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

RECONHECIMENTO DE ENTIDADES NOMEADAS E RELAÇÕES NO DOMÍNIO DE PRIVACIDADE E RESPONSABILIZAÇÃO

MÍRIAN BRUCKSCHEN

Dissertação de Mestrado apresentada como requisito para obtenção do título de Mestre em Ciência da Computação pelo Programa de Pós-graduação da Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Renata Vieira

Porto Alegre, Brasil
2010

Dados Internacionais de Catalogação na Publicação (CIP)

B888r Bruckschen, Mírian.
Reconhecimento de entidades nomeadas e relações no domínio de privacidade e responsabilização / Mírian Bruckschen. – Porto Alegre, 2010.
115 p.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profa. Dra. Renata Vieira

1. Informática. 2. Linguística Computacional. 3. Processamento da Linguagem Natural. 4. Ontologia. 5. Análise Semântica (Programação). 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 **"Reconhecimento de Entidades Nomeadas e Relações no Domínio de Privacidade e Responsabilização"**, apresentada por Mírian Bruckschen, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional , aprovada em 20/12/2010 pela Comissão Examinadora:

Profa. Dra. Renata Vieira – Orientadora — PPGCC/PUCRS

Profa. Dra. Vera Lúcia Strube de Lima - — PPGCC/PUCRS

Profa. Dra. Aline Villavicencio – — UFRGS

Homologada em 10/05/11, conforme Ata No. 007 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti,
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

"(...) parece que não vês que as palavras são rótulos que se pegam às cousas, não são as cousas, nunca saberá como são as cousas, nem sequer que nomes são na realidade os seus, porque os nomes que lhes deste não são mais que isso, os nomes que lhes deste,"

– José Saramago, "As intermitências da morte".

"What's in a name? that which we call a rose
By any other name would smell as sweet;"

– William Shakespeare, "The tragedy of Romeo and Juliet".

AGRADECIMENTOS

À minha família querida: meus pais Aureo e Margareth, e meus irmãos Gabriel e Daila. Obrigada por fazerem parte da minha vida, por me apoiarem, por se orgulharem de mim e de minhas conquistas e por acreditarem que eu podia fazer sempre mais e melhor. Obrigada pela paciência, pelo carinho e por entenderem todos os meus sumiços sem prazo definido para o escritório. Nada do que eu consegui realizar teria sido possível sem o amor, o suporte e o incentivo de vocês.

Ao companheiro que escolhi para compartilhar minha vida, com seus percalços e alegrias. Ao meu tão amado Leandro, por toda a paciência, incentivo e apoio incondicional às minhas decisões. Por me obrigar a programar, a anotar entidades nomeadas e, por fim, a escrever muito – mesmo quando eu não queria. Obrigada por passar vários dias comigo em silêncio enquanto eu trabalhava, e pelos muito bem-vindos intervalos. Impensável ter conseguido chegar tão longe sem ele ao meu lado o tempo todo, me amando tanto e me fazendo trabalhar.

Ao restante da minha família (inclusive minha família por extensão que entrou na minha vida junto com o Leandro, especialmente os seus pais Leni e Zé), agradeço pelo apoio sempre presente, pelas alegrias que me proporcionaram nos momentos que eu podia estar por perto, e por entenderem e me apoiarem quando eu não podia. Agradeço a todos vocês e prometo que estarei mais presente a partir de agora.

Aos meus amigos, tanto os que conheci nesta caminhada, quanto os de mais tempo e que sempre me apoiaram. Amigos que me ouviram quando eu precisava falar e me falaram quando eu precisava ouvir. Me fizeram rir e me ouviram choramingar outras tantas vezes. Aos queridos Clarissa, Douglas, Fabi, Lari, Leo, Marlo, Pati, Paulo, Sandrinha e Tati, meu muito obrigada. Vocês são uma parte importante da minha vida, e eu tenho sorte por tê-los como amigos.

À minha professora orientadora Renata Vieira, que me iniciou na pesquisa e na academia. Obrigada por ter me dado liberdade para explorar as possibilidades que eu achava mais interessantes dentro da minha pesquisa, e por ter me guiado na direção certa quando eu precisava de orientação e de um conselho mais experiente na nossa área.

Aos colegas do laboratório de PLN, muito obrigada pelas trocas, discussões e trabalhos realizados em conjunto desde os meses antes do mestrado até agora. Ainda aos professores, colegas de aulas e pesquisa e funcionários da FACIN, do CPCA, agradeço por todo apoio e coleguismo durante estes dois anos de mestrado. Um agradecimento especial ao Tomas e ao Kieran por participarem da avaliação e pelos valiosos comentários. Agradeço também aos meus novos colegas e amigos na Prefeitura Municipal de Novo Hamburgo, pela convivência e companheirismo nos últimos meses.

À empresa Hewlett-Packard pelo suporte financeiro e pela oportunidade de trabalhar no projeto *Privacy/APAO*.

RECONHECIMENTO DE ENTIDADES NOMEADAS E RELAÇÕES NO DOMÍNIO DE PRIVACIDADE E RESPONSABILIZAÇÃO

RESUMO

O gerenciamento de grandes volumes de informação é uma área de crescente interesse e pesquisa, tanto na academia quanto na indústria. Diferentes mecanismos já foram propostos com o objetivo de facilitar a criação, gerenciamento e manutenção de bases de conhecimento, e recentemente ontologias têm despontado como um forte candidato para tal função. Ontologias são o principal mecanismo para representação do conhecimento em contextos tecnológicos atuais como o da *Web* Semântica. Entretanto, a construção manual destas ontologias é custosa, dado o montante de informação a ser processada para a execução desta tarefa.

Com esta motivação, este trabalho propõe que a confecção de ontologias, mais especificamente a sua população, pode ser automatizada pela tarefa de Reconhecimento de Entidades Nomeadas (REN). O trabalho compreende diferentes tarefas da área de Processamento de Linguagem Natural: Reconhecimento de Entidades Nomeadas, Reconhecimento de Relações e Aprendizado de Ontologias.

Para a execução da tarefa de população de ontologias, foi construída manualmente uma ontologia do domínio de privacidade e posteriormente desenvolvido um método para executar a sua população através da tarefa de REN. Este método compreende a população da ontologia com instâncias e relações. Para validar este método, foi desenvolvido um sistema que o implementa. Este sistema foi testado sobre um *corpus* montado pela autora deste trabalho. Este *corpus* é composto por documentos da área de privacidade e responsabilização, e da legislação associada a este tema.

São apresentados neste trabalho o método, o sistema desenvolvido, as avaliações a que este trabalho foi submetido e suas conclusões.

Palavras-chave: Processamento de Linguagem Natural; Reconhecimento de Entidades Nomeadas; Relações Semânticas; Aprendizado de Ontologias.

NAMED ENTITY AND RELATIONS RECOGNITION IN PRIVACY AND ACCOUNTABILITY DOMAIN

ABSTRACT

Management of large masses of information is an area growing in interest and research, both in the academic environment and in the industry. Several mechanisms have already been proposed aiming the ease of creation, management and maintenance of knowledge bases, and recently ontologies have been considered as serious candidates for this task. Ontologies are the main mechanism for knowledge representation in technological contexts as the Semantic Web. However, the manual construction of these ontologies is very expensive, due to the amount of information to be processed for the execution of this task.

With this motivation, this work proposes that ontology construction, more specifically their population, can be automatized through the task of Named Entity Recognition (NER). The work comprehends different tasks in Natural Language Processing area: Named Entity Recognition, Relations Recognition and Ontology Learning.

For the execution of the ontology population task, we developed an ontology on the privacy domain and, after that, a method to populate this ontology using NER. This method comprehends population of the ontology with instances and relations. In order to validate this method, we developed a system that implements it. This system was tested over a *corpus* assembled by the author of this dissertation. This *corpus* is composed by documents of privacy and accountability area, and by legislation associated to this subject.

In this dissertation we present the method, the developed system, the evaluations carried on for this work and final conclusions on the obtained results.

Keywords: Natural Language Processing; Named Entity Recognition; Semantic Relations; Ontology Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ACE	*Automatic Content Extraction*
ACL	*Association for Computational Linguistics*
CPCA	Centro de Pesquisa em Computação Aplicada
EI	Extração de Informação
EN	Entidades Nomeadas
ER	Entidade-Relacionamento
GPE	*Geo-Political Entity*
HP	*Hewlett-Packard*
IA	Inteligência Artificial
MUC	*Message Understanding Conference*
OWL	*Web Ontology Language*
PEP	Plano de Estudo e Pesquisa
PLN	Processamento de Linguagem Natural
POS	*Part-Of-Speech*
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul
RC	Representação do Conhecimento
RDF	*Resource Description Framework*
REN	Reconhecimento de Entidades Nomeadas
ReRelEM	Reconhecimento de Relações entre Entidades Mencionadas
UIMA	*Unstructured Information Management Architecture*
W3C	*World Wide Web Consortium*
WSJ	*Wall Street Journal*

SUMÁRIO

1. Introdução

> "Todo começo é involuntário."
> – Fernando Pessoa, em "Mensagem".

> "Tudo é loucura ou sonho no começo. Nada do que o homem fez no mundo teve início de outra maneira – mas já tantos sonhos se realizaram que não temos o direito de duvidar de nenhum."
> – Monteiro Lobato, em "Mundo da Lua".

O fenômeno da *Web* e, mais recentemente, a proposta de uma extensão semântica desta têm despertado o interesse de pesquisadores de diversas áreas dentro da Computação. Várias novas aplicações surgem no intuito de processar e compreender informação disposta na forma textual, alimentando bases de dados, acrescentando significado à informação e servindo de suporte aos usuários, dentre outras aplicações.

Um ramo atuante neste domínio é composto pelos grupos que trabalham com Extração de Informação (EI). Através de técnicas variadas, pesquisas nesta linha buscam identificar informação relevante em alguma mídia, muitas vezes texto. Frequentemente, estas pesquisas são auxiliadas ou embasadas em técnicas de Processamento de Linguagem Natural (PLN).

Além do processamento, outra preocupação dos pesquisadores envolvidos nesta área é o armazenamento e a representação desta informação. No contexto da *Web* Semântica, ontologias são o principal mecanismo para representação de conhecimento, e têm sido um padrão bastante adotado recentemente. Ontologias provêem semântica bem-definida e estrutura flexível para a representação de virtualmente qualquer domínio do conhecimento humano. Além disso, existe uma variedade de ferramentas e ambientes de desenvolvimento voltados para a sua construção, manutenção e utilização em aplicações diversas.

Entretanto, as ontologias em si ainda são de difícil e custosa construção, uma vez que demandam atenção direta de especialistas do domínio que se deseja representar. Adicionalmente, sua construção automática ainda não é uma realidade, levando muitos pesquisadores a investirem nesta linha de trabalho.

Nossa pesquisa, detalhada nesta dissertação de mestrado, propõe que tarefas das áreas de EI e PLN possam também auxiliar na automatização do aprendizado de ontologias. Esta é a motivação principal deste trabalho, cujo contexto e escopo são apresentados nas próximas seções.

1.1 Contextualização

No domínio de privacidade de dados, cada vez mais são percebidos esforços para uma representação expressiva e não-ambígua dos principais conceitos envolvidos, motivados pela crescente importância deste requisito em serviços e soluções fornecidas pela indústria a seus clientes. Uma

das dificuldades inerentes à tarefa de Engenharia do Conhecimento nesse domínio é a falta de padronização e consenso que existe na documentação normativa e não-normativa sobre privacidade.

Embora possa ser esperado que leis devam ser claras e não-ambíguas, a interpretação de cada uma depende largamente da operação executada, do tipo de dado envolvido e do contexto específico. Cada situação de manipulação de dados pode requerer diferentes ações a serem realizadas pelas organizações a fim de preservar a privacidade dos dados dos seus clientes. A definição de quais destas ações são necessárias, desejáveis e opcionais em cada situação é uma tarefa difícil e custosa, e requer conhecimento de especialistas em privacidade.

Estes assuntos têm se tornado cada vez mais importantes em todos os setores da sociedade. Entretanto, apesar do número de soluções propostas especialmente ao longo dos últimos anos, ainda existem muitas questões em aberto acerca da representação do conhecimento para a área.

Algumas das questões mais pertinentes em aberto referem-se ao tratamento e análise das diversas regulamentações sobre privacidade, de forma a automatizar a verificação do cumprimento dos requisitos de privacidade de dados dos indivíduos impostos por estas regulamentações.

Diversas tecnologias têm sido exploradas para automatizar a construção de modelos que representem o domínio da privacidade, suas regras e particularidades, e garantam o seu cumprimento em aplicações computacionais. Para tanto, estes modelos devem prover uma representação concisa e eficiente do domínio, e possibilitar a realização de consultas e verificações.

Devido a tais requisitos, ontologias se apresentam como a escolha natural para a representação deste conhecimento. Entretanto, muitos desafios ainda existem nas tarefas de construção e população de tais ontologias, como a sua criação e população.

Este é o tema principal do projeto de pesquisa onde este trabalho de mestrado se encaixa, o *Privacy/APAO*, conduzido em uma parceria PUCRS/HP.

1.2 Definição do objeto de pesquisa

A partir do contexto apresentado, percebe-se a importância da automatização de tarefas relacionadas à engenharia de ontologias, desde as etapas de criação até sua população e manutenção.

O escopo deste trabalho é a população automática de ontologias com instâncias e relações a partir do processamento de um *corpus* de domínio.

Nesta dissertação de mestrado, apresentamos um método que faz uso da tarefa de Reconhecimento de Entidades Nomeadas (REN), clássica da área de EI, na população de uma ontologia do domínio legal e jurídico, voltada para a área de privacidade e responsabilização na indústria de *software*.

Para validação deste método, apresentamos um sistema que o implementa, avaliações a que este sistema foi submetido e seus resultados.

1.3 Organização do trabalho

O restante desta dissertação está organizado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica deste trabalho, com conceitos e exemplos importantes para a compreensão e contextualização do restante do texto. Trabalhos relacionados a este trabalho, nas suas três principais áreas (Reconhecimento de Entidades Nomeadas, Reconhecimento de Relações e Aprendizado de Ontologias) são apresentados no Capítulo 3.

As principais contribuições deste trabalho são descritas nos Capítulos seguintes: o Capítulo 4 apresenta os recursos utilizados e desenvolvidos, e o Capítulo 5 apresenta o método de cinco etapas proposto para a população de ontologias através do Reconhecimento de Entidades Nomeadas e Relações. O Capítulo 6 detalha o sistema desenvolvido com a finalidade de validação deste método. As avaliações a que este trabalho foi submetido são apresentadas e analisadas no Capítulo 7.

Finalmente, o documento é encerrado no Capítulo 8, com considerações finais sobre o trabalho aqui apresentado, que também relaciona suas contribuições e introduz possibilidades de trabalhos futuros.

2. Fundamentação Teórica

"Some books are to be tasted, others to be swallowed, and some few to be chewed and digested."
– Francis Bacon, em *"Essays, Civil and Moral (Of Studies)"*.

"Science is a human subject, developed by people who step on each other's toes at least as often as they stand on each other's shoulders."
– John Sowa, em *"Knowledge representation"*.

Neste Capítulo, abordamos a fundamentação teórica deste trabalho. Esta fundamentação é composta pelos assuntos a seguir, que serão definidos e ilustrados nas próximas seções: Reconhecimento de Entidades Nomeadas e Relações; e Aprendizado e População de Ontologias.

2.1 Reconhecimento de Entidades Nomeadas

A tarefa de Reconhecimento de Entidades Nomeadas (REN) trata da identificação de entidades nomeadas em textos e sua posterior classificação em alguma taxonomia predefinida.

Segundo Nadeau e Sekine [NS07], entidades nomeadas são aquelas que possuem um ou mais designadores rígidos (conforme definido por Saul Kripke [Kri81]). Desta forma, exemplos de entidades nomeadas são nomes próprios e alguns termos que referenciam espécies biológicas ou substâncias, mas não descrições definidas. Comumente inclui-se entre estas entidades referências precisas a datas ("Dezembro de 2009", mas não "Dezembro" ou "Dezembro passado"), números e unidades monetárias ("R$ 200", "30 milhões") [NS07, SC07a].

REN é uma tarefa fundamental na área de Extração de Informação (EI). Já faz parte de avaliações conjuntas da área desde as primeiras edições destas, como o *Message Understanding Conference* (MUC) [GS96]. Nesta avaliação em particular, a tarefa de REN foi proposta pela primeira vez com a ideia de apresentar um desafio que pudesse apresentar bons resultados em um curto espaço de tempo.

Posteriormente, outras avaliações foram propostas e a tarefa tem crescido em importância e abrangência, incluindo novos tipos de entidade a depender do domínio estudado, novas linguagens e técnicas. Atualmente, é realizado o *Automatic Content Extraction* (ACE), cujas tarefas geralmente incluem o reconhecimento de entidades específicas de interesse [NIS08a].

Em língua portuguesa, é realizado o HAREM (HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas[1]) [SSCV06, SC07b, MS08]. O HAREM realizou sua segunda edição em 2008.

A categorização de Entidades Nomeadas (EN) é um componente indissociável da tarefa de REN, e tem feito parte de sua definição desde as primeiras tentativas em sua direção [CBFR99, GS96]. Desde

[1] Nesta avaliação, são reconhecidas o que os organizadores chamam de Entidades Mencionadas, que também atendem à definição dada anteriormente.

então, várias taxonomias foram propostas para a classificação das EN identificadas, variando em abrangência e rigidez nos critérios de classificação. Algumas destas classificações são apresentadas e comentadas nas subseções que seguem.

A tarefa de classificação das entidades nomeadas identificadas depende da aplicação a que a tarefa se destina. Tipicamente, em avaliações conjuntas é definido um conjunto pequeno de classes, enquanto em aplicações para domínios específicos estas classes são estendidas ou especializadas de acordo com as necessidades e potencialidades do domínio.

No caso do presente trabalho, por exemplo, a condução da tarefa dá-se baseada fortemente no domínio estudado. O mesmo não acontece em avaliações conjuntas, por exemplo, onde as classificações propostas são bastante abrangentes, mas pouco especializadas, como as do HAREM e do ACE [SC07a, NIS08a, Lin08a].

2.1.1 Classificação do MUC (1996)

Um trabalho feito por Grishman e Sundheim [GS96], publicado como um encerramento do MUC-6, descreve uma visão histórica das edições do MUC até então. O MUC foi a primeira avaliação onde figurou a tarefa de REN. As classes constituíam-se em duas apenas (ENAMEX, "*entity name expression*", e NUMEX, "*numeric expression*"), mas eram divididas em subtipos.

A Tabela 2.1 apresenta as classes e subtipos propostos para classificação na tarefa durante a avaliação.

Tabela 2.1 – Classes do MUC-6 para categorização de entidades nomeadas.

Classe	Subtipos	Exemplo de marcação
ENAMEX	PERSON, ORGANIZATION, LOCATION	`<ENAMEX TYPE="PERSON">Dooner </ENAMEX>`
NUMEX	MONEY, PERCENT, TIME	`<NUMEX TYPE="MONEY">400 million</ENAMEX>`

Posteriormente, esta classificação evoluiu na que foi apresentada para a resolução da tarefa em 1999 [CBFR99], que é apresentada na Tabela 2.2. Foram acrescentados, nesta edição, subtipos que dificilmente teriam sido considerados como entidades nomeadas pelas definições clássicas apresentadas até então. Entretanto, houve acordo na comunidade de que são subtipos importantes e que deveriam ser reconhecidos e incluídos na tarefa.

2.1.2 Classificação de Brunstein (2002)

Ada Brunstein [Bru02] apresenta uma categorização com vistas à resolução de uma tarefa de mais alto nível, que é a de perguntas e respostas. A utilização de REN nesta tarefa não é uma técnica nova, e neste documento onde consta a categorização proposta, a autora também apresenta instruções para o relacionamento das entidades identificadas com as possíveis respostas que poderiam

Tabela 2.2 – Classes do MUC-9 para categorização de entidades nomeadas.

Classe	Subtipos	Exemplo de marcação
ENAMEX	PERSON, ORGANIZATION, LOCATION	<B_ENAMEX TYPE="ORGANIZATION"> U.S. Fish and Wildlife Service<E_ENAMEX>
TIMEX	DATE, TIME, DURATION	<B_TIMEX TYPE="DURATION">five years<E_TIMEX>
NUMEX	MONEY, MEASURE, PERCENT, CARDINAL	<B_NUMEX TYPE="MONEY">180 million Canadian dollars<E_NUMEX>

incluí-las. Um exemplo disto é a descrição de entidades do tipo Date, onde a autora informa que estas referem-se a questões do tipo *"how long"* e *"when"*.

A taxonomia proposta é uma extensão daquela já trabalhada em versões anteriores do MUC, com a adição de categorias presentes na literatura da área de sistemas de perguntas e respostas e outras decorrentes da análise do *corpus* utilizado (*Wall Street Journal* (WSