Language Processing, Corpus Construction, Language Resources, Information Extraction, Named Entities Classification.

SUMÁRIO

1. INTRODUÇÃO

A Wikipedia, desde sua criação, vem sendo estudada como um recurso de linguagem para diferentes tarefas de processamento da linguagem natural. Nas tarefas de extração da informação, a Wikipedia já foi utilizada para reconhecimento de entidades nomeadas [NRR+13], extração de relações [NMI07] e desambiguação de entidades nomeadas [BP06]. Dois elementos estruturais básicos da Wikipedia, os títulos e os wikilinks, são explorados como anotações de entidades nomeadas para treino de classificadores, mas não são suficientes. Os wikilinks delimitam termos que podem ser entidades nomeadas, mas a ausência de uma categoria semântica consistente de artigos de acordo com o assunto principal do artigo requer a observação de outras características presentes na estrutura dos artigos da Wikipedia. Diferentes autores exploraram outros elementos estruturais para atribuição de uma categoria semântica aos artigos, e consequente uso destas categorias como anotações das palavras que compõem os wikilinks para formação de *corpora* anotados. Dentre as diferentes abordagens para classificação dos artigos, é bastante comum o uso das categorias e suas hierarquias, infoboxes, redirecionamentos e desambiguações. De forma menos comum, as regras implícitas para formação de nomes de categorias e padrões de escrita de artigos também foram exploradas para extração de categorias semânticas.

No entanto, se atentarmos à motivação de projetos de construção de bases de conhecimento a partir da Wikipedia, como a DBpedia e o YAGO2, vemos que estes projetos utilizam os mesmos elementos estruturais acima citados para a mesma finalidade de atribuir uma categoria semântica consistente aos artigos da Wikipedia. Especificamente no caso da DBpedia, há uma comunidade de usuários mantenedores responsáveis por expandir e aprimorar as ferramentas para classificação dos artigos a cada nova versão da DBpedia. Mais recentemente, [PB14] apresentaram modelos estatísticos aplicados à Wikipedia e à DBpedia que aumentaram o número de artigos classificados na DBpedia Ontology sem a necessidade de infoboxes, assim expandindo o número de artigos classificados na DBpedia sem depender de edições na Wikipedia.

Trabalhos anteriores de geração de corpus anotado a partir da Wikipedia como [NRR+13] já exploraram o uso da Wikipedia como fonte para a geração de *corpora* anotados para reconhecimento de entidades nomeadas. No entanto, estes trabalhos concentraram esforços na identificação da categoria semântica de artigos para anotação dos wikilinks, na etiquetagem de wikilinks com a categoria semântica correspondente, na identificação outras menções a entidades presentes nos artigos, e na seleção das sentenças para inclusão no corpus. E se ao invés de identificar a categoria semântica através de heurísticas aplicadas às estruturas de artigos, utilizarmos uma ontologia como a DBpedia? A DBpedia, por identificar artigos da Wikipedia como entidades, e descrevê-los como instâncias de classes da ontologia, pode substituir o esforço de identificação da categoria semântica dos artigos, enriquecendo a abrangência de instâncias devido aos aprimoramentos dos mecanismos de extração? São essas as questões que nortearam esta pesquisa, que foi organizada da seguinte forma:

1. **Introdução** (este capítulo);
2. **Fundamentação Teórica**, onde estão descritos os recursos utilizados e conceitos fundamentais para o desenvolvimento do trabalho;
3. **Trabalhos relacionados**, que descreve três trabalhos que motivaram esta pesquisa;
4. **Geração de *corpus* para o reconhecimento de entidades nomeadas a partir da Wikipedia**, onde está descrito o procedimento para construção de *corpora* de acordo com as questões acima;
5. **Implementação**, que detalha o *software* construído para a geração de *corpus*;
6. **Experimentos**, que descreve os experimentos realizados para avaliação da aplicabilidade dos *corpora* gerados automaticamente na tarefa de reconhecimento de entidades nomeadas;
7. **Conclusão e Trabalhos Futuros**, onde se encontram as respostas para as questões apresentadas acima, obtidas a partir do processo, *software* e avaliações construídas neste trabalho.

Como principais contribuições deste trabalho, podemos destacar:

- A geração automática de *corpora* a partir da Wikipedia e DBpedia em Português, com possibilidade de seleção dos artigos da Wikipedia a partir das Categorias da Wikipedia, anotado com categorias semânticas escolhidas a partir das classes da ontologia da DBpedia;
- A disponibilização de todos os artefatos de software necessários para a geração automática de *corpora*, na forma de Software Livre;
- Um processo para avaliação do desempenho de diferentes *corpora* com objetivo de identificar quais ações aplicadas na geração e anotação automática foram efetivas em comparação à ações anteriores.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Recursos de Linguagem

Recursos de linguagem são as ferramentas e dados necessários para o desenvolvimento de aplicações de processamento de linguagem natural. Os recursos de dados são necessários para o estudo de fenômenos linguísticos, avaliação de sistemas e para o treino e teste de componentes de aprendizagem de máquina utilizados no processamento de linguagem natural. Corpora é um recurso de dados formado por conjunto de corpus com anotações de classes gramaticais, categorias semânticas, entre outras. Gazetteers são recursos de dados compostos por listas de nomes de entidades pertencentes a categorias semânticas como Localizações e Organizações, por exemplo. Bases de conhecimento, por sua vez, são recursos de dados e ferramentas que armazenam informações que descrevem conhecimento sobre entidades e fatos. As ferramentas, por sua vez, são os recursos como tokenizers, analisadores sintáticos e etiquetadores de classes gramaticais (part-of-speech taggers), que possibilitam a construção de aplicações de processamento de linguagem natural através do reuso destas ferramentas. Dentre estas duas formas de recursos, este trabalho têm como principal objetivo o uso de recursos de dados.

Segundo [HNP13], recursos estruturados são recursos legíveis por máquinas e que codificam relações de vários tipos de acordo com o nível de informação. Recursos estruturados são de alta qualidade pois são construídos a partir do conhecimento de especialistas de domínio, lexicógrafos e linguistas. No entanto, recursos estruturados são escassos pois demandam grandes esforços na criação e atualização. Por serem construídos manualmente, dependem da disponibilidade de especialistas para ampliar sua cobertura e para mantê-los atualizados frente a acontecimentos recentes. Além disto, o conhecimento codificado em uma língua não é transferível para outras, demandando a recodificação em outras línguas. Os recursos estruturados mais comuns são:

- Tesauros: coleções de termos relacionados;
- Taxonomias: estruturas hierárquicas de classificação de termos;
- Ontologias: modelos de conhecimento que incluem conceitos, relações de diversos tipos, regras e axiomas.

Recursos não estruturados, por sua vez, são coleções de textos que não possuem conhecimento formalizado e são legíveis por máquinas apenas como sequências de caracteres e palavras. Diferentes modelos estatísticos são capazes de extrair conhecimento de coleções não estruturadas, e a grande quantidade de textos disponíveis na WWW possibilita a construção de bases de conhecimento de grande cobertura. Mas são limitados pela carência de textos que explicitem conhecimento de senso comum ([HNP13]). Também, modelos estatísticos não são capazes de emitir conhecimento com qualidade equivalente aos recursos construídos por especialistas.

Os autores entendem que as limitações de recursos não estruturados são complementares às limitações dos recursos estruturados. Enquanto que recursos não estruturados possibilitam ampla cobertura com baixa qualidade, recursos estruturados possuem alta qualidade mas reduzida cobertura. Recursos semiestruturados construídos de forma colaborativa na WWW codificam o conhecimento voluntariamente disponibilizado por usuários destes recursos, cobrindo diferentes áreas de conhecimento e possuindo qualidade equiparável àquela obtida a partir de especialistas. Os recursos semiestruturados mencionados pelos autores são o Wiktionary, Flickr, Twitter, Yahoo! Answers e, com maior ênfase, a Wikipedia.

Em [GK13] são apresentados os resultados de pesquisas que utilizaram com sucesso recursos de linguagem construídos colaborativamente como substitutos de recursos de linguagem construídos de forma convencional. Os exemplos mais proeminentes são a Wikipedia – uma enciclopédia construída colaborativamente, o Wiktionary – que é um dicionário construído colaborativamente, e recursos construídos a partir de ferramentas de inteligência coletiva como o Mechanical Turk e Games with a Purpose.

2.1.1 A Wikipedia

[Las03] descreve a Wikipedia como uma enciclopédia *online* livre onde todos os leitores podem atualizar seu conteúdo através da inclusão e edição de seus artigos. Ao invés de seguir um processo de *peer review* por especialistas, os artigos da Wikipedia estão disponíveis para correções e aprimoramentos por parte dos leitores.

Em sua revisão bibliográfica sobre o uso da Wikipedia em pesquisas de Ciências da Computação, [MMLW09] define a Wikipedia como um recurso valioso com amplas funcionalidades. Sua revisão demonstra como pesquisadores desenvolveram técnicas sofisticadas para extração de conhecimento a partir de diferentes perspectivas:

- A Wikipedia como uma Enciclopédia;
- A Wikipedia como um *corpus*;
- A Wikipedia como um tesauro;
- A Wikipedia como uma base de dados;
- A Wikipedia como uma ontologia;
- A Wikipedia como um grafo.

Numa revisão mais recente feita por [HNP13], os autores apresentam a Wikipedia como um recurso de semiestruturado.

A Wikipedia é segmentada em línguas, sendo que cada língua possui um subdomínio único dentro do domínio wikipedia.org. A atualização de estatísticas de 30 de Setembro de 2014 [1] lista 250 línguas ativas.

Artigos são escritos utilizando texto livre, mas alguns recursos estruturados estão disponíveis para organização dos artigos em categorias, para conexão entre diferentes artigos, bem como para apresentação de propriedades relevantes do assunto descrito no artigo.

Títulos

Cada artigo da Wikipedia possui um nome, o título, que é a forma mais comum de identificação do conceito ou entidade descrita neste artigo. As classes mais comuns descritas na Wikipedia são pessoas, organizações, lugares, obras, acontecimentos e espécies de seres vivos. O título deve ser único dentro do conjunto de artigos da Wikipedia para uma língua.

Wikilinks

A garantia da unicidade do título possibilita referenciar um artigo através de seu título, e a Wikipedia explora esta possibilidade através das ligações internas (*Wikilinks*). Sendo ela uma ferramenta online, seus *Wikilinks* são referências ativas que permitem a navegação entre artigos em uma rede de ligações internas construídas pelos editores dos artigos. A Wikipedia recomenda que apenas a primeira ocorrência de uma referência a outro artigo seja ligada através de *wikilink*. Seguindo esta recomendação, as demais ocorrências de referências serão apresentadas em texto simples, cabendo ao leitor interpretá-las como menções a outros assuntos.

Ao referenciar um artigo através de *wikilink* o editor pode optar por uma diferente forma textual que identifique o assunto do artigo referenciado. Um exemplo é o filme "Pink Floyd The Wall", produzido pelo diretor britânico Alan Parker e escrito pelo músico Roger Waters. Apesar do nome do filme ser "Pink Floyd The Wall" e ser referenciado desta forma, outros *wikilinks* presentes na Wikipedia mencionam somente "The Wall" mas referenciam o mesmo artigo "Pink Floyd The Wall", tal como ocorre no artigo do músico Roger Waters[2].

Desambiguações

É comum que diferentes conceitos ou entidades homônimas, o que violaria esta restrição de unicidade. Nestes casos, o artigo que define o conceito mais conhecido permanece com o nome, e os demais títulos devem possuir um sufixo para desambiguação. O sufixo de desambiguação deve apresentar um detalhe que possibilite a diferenciação de um artigo dos demais. Sugere-se a criação de uma página específica de desambiguação que liste os diferentes artigos homônimos com ligações internas para seus conteúdos. Quando não for possível determinar qual dos conceitos é o mais

[1] http://stats.wikimedia.org/PT/Sitemap.htm, acessado em 13-11-2014
[2] http://en.wikipedia.org/wiki/Roger_Waters (em inglês), acessado em 10-11-2014

conhecido, a página de desambiguação possui o nome e todas as demais páginas possuem o sufixo de desambiguação. Um exemplo da Wikipedia em Português é o nome Paradise Lost, que pode ser:

- a banda inglesa de heavy metal,
- o álbum da banda estadunidense Symphony X, ou
- o poema do século XVII escrito pelo poeta inglês John Milton.

Neste exemplo acima não há como determinar o quão conhecida é cada entidade, portanto a página de desambiguação possui o título "Paradise Lost"[3]. O artigo da banda possui o título "Paradise Lost (banda)", e o artigo do álbum possui o título "Paradise Lost (álbum de Symphony X)". Já o nome "metal" seguramente mais conhecido se refere ao elemento. Neste caso o artigo de título "Metal"[4] descreve o elemento, enquanto que a página de desambiguação possui o título "Metal (desambiguação)"[5]. A página de desambiguação lista referências para o elemento "metal" e para o estilo musical "Heavy Metal".

Redirecionamentos

Outra variação de *Wikilinks* são os redirecionamentos, empregados quando um conceito ou entidade é conhecida por diferentes formas textuais. Esta situação conflitaria com a restrição de unicidade de títulos, impondo a repetição do conteúdo de um artigo com diferentes títulos. Esta situação ocorre com pessoas, quando esta é conhecida tanto por seu nome completo quanto por parte de seu nome ou apelido. É o caso do escritor americano H. P. Lovecraft, bastante conhecido apenas pelo seu sobrenome, Lovecraft. Também ocorre com organizações conhecidas por seu nome e pela sigla do nome, como é o caso da Pontifícia Universidade Católica, também conhecida por sua sigla PUC. Compreende ainda os casos em que uma entidade muda de nome, como é o caso da empresa BRF, que antes era chamada de Brasil Foods S.A. A Wikipedia soluciona estes problemas com páginas de redirecionamento. Estas páginas possuem apenas um titulo e conteúdo especial que redireciona o navegador para o artigo especificado. O recurso de redirecionamento possibilita também que usuários consigam navegar para páginas cujo título foi alterado, pois a Wikipedia não possui recursos para identificar *Wikilinks* para artigos que tiveram seu título modificado. O artigo da Wikipedia em Inglês para a banda americana "Mr. Big" exemplifica este caso, pois inicialmente o título era "Mr. Big (band)", e posteriormente foi modificado para "Mr. Big (american band)" por existir uma banda inglesa com mesmo nome. A seguir, a tabela 2.1 ilustra alguns exemplos de redirecionamento previamente citados.

[3] http://pt.wikipedia.org/wiki/Paradise_Lost, acessado em 10-11-2014.
[4] http://pt.wikipedia.org/wiki/Metal, acessado em 10-11-2014.
[5] http://pt.wikipedia.org/wiki/Metal_(desambiguaç~ao), acessado em 10-11-2014.

Tabela 2.1: Exemplos de redirecionamentos da Wikipedia

6

Origem	**Destino**
https://pt.wikipedia.org/wiki/Lovecraft	https://pt.wikipedia.org/wiki/H._P._Lovecraft
https://pt.wikipedia.org/wiki/PUC	https://pt.wikipedia.org/wiki/Pontifícia_Universidade_Católica
http://en.wikipedia.org/wiki/Mr._Big_(band)	https://en.wikipedia.org/wiki/Mr._Big_(American_band)

Infoboxes

De acordo com a documentação de ajuda da Wikipedia[7], *infoboxes* são tabelas de formato fixo que apresentam de forma resumida aspectos relevantes de um artigo. É um recurso opcional mas que, quando aplicado, apresenta atributos comuns entre assuntos diferentes. A Wikipedia recomenda o uso de modelos de *infobox* predefinidos, pois estes já possuem nomes conhecidos e sugestão de atributos comuns por categorias. Ao empregar *infoboxes* predefinidos em um artigo, o mecanismo de apresentação visual de artigos aplica formatações especiais que enriquecem a formatação da tabela. Estas predefinições também são utilizadas como metadados por projetos como por exemplo a DBpedia. A figura 2.1 mostra a infobox do artigo sobre o músico canadense Devin Townsend[8]. Junto é exibida a predefinição "Info/Música/artista"[9], recomendada para uso em artigos sobre músicos, cantores, bandas, conjuntos e grupos musicais. Esta figura mostra que, para o exemplo do músico Devin Garrett Townsend, nem todos atributos foram informados. Há também atributos que não são visíveis, tal como o atributo "fundo", que de acordo com seu valor define se a *infobox* é sobre um músico ou banda/grupo e aplica um estilo visual distinto conforme o valor do atributo.

Categorias

Todo artigo da Wikipedia deve ter ao menos uma categoria. Categorias são coleções que identificam tópicos da enciclopédia. As categorias fazem parte de uma estrutura de informação hierárquica, com categorias mais específicas compondo categorias mais abrangentes. No entanto, esta estrutura não é uma taxonomia, visto que uma categoria pode pertencer a mais de uma categoria. Categorias são estruturas diferentes de artigos, podendo receber um título já utilizado em artigo. No entanto, nomes de categorias também devem ser únicos dentro de uma mesma língua.

O já mencionado artigo sobre o músico canadense Devin Townsend tem, entre outras categorias, as categorias "Cantores do Canadá", "Guitarristas do Canadá", "Guitarristas de heavy metal", "Músicos de metal progressivo" e "Naturais de New Westminster". Estas categorias são associadas com outras categorias mais abrangentes em uma relação de subsunção, sendo que "Cantores do Canadá" e "Guitarristas do Canadá" são subcategorias da categoria "Músicos do Canadá". "Guitarristas de heavy metal" é uma subcategoria da categoria "Músicos de heavy metal". "Músicos de metal progressivo" é, assim como "Músicos de heavy metal", subcategoria de "Músicos de metal".

[7] https://en.wikipedia.org/wiki/Help:Infobox, acessado em 10-11-2014.
[8] https://pt.wikipedia.org/wiki/Devin_Townsend, acessado em 11-11-2014.
[9] Acessado em 11-11-2014.

Informação geral	
Nome completo	Devin Garrett Townsend
Nascimento	5 de Maio de 1972
Origem	New Westminster, Colúmbia Britânica
País	Canadá
Gênero(s)	Metal progressivo, metal extremo, metal industrial, música ambiente, thrash metal, death metal, grindcore, rock progressivo, punk rock, new age, música eletrônica, música clássica, country
Instrumento(s)	Guitarra, baixo, teclado, vocal, banjo
Modelos de instrumentos	Peavey, Framus, Sadowsky, ESP, Fender e Gibson
Período em atividade	1993–atualmente
Gravadora(s)	Hevy Devy Records, InsideOut Music, Century Media
Afiliação(ões)	The Devin Townsend Band, Strapping Young Lad, Steve Vai, Punky Brüster, IR8, Grey Skies, Caustic Thought, Noisescapes, Devin Townsend Project, Casualties Of Cool
Página oficial	www.hevydevy.com www.ziltoid.com

(a) Infobox

```
{{Info/Música/artista
 |nome                  =
 |imagem                =
 |imagem_tamanho        =
 |imagem_legenda        =
 |fundo                 =
 |nome completo         =
 |apelido               =
 |nascimento_data       =
 |nascimento_cidade     =
 |nascimento_país       =
 |origem                =
 |país                  =
 |morte_data            =
 |morte_local           =
 |nacionalidade         =
 |gênero                =
 |ocupação              =
 |instrumento           =
 |instrumentos notáveis =
 |modelos               =
 |tipo vocal            =
 |período               =
 |outras ocupações      =
 |gravadora             =
 |afiliações            =
 |influências           =
 |influenciados         =
 |website               =
 |assinatura            =
}}
```

(b) Predefinição

Figura 2.1: Exemplo de Infobox do artista Devin Garret Townsend, ao lado da predefinição de infobox para artistas.

2.2 *Linked Data*

[BHBL09] definem *Linked Data* como a conexão entre itens presentes em diferentes fontes de dados em um único espaço de dados global. *Linked Data* utiliza padrões da World Wide Web em um modelo de dados comum que possibilita a criação de aplicações capazes de operar neste espaço de dados. Entidades em *Linked Data* devem utilizar URIs no formato HTTP de forma que possam ser acessadas pela Web e ao serem acessadas retornem informações úteis sobre esta entidade utilizando padrões Web como o formato RDF. Quando uma entidade estiver relacionada à outra entidade publicada em *Linked Data*, deve-se utilizar a URI HTTP desta outra entidade, possibilitando a navegação entre estas duas entidades. Dentre as fontes de dados que compõe a *Linked Open Data*, a DBpedia se sobressai como um núcleo de interligação entre as diferentes fontes.

2.2.1 DBpedia

A DBpedia [LIJ+14] é um projeto que extrai conhecimento multilíngue e estruturado da Wikipedia e publica este conhecimento utilizando padrões da Web Semântica e *Linked Data*. Um de seus objetivos é de fornecer capacidades de pesquisa e consulta aos dados da Wikipedia, extraindo dados estruturados que podem ser utilizados para responder consultas expressivas. Os dados são extraídos de diferentes partes de páginas da Wikipedia na forma de sentenças RDF, e expostos para consultas através de endpoints SPARQL.

Do processo de construção da DBpedia resulta um conjunto de entidades correspondente aos artigos da Wikipedia e classificadas de acordo com a DBpedia Ontology. A DBpedia Ontology é uma ontologia multi-domínio que em sua versão *(release 2014)*[10] cobre um conjunto de 685 diferentes classes composta por uma hierarquia de subsunção de até oito níveis de profundidade. Em seus níveis superiores estão conceitos como Pessoa, Organização, Localização e Eventos – conceitos que correspondem às categorias semânticas comumente requisitadas em tarefas de Classificação de Entidades Nomeadas. Nos níveis mais baixos estão conceitos mais específicos tais como Banda, uma especialização de Organização, ou Escritor, especialização de Pessoa.

O conteúdo da DBpedia é construído a partir da Wikipedia através de extratores, que são programas escritos para converter partes específicas de artigos da Wikipedia em sentenças RDF. Em [LIJ+14] os autores propõe a divisão nas seguintes quatro categorias: Mapping-Based Infobox Extraction, Raw Infobox Extraction, Feature Extraction e Statistical Extraction.

Mapping-Based Infobox Extraction

Esta é a principal categoria de extratores. Estes extratores utilizam regras de mapeamento criadas e mantidas pela comunidade para mapear dados apresentados nas *infoboxes* de artigos da Wikipedia em instâncias da DBpedia Ontology. Como a maioria das *infoboxes* são baseadas em modelos predefinidos, a comunidade construiu mapeamentos distintos destas predefinições para as classes das ontologias. Com isto cada artigo que utiliza *infobox* baseada em predefinição passa a ser uma instância de classe. Outro esforço realizado pela comunidade consiste na normalização de diferentes variações de nomes de propriedades presentes nas *infoboxes*, bem como no uso de diferentes *infoboxes* para artigos de assuntos comuns. Cada mapeamento é também uma página da DBpedia Wiki, possibilitando a criação e edição destes mapeamentos da mesma forma que páginas da Wikipedia.

O exemplo da tabela 2.2 abaixo explica o mapeamento do modelo predefinido "Info/Música/artista"[11] para as classes Banda ou Artista Musical. Este exemplo é relevante pois o mesmo modelo predefinido é válido para duas classes distintas.

[10] http://wiki.dbpedia.org/Ontology, acessado em 11-11-2014.
[11] http://mappings.dbpedia.org/index.php/Mapping_pt:Info/Música/artista, acessado em 12-11-2014.

Tabela 2.2: Regra de mapeamento de um *infobox template* para a classes da DBpedia Ontology. As regras são aplicadas na ordem em que se encontram, e a classe é definida primeira regra válida.

Condição	**Classe**
infobox.integrantes.isSet	Banda
infobox.exintegrantes.isSet	Banda
infobox.fundo $==$ *'grupo_ou_banda'*	Banda
infobox.fundo $==$ *'banda_cover'*	Banda
infobox.fundo $==$ *'grupo_clássico'*	Banda
senão	Artista Musical

Dados extraídos por meio da categoria *Mapping-Based Infobox Extraction* são dados considerados de maior qualidade, por serem resultado de consenso entre membros da comunidade. Também apresentam maior qualidade pois o extrator atribui tipos de dados adequados a cada atributo. No entanto, pela Wikipedia possibilitar edições das *infoboxes* sem reforçar a aplicação das recomendações de uso de *templates*, apenas um subconjunto mais comum de atributos é mapeado por estes extratores. Estas sentenças são importadas em um banco de dados semântico com capacidades de execução de consultas SPARQL e geração dinâmica de páginas para exibição do conteúdo relacionado a um recurso. Os identificadores dos recursos na DBpedia são idênticos aos nomes dos artigos equivalentes na Wikipedia.

A DBpedia armazena também relações expressas como hyperlinks, redirecionamentos e desambiguações na Wikipedia. As relações de hyperlinks, no entanto, não registram a forma utilizada no texto da Wikipedia, e por isto não podem ser consideradas como substitutas para os hyperlinks presentes nos artigos da Wikipedia. As relações de redirecionamento e desambiguação, por sua vez, são fiéis aos seus originais na Wikipedia.

2.2.2 Wikidata

[Vra13] apresenta o projeto Wikidata como um projeto livre, colaborativo, multilíngue, de coleta de dados estruturados para suportar a Wikipedia e outras iniciativas. A Wikidata difere da Wikipedia, pois na Wikipedia o conteúdo dos artigos é escrito em prosa, enquanto que o propósito da Wikidata é coletar dados estruturados que possam ser reutilizados por outros sistemas, como a própria Wikipedia. O principal objetivo da Wikidata na sua concepção foi servir de base de conhecimento para as diferentes edições (idiomas) da Wikipedia. O autor cita como exemplo os dados demográficos presentes nos artigos sobre locais: na Wikipedia, atualização da população de uma cidade deve ser feita manualmente no texto do artigo, para cada idioma. Com a Wikidata, os artigos da Wikipedia podem referenciar a propriedade "população" da cidade, sendo automaticamente sincronizada.

Todos os itens da Wikidata são identificados por URIs utilizando padrões da Web Semântica e Linked Data, retornando os dados nos formatos RDF e JSON. Douglas Adams, escritor

britânico, é identificado pela URI https://www.wikidata.org/entity/Q42. Wikidata está conectada a datasets externos utilizando URIs de repositórios como o IMDb (Internet Movie Database) e MusicBrainz, e também padrões como ISBN para identificação de livros, ICD-9 para doenças, e ISO 639 para idiomas. As propriedades de uma entidade possibilitam a adição de fontes externas que corroboram a verificabilidade de tal informação. No exemplo de Douglas Adams, a propriedade "alma mater" possui o valor University of Cambridge[12], referenciando como uma das fontes desta informação a página http://www.nndb.com/people/731/000023662. Os diferentes nomes pelos quais uma entidade pode ser referenciada são valores da propriedade "Also known as", que podem ser utilizados em tarefas de Entity Disambiguation/Linking. Todos os artigos da Wikipedia em diferentes idiomas que descrevem um mesmo item são relacionados na Wikidata, facilitando a obtenção de conteúdo textual multilíngue.

2.3 Reconhecimento de Entidades Nomeadas

A tarefa de Reconhecimento de Entidades Nomeadas (REN) foi definida em [GS96], como a tarefa "a qual basicamente envolve identificar os nomes de todas as pessoas, organizações e localizações geográficas em um texto", envolvendo também a identificação de expressões de data, hora, valores monetários e percentuais. Distingue-se o Reconhecimento de Entidades Nomeadas da sua classificação, mas para fins de clareza trataremos ambas como o Reconhecimento de Entidades Nomeadas (REN).

Em sua forma mais comum a tarefa de REN reconhece um número predefinido de categorias semânticas, tais como as definidas em [GS96]. Mas também foi aplicada com sucesso em domínios específicos como biologia [CMO12] e geologia [NMG10], onde se utiliza um número maior de categorias relacionadas ao domínio. Uma forma bastante comum e estudada é o uso de regras linguísticas para o reconhecimento das entidades. Nesta abordagem as regras são manualmente codificadas a partir de conhecimento gramatical e de domínio, requerendo especialização em ambos para obtenção de bons resultados [NS07]. Ainda, o uso de regras linguísticas restringe seu uso a documentos escritos na língua para a qual as regras foram codificadas, impossibilitando seu uso em outras línguas.

Outra abordagem bastante comum de REN é através do uso de aprendizagem de máquina [NS07]. Algoritmos devem ser treinados com textos previamente anotados para que aprendam através destas anotações como identificar se determinada sequência de palavras pertence a uma das categorias anotadas. As anotações identificam para cada palavra qual a sua classe gramatical e, caso seja um nome de entidade, o seu tipo. O REN em um texto passa por duas fases: a anotação das classes gramaticais do texto e a efetiva anotação dos nomes com a categoria semântica.

A qualidade do algoritmo depende da capacidade do anotador para identificar corretamente os nomes, e está limitada aos tipos de entidades utilizados no corpus. Em algoritmos de aprendizagem

[12] https://www.wikidata.org/wiki/Q42, acessado em 04-01-2014.

de máquina supervisionados, espera-se de ele seja capaz de generalizar características encontradas no texto de treino – conhecido como Gold Corpus – na forma de padrões que quando identificados em outros textos levem à correta classificação das entidades. Para que um algoritmo seja capaz de aprender estes padrões ele precisa de um corpus suficientemente grande e diverso. O tamanho do corpus interfere, pois o algoritmo precisa de repetições de um padrão para ser capaz de reconhecê-lo. Conforme [Dom12], a diversidade nos conjuntos de treino e testes é necessária, pois se o algoritmo for exposto a um número muito grande de ocorrências de um mesmo padrão ele não será capaz de identificar pequenas variações neste padrão e nem será capaz de diferenciar classes distintas com padrões similares.

[RR09] defende que REN é um problema de previsão sequencial, solucionável com modelos típicos para esta classe de problemas como Hidden Markov Models, Conditional Random Fields, e Perceptrons. Também apresenta um conjunto de quatro questões chave que devem ser consideradas no desenvolvimento de um sistema de REN: Como representar sequências de palavras em um REN? Qual algoritmo de inferência utilizar? Como modelar as dependências não locais? Como utilizar recursos externos de conhecimento no REN?

Um algoritmo supervisionado bastante estudado para REN é o Conditional Random Fields (CRF) [LMP01] [ML03], cujo sucesso é atribuído à sua capacidade de reconhecer padrões sequenciais que cercam os elementos anotados como entidades nomeadas.

[AV13] realizam a tarefa de REN para a Língua Portuguesa utilizando o método CRF, obtendo boa precisão no corpus do Segundo HAREM [COM+08]. Neste trabalho as autoras ressaltam que o número reduzido de exemplos no Gold Standard para algumas categorias semânticas resultou em um modelo menos abrangente para estas categorias, afetando a Medida-F utilizada como métrica para comparação entre os métodos de REN que participaram na avaliação conjunta do Segundo HAREM [MS08].

3. TRABALHOS RELACIONADOS

Um dos primeiros trabalhos que explora a Wikipedia em tarefas de classificação e desambiguação de entidades nomeadas, [BP06] destaca quais características da Wikipedia são importantes recursos de linguagem nestas tarefas. Neste trabalho os autores descrevem de que forma elementos estruturados tais como títulos, categorias, redirecionamentos, desambiguações e *wikilinks* foram utilizados para identificar artigos da Wikipedia cujo assunto é uma entidade nomeada, e como a combinação entre categorias, redirecionamentos e desambiguações foi utilizada para atribuir uma categoria semântica para estas entidades. A maior contribuição deste trabalho está na identificação dos elementos estruturados da Wikipedia que contribuem para classificar os artigos. Importante ressaltar que este trabalho precede a criação de bases de conhecimento como a DBpedia, YAGO e Wikidata.

Posteriormente, [NRR+13] estendeu as ideias de [BP06], [Cuc07] e [RS08] para utilizar a Wikipedia na criação de um corpus multilíngue anotado. Os autores atribuem ao alto custo da anotação de corpus por especialistas a carência de métodos de alta performance para Reconhecimento de Entidades Nomeadas para a maioria dos idiomas e domínios, e apresenta uma alternativa baseada na Wikipedia para criação de um corpus multilíngue anotado. Esta alternativa baseia-se em uma recomendação de estilo de escrita para a Wikipedia que requer que ao menos a primeira menção a qualquer termo ou nome em um artigo da Wikipedia tenha um link para o artigo correspondente. Este método transforma estes *links* em anotações de Entidades Nomeadas.

O sistema de processamento da Wikipedia executa os seguintes passos:

1. Classificação de cada artigo em uma categoria semântica;
2. Propagação da classificação para os diferentes idiomas;
3. Seleção dos artigos com *links* para outros artigos;
4. Anotação de cada *link* conforme a categoria semântica do artigo alvo;
5. Mapeamento da ontologia de entidades para um esquema de Entidades Nomeadas;
6. Ajuste dos limites da entidade para adequar ao esquema de Entidades Nomeadas;
7. Seleção de trechos do artigo para inclusão em um corpus.

Para classificação dos artigos em categorias semânticas, os autores utilizam diversas características de um artigo, como o texto, a estrutura do documento, o título, os *links*, as categorias, os templates, os infoboxes e as páginas de desambiguação. O uso destas características em um modelo estatístico possibilita alta precisão na escolha da categoria semântica para classificação da entidade, e por não depender de regras linguísticas é aplicável a diferentes idiomas. As categorias semânticas utilizadas compreendem uma ontologia de 154 tipos divididos em três níveis.

Após classificar o artigo, os *links* para a mesma entidade em outros idiomas são utilizados para propagar a classificação em outros idiomas, criando um corpora multilíngue cobrindo oito idiomas além do inglês (conjunto que compreende os nove idiomas mais populares da Wikipedia). Todos os artigos com texto contendo *links* para outros artigos são selecionados para anotação destes *links* conforme a categoria semântica do artigo alvo. Para utilização do corpus para treino de classificadores para tarefas de conferências como MUC-6/7 e CoNLL, a ontologia de categorias semânticas deve ser mapeada para os schemas utilizados nestas tarefas.

A inclusão de trechos dos artigos no corpus de treino do algoritmo utiliza um critério de seleção que inclui somente sentenças onde todas as palavras capitalizadas são *links* para artigos com categoria semântica atribuída, e a sentença deve possuir ao menos uma entidade anotada. Como o estilo de escrita da Wikipedia dita que somente a primeira menção deve possuir link para o artigo mencionado, os autores desenvolveram um algoritmo para identificação de formas alternativas no mesmo artigo onde menções possuem link. São três os tipos de formas alternativas tratados:

1. O título do artigo e o título de todas as páginas de *redirect* para o artigo. Neste tipo de forma alternativa não são consideradas expressões do título posicionadas após vírgula ou entre parênteses.

2. Título de páginas de desambiguação e o texto do *link* para o artigo conforme escrito na página de desambiguação.

3. Todas as diferentes formas de texto utilizadas em *links* para o artigo.

Os autores avaliaram os *corpora* obtidos através deste método comparando o modelo construído a partir da Wikipedia com modelos construídos através de *Gold Standard* e identificaram que o resultado foi tão bom quanto estes modelos ao compará-los em conjuntos de testes não correspondentes aos *corpora* de treino. Constataram também que o resultado foi consistente em todos os idiomas. Os melhores resultados de classificação foram nas categorias semânticas de *Person*, *Location (Place)*, e *Disambiguation*, enquanto que os mais difíceis foram *Organisation* e *Misc*, pois a grafia dos nomes de entidades destas duas últimas categorias é irregular em comparação com a grafia das entidades das categorias *Person* e *Location*. Esta observação também foi consistente em todos os idiomas.

Outra razão para categorias onde a performance não foi boa está relacionada a pouca quantidade de instâncias para treino, à diversidade destas instâncias, e à ausência das características utilizadas no treino do modelo, tais como as categorias e os infoboxes.

A relevância deste trabalho está no uso da própria Wikipedia como *corpus* para treino de classificadores. Outra contribuição importante deste trabalho está nas conclusões como utilizar apenas as sentenças que possuem entidades anotadas, e seleção das sentenças apenas de artigos que mencionam outros artigos que foram identificados como entidades nomeadas. No entanto, assim como nos trabalhos anteriores mencionados, este trabalho utiliza somente a Wikipedia para identificar quais artigos descrevem entidades nomeadas, apesar da existência de bases de conhecimento relacionadas à Wikipedia.

O uso da DBpedia no processo de classificação de entidades nomeadas iniciou com o uso das informações sobre tipos das entidades da *Linked Open Data* (LOD) para melhorar métodos de classificação de entidades nomeadas [NZQW10]. Os autores consideram que os métodos de classificação de entidades nomeadas baseados em aprendizagem de máquina utilizam características linguísticas extraídas de textos de treino sendo, portanto, dependentes de contexto. Modelos são construídos a partir destas características para obtenção de conhecimento a posteriori. Características extraídas da LOD são independentes deste contexto, sendo assim conhecimento disponível a priori.

Para as características obtidas da LOD os autores criaram um mecanismo de pontuação que indica a possibilidade de uma entidade ser de um determinado tipo, utilizando as entidades, as diferentes formas de mencionar estas entidades e os diferentes tipos que uma entidade pode representar. Para identificação dos tipos correspondentes os autores utilizam uma ontologia própria que contempla os tipos utilizados com mais frequência nos datasets escolhidos.

Os diferentes nomes de uma entidade são obtidos de predicados específicos para nomear entidades na LOD, rdfs:label e foaf:name. Os tipos são obtidos do predicado rdf:type. Nomes alternativos são obtidos a partir de relacionamentos comuns na LOD, que são os redirects e disambiguates da Wikipedia registrados na DBpedia, mais os predicados owl:sameAs, utilizados para declarar que diferentes URIs referem-se à mesma coisa. Estes dados são consolidados em um mecanismo de armazenamento onde todos os pares $< nome, tipo >$ são armazenados em uma estrutura de índice inversa. Considerando um exemplo mais complexo que envolve diferentes tipos, ambiguidade, e redirecionamentos, o nome Garibaldi possui as seguintes desambiguações na DBpedia [1]:

Nome	**Origem**	**Classes da DBpedia**
Giuseppe Garibaldi Garibaldi	label, name disambiguates:name	Thing, Agent, Person
Garibaldi Alves Filho Garibaldi Alves Garibaldi	label, name redirect:label disambiguates:name	Thing, Agent, Person, Politician
Garibaldi	label, name	Thing, Place, PopulatedPlace, Settlement, City
4317 Garibaldi Garibaldi	label name	Thing, CelestialBody, Asteroid
Garibaldi (Igrejinha) Garibaldi	label disambiguates:name	Thing
Garibaldi (Oregon) Garibaldi	label name	Thing, Place, PopulatedPlace
Garibaldi (Santa Fé) Garibaldi	label disambiguates:name	Thing
Associação Garibaldi de Esportes Garibaldi	label name	Thing, Agent, Organization, SportsTeam, SoccerClub

Tabela 3.1: Diferentes entidades, de diferentes classes, possuem o mesmo nome.

[1] http://pt.dbpedia.org/page/Garibaldi_(desambiguaç~ao), acessado em 26/12/2013.

Neste exemplo somente o nome Garibaldi possui 13 tipos distintos. Garibaldi Alves Filho, por sua vez, possui três nomes diferentes para quatro tipos diferentes. Os autores consideram que cada nome corresponde a uma instância, então o nome Garibaldi possui oito instâncias diferentes, com cada instância variando entre um e cinco tipos.

Para cálculo da pontuação é necessário consultar o nome da entidade no mecanismo de armazenamento e identificar os tipos correspondentes ao tipo alvo. A consulta da entidade é feita a partir do nome reconhecido no texto, buscando este nome no conjunto de nomes armazenados.

No exemplo da tabela 3.1, assumindo que Organization, Place e Person são disjuntos, a consulta ao nome Garibaldi e tipo alvo Person atribuirá -1 à cidade de Garibaldi e ao clube de futebol, e $+1$ para Giuseppe Garibaldi e Garibaldi Alves Filho. Considerando que Person está a dois passos de distância de Thing, Garibaldi (Santa Fé) e Garibaldi (Igrejinha) receberão pontuação de 0.5. 4317 Garibaldi, por sua vez, receberá pontuação de 0.25, pois Asteroid está a quatro passos de distância de Person utilizando Thing como ancestral comum e, não há relação de disjunção estabelecida com Person. A estratégia agressiva atribuirá a pontuação de $+1$ para o nome Garibaldi e tipo alvo Person e as características fornecidas para o método de aprendizagem de máquina serão $+1$ para a probabilidade de o tipo alvo estar em conformidade com a LOD e zero para a probabilidade de o tipo alvo estar em conflito com a LOD.

Uma tentativa (possivelmente errônea) de classificar Garibaldi com a categoria semântica de Event, disjunta de Person, Place e Organization, resultará em zero para a probabilidade do tipo alvo estar em conformidade e $+1$ para o tipo alvo estar em conflito com a LOD, pois a maioria dos pontos calculados terá valor de -1.

Apesar deste trabalho não aplicar as asserções da DBpedia diretamente no *corpus* de treino, sua contribuição no uso das classes de ontologias tais como a DBpedia como *features* no processo de classificação, e as diferentes formas de mencionar as instâncias de classes a partir de atributos da DBpedia presentes também na Wikipedia, sugerem a possibilidade de conectar as classes da ontologia diretamente nos textos da Wikipedia, utilizando os wikilinks e títulos da Wikipedia. Os títulos e wikilinks são a origem dos nomes encontrados em propriedades de instâncias da DBpedia como *label*, *name*, *redirect:label* e*disambiguates:name*, possibilitando a conexão entre as menções presentes nos wikilinks e as classes presentes na DBpedia.

4. GERAÇÃO DE *CORPUS* PARA O RECONHECIMENTO DE ENTIDADES NOMEADAS A PARTIR DA WIKIPEDIA

Para ilustrar o propósito deste trabalho, tomemos como exemplo a figura 4.1 abaixo, contendo a primeira sentença do artigo sobre o Brasil na Wikipedia em Português[1].

Brasil, oficialmente **República Federativa do Brasil**, é o maior país da América do Sul e da região da América Latina, sendo o quinto maior do mundo em área territorial (equivalente a 47% do território sul-americano) e população (com mais de 201 milhões de habitantes).

Figura 4.1: Exemplo extraído do primeiro parágrafo do artigo sobre o Brasil.

Este primeiro parágrafo apresenta ligações (wikilinks) com outros 5 artigos da Wikipedia mencionados pelas seguintes formas textuais:

- país, referencia https://pt.wikipedia.org/wiki/País
- América do Sul, referencia https://pt.wikipedia.org/wiki/América_do_Sul
- América Latina, referencia https://pt.wikipedia.org/wiki/América_Latina
- área territorial, referencia https://pt.wikipedia.org/wiki/Lista_de_países_e_territórios_por_área
- população, referencia https://pt.wikipedia.org/wiki/Lista_de_países_por_populaç~ao

Em um *corpus* para classificação de entidades nomeadas, espera-se do parágrafo ilustrado acima que ele apresente uma anotação de Lugar para as entidades Brasil, República Federativa do Brasil, América do Sul e América Latina, pois todas elas são lugares. Destas, América do Sul e América Latina possuem ligação com outros artigos, enquanto que Brasil e República Federativa do Brasil não possuem ligação por serem o assunto do próprio artigo. Por outro lado, os termos **país** , **área territorial** e **população** não devem figurar com anotações por não serem entidades. Infelizmente os wikilinks não possuem propriedades semânticas que possibilitem esta distinção, mas é possível consultar uma base de conhecimento como a DBpedia para filtrar as entidades e anotá-las com a categoria semântica adequada. Como a DBpedia é construída com base na Wikipedia, há uma relação única e bidirecional entre cada recurso da DBpedia e seu correspondente original na Wikipedia. A tabela 4.1 abaixo apresenta esta relação para os wikilinks em questão:

Todo recurso da DBpedia é instância de alguma classe, seja esta uma classe implícita (a classe raiz *Thing*) ou explícita - declarada a partir do extrator *Mapping-Based Infobox Extraction*. A tabela 4.2 abaixo apresenta a classe declarada para cada uma das instâncias previamente introduzidas:

Recursos que são instâncias somente da classe raiz *Thing* devem ser desconsiderados por não pertencerem a nenhuma das categorias semânticas de forma explícita. Desta forma, os wikilinks

[1] http://pt.wikipedia.org/wiki/Brasil, acessado em 08-04-2014.

Tabela 4.1: Relação entre os artigos da Wikipedia e os recursos correspondentes na DBpedia

Wikipedia	**DBpedia**[2]
pt.wikipedia.org/wiki/País	dbpedia:País
pt.wikipedia.org/wiki/América_do_Sul	dbpedia:América_do_Sul
pt.wikipedia.org/wiki/América_Latina	dbpedia:América_Latina
pt.wikipedia.org/wiki/Lista_de_países_e_territórios_por_área	dbpedia:Lista_de_países_e_territórios_por_área
pt.wikipedia.org/wiki/Lista_de_países_por_populaç~ao	dbpedia:Lista_de_países_por_populaç~ao

Tabela 4.2: Recursos da DBpedia e suas classes

Instância	**Classe**
http://pt.dbpedia.org/resource/País	*Thing*
http://pt.dbpedia.org/resource/América_do_Sul	*AdministrativeRegion*
http://pt.dbpedia.org/resource/América_Latina	*AdministrativeRegion*
http://pt.dbpedia.org/resource/Lista_de_países_e_territórios_por_área	*Thing*
http://pt.dbpedia.org/resource/Lista_de_países_por_populaç~ao	*Thing*

país, área territorial e população serão tratados como palavras sem semântica associada. É comum na Wikipedia encontrarmos nomes de entidades sem menção a artigos que definam estas entidades. Esta situação pode ocorrer quando não existe um artigo que defina a entidade naquela linguagem ou quando o autor do artigo não introduziu o wikilink para o artigo, como ocorre com "Brasil" e "República Federativa do Brasil" . Sem a aplicação de uma fase de reconhecimento destas entidades nos textos dos artigos não é possível anotá-los como entidades nomeadas. Esta fase está fora do escopo deste trabalho.

Os outros recursos restantes, América do Sul e América Latina, pertencem ambos à classe *AdministrativeRegion*. Conforme apresentado na figura 4.2 abaixo, que representa um subconjunto da relação entre classes da DBpedia 3.9, a classe *AdministrativeRegion* é uma subclasse da classe *Place*. Desta forma, América do Sul e América Latina devem ser anotadas com a categoria semântica *Place*.

Utilizando como exemplo o formato de anotação definido no CoNLL 2003, o trecho "é o maior país da América do Sul" será anotado conforme exemplo abaixo, servindo como *corpus* de treino para sistemas de classificação que suportem este formato de entrada.

```
é       O
o       O
maior   O
país    O
da      O
América PLACE
do      PLACE
Sul     PLACE
```

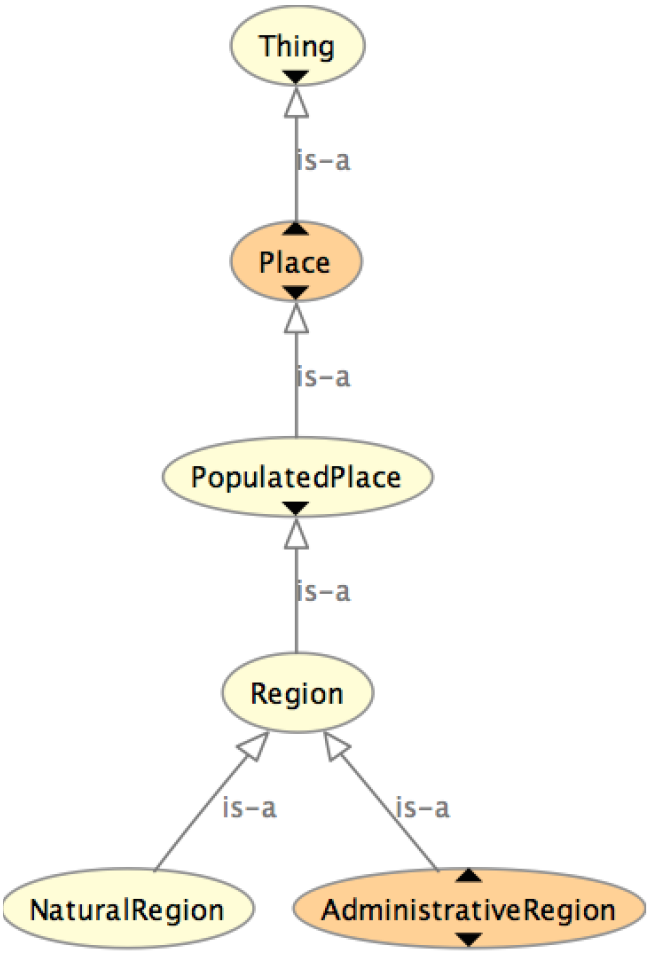

Figura 4.2: Hierarquia parcial de classes para Lugares (*Place*), destacando a relação de subsunção entre a classe *Place* e a classe *AdministrativeRegion*

5. IMPLEMENTAÇÃO

A construção de um *corpus* de treino para classificação de entidades nomeadas a partir da Wikipedia e da DBpedia é composta por uma sequência de passos que visam relacionar títulos e wikilinks da Wikipedia com instâncias e classes da DBpedia para permitir a seleção de sentenças que possuem exemplos de entidades pertencentes às categorias semânticas de interesse. Sendo a Wikipedia uma base de dados textual e a DBpedia uma base de dados estruturada, é necessário ligá-las de uma forma que preserve a estrutura das sentenças da Wikipedia mas que possibilite a aplicação de filtros e relacionamentos com elementos da DBpedia. Dado que os artigos da Wikipedia são escritos com uma linguagem de marcação, é necessário processar cada artigo para extração do texto preservando somente marcações de wikilinks. Além disso, os wikilinks referenciam os artigos da Wikipedia através do seu título, apesar de cada artigo possuir uma URL única na linguagem em que foi escrito na Wikipedia. Tanto as referências dos wikilinks quanto os títulos devem ser convertidos para o mesmo formato de IRI (Internationalized Resource Identifier) da DBpedia. Como o objetivo final é a extração de sentenças que possuem exemplos das categorias semânticas desejadas, o algoritmo projetado neste trabalho prioriza a extração do subconjunto das instâncias da DBpedia cuja classe equivale a uma das categorias semânticas desejadas, descartando qualquer relação que não possua exemplos de entidades destas categorias. Durante toda a implementação foram utilizadas somente linguagens de programação, bibliotecas e *frameworks* disponíveis sob licença *open source*, assegurando a disponibilidade de todos os recursos necessários para reprodução da construção do *corpus*. O código-fonte do projeto está disponível em https://github.com/crisweber/ner-corpus-construction, sem restrições para uso.

A relação abaixo sumariza os passos necessários, e as seções seguintes detalham a implementação destes passos.

1. Obter os *datasets* da Wikipedia e da DBpedia.

2. Selecionar os artigos relevantes de acordo com as categorias semânticas desejadas.

3. Transformar o código fonte dos artigos em textos sem marcação.

4. Substituir os wikilinks por anotação da instância da DBpedia correspondente.

5. Remover wikilinks de palavras que não referenciam instâncias das categorias semânticas desejadas.

6. Substituir as referências às instâncias pela categoria semântica.

7. Remoção de sentenças sem anotações.

5.1 Obtenção dos *datasets* da Wikipedia e da DBpedia

A construção do *corpus* apresentado neste trabalho teve como base a versão 3.9 da DBpedia e uma cópia dos artigos da *Portuguese Wikipedia* em formato XML com data de geração de 05/10/2013. A versão 3.9 da DBpedia foi gerada a partir de cópia dos artigos da Wikipedia exportados entre Março e Abril de 2013. Não foi possível obter cópia dos artigos deste período porque a Wikimedia disponibiliza somente as cópias mais recentes da Wikipedia, e na ocasião do início deste trabalho a cópia mais próxima foi a de 02/05/2013. Uma cópia foi preservada na Web, no serviço de armazenamento durável Amazon S3 [1] para assegurar a repetição da construção do *corpus*, e está disponível através da seguinte URL: https://s3.amazonaws.com/ner-silvercorpus/pt/wikimedia/ptwiki-20131005-pages-articles.xml.

5.1.1 Pré-processamento do formato de exportação da Wikipedia

O formato XML empregado nas cópias dos artigos da Wikipedia, apesar de seus benefícios para interoperabilidade, dificulta o processamento do conteúdo do artigo. O formato ideal para a construção requer somente o título do artigo e o conteúdo, enquanto que no formato XML há elementos descrevendo metadados do artigo e elementos que possibilitam agrupar todos os artigos em um único XML. Além disto, o agrupamento de todos os artigos em um único arquivo XML impõe limitações na manipulação deste arquivo em função do tamanho. Para transformação do formato XML em um formato mais adequado, e também para contornar as limitações no processamento de arquivos XML grandes, o projeto Apache Mahout [2] disponibiliza o utilitário **seqwiki**, que converte os formatos XML de exportação da Wikimedia em arquivos no formato SequenceFile [3]. O formato SequenceFile resultante é composto por todos os artigos presentes no XML, organizados em pares chave e valor, onde a chave é o título do artigo e o valor é o código-fonte do artigo. Apesar de ser um único arquivo, seu formato é suportado por ambientes de processamento distribuído como o Apache Hadoop [4] e o Apache Spark [5]. O utilitário foi utilizado da seguinte forma:

```
mahout seqwiki -all -i <caminho do XML> -o <caminho do SequenceFile>
```

Com estes parâmetros, o utilitário extrai todos os artigos presentes no arquivo XML no caminho informado no parâmetro '-i' e gera o SequenceFile no caminho informado no parâmetro '-o'. Este utilitário deve ser executado no ambiente de processamento distribuído do Apache Hadoop, e

[1] http://aws.amazon.com/pt/s3/, acessado em 17-12-2014
[2] http://mahout.apache.org, acessado em 22-09-2014
[3] http://hadoop.apache.org/docs/current/api/org/apache/hadoop/io/SequenceFile.html, acessado em 22-09-2014
[4] http://hadoop.apache.org, acessado em 22-09-2014
[5] http://spark.apache.org, acessado em 22-09-2014

os caminhos do XML e do SequenceFile devem ser caminhos em sistemas de arquivos suportados pelo Hadoop, como o HDFS [6] ou o S3.

5.1.2 *Datasets* da DBpedia

Os seguintes *datasets* da DBpedia são necessários para geração do *corpus*. É importante que o formato dos arquivos seja o Turtle[7] (ttl), pois o formato é interpretado diretamente no processo de geração do *corpus*, sem suporte a outros formatos de RDF.

- O *dataset* `instance_types_pt.ttl` [8] possui a relação das instâncias da DBpedia com as suas classes. Deste *dataset* são extraídas as classes para anotação das entidades nomeadas do *corpus* da Wikipedia.
- O *dataset* `page_links_pt.ttl` [9] possui a relação entre instâncias presentes na Wikipedia na forma de wikilinks entre os artigos correspondentes. Neste *dataset*, no entanto, os artigos são identificados através da IRI da instância na DBpedia. Este *dataset* é utilizado em conjunto com o `instance_types_pt.ttl` para remover os artigos que não fazem referência às instâncias das classes correspondentes às categorias semânticas desejadas.
- O *dataset* `redirects_transitive_pt.ttl` [10] possui a relação de redirecionamento de artigos da Wikipedia, identificados através da IRI das instâncias. Este *dataset* é utilizado para identificar a classe das instâncias que são redirecionamentos para outras instâncias.
- O *dataset* `wikipedia_links_pt.ttl` [11] possui a relação da URL da Wikipedia de cada instância da DBpedia, e é utilizado na identificação das instâncias da DBpedia de acordo com a URL da Wikipedia.

5.2 Seleção dos artigos relevantes

A seleção dos artigos relevantes ocorre através da combinação das relações presentes nos *datasets* da DBpedia expostos na seção anterior. Neste ponto, o uso da DBpedia elimina a necessidade de processar o conteúdo dos artigos para decidir quais são relevantes, reduzindo a necessidade de processamento dos artigos para somente aqueles que possuem exemplos das categorias semânticas desejadas. A sequência de operações relacionais abaixo descreve a seleção dos artigos,

[6] http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html, acessado em 22-09-2014
[7] http://www.w3.org/TR/2012/WD-turtle-20120710/, acessado em 22-09-2014
[8] http://wiki.dbpedia.org/Downloads39?v=2zd#mapping-based-types
[9] http://wiki.dbpedia.org/Downloads39?v=2zd#wikipedia-pagelinks
[10] http://wiki.dbpedia.org/Downloads39?v=2zd#transitive-redirects
[11] http://wiki.dbpedia.org/Downloads39?v=2zd#links-to-wikipedia-article

sendo que as relações InstanceTypes, PageLinks, RedirectsTransitive e WikipediaLinks são, respectivamente, `instance_types_pt.ttl`, `page_links_pt.ttl`, `redirects_transitive_pt.ttl` e `wikipedia_links_pt.ttl`. Todas as relações são compostas por $(subject, predicate, object)$.

Esta sequência de operações relacionais realiza a seguinte série de seleções e filtros sobre os *datasets* da DBpedia:

1. seleciona o subconjunto *InstancesSelectedClasses* das instâncias que pertencem às classes *Organisation*, *Person* ou *Place*. Importante observar a grafia de *Organisation* segundo o inglês europeu, utilizado na DBpedia.

2. seleciona o subconjunto *RedirectClasses* de redirecionamentos que referenciam este subconjunto de instâncias.

3. cria uma nova relação *InstancesClasses* a partir da união dos subconjuntos de instâncias (*InstancesSelectedClasses*) e redirecionamentos (*RedirectClasses*).

4. seleciona o subconjunto *ArticlesWithSelectedClasses* de instâncias cujos artigos na Wikipedia (*PageLinks*) referenciam instâncias do subconjunto anterior (*InstancesClasses*).

Apesar da existência de *parsers* e bibliotecas de consulta sobre RDF, a implementação destas operações relacionais foi realizada com operações de transformação sobre coleções na interface de programação do ambiente de processamento distribuído Apache Spark [ZCD+12]. Esta decisão se justifica ao utilizar o mesmo ambiente distribuído para os demais passos da geração do *corpus*, onde as relações *InstancesClasses* e *ArticlesWithSelectedClasses* são utilizadas. No Apache Spark, cada *dataset* da DBpedia foi carregado em uma área de memória compartilhada distinta, chamada *RDD* (*Resilient Distributed Dataset*), onde as operações subsequentes resultaram em *RDDs* intermediários até a obtenção dos *datasets* finais, *InstancesClasses* e *ArticlesWithSelectedClasses*. Estes dois *datasets* permanecem disponíveis nas áreas de memória compartilhada para uso posterior na transformação dos artigos remanescentes e anotação das entidades nomeadas.

5.3 Transformação do código-fonte dos artigos

Para a transformação do código-fonte dos artigos em textos com marcações dos wikilinks é necessário aplicar um *parser* que converta o texto com marcações da *Markup Language* da Wikipedia em uma hierarquia de objetos tipados de acordo com as marcações, para assim selecionar as marcações que devem permanecer - os wikilinks - e remover todas as demais marcações. Elementos com apelo visual tais como tabelas, listas, códigos-fonte em outras linguagens, e expressões matemáticas são descartados por serem elementos que adicionam ruído no treino de classificadores [NRR+13]. A biblioteca *Bliki engine* [12] possui conversores da *Markup Language* para HTML e para texto sem marcações. Esta biblioteca foi construída sobre um mecanismo extensível de renderização

[12] https://bitbucket.org/axelclk/info.bliki.wiki, acessado em 17-12-2014

baseado nestes conversores, que recebem a hierarquia de objetos criada pelo *parser* e retornam o texto final convertido. O conversor para texto sem marcações serviu como base para criação de um conversor similar que remove todas as marcações, exceto as de wikilinks.

A operação tem como entrada o arquivo SequenceFile resultante do pré-processamento do arquivo XML contendo os artigos da Wikipedia, o *RDD ArticlesWithSelectedClasses*, e o *RDD InstancesClasses*. O SequenceFile é carregado em um RDD do Spark, filtrado para descartar páginas de redirecionamento, e então é gerado um novo *RDD* intermediário, com o subconjunto dos artigos que estão contidos no *RDD ArticlesWithSelectedClasses*. Após, o conversor descrito acima é aplicado no código-fonte dos artigos para extração do texto com marcações de *wikilinks*, resultando em um *RDD* chamado *PlainArticles*.

Nesta operação de transformação do código-fonte dos artigos ocorre também a transformação dos títulos dos artigos em IRIs da DBpedia, seguindo as regras de conversão definidas para a DBpedia [13]. O título do artigo é transformado para relacionar com o *RDD ArticlesWithSelectedClasses* e obter o subconjunto dos artigos cujas sentenças farão parte do *corpus* anotado. No momento da geração do novo marcador, os títulos dos artigos referenciados nos wikilinks são transformados nas IRIs correspondentes da DBpedia.

5.4 Anotação das entidades

Na etapa final, de anotação das entidades, um novo *RDD* intermediário é construído a partir dos textos presentes no *RDD PlainArticles*. Neste *RDD*, chamado *Mentions*, cada texto da Wikipedia é substituído por uma tupla contendo este texto e o conjunto de todos os wikilinks presentes no texto. Estes conjuntos são relacionados com o *RDD InstanceClasses* para criação de um dicionário cujas chaves são os wikilinks presentes nos textos selecionados e o valor é a entidade nomeada anotada com a classe encontrada no *RDD InstanceClasses*. O dicionário é combinado com o *RDD Mentions* para substituição dos wikilinks pelas entidades nomeadas anotadas. Wikilinks sem correspondência são removidos, permanecendo somente o texto do wikilink. Finalmente, cada texto anotado é convertido numa lista de sentenças utilizando o Stanford NLP [MSB+14]. Esta lista é filtrada para remoção de sentenças sem entidades nomeadas anotadas, e as demais sentenças são combinadas e escritas em um *dataset* que pode ser utilizado como *corpus* anotado para treino de classificadores de entidades nomeadas.

5.5 Execução

Todo o processamento da construção do *corpus* ocorreu em servidores cloud na Amazon AWS. Este processo utilizou 5 instâncias do tipo m3.xlarge, dotadas de 4 unidades de processamento e 15GiB de memória RAM. A etapa inicial, de pré-processamento do *dump* da Wikipedia, durou 15

[13] http://wiki.dbpedia.org/URIencoding, acessado em 22-09-2014

minutos. As etapas de seleção dos artigos, *parse* da Markup Language e anotação das entidades, combinadas, duraram 5 minutos.

Deste processamento resultou um arquivo com pouco mais de 1,5 milhões de sentenças anotadas, contendo aproximadamente 2 milhões de *tokens* anotados como entidades do tipo *PLACE*, 1,25 milhões de *tokens* de entidades do tipo *PERSON*, e mais de 760 mil entidades do tipo *ORGANISATION*.

Em números exatos:

- 1.508.524 sentenças anotadas
- 2.215.575 *tokens* anotados como *Place*
- 1.252.180 *tokens* anotados como *Person*
- 768.642 *tokens* anotados como *Organisation*

6. EXPERIMENTOS

6.1 Introdução

Neste capítulo relatamos os experimentos realizados para avaliar a aplicabilidade do *corpus* da Wikipedia anotado automaticamente na tarefa de classificação de entidades nomeadas. A avaliação foi dividida em três experimentos:

- Criação de *corpora* para treino de classificadores,
- Comparação dos classificadores com o *corpus* do Primeiro Harem,
- Investigação do efeito do uso de diferentes estilos de escrita nos *corpora* de treino e de teste dos classificadores.

Em função do elevado número de sentenças anotadas produzidas pelo método proposto de geração de *corpus* anotado, foi necessário estabelecer um número máximo de sentenças para o *corpus*. No entanto, o número de sentenças pode ser determinante para a qualidade do *corpus* na tarefa de treino do classificador, pois o número de exemplos presentes em um conjunto de treino afeta o resultado [Dom12]. Além disso, o conteúdo das sentenças selecionadas também influencia na qualidade do classificador. Desta forma, foi necessário um experimento para seleção das sentenças que farão parte de um *corpus*.

Para validação da aplicabilidade de um *corpus* anotado no treino de um classificador de entidades nomeadas, é necessário utilizar um *corpus* de testes para comparação do desempenho da classificação produzida por este corpus com a classificação produzida a partir de outro *corpus* alternativo de treino. Para a língua portuguesa existem dois *corpora* amplamente estudados e que são aplicáveis para este experimento: o *corpus* do Primeiro Harem [SC07] e o *corpus* do Segundo Harem [MS08]. O *corpus* do Segundo Harem foi utilizado como *corpus* de teste, e o *corpus* do Primeiro Harem foi utilizado como *corpus* de treino para o classificador alternativo.

Por fim, o efeito do estilo de escrita empregado nos dois *corpora* também foi investigado em um experimento. O efeito do estilo de escrita do *corpus* na tarefa de classificação foi identificado por [PK01], e os estilos utilizados no Primeiro e Segundo Harem diferem dos estilos utilizados na Wikipedia, sendo necessário avaliar o seu impacto. Este impacto foi medido comparando o desempenho de dois classificadores, treinados com o Primeiro Harem e com o *corpus* da Wikipedia, testados utilizando o *corpus* do Segundo Harem e um segundo *corpus* extraído da Wikipedia. Concluindo os experimentos, foram realizadas avaliações de robustez entre os diferentes estilos de escrita através de combinações entre os dois *corpora* de treino.

Em todos os experimentos, as categorias semânticas avaliadas foram Pessoa (*PERSON*), Organização (*ORGANISATION*) e Lugar (*PLACE*). A escolha destas três categorias foi influenciada pela abrangência destas categorias tanto na Wikipedia quanto nos *corpora* do Primeiro Harem e do

Segundo Harem. Outro fator importante foi que categorias presentes no Harem como Tempo e Valor não possuem correspondente anotado com wikilinks na Wikipedia, enquanto que categorias como Abstração e Outro demandariam um estudo mais aprofundado das classificações realizadas pelo extrator da DBpedia, não sendo objetivo deste trabalho. As tabelas e figuras estão em inglês por restrições encontradas no uso do português na ferramenta utilizada durante as análises, a linguagem R ([R C14]). Os termos classificador, modelo, *corpus*/*corpora* e *datasets*, embora tenham significados diferentes, são utilizados de forma intercambiável durante a descrição dos experimentos. Isto ocorre devido ao fato do *corpus* (ou *dataset*) ser o conjunto de treino dos classificadores para geração de um modelo que, combinado com o algoritmo de classificação, é utilizado para classificar outros *datasets*.

6.2 Experimento I - Criação de *corpora* para treino de classificadores

6.2.1 *Corpora*

Para determinar o número de sentenças que farão parte do *corpus*, foram criados 16 diferentes *datasets*, cada um com 20 *corpora* gerados com pseudo aleatorização para possibilitar a replicação do experimento. Cada *dataset* possui um número fixo de sentenças por *corpus*, partindo de 500 até 8000 sentenças, com incremento de 500 entre cada *dataset*. O script para geração das amostras pode ser encontrado no ANEXO A. A tabela 6.1 apresenta o sumário da quantidade de *tokens* de cada conjunto gerado. A proximidade entre os mínimos e máximos, bem como a proximidade entre as médias e medianas, mostram uma distribuição normal onde há pouca variação na quantidade de *tokens* por sentença nos *corpora* gerados.

A figura 6.1 abaixo ilustra a distribuição da quantidade de classes por conjunto em um gráfico de BoxPlot. Neste gráfico é possível observar que a proporção de entidades nomeadas por classe mantém-se através dos conjuntos e *corpora*. Cada par de dois gráficos possui escala diferente no eixo X para dar destaque à proporção. Importante ressaltar o achatamento dos gráficos, indicando que há pouca variação na quantidade de instâncias por classe dentro dos *corpora* de cada conjunto. As maiores variações observadas estão na classe *PLACE*. Os pontos pretos representam *outliers* dentro de cada classe.

A figura 6.2 ilustra o crescimento linear do número de instâncias por classe de acordo com o aumento no número de sentenças no *corpora*. Percebe-se que dentre as entidades anotadas nos *corpora* da Wikipedia há um maior número de exemplos de Lugares, seguido por Pessoas e com um menor número de menções à Organizações. O grau de crescimento do número de exemplos em função do número de sentenças aumenta de forma mais acentuada para Lugares, também seguido de Pessoas, e com um crescimento menos acentuado do número de menções à Organizações. A figura 6.3 apresenta, em forma de equação, as relações lineares observadas. Nos três casos a quantidade

Tabela 6.1: Sumário da quantidade de *tokens* por conjunto. Cada conjunto consiste em 20 diferentes grupos de sentenças extraídas da Wikipedia de forma aleatória. Os valores apresentados para as medidas de mínimo, média, mediana e máximo são relativos a cada conjunto de 20 grupos de sentenças

Dataset	*Min*	*Average*	*Median*	*Max*
500	14.354	15.374	15.404	16.762
1000	29.179	30.724	30.610	32.904
1500	44.223	45.811	45.603	49.071
2000	59.363	60.839	60.758	62.118
2500	74.074	76.378	76.448	78.477
3000	88.143	91.736	91.258	97.519
3500	104.783	106.790	106.344	109.440
4000	119.792	121.900	121.922	125.046
4500	135.183	136.848	136.766	139.925
5000	149.934	152.860	152.780	155.476
5500	165.220	168.242	167.962	174.274
6000	179.385	182.668	182.538	186.451
6500	195.206	198.610	198.339	201.910
7000	209.429	213.090	213.076	216.763
7500	225.585	229.142	228.837	233.418
8000	242.036	244.899	244.200	248.058

de sentenças é uma variável preditora para a quantidade de exemplos, com $p - value < 0.001$, o que sugere que esta distribuição se mantém no conjunto de sentenças da Wikipedia.

6.2.2 Treino dos classificadores

O classificador utilizado nos experimentos é o Stanford NER[1] [FGM05]. Este classificador é uma implementação de CRF de cadeias lineares (*linear chains Conditional Random Fields*) que já foi aplicado em diferentes linguagens, como Alemão, Chinês, Espanhol e Inglês. O Stanford NER suporta formatos similares ao da *CoNLL 2003 NER Shared Task* [TD03], sendo o formato utilizado nestes experimentos uma versão reduzida do formato original. O formato reduzido utilizado nos experimentos é composto por dois campos: a palavra e a categoria semântica. A notação utilizada para descrever a categoria semântica compreende as três categorias semânticas avaliadas e "O" para categorizar palavras e símbolos que não são entidades nomeadas. Esta notação é mais simples que outras notações comumente empregadas, como BIO e BILOU [RR09], pois em textos da língua portuguesa são raros os casos em que diferentes exemplos de uma mesma categoria semântica aparecem em sequência, sem outro símbolo separando os exemplos. As características (*features*) utilizadas no treino dos classificadores estão no ANEXO C.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml, acessado em 02/11/2014

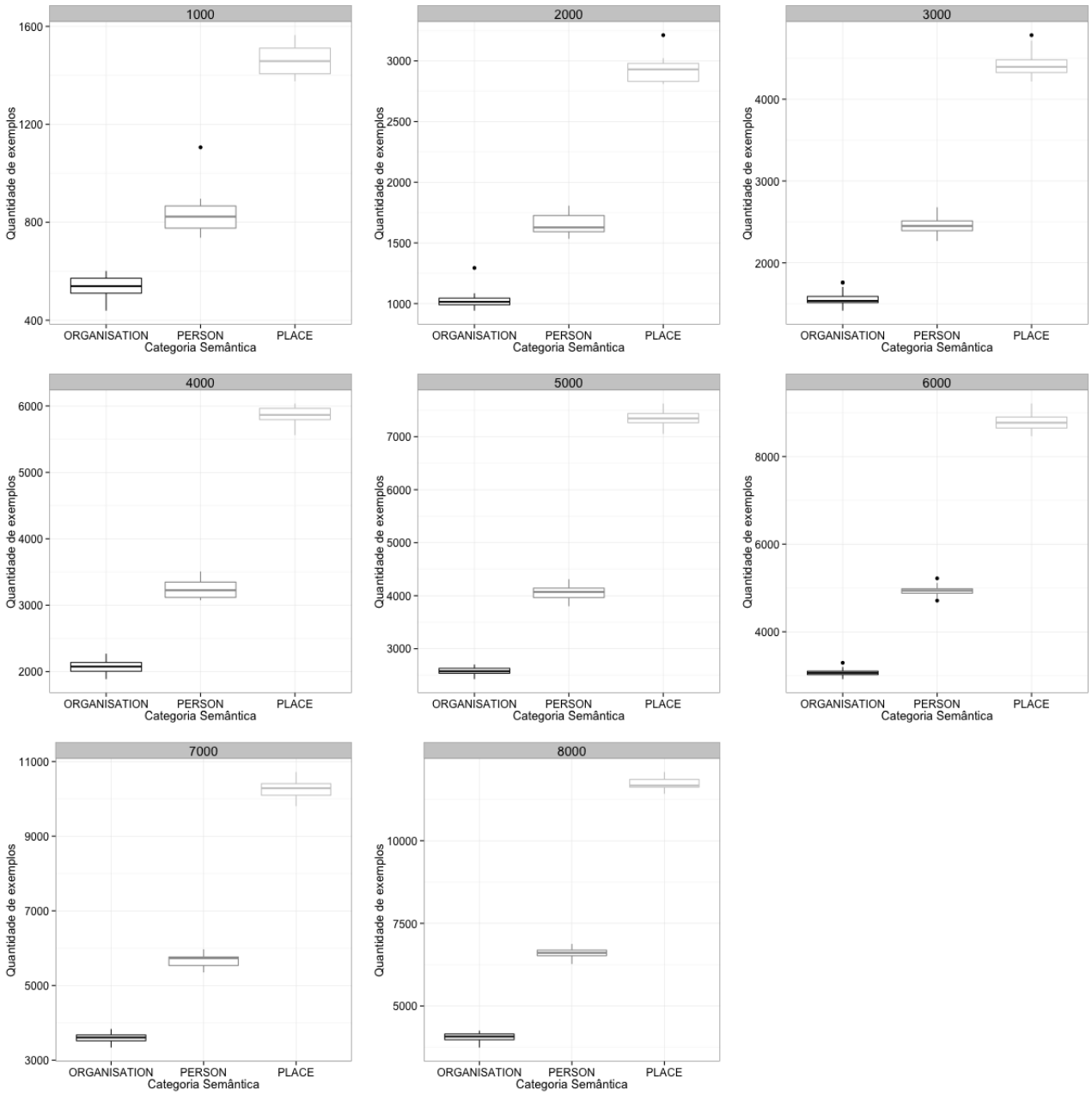

Figura 6.1: Quantidade de exemplos de categorias semânticas, por conjunto. Cada gráfico apresenta o boxplot da quantidade de exemplos de cada uma das três categorias semânticas pela quantidade de sentenças do conjunto, variando de 1000 até 8000 sentenças em incrementos de 1000. Cada conjunto é formado por 20 diferentes grupos de sentenças, todas extraídas da Wikipedia de forma aleatória. Cada sentença possui ao menos uma entidade nomeada anotada na forma de wikilink. Os gráficos foram posicionados lado a lado para destacar que a proporção entre o número de exemplos se mantém com o aumento do número de sentenças.

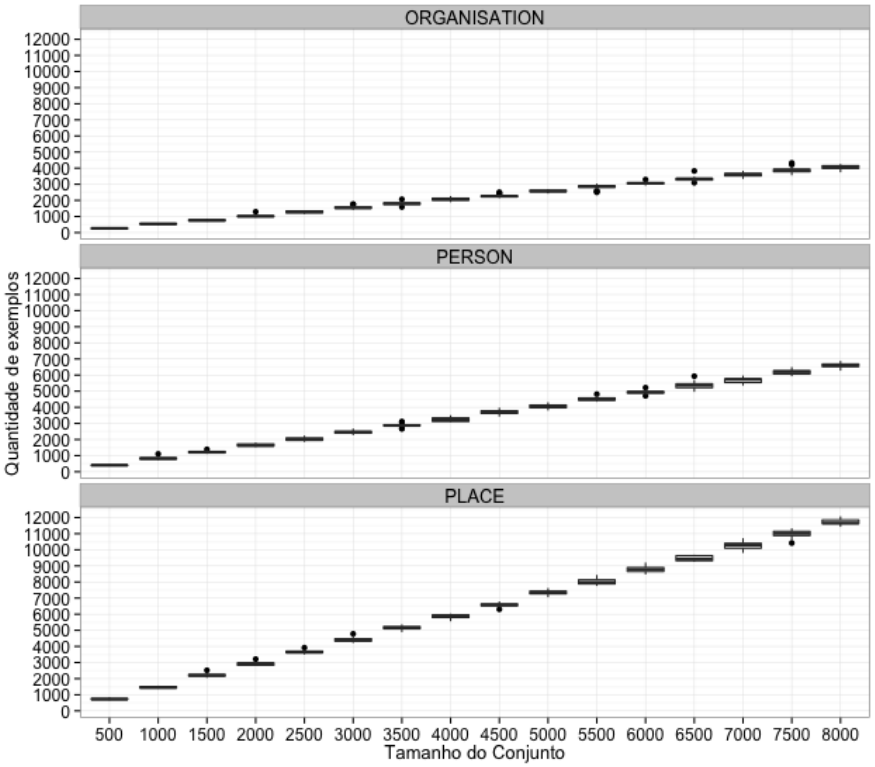

Figura 6.2: Crescimento do número de exemplos das categorias semânticas pelo tamanho do conjunto, para cada categoria semântica, em conjuntos de sentenças obtidos da Wikipedia com entidades nomeadas anotadas na forma de wikilinks de forma aleatória. O crescimento do número de exemplos de acordo com o número de sentenças sugere uma relação linear entre estas duas medidas.

$$numberOfOrganisations = 0.5131 \cdot NumberOfSentences$$
$$numberOfPersons = 0.8208 \cdot NumberOfSentences$$
$$numberOfPlaces = 1.4646 \cdot NumberOfSentences$$

Figura 6.3: Modelos das relações lineares que estimam o número de exemplos de uma categoria semântica a partir do número de sentenças. Esta relação foi observada em conjuntos variando de 500 a 8000 sentenças da Wikipedia, extraídas aleatoriamente, e contendo entidades nomeadas anotadas na forma de wikilinks. Os coeficientes confirmam o desbalanceamento da quantidade de exemplos de cada classe.

O treino dos classificadores foi realizado em servidores *cloud* na Amazon AWS, em instâncias do tipo **c3.8xlarge**. Estas instâncias são otimizadas para computação, possuindo CPU da família Intel Xeon E5-2680 v2, mais rápida que os demais tipos e contando com 32 *threads* de processamento, unidades de disco SSD para operações de leitura e escrita de arquivos mais rápida, e 60GiB de memória RAM. Uma única instância treinou um classificador para cada arquivo de treino de cada um dos 20 *corpora* nos 16 conjuntos descritos acima, resultando em um total 320 classificadores para avaliação. Após algumas execuções do Stanford NER foi possível observar que este processo utiliza uma única *thread* durante boa parte do processo de treino do classificador, apesar da eventual alocação de mais threads em alguns momentos. Também foi possível observar que os conjuntos com mais de 5000 sentenças precisam de quantidade muito maiores de memória para o processo. Desta forma, a primeira alocação de recursos para processamento em paralelo dos classificadores foi de acordo com a tabela 6.2. Nesta tabela também estão presentes o tempo médio para treino de um classificador com o número de sentenças do conjunto, e a duração total para treino de 20 classificadores pertencentes ao conjunto.

Tabela 6.2: Treino dos classificadores - primeira alocação. Esta distribuição possibilita o treino dos classificadores em um único computador com 60GiB de memória. O tempo total de uso deste computador para treino dos classificadores desta alocação corresponde à duração total do treino de 20 classificadores a partir de 20 diferentes grupos de 8000 sentenças da Wikipedia.

Number of Sentences	*Memory*	*Duration (Average)*	*Duration (Total)*
1000	2GB	1min 50sec	36min 40sec
2000	4GB	4min 3sec	1h 21min
3000	4GB	7min 45sec	2h 35min
4000	4GB	11min 3sec	3h 41min
5000	8GB	14min 34sec	4h 51min
6000	8GB	17min 53sec	5h 57min
7000	12GB	22min	7h 20min
8000	12GB	25min 28sec	8h 30min

A segunda alocação, para os demais 8 conjuntos, seguiu a distribuição de memória apresentada na tabela 6.3 abaixo.

Conforme a figura 6.4, há uma relação linear entre o número de sentenças e o tempo médio aproximado para treino do classificador. Esta relação e expressa pela equação 6.1. Esta relação é válida desde que a memória por processo também aumente de acordo com o tamanho da entrada, de acordo com o mostrado nas tabelas 6.2 e 6.3 acima.

$$executionTime = 0.2033 \cdot numberOfSentences - 122.5750 \tag{6.1}$$

Tabela 6.3: Treino dos classificadores - segunda alocação. Esta distribuição possibilita o treino dos classificadores em um único computador com 60GiB de memória. O tempo total de uso deste computador para treino dos classificadores desta alocação corresponde à duração total do treino de 20 classificadores a partir de 20 diferentes grupos de 7500 sentenças da Wikipedia.

Number of Sentences	*Memory*	*Duration (Average)*	*Duration (Total)*
500	2GB	41sec	13min
1500	4GB	3min 18sec	1h 46min
2500	4GB	6min 28sec	2h 29min
3500	6GB	9min 18sec	3h 6min
4500	10GB	12min 44sec	4h 54min
5500	12GB	16min 22sec	5h 47min
6500	14GB	20min 31sec	6h 50min
7500	16GB	23min 47sec	7h 35min

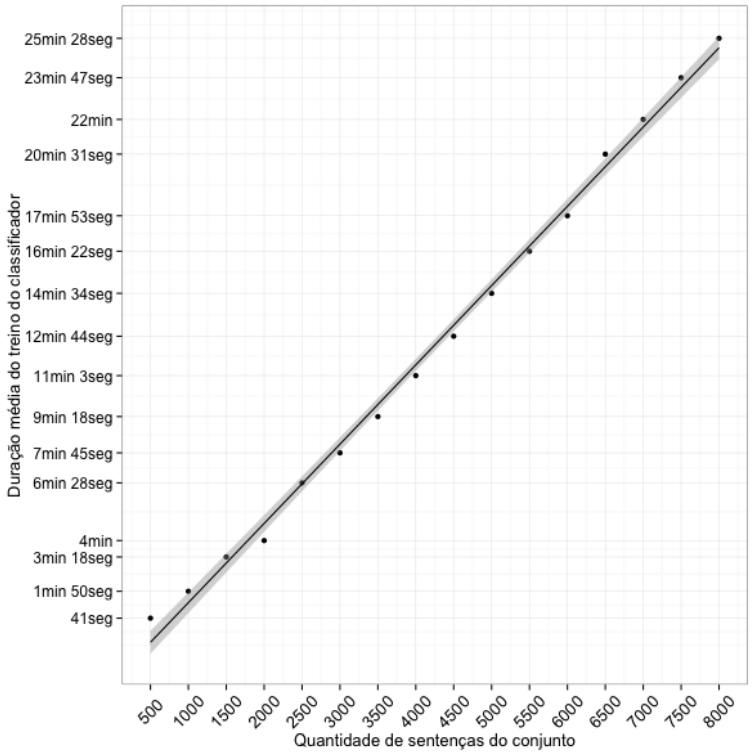

Figura 6.4: Relação entre a quantidade de sentenças no conjunto de treino e o tempo médio para treino do classificador utilizando o Stanford NER. Esta relação linear assume que memória suficiente foi alocada para o treino de cada classificador. As tabelas 6.2 e 6.3 apresentam a quantidade de memória necessária para treino de classificadores com entradas cujo tamanho varia de 500 até 8000 sentenças.

6.2.3 Teste dos Classificadores

O teste dos 320 classificadores treinados foi realizado com o *corpus* do Segundo Harem para avaliação das seguintes medidas:

- Precisão (*PRECISION*), que representa a proporção entre o número de entidades reconhecidas e classificadas de acordo com o *corpus* de testes, em relação ao número de entidades reconhecidas;
- Cobertura (*RECALL*), que representa a proporção entre o a quantidade de entidades existentes no *corpus* de testes que foram reconhecidas, em relação ao número de entidades existentes neste *corpus*;
- Medida-F (*F-MEASURE*), a média harmônica entre precisão e cobertura.

Visto que o formato XML do *corpus* do Segundo Harem não é suportado pelo Stanford NER, foi necessário convertê-lo para o formato apropriado. Para tal, foi utilizado o utilitário "corpus-processor"[2]. Nesta conversão, somente as entidades de tipo ORGANISATION, PERSON e PLACE permanecem anotadas. As demais categorias semânticas receberam a etiqueta O (de Outros), assim como as palavras que não são exemplos de nenhuma destas categorias.

Coeficiente de variação de Pearson

O primeiro teste foi o do coeficiente de variação de Pearson das três medidas nos 16 diferentes conjuntos. Este teste foi importante para avaliar a consistência dos resultados dos 20 classificadores de cada conjunto, determinando o quão diferentes são as medidas obtidas pelos classificadores que compartilham do mesmo tamanho de entrada. Conhecer o coeficiente de variação entre diferentes classificadores de mesmo tamanho de entrada permite determinar qual intervalo de tamanhos de entrada produzem resultados mais consistentes, contribuindo para a escolha do tamanho para treino do classificador.

Os resultados são apresentados na figura 6.5 abaixo, com destaque para a contribuição da variância de cada uma das três categorias semânticas. A Precisão apresenta variação muito menor que a Cobertura, e a Medida-F é bastante influenciada pela variação da Cobertura. É perceptível que nos conjuntos a partir de 2000 sentenças a diferença de variação com os demais conjuntos reduz, indicando uma relativa estabilidade nas medidas. Esta relativa estabilidade pode ser interpretada como maior probabilidade de que um classificador treinado com este número de sentenças, selecionadas aleatoriamente, apresente medidas similares aos demais classificadores treinados com amostras de mesmo tamanho, obtidas da mesma população (a Wikipedia em Português).

Percebe-se também que a maior variação ocorre na classificação de organizações (*ORGANISATION*) enquanto que lugar (*PLACE*) é a categoria com menor variação. O tamanho da

[2] https://github.com/dasdad/corpus-processor, acessado em 26-05-2014

variação entre as categorias semânticas possui uma correlação inversa à quantidade de entidades de cada categoria semântica. Quanto menor o número de entidades de uma categoria, maior a variação nas medidas de desempenho desta categoria. Utilizando como exemplo a categoria Organização, que possui o menor número de exemplos, é possível observar uma maior variação nas três medidas de desempenho. nos conjuntos com até 2000 sentenças. Cabe ressaltar que apenas com a correlação não é possível determinar se a quantidade de entidades de uma categoria semântica é responsável pelo coeficiente de variação das medidas.

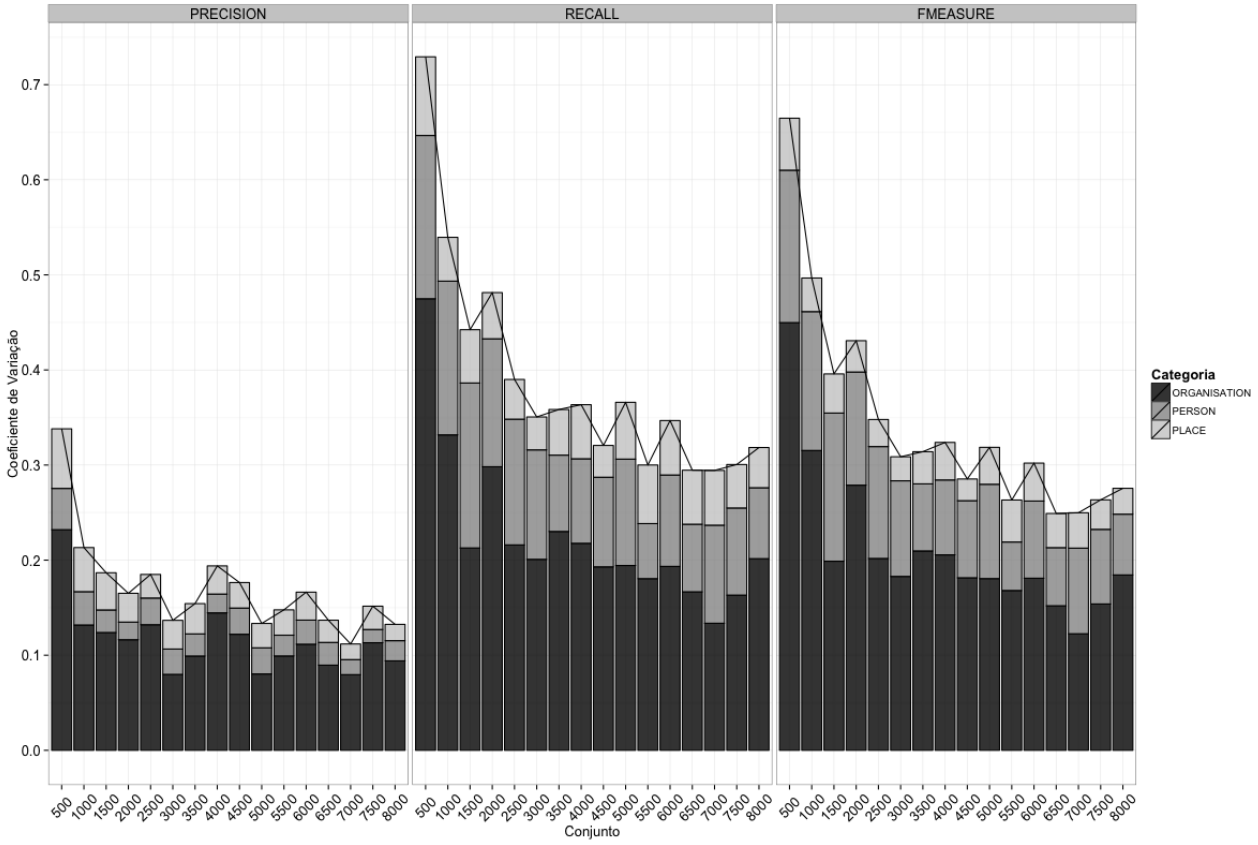

Figura 6.5: Coeficiente de Variação pelos tamanhos dos conjuntos. Cada conjunto representa um grupo de 20 diferentes classificadores treinados com um número de sentenças onde a quantidade de sentenças varia de 500 a 8000, em incrementos de 500 sentenças. Toda sentença deve ter ao menos um exemplo de categoria semântica na forma de wikilink. Estes classificadores foram testados utilizando o *corpus* do Segundo Harem, e o tamanho da barra indica o coeficiente de variação entre os classificadores pertencentes ao mesmo conjunto. Um coeficiente de variação menor indica que há uma menor dispersão no resultado das medidas de desempenho dos classificadores deste grupo.

Análise das medidas de desempenho

Em seguida, as medidas de Precisão, Cobertura e Medida-F foram analisadas de forma combinada para identificar quais classificadores apresentaram os melhores resultados globais. Foi necessário combinar as medidas das três categorias semânticas em uma medida única, pois o melhor resultado individual de uma categoria pode apresentar resultados muito inferiores para as demais. As medidas foram combinadas através do recálculo de Precisão, Cobertura e Medida-F utilizando

os somatórios de Verdadeiros Positivos (*True Positive*), Falsos Positivos (*False Positive*) e Falsos Negativos (*False Negative*) conforme as equações abaixo.

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \tag{6.2}$$

$$Recall = \frac{\sum TP}{\sum TP + \sum FN} \tag{6.3}$$

$$FMeasure = \frac{2 \cdot \sum TP}{2 \cdot \sum TP + \sum FP + \sum FN} \tag{6.4}$$

As tabelas 6.4, 6.5 e 6.6 abaixo apresentam os 10 melhores resultados combinados para as medidas de Precisão, Cobertura e Medida-F, respectivamente. O número de sentenças de cada conjunto está dentro do intervalo de coeficientes de variação mais estáveis. Não há interseção entre os melhores resultados de Precisão e os melhores resultados de Cobertura ou Medida-F, enquanto que os melhores resultados de Cobertura e Medida-F são os mesmos.

Tabela 6.4: Classificadores de melhor desempenho segundo a medida de precisão. Nesta tabela, Conjunto identifica a quantidade de sentenças no *corpus* de treino, e Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de precisão, e representam as medidas de desempenho dos classificadores quando testados com o *corpus* do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
5000	15	0.7266	0.2019	0.3160
3000	11	0.7206	0.1692	0.2740
3500	12	0.7193	0.1805	0.2886
7500	9	0.7165	0.1996	0.3122
8000	6	0.7162	0.1888	0.2988
6500	17	0.7159	0.2030	0.3164
8000	20	0.7156	0.1784	0.2856
3500	4	0.7155	0.1777	0.2847
6000	7	0.7152	0.1832	0.2917
5000	1	0.7148	0.2007	0.3134

A ausência de classificadores com bons resultados nas três medidas sugere uma correlação negativa entre os grupos. Como a Medida-F é uma combinação da Precisão com a Cobertura, o teste da correlação negativa foi realizado entre as medidas de Precisão e Cobertura, a partir dos 10 melhores resultados de cada medida. O cálculo da correlação entre estes dois conjuntos de 20 observações resulta em -0.708. Esta correlação negativa, para o tamanho da amostra, possui significância estatística ($p < .01$). Desta forma, a escolha do melhor classificador depende da escolha de qual medida de desempenho se pretende utilizar. As tabelas de 6.7 até 6.12 apresentam os resultados das medidas de desempenho e matrizes de confusão para os três melhores classificadores. Dentre estes três, a escolha foi reduzida para as medidas de Precisão e Medida-F.

Tabela 6.5: Classificadores de melhor desempenho segundo a medida de cobertura. Nesta tabela, Conjunto identifica a quantidade de sentenças no *corpus* de treino, e Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de cobertura, e representam as medidas de desempenho dos classificadores quando testados com o *corpus* do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
6500	12	0.6783	0.2155	0.3270
6000	15	0.6984	0.2143	0.3280
7000	17	0.6901	0.2143	0.3271
7500	15	0.6994	0.2116	0.3248
6500	15	0.6811	0.2099	0.3210
7500	19	0.7008	0.2093	0.3223
6000	8	0.6946	0.2083	0.3205
8000	4	0.7101	0.2081	0.3219
3500	17	0.7027	0.2079	0.3208
7000	8	0.6946	0.2079	0.3200

Tabela 6.6: Classificadores de melhor desempenho segundo a medida-F. Nesta tabela, Conjunto identifica a quantidade de sentenças no *corpus* de treino, e Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de medida-F, e representam as medidas de desempenho dos classificadores quando testados com o *corpus* do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
6000	15	0.6984	0.2143	0.3280
7000	17	0.6901	0.2143	0.3271
6500	12	0.6783	0.2155	0.3270
7500	15	0.6994	0.2116	0.3248
7500	19	0.7008	0.2093	0.3223
8000	4	0.7101	0.2081	0.3219
6500	15	0.6811	0.2099	0.3210
3500	17	0.7027	0.2079	0.3208
6000	8	0.6946	0.2083	0.3205
7000	8	0.6946	0.2079	0.3200

Tabela 6.7: Medidas de desempenho do classificador de melhor precisão média, chamado de "(WP) Melhor Precisão", quando testado com o *corpus* do Segundo Harem. Este classificador corresponde ao Conjunto 5000, Amostra 15.

Category	*PRECISION*	*RECALL*	*FMEASURE*
ORGANISATION	0.5077	0.0675	0.1191
PERSON	0.8760	0.1570	0.2663
PLACE	0.6884	0.3750	0.4855

Tabela 6.8: Matriz de Confusão para o Classificador "(WP) Melhor Precisão" ao testá-lo com o corpus do Segundo Harem.

	Reference			
Prediction	*OTHER*	*ORGANISATION*	*PERSON*	*PLACE*
OTHER	80521	1657	3408	1398
ORGANISATION	38	125	65	8
PERSON	28	15	770	10
PLACE	90	118	80	665

Tabela 6.9: Medidas de desempenho do classificador de melhor cobertura média, chamado de "(WP) Melhor Cobertura", quando testado com o *corpus* do Segundo Harem. Este classificador corresponde ao Conjunto 6500, Amostra 12.

Category	*PRECISION*	*RECALL*	*FMEASURE*
ORGANISATION	0.4483	0.0798	0.1354
PERSON	0.8247	0.1614	0.2699
PLACE	0.6542	0.4043	0.4998

Tabela 6.10: Matriz de Confusão para o Classificador "(WP) Melhor Cobertura" ao testá-lo com o corpus do Segundo Harem.

	Reference			
Prediction	*OTHER*	*ORGANISATION*	*PERSON*	*PLACE*
OTHER	80479	1628	3294	1324
ORGANISATION	52	142	86	15
PERSON	36	9	838	11
PLACE	110	136	105	731

Tabela 6.11: Medidas de desempenho do classificador de melhor medida-F média, chamado de "(WP) Melhor Medida-F", quando testado com o *corpus* do Segundo Harem. Este classificador corresponde ao Conjunto 6000, Amostra 15.

Category	*PRECISION*	*RECALL*	*FMEASURE*
ORGANISATION	0.4437	0.0685	0.1187
PERSON	0.8195	0.1667	0.2770
PLACE	0.6820	0.4005	0.5046

Tabela 6.12: Matriz de Confusão para o Classificador "(WP) Melhor Medida-F" ao testá-lo com o corpus do Segundo Harem.

	Reference			
Prediction	*OTHER*	*ORGANISATION*	*PERSON*	*PLACE*
OTHER	80489	1664	3310	1321
ORGANISATION	49	121	73	17
PERSON	49	19	852	27
PLACE	90	111	88	716

Através da figura 6.6 percebe-se uma grande diferença da Medida-F entre as categorias semânticas Organização, Pessoa e Lugar. Esta diferença aparenta ser proporcional à diferença entre o número de exemplos de cada categoria nos conjuntos de treino apresentada nas figuras 6.1 e 6.2. A relação entre o número de instâncias de uma classe e a medida de Cobertura ou Medida-F, utilizando o coeficiente linear de Pearson, é estatisticamente significante com $r(318) = .639, p < .01$ para Organizações, $r(318) = .708, p < .01$ para Pessoas e $r(318) = .738, p < .01$ para Lugares.

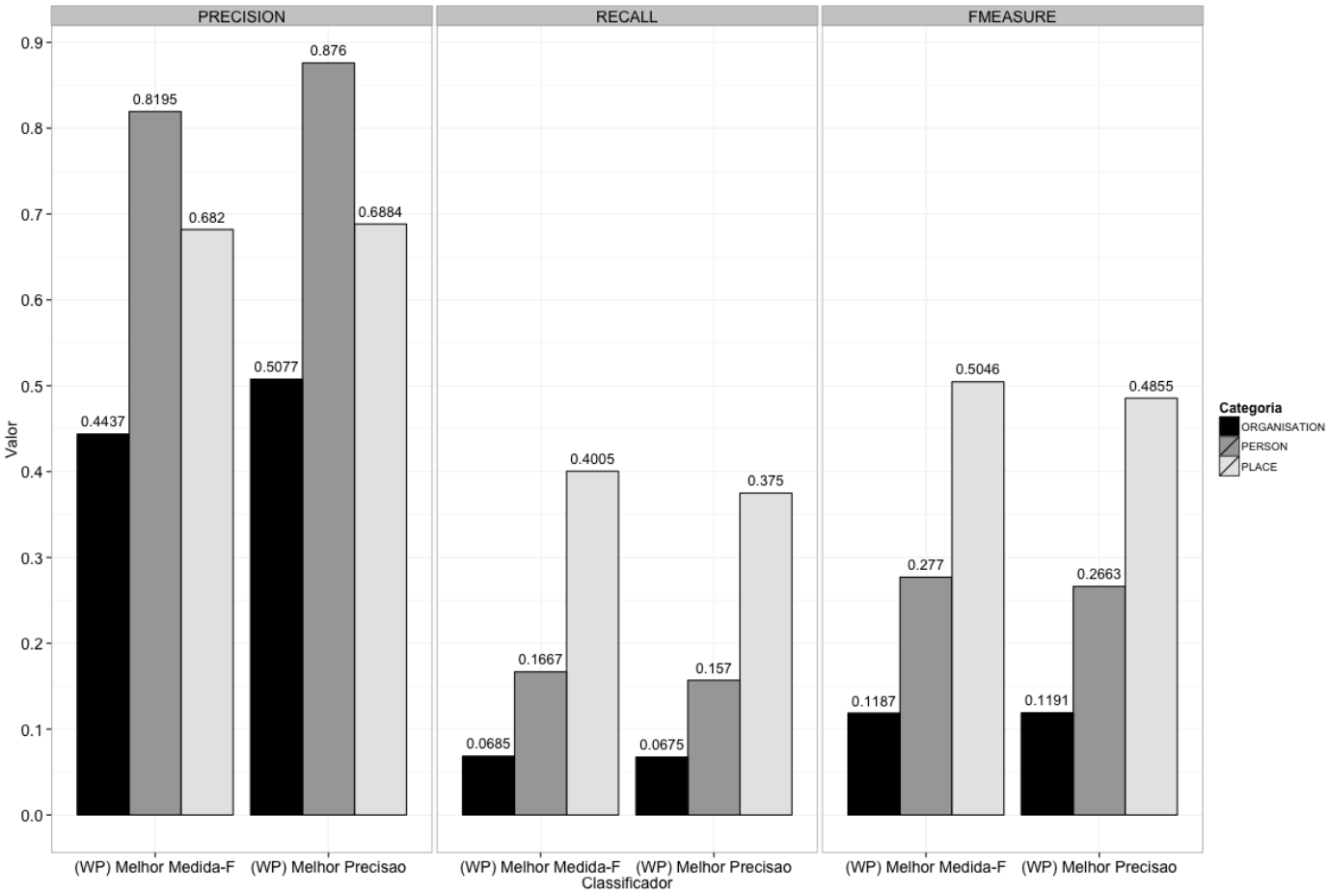

Figura 6.6: Comparativo do melhor desempenho obtido pelos classificadores treinados com *corpus* extraído da Wikipedia nas medidas de Precisão, Cobertura e Medida-F, quando testados com o *corpus* do Segundo Harem.

6.3 Experimento II - Comparação dos classificadores com o *corpus* do Primeiro Harem

Neste experimento, os dois classificadores selecionados no primeiro experimento - os de melhor Precisão e melhor Medida-F - foram comparados com um terceiro classificador treinado a partir da Coleção Dourada do Primeiro Harem. Este experimento serviu para medir o impacto do uso de um *corpus* anotado manualmente por especialistas em comparação aos classificadores treinados a partir de *corpus* gerado com base em anotações obtidas da Wikipedia. O classificador foi treinado utilizando as mesmas configurações do Stanford NER aplicadas aos classificadores treinados com a Wikipedia. A Coleção Dourada do Primeiro Harem tem, em sua composição, 95.711 *tokens*. Ela é aproximada, em quantidade de *tokens*, aos *corpora* da Wikipedia gerados com combinações de 3.000 sentenças. A tabela 6.13 apresenta as quantidades de entidades nomeadas nos diferentes *corpora*, por categoria semântica. Percebe-se que a distribuição na Coleção Dourada do Primeiro Harem é mais homogênea que a dos conjuntos de tamanho similar extraídos da Wikipedia. Já em comparação com os dois classificadores selecionados no experimento anterior (tabela 6.14), o número de entidades nomeadas é muito menor.

Tabela 6.13: Comparativo entre a quantidade de exemplos de cada categoria semântica no *corpus* do Primeiro Harem e a média da quantidade de exemplos destas categorias semânticas nos diferentes *corpora* extraídos da Wikipedia cujo número de *tokens* é próximo do número de *tokens* do *corpus* do Primeiro Harem.

Category	*First Harem*	*Wikipedia average (3000 sentences)*
ORGANISATION	2169	1560
PERSON	2242	2445
PLACE	2038	4427

Tabela 6.14: Comparativo da quantidade de exemplos de cada categoria semântica nos *corpora* de treino dos três classificadores avaliados neste experimento. (WP) Melhor Precisão e (WP) Melhor Medida-F são os classificadores selecionados no experimento anterior.

Category	*First Harem*	*(WP)Best Precision*	*(WP)Best FMeasure*
ORGANISATION	2169	2554	3044
PERSON	2242	4113	4911
PLACE	2038	7541	8508

O tempo médio necessário para treino do classificador utilizando este *corpus*, após 5 execuções, foi de aproximadamente 4 minutos e 20 segundos. Este tempo é similar ao tempo necessário para treino de um classificador a partir de conjuntos de 2000 sentenças da Wikipedia.

Tabela 6.15: Melhores resultados obtidos nas medidas de desempenho

Classifier	*Category*	*PRECISION*	*RECALL*	*FMEASURE*
(WP)Best FMeasure	*ORGANISATION*	0.4437	0.0685	0.1187
(WP)Best FMeasure	*PERSON*	0.8195	0.1667	0.2770
(WP)Best FMeasure	*PLACE*	0.6820	0.4005	0.5046
(WP)Best Precision	*ORGANISATION*	0.5077	0.0675	0.1191
(WP)Best Precision	*PERSON*	0.8760	0.1570	0.2663
(WP)Best Precision	*PLACE*	0.6884	0.3750	0.4855
First Harem	*ORGANISATION*	0.4958	0.4192	0.4543
First Harem	*PERSON*	0.7049	0.5101	0.5919
First Harem	*PLACE*	0.6155	0.5756	0.5949

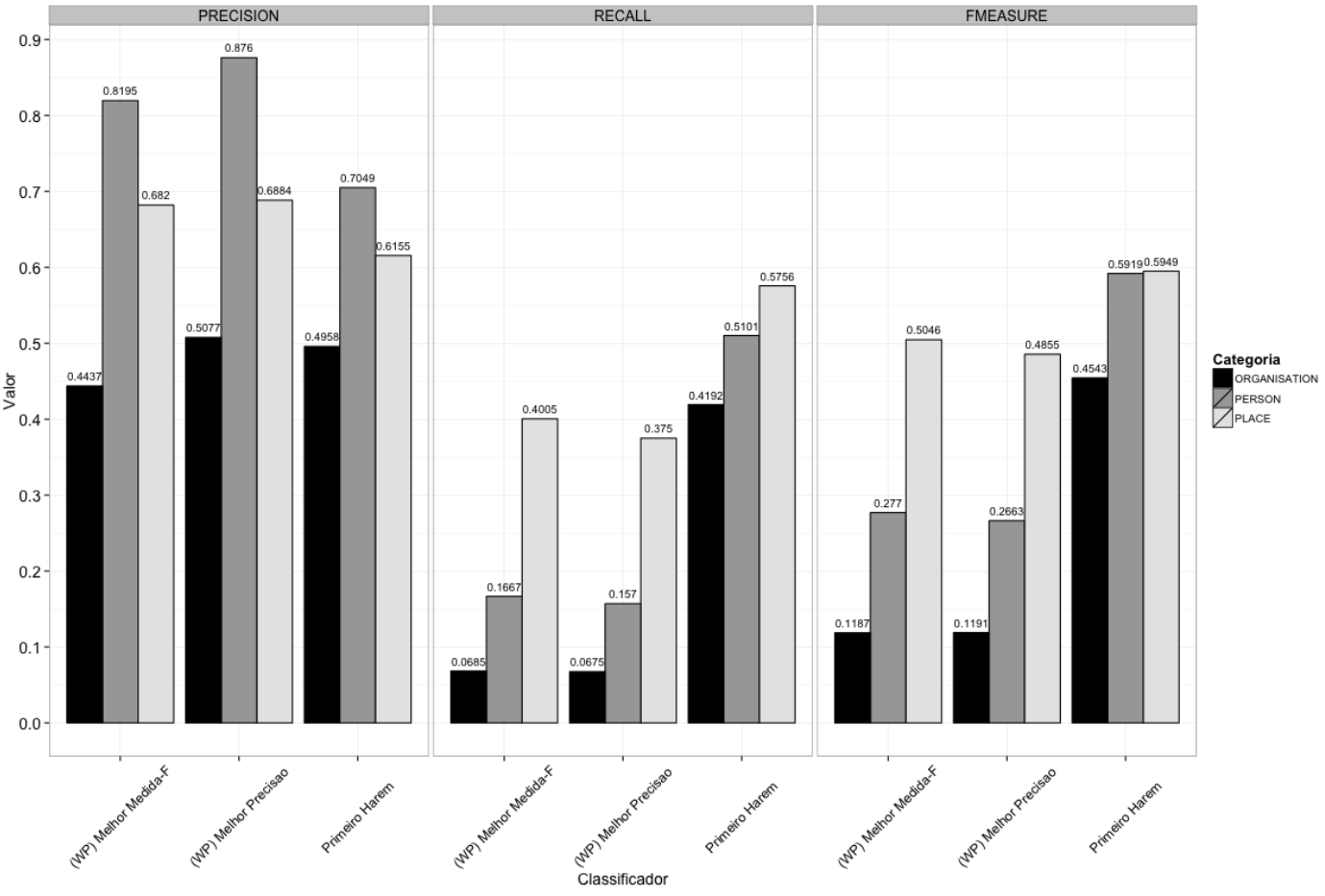

Figura 6.7: Comparativo entre melhores classificadores da Wikipedia e classificador do Primeiro Harem quando testados com o *corpus* do Segundo Harem.

Conforme apresentado na tabela 6.15 e na figura 6.7, o desempenho do classificador treinado com o *corpus* do Segundo Harem apresentou uma Medida-F muito melhor no teste com o *corpus* do Primeiro Harem. Em comparação com os classificadores treinado apenas com a Wikipedia, o bom resultado na Medida-F pode ser explicado pela maior Cobertura. Enquanto que no classificador de melhor Medida-F treinado com a Wikipedia a média da Cobertura foi de 0.2143, no classificador treinado com o *corpus* do Segundo Harem a média da Cobertura foi de 0.5092. Por outro lado, a melhor Precisão média obtida com classificadores treinados com a Wikipedia foi de 0.7266, enquanto que a Precisão média obtida pelo classificador treinado com o *corpus* do Segundo Harem foi de 0.6254.

6.4 Experimento III - Investigação do efeito do uso de diferentes estilos de escrita nos *corpora* de treino e de teste dos classificadores

Para melhor compreender as diferenças nos resultados obtidos com os diferentes classificadores apresentados, foi avaliado o desempenho destes classificadores com um *corpus* de diferente estilo de escrita. É sabido de [SC07] e [MS08] que os *corpora* do Primeiro e Segundo Harem são compostos por textos de diferentes estilos. A tabela 6.16 apresenta a proporção dos 8 estilos que juntos compõe o *corpus* do Primeiro Harem em relação ao número de palavras. A tabela 6.17, por sua vez, apresenta a proporção dos 14 tipos de texto que compõe o *corpus* do Segundo Harem, também em relação ao número de palavras. Apesar da publicação do Primeiro Harem não identificar claramente a origem de cada tipo de texto, é possível observar que no Segundo Harem o estilo de escrita jornalístico predomina com 35% de participação, seguido pelo estilo didático, com $22,5\%$ de participação. Em [MS08] consta também que a participação de textos obtidos da Wikipedia corresponde a $12,34\%$ do número de palavras do *corpus* do Segundo Harem.

Tabela 6.16: Composição do Primeiro Harem, obtido de [SC07]

Estilo	Participação (em %)
Entrevista	$31,54$
Web	$18,29$
Jornalístico	$12,73$
Literário	$11,52$
Expositivo	$8,22$
Politico	$6,57$
Correio electrónico	$6,53$
Técnico	$4,61$

Tabela 6.17: Composição do Segundo Harem, obtido de [MS08]

Estilo	Participação (em %)
Notícia	$21,00$
Didático	$20,84$
Perguntas	$11,62$
Opinião	$9,46$
Blogue jornalístico	$8,72$
Blogue pessoal	$8,36$
Ensaio	$7,83$
Entrevista	$3,93$
Perguntas faq	$2,46$
Blogue humorístico	$2,07$
Legislativo	$1,30$
Promocional	$1,08$
Literário	$0,81$
Texto privado manuscrito	$0,51$

6.4.1 Teste dos classificadores utilizando sentenças da Wikipedia

O objetivo deste teste foi avaliar qual o impacto de um estilo de escrita diferente no desempenho do classificador. Para avaliar o desempenho dos classificadores foi necessário selecionar um conjunto de sentenças da Wikipedia que não foram utilizadas no treino dos classificadores, eliminando a possibilidade de influência no desempenho dos classificadores por terem sido previamente expostos às sentenças. Dentre as 320 conjuntos de sentenças anotadas gerados a partir da Wikipedia, dois deles não possuem sentenças em comum com os dois conjuntos de treino que apresentaram melhor desempenho nos testes com o *corpus* do Primeiro Harem. Estes dois conjuntos possuem 500 sentenças cada, e não há sentenças em comum entre os dois conjuntos. O script para identificação dos conjuntos pode ser visto no ANEXO B. A tabela 6.18 apresenta as características destes dois conjuntos.

Tabela 6.18: Características dos conjuntos para teste dos classificadores utilizando a Wikipedia

Dataset	*Sample*	*Number of Tokens*	*ORGANISATION*	*PERSON*	*PLACE*
500	4	14668	257	379	724
500	5	15115	218	446	681

Para este teste foi criado mais um classificador, treinado com o *corpus* do Segundo Harem. A tabela 6.19 apresenta a média de desempenho dos 4 classificadores quando testados com o *corpus* de teste composto por 500 sentenças anotadas da Wikipedia (amostra 4). Nesta tabela é possível observar que o melhor resultado de Medida-F em classificador treinado com sentenças da Wikipedia

foi 58% maior que o melhor resultado de mesma medida obtido por classificador treinado com *corpus* do Harem. A tabela 6.20 apresenta o resultado dos mesmos 4 classificadores quando testados com outro *corpus* de teste, também composto por 500 sentenças anotadas da Wikipedia (amostra 5). Neste, a melhor Medida-F obtida por um classificador treinado com *corpus* da Wikipedia foi 43% maior que a melhor Medida-F obtida por um classificador treinado com *corpus* do Harem.

Tabela 6.19: Medidas de desempenho de classificadores testados com sentenças da Wikipedia. Para esta comparação foi utilizado o *corpus* de teste composto por 500 sentenças da Wikipedia, amostra 4.

Classifier	*PRECISION*	*RECALL*	*FMEASURE*
(WP)Best Precision	0.7383	0.5429	0.6257
(WP)Best F-Measure	0.7632	0.6042	0.6744
Second Harem	0.3655	0.5110	0.4262
First Harem	0.3561	0.4596	0.4013

Tabela 6.20: Medidas de desempenho de classificadores testados com sentenças da Wikipedia. Para esta comparação foi utilizado o *corpus* de teste composto por 500 sentenças da Wikipedia, amostra 5

Classifier	*PRECISION*	*RECALL*	*FMEASURE*
(WP)Best Precision	0.7216	0.5889	0.6485
(WP)Best F-Measure	0.7147	0.5938	0.6487
Second Harem	0.3601	0.5543	0.4366
First Harem	0.3892	0.5444	0.4539

6.4.2 Testes com classificadores combinados

A partir dos experimentos anteriores foi possível identificar, nos testes com diferentes estilos de escrita, o impacto negativo no desempenho de classificadores treinados com textos da Wikipedia ao testá-los com os *corpora* do Harem. Também foi possível identificar o impacto das regras de classificação empregadas no Harem ao testar com sentenças da Wikipedia os classificadores treinados com *corpus* do Harem. Um último teste foi realizado para avaliar se classificadores treinados com a combinação de sentenças da Wikipedia com o *corpus* do Primeiro Harem apresentam melhores desempenho nos testes com o *corpus* do Segundo Harem e também nos testes com sentenças da Wikipedia. Para este teste, todos os 360 *corpora* extraídos da Wikipedia foram combinados com o *corpus* do primeiro Harem.

A primeira avaliação realizada foi do coeficiente de variação de classificadores treinados com os 360 novos *corpora* e testados com o *corpus* do Segundo Harem. O resultado desta avaliação

está na figura 6.8. A partir da figura é possível perceber que a medida de precisão está estável ao longo dos 20 diferentes conjuntos. A medida de cobertura é mais estável entre os conjuntos compostos por 3000 e 7500 sentenças, ambos acrescidos ao $corpus$ do Primeiro Harem.

Em comparação com os coeficientes de variação dos *corpora* compostos somente com sentenças da Wikipedia, apresentados na figura 6.5, podemos perceber que a variação reduziu em média 50% ao longo das medidas de Precisão, Cobertura, e Medida-F. Além disto, o resultado da Medida-F nos testes com os classificadores treinados com *corpus* combinado possuem menor correlação com a medida de cobertura. A correlação entre estas duas medidas no experimento com *corpora* combinado foi de 0.9905. A categoria Organização, que na figura 6.5 apresentava maior variação nas três medidas de desempenho, teve sua variação reduzida após a inclusão das sentenças do *corpus* do Primeiro Harem. Conforme visto nas tabelas 6.13 e 6.14, o *corpus* do Primeiro Harem possui um número próximo de exemplos de cada categoria semântica, sugerindo que o baixo desempenho para classificação de entidades da categoria Organização se deu pela baixa quantidade de exemplos desta categoria.

A segunda avaliação teve por objetivo identificar, dentre os 360 classificadores treinados com os novos *corpora*, quais obtiveram as melhores medidas de desempenho nos testes com o *corpus* do Segundo Harem. Novamente, o conjunto dos 10 classificadores de melhor Precisão é disjunto dos conjuntos dos 10 classificadores de melhor Cobertura e de melhor Medida-F. Além disto, conforme já evidenciado em experimentos anteriores, os classificadores de melhor Medida-F são influenciados pela baixa cobertura obtida pelos classificadores quando testados com o *corpus* do Segundo Harem, resultando na presença dos mesmos classificadores nos dois conjuntos, ainda que em diferentes posições segundo o critério de melhor desempenho nas medidas. Os 10 classificadores de melhor Precisão estão relacionados na tabela 6.21. Os classificadores de melhor Cobertura estão relacionados na tabela 6.22; os de melhor Medida-F, na tabela 6.23. Novamente, a escolha de um classificador depende da medida de desempenho.

Através da tabela 6.24 é possível comparar o desempenho dos três melhores classificadores em cada medida de desempenho quando testados com o *corpus* do Segundo Harem, com os resultados divididos nas três categorias semânticas que os classificadores foram treinados para reconhecer. Nela, o classificador "(Mixed)Best F-Measure" corresponde ao classificador treinado com o *corpus* de 1000 sentenças, amostra 19, combinadas com o *corpus* do Primeiro Harem. O classificador "(Mixed)Best Recall" é o classificador treinado com o *corpus* de 2000 sentenças, amostra 8, combinadas com o *corpus* do Primeiro Harem, enquanto que o classificador "(Mixed)Best Precision" é o classificador treinado com o *corpus* de 7500 sentenças, amostra 4, combinadas com o *corpus* do Primeiro Harem. Também está presente o resultado do classificador treinado com o *corpus* do Primeiro Harem e treinado com o *corpus* do Segundo Harem. Nestes testes com o *corpus* do Segundo Harem, os classificadores treinados com a combinação de sentenças da Wikipedia com o *corpus* do Primeiro Harem apresentaram melhor medida de Precisão quando comparados com o desempenho do classificador treinado com o *corpus* do Primeiro Harem sozinho. No entanto, o

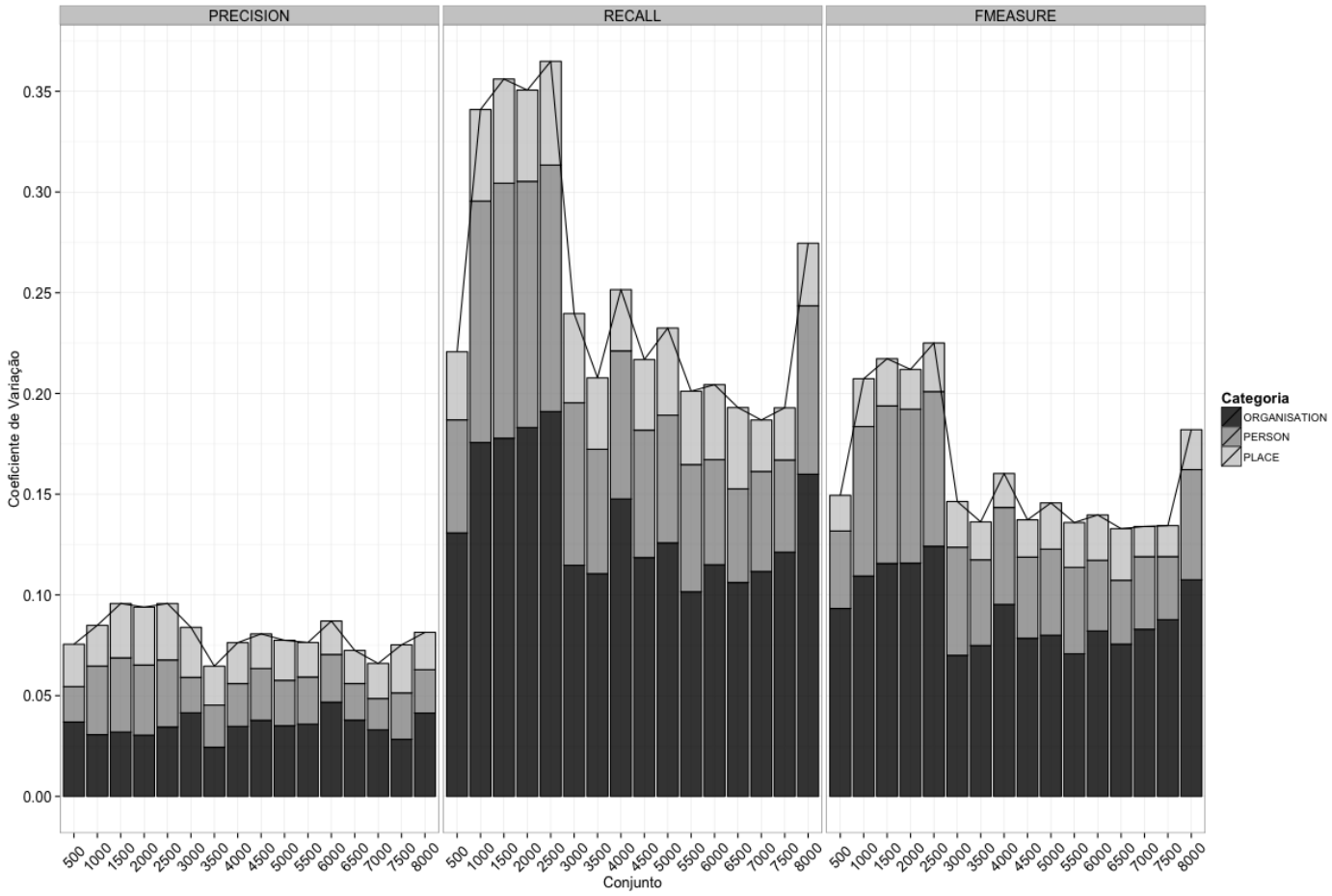

Figura 6.8: Coeficiente de Variação pelos tamanhos dos conjuntos. Cada conjunto representa um grupo de 20 diferentes classificadores treinados com uma composição das mesmas sentenças apresentadas na figura 6.5, acrescidas do *corpus* do Primeiro Harem. Estes classificadores foram testados utilizando o *corpus* do Segundo Harem, e o tamanho da barra indica o coeficiente de variação entre os classificadores pertencentes ao mesmo conjunto. Um coeficiente de variação menor indica que há uma menor dispersão no resultado das medidas de desempenho dos classificadores deste grupo.

Tabela 6.21: Classificadores de melhor desempenho segundo a medida de precisão. Nesta tabela, Conjunto identifica a quantidade de sentenças no corpus de treino, combinado com as sentenças do corpus do Primeiro Harem. Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de precisão, e representam as medidas de desempenho dos classificadores quando testados com o corpus do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
7500+Harem	4	0.7268	0.3375	0.4609
6000+Harem	12	0.7199	0.3301	0.4527
4500+Harem	2	0.7176	0.3416	0.4629
7000+Harem	15	0.7170	0.3446	0.4655
3000+Harem	9	0.7166	0.3248	0.4470
5500+Harem	18	0.7148	0.3474	0.4675
5500+Harem	12	0.7145	0.3347	0.4559
7000+Harem	12	0.7145	0.3301	0.4516
4500+Harem	9	0.7140	0.3494	0.4692
6500+Harem	15	0.7132	0.3446	0.4647

Tabela 6.22: Classificadores de melhor desempenho segundo a medida de cobertura. Nesta tabela, Conjunto identifica a quantidade de sentenças no corpus de treino, combinado com as sentenças do corpus do Primeiro Harem. Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de cobertura, e representam as medidas de desempenho dos classificadores quando testados com o corpus do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
2000+Harem	8	0.6561	0.4392	0.5262
1000+Harem	19	0.6682	0.4372	0.5285
2000+Harem	15	0.6552	0.4369	0.5242
1500+Harem	8	0.6628	0.4362	0.5262
1500+Harem	15	0.6511	0.4360	0.5223
2500+Harem	15	0.6650	0.4346	0.5257
1000+Harem	4	0.6533	0.4342	0.5216
2000+Harem	4	0.6589	0.4339	0.5232
1500+Harem	4	0.6570	0.4335	0.5223
2000+Harem	19	0.6563	0.4335	0.5221

Tabela 6.23: Classificadores de melhor desempenho segundo a medida-F. Nesta tabela, Conjunto identifica a quantidade de sentenças no corpus de treino, combinado com as sentenças do corpus do Primeiro Harem. Amostra identifica qual dentre os 20 grupos de sentenças foi utilizado no treino. Os resultados estão dispostos em ordem decrescente de medida-F, e representam as medidas de desempenho dos classificadores quando testados com o corpus do Segundo Harem.

Dataset	*Sample*	*PRECISION*	*RECALL*	*FMEASURE*
1000+Harem	19	0.6682	0.4372	0.5285
1500+Harem	8	0.6628	0.4362	0.5262
2000+Harem	8	0.6561	0.4392	0.5262
1500+Harem	19	0.6710	0.4323	0.5258
2500+Harem	15	0.6650	0.4346	0.5257
2500+Harem	11	0.6711	0.4312	0.5250
2000+Harem	15	0.6552	0.4369	0.5242
2000+Harem	4	0.6589	0.4339	0.5232
1500+Harem	15	0.6511	0.4360	0.5223
1500+Harem	4	0.6570	0.4335	0.5223

classificador treinado com o *corpus* do Primeiro Harem apresentou melhor desempenho nas medidas de Cobertura e Medida-F.

Finalmente, a tabela 6.25 sumariza os resultados dos diferentes classificadores treinados nos experimentos, exibindo um comparativo dos testes realizados com o *corpus* do Segundo Harem e com um *corpus* composto por 500 sentenças da Wikipedia. Nesta tabela é possível observar que os dois classificadores de melhor Precisão treinados com sentenças da Wikipedia obtiveram resultado similar no treino com o *corpus* do Segundo Harem, tanto com quanto sem o *corpus* do Primeiro Harem combinado. O resultado destes dois classificadores é significativamente maior que o resultado obtido no teste do classificador treinado com o *corpus* do Primeiro Harem, aumentando a Precisão de 62% para 72%. Contudo, a Cobertura alcançada pelos classificadores treinados com sentenças da Wikipedia no teste com o *corpus* do Segundo Harem foi muito mais baixa que a obtida no teste com o classificador treinado com o *corpus* do Primeiro Harem, e mesmo a combinação do *corpus* do Primeiro Harem com um número baixo de sentenças da Wikipedia no *corpus* "(Mixed)Best F-Measure" reduziu a Cobertura de 50.9% para 43.3%, fazendo com que o desempenho medido na Medida-F também fique abaixo do desempenho obtido com o classificador treinado com o *corpus* do Primeiro Harem. Não há comparação entre os classificadores treinados com os *corpora* mistos em teste com sentenças da Wikipedia pois não há uma combinação entre os *corpora* gerados e que não possua uma sentença em comum com os quatro *corpora* que contém sentenças da Wikipedia.

6.5 Interpretação dos resultados dos experimentos

Os conjuntos de sentenças extraídas da Wikipedia apresentam quantidades similares de *tokens*, números similares de exemplos das categorias semânticas avaliadas, e medidas de desempe-

Tabela 6.24: Comparativo dos melhores resultados obtidos nas medidas de desempenho em testes com o *corpus* do Segundo Harem. O classificador *(Mixed)Best F-Measure* utilizou, como *corpus* de treino, a combinação do *corpus* do Primeiro Harem com a amostra 19 do conjunto de 1000 sentenças da Wikipedia. O classificador *(Mixed)Best Recall* utilizou como *corpus* de treino a combinação do *corpus* do Primeiro Harem com a amostra 8 do conjunto de 2000 sentenças da Wikipedia. Já o classificador *(Mixed) Best Precision* foi treinado com a combinação do *corpus* do Primeiro Harem com a amostra 9 do conjunto de 7500 sentenças da Wikipedia.

Classifier	*Category*	*PRECISION*	*RECALL*	*FMEASURE*
(Mixed)Best F-Measure	*ORGANISATION*	0.5438	0.2669	0.3580
(Mixed)Best F-Measure	*PERSON*	0.7645	0.3686	0.4974
(Mixed)Best F-Measure	*PLACE*	0.6575	0.4799	0.5549
(Mixed)Best Recall	*ORGANISATION*	0.5614	0.2526	0.3484
(Mixed)Best Recall	*PERSON*	0.7487	0.3541	0.4808
(Mixed)Best Recall	*PLACE*	0.6538	0.4954	0.5637
(Mixed)Best Precision	*ORGANISATION*	0.5333	0.2372	0.3284
(Mixed)Best Precision	*PERSON*	0.7997	0.3087	0.4455
(Mixed)Best Precision	*PLACE*	0.6835	0.4915	0.5718
First Harem	*ORGANISATION*	0.4958	0.4192	0.4543
First Harem	*PERSON*	0.7049	0.5101	0.5919
First Harem	*PLACE*	0.6155	0.5756	0.5949

Tabela 6.25: Resultados dos diferentes classificadores treinados nos experimentos, exibindo um comparativo dos testes realizados com o corpus do Segundo Harem e com um corpus composto por 500 sentenças da Wikipedia.

	Second Harem Test Set			**Wikipedia 500 sentences Test Set**		
Training Set	PRECISION (%)	RECALL (%)	FMEASURE (%)	PRECISION (%)	RECALL (%)	FMEASURE (%)
First Harem	62.54	50.92	56.14	35.61	45.96	40.13
(WP)Best Precision	72.66	20.19	31.60	73.83	54.29	62.57
(WP)Best F-Measure	69.84	21.43	32.80	76.32	60.42	67.44
(Mixed)Best Precision	72.68	33.75	46.09			
(Mixed)Bext F-Measure	66.82	43.72	52.85			

nho próximas. A partir do número de sentenças extraídas é possível determinar o número de *tokens*, o número de exemplos de cada uma das diferentes categorias semânticas, e a duração do treino do classificador. A qualidade do classificador, contudo, depende de quais sentenças foram selecionadas, como foi possível observar pela grande dispersão no desempenho da Cobertura em classificadores criados a partir de conjuntos de mesma quantidade de sentenças.

Espera-se uma diferença entre 10% e 20% na medida de Precisão dos classificadores de um mesmo conjunto, sendo que a variação é menor que 5% para as categorias semânticas Pessoa (*PERSON*) e Lugar (*PLACE*). Já para a medida de Cobertura a variação do desempenho dos classificadores de um mesmo conjunto esteve entre 30% e 40% nos conjuntos com mais de 2000 sentenças. Esta grande diferença na medida de Cobertura indica que há combinações de sentenças que produzem resultados muito melhores que outras combinações, enquanto que há pouca variação para a Precisão. Identificar quais são estas sentenças, portanto, pode levar a classificadores com melhor desempenho da Medida-F no teste com o *corpus* do Segundo Harem.

Através das medidas de desempenho e matriz de confusão dos três classificadores com melhor resultado no teste com o *corpus* do Segundo Harem, foi possível perceber que um grande número de entidades nomeadas com categoria semântica associada foram classificadas como palavras sem categoria semântica. Esta é a razão para a baixa Cobertura obtida por estes classificadores, e o motivo está na ausência de wikilink. Esta ausência é resultado da recomendação de escrita de artigos proposto para a Wikipedia, que recomenda a aplicação de wikilink somente na primeira ocorrência da entidade, deixando as demais ocorrências sem wikilink.

Um número significativo de entidades nomeadas pertencentes à categoria semântica Organização foram identificadas como pertencentes à categoria semântica Lugar, afetando ainda mais a classificação deste tipo de entidade. Uma possível razão para este erro de classificação é o fato dos *corpora* de treino possuírem poucos exemplos de Organizações e muitos exemplos de Lugares. Conforme visto nas figuras 6.1, 6.2 e 6.3, o número de exemplos de entidades nomeadas pertencentes à categoria semântica Lugar é quase três vezes maior que o número de exemplos da categoria semântica Organização.

A categoria semântica Pessoa teve uma medida de Cobertura ruim pois muitas entidades desta categoria presentes no *corpus* de teste não foram reconhecidas pelos classificadores. Esta baixa taxa de acerto é consequência do maior número de ocorrência de nomes de pessoas sem wikilink nos artigos da Wikipedia, bem como de nomes de pessoas com wikilink mas cujo artigo não foi classificado como Pessoa na DBpedia.

Estas conclusões foram validadas através da comparação do resultado das medidas de desempenho obtidas pelos classificadores treinados com a Wikipedia com o resultado das medidas de desempenho obtidas pelo classificador treinado com o *corpus* do Primeiro Harem. Este *corpus* possui um número mais próximo de exemplos de cada uma das categorias semânticas avaliadas, e por ter sido anotado manualmente, não sofre dos problemas relacionados à Wikipedia (recomendação de escrita) e da DBpedia (ausência da classe para a instância). Classificadores treinados com *corpus* composto pelo *corpus* do Primeiro Harem mais um conjunto de sentenças da Wikipedia apresentaram

melhora na Cobertura e Medida-F. Estas medidas foram menos impactadas pela recomendação de escrita da Wikipedia, mas ainda foram impactadas pela diferença entre o número de exemplos de cada categoria semântica.

No entanto, dois classificadores treinados com o *corpus* do Primeiro Harem e do Segundo Harem tiveram desempenho ruim quando testados com grupos de sentenças da Wikipedia. O mau desempenho reforça o impacto que o estilo de escrita dos *corpora* de treino e de teste exerce sobre o desempenho dos classificadores. Nestes experimentos, classificadores treinados e testados com *corpus* obtido a partir da Wikipedia apresentaram melhores resultados que aqueles treinados com *corpus* de diferente estilo de escrita.

7. CONCLUSÃO E TRABALHOS FUTUROS

7.1 Conclusões

Através da anotação automática de entidades nomeadas em um grande número de sentenças da Wikipedia, utilizando a DBpedia como base de dados de entidades nomeadas, criamos um grande *corpus* anotado para treino de classificadores de entidades nomeadas. A combinação de um processo de anotação automática, uso das sentenças da Wikipedia, e uso da ontologia na DBpedia apresenta algumas vantagens, listadas abaixo:

- Possibilidade de selecionar sentenças segundo critérios específicos, como a presença de casos de difícil classificação, exemplos de relações entre entidades, entre outros, que podem ser aplicados para outras tarefas de *Information Extraction*, bem como para complementar outros *corpora* existentes;
- Anotação de subcategorias específicas (pessoas: músicos, organizações: bandas, lugares: casas de show, estádios), presentes na Wikipedia;
- Anotação de entidades e artigos de domínios específicos, como Biologia, Geologia e Política, presentes na Wikipedia;
- A evolução do conteúdo da Wikipedia e dos processos de extração da DBpedia beneficiam a qualidade do das anotações.

Além das vantagens listadas acima, destacamos as seguintes contribuições desta pesquisa:

- A criação do processo, capaz de extrair e anotar as sentenças dos originais dos artigos da Wikipedia anotados com a *Markup Language*, selecionar artigos a partir dos artigos relacionados, e substituir as anotações de wikilinks por anotações de categorias semânticas de entidades nomeadas. Os códigos-fonte dos artefatos desenvolvidos para este processo estão disponíveis sob licença *Open Source*, e o processo foi apresentado em [WV14];
- O conjunto de avaliações e testes realizados nos experimentos podem ser utilizados como instrumento para medir o resultado de mudanças feitas no processo de geração do *corpus* da Wikipedia;
- O bom resultado obtido na medida de Precisão é significativo e mostramos que a combinação com outro *corpus* pode melhorar a medida de Precisão.

Para avaliação da qualidade do processo de anotação, geramos 360 diferentes *corpora* com quantidade de sentenças variando de 500 até 8000, selecionadas de forma aleatória de um original de pouco mais de $1,5$ milhões de sentenças. Estes *corpora* foram utilizados para no treino de classificadores utilizando a ferramenta Stanford NER. Todos os classificadores foram testados

com o *corpus* do Segundo Harem. Os dois classificadores melhor posicionados nos testes segundo as medidas de desempenho Precisão e Medida-F foram selecionados para comparação com outro classificador, este treinado com o *corpus* do Primeiro Harem. Nesta avaliação, observamos que os classificadores treinados com sentenças da Wikipedia apresentaram melhor precisão que o classificador treinado com o *corpus* do Primeiro Harem, principalmente na classificação de entidades do tipo *Person* (pessoas). A medida de cobertura obtida pelos classificadores treinados com sentenças da Wikipedia, no entanto, ficou muito abaixo da medida de cobertura obtida pelo classificador treinado com o *corpus* do Primeiro Harem, principalmente na classificação de entidades do tipo *Organisation* (organizações). Esta diferença influenciou no baixo desempenho aferido na Medida-F. Por outro lado, na comparação destes classificadores utilizando *corpora* de testes composto por sentenças da Wikipedia, os classificadores treinados com sentenças da wikipedia apresentaram melhor desempenho nas três medidas - Precisão, Cobertura e Medida-F.

Estas observações estão de acordo com observações de trabalhos anteriores, como [AV13], que já identificou a relação entre o baixo número de exemplos de uma categoria semântica e o mau desempenho na medida de cobertura. Em [NRR+13] o autor destacou que entidades pertencentes à categoria semântica Organização são mais difíceis de identificar pois a grafia dos nomes de organizações é irregular em comparação com a grafia das entidades de outras categorias, como pessoa e lugar. Outra observação consistente com trabalhos anteriores diz respeito ao efeito do estilo de escrita entre o conjunto de treino e o conjunto de teste, relatada por [PK01]. Observamos que o classificador treinado com *corpus* do Primeiro Harem apresentou melhor desempenho no teste com o *corpus* do Segundo Harem, porém seu desempenho foi baixo ao testar com *corpora* formados por sentenças da Wikipedia. O mesmo foi observado em classificadores treinados com *corpora* da Wikipedia, que tiveram bom desempenho quando testados com *corpora* da Wikipedia e pior desempenho em testes com os *corpora* do Primeiro Harem e do Segundo Harem.

Apesar do bom desempenho obtido na medida de precisão, o baixo desempenho na medida de cobertura indica que devemos melhorar o processo de anotação automática para obter mais exemplos das categorias semânticas selecionadas. O desbalanceamento do número de exemplos das três categorias, proporcional ao desempenho observado nas três categorias, parece agravar o desempenho nos testes com os *corpora* do Harem, que é balanceado.

7.2 Trabalhos futuros

Para produzir *corpora* capazes de melhores resultados nas medidas de desempenho, é necessário aumentar o número de exemplos de entidades nomeadas, desta forma obtendo uma maior diversidade de sentenças com entidades nomeadas que possibilite a generalização do classificador para detectar entidades em diferentes estruturas gramaticais. A presença de entidades anotadas com wikilinks nas sentenças da Wikipedia é mais comum nas primeiras sentenças dos artigos, reduzindo a diversidade. É possível aumentar o número de exemplos de três formas: (1) a anotação das ocorrências de entidades sem wikilink quando existirem ocorrências anteriores com wikilink; (2) a

anotação de formas alternativas para mesma entidade - apenas nome, apenas sobrenome, apelidos - quando estas formas alternativas possuírem wikilink em outros artigos; (3) o uso da classificação da DBpedia em outras línguas quando a DBpedia em Português classifica como Thing.

Devido à variação do desempenho dos classificadores treinados com *corpora* formados pelo mesmo número de sentenças, consideramos que algumas sentenças podem melhorar a qualidade do classificador, enquanto que outras sentenças podem prejudicar o resultado. Como identificar bons exemplos de sentenças? Uma opção é otimizar a seleção das sentenças que farão parte do *corpus* de treino através de uma função multiobjetivo que busque um bom desempenho com *corpus* de testes de diferentes estilos de escrita - um problema de Otimização Combinatória.

REFERÊNCIAS BIBLIOGRÁFICAS

[AV13] Amaral, D. O. F.; Vieira, R. "O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa". In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013, pp. 59–68.

[BHBL09] Bizer, C.; Heath, T.; Berners-Lee, T. "Linked Data - The Story So Far", *International Journal on Semantic Web and Information Systems*, vol. 5–3, Jan 2009, pp. 1–22.

[BP06] Bunescu, R.; Pasca, M. "Using Encyclopedic Knowledge for Named Entity Disambiguation". In: 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 9–16.

[CMO12] Campos, D.; Matos, S.; Oliveira, J. L. "Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools". In: *Theory and Applications for Advanced Text Mining*, Sakurai, S. (Editor), InTech, 2012, cap. 8, pp. 175–195.

[COM+08] Carvalho, P.; Oliveira, H. G.; Mota, C.; Santos, D.; Freitas, C. "Segundo HAREM: Modelo geral, novidades e avaliação". In: *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, Mota, C.; Santos, D. (Editores), Linguateca, 2008, cap. 1, pp. 11–31.

[Cuc07] Cucerzan, S. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data". In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 708–716.

[Dom12] Domingos, P. "A few useful things to know about machine learning", *Communications of the ACM*, vol. 55–10, Out 2012, pp. 78–87.

[FGM05] Finkel, J. R.; Grenager, T.; Manning, C. "Incorporating non-local information into information extraction systems by Gibbs sampling". In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05, 2005, pp. 363–370.

[GK13] Gurevych, I.; Kim, J. "The People's Web Meets NLP". Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[GS96] Grishman, R.; Sundheim, B. "Message Understanding Conference-6: A Brief History", *Proceedings of the 16th conference on Computational linguistics*, vol. 1, 1996, pp. 466–471.

[HNP13] Hovy, E.; Navigli, R.; Ponzetto, S. P. "Collaboratively built semi-structured content and Artificial Intelligence: The story so far", *Artificial Intelligence*, vol. 194, Jan 2013, pp. 2–27.

[Las03] Laslie, M. "The People's Encyclopedia", *Science*, vol. 301–September, 2003, pp. 1299.

[LIJ+14] Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; Bizer, C. "DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia", *Semantic Web Journal*, vol. 1, 2014, pp. 1–29.

[LMP01] Lafferty, J.; McCallum, A.; Pereira, F. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, vol. 2001–Icml, 2001, pp. 282–289.

[ML03] McCallum, A.; Li, W. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 2003, pp. 188–191.

[MMLW09] Medelyan, O.; Milne, D.; Legg, C.; Witten, I. H. "Mining meaning from Wikipedia", *International Journal of Human-Computer Studies*, vol. 67–9, Set 2009, pp. 716–754.

[MS08] Mota, C.; Santos, D. (Editores). "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM". Linguateca, 2008, 1a edição ed..

[MSB+14] Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; McClosky, D. "The Stanford CoreNLP natural language processing toolkit". In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.

[NMG10] N.V, S.; Mitra, P.; Ghosh, S. "Conditional Random Field Based Named Entity Recognition in Geological text", *International Journal of Computer Applications*, vol. 1–3, 2010, pp. 143–147.

[NMI07] Nguyen, D. P. T.; Matsuo, Y.; Ishizuka, M. "Subtree Mining for Relation Extraction from Wikipedia", *Computational Linguistics*, vol. 22, 2007, pp. 125–128.

[NRR+13] Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; Curran, J. R. "Learning multilingual named entity recognition from Wikipedia", *Artificial Intelligence*, vol. 194, Jan 2013, pp. 151–175.

[NS07] Nadeau, D.; Sekine, S. "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, vol. 30–1, Jan 2007, pp. 3–26.

[NZQW10] Ni, Y.; Zhang, L.; Qiu, Z.; Wang, C. "The Semantic Web – ISWC 2010". Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, *Lecture Notes in Computer Science*, vol. 6496.

[PB14] Paulheim, H.; Bizer, C. "Improving the Quality of Linked Data Using Statistical Distributions", *International Journal on Semantic Web and Information Systems*, vol. 10–2, 2014, pp. 63–86.

[PK01] Poibeau, T.; Kosseim, L. "Proper name extraction from non-journalistic texts". In: In Computational Linguistics in the Netherlands, 2001, pp. 144–157.

[R C14] R Core Team. "R: A Language and Environment for Statistical Computing". R Foundation for Statistical Computing, Vienna, Austria, 2014, Capturado em: http://www.R-project.org/.

[RR09] Ratinov, L.; Roth, D. "Design Challenges and Misconceptions in Named Entity Recognition". In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), 2009, pp. 147–155.

[RS08] Richman, A. E.; Schone, P. "Mining Wiki Resources for Multilingual Named Entity Recognition", *Proceedings of ACL-08*, –June, 2008, pp. 1–9.

[SC07] Santos, D.; Cardoso, N. "Reconhecimento de entidades mencionadas em português". Linguateca, 2007, 1a edição ed..

[TD03] Tjong Kim Sang, E. F.; De Meulder, F. "Introduction to the CoNLL-2003 shared task". In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -, 2003, pp. 142–147.

[Vra13] Vrandecic, D. "The Rise of Wikidata", *IEEE Intelligent Systems*, vol. 28–4, Jul 2013, pp. 90–95.

[WV14] Weber, C.; Vieira, R. "Building a Corpus for Named Entity Recognition using Portuguese Wikipedia and DBpedia". In: I Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish, 2014, pp. 9–15.

[ZCD+12] Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauley, M.; Franklin, M. J.; Shenker, S.; Stoica, I. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing". In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, 2012, pp. 2–2.

ANEXO A – Script R para geração de amostras de sentenças

```
setwd("~")
if(!file.exists("samples")) {
  dir.create("samples")
}
f1 = readLines(con=file("wp_noquote_all_sentences.txt", open = "r"))

setwd("samples")

aSeed = 1
for(aSize in seq(from = 500, to = 8000, by = 500)) {
  dirName = as.character(aSize)
  if(!file.exists(dirName)) {
    dir.create(dirName)
  }

  setwd(dirName)
  for(aSample in 1:20) {
    set.seed(aSeed)
    sentences = sample(f1, size = aSize, replace = FALSE)
    fileName = paste("sample", aSize, aSample, sep = "_")
    writeLines(sentences, fileName)
    aSeed = aSeed + 1
  }
  setwd("..")
}
```

srcCodes/createSampleFiles.R

ANEXO B – Script R para identficar conjuntos de sentenças sem interseção

```
size = seq(from = 500, to = 8000, by = 500)
seed = seq(from = 1, to = 20)

wpSentences = seq(from = 1, to = 1508524)

f = function(aSize, aSeed) {
  set.seed(aSeed)
  sentences = sample(wpSentences, size = aSize, replace = FALSE)
}

combinations = expand.grid(aSize = size, aSeed = seed)

rownames(combinations) = paste(combinations$aSize, combinations$aSeed,
    sep=".")

samples = apply(combinations, 1, function(x) {do.call(f, as.list(x))})

samples.by.size = combinations[order(combinations$aSize), ]

intersect.precision = sapply(rownames(head(samples.by.size, 100)),
    function(x) length(intersect(samples[[x]], samples$`5000.15`)))

no.intersect.precision = which(intersect.precision == 0)

intersect.fmeasure = sapply(rownames(head(samples.by.size, 100)),
    function(x) length(intersect(samples[[x]], samples$`6000.15`)))

no.intersect.fmeasure = which(intersect.fmeasure == 0)

samples.to.test = intersect(names(no.intersect.precision), names(no.
    intersect.fmeasure))
print(samples.to.test)
print(length(intersect(samples$`500.4`, samples$`500.5`)))
```

srcCodes/findNoIntersection.R

ANEXO C – Configuração do Stanford NER para treino dos classificadores

```
map = word=0,answer=1
saveFeatureIndexToDisk = true

useClassFeature=true
useWord=true
useNGrams=true
noMidNGrams=true
useLongSequences=true

useDisjunctive=true
disjunctionWidth=5
maxNGramLeng=6
usePrev=true
useNext=true
useSequences=true
usePrevSequences=true
maxLeft=1

useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=dan2useIC

useOccurrencePatterns=true
useLastRealWord=true
useNextRealWord=true

type=crf

useObservedSequencesOnly=true

useQN = true
QNsize = 25
featureDiffThresh=0.05
```

srcCodes/mixed–sources.prop

Property Name	Description
useWord	Gives you feature for w
useNGrams	Make features from letter n-grams, i.e., substrings of the word
usePrev	Gives you feature for (pw,c), and together with other options enables other previous features, such as (pt,c) [with useTags)
useNext	Gives you feature for (nw,c), and together with other options enables other next features, such as (nt,c) [with useTags)
wordShape	dan2useIC : Caracteres maiúsculos consecutivos tornam-se um único X. Minúsculos, um x. Dígitos, um d. A forma para palavras menores que 4 letras (símbolos) recebe um sufixo composto por ":"seguido do tamanho da palavra (1 a 3).
useSequences	Does not use any class combination features if this is false
usePrevSequences	Does not use any class combination features using previous classes if this is false
useLongSequences	Use plain higher-order state sequences out to minimum of length or maxLeft
noMidNGrams	Do not include character n-gram features for n-grams that contain neither the beginning or end of the word
maxNGramLeng	If this number is positive, n-grams above this size will not be used in the model
useDisjunctive	Include in features giving disjunctions of words anywhere in the left or right disjunctionWidth words (preserving direction but not position)
disjunctionWidth	The number of words on each side of the current word that are included in the disjunction features
useClassFeature	Include a feature for the class (as a class marginal). Puts a prior on the classes which is equivalent to how often the feature appeared in the training data.
useOccurrencePatterns	This is a very engineered feature designed to capture multiple references to names. If the current word isn't capitalized, followed by a non-capitalized word, and preceded by a word with alphabetic characters, it returns NO-OCCURRENCE-PATTERN. Otherwise, if the previous word is a capitalized NNP, then if in the next 150 words you find this PW-W sequence, you get XY-NEXT-OCCURRENCE-XY, else if you find W you get XY-NEXT-OCCURRENCE-Y. Similarly for backwards and XY-PREV-OCCURRENCE-XY and XY-PREV-OCCURRENCE-Y. Else (if the previous word isn't a capitalized NNP), under analogous rules you get one or more of X-NEXT-OCCURRENCE-YX, X-NEXT-OCCURRENCE-XY, X-NEXT-OCCURRENCE-X, X-PREV-OCCURRENCE-YX, X-PREV-OCCURRENCE-XY, X-PREV-OCCURRENCE-X.
useTypeySequences	Some first order word shape patterns.
maxLeft	The number of things to the left that have to be cached to run the Viterbi algorithm: the maximum context of class features used.
useTypeSeqs	Use basic zeroeth order word shape features.
useTypeSeqs2	Add additional first and second order word shape features
useLastRealWord	Iff the prev word is of length 3 or less, add an extra feature that combines the word two back and the current word's shape.
useNextRealWord	Iff the next word is of length 3 or less, add an extra feature that combines the word after next and the current word's shape.

Tabela ANEXO C.1: Propriedades do arquivo de configuração que são utilizadas como *features* para o CRF, e suas finalidades, obtida de http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/ie/NERFeatureFactory.html (acessada em 05-01-2015).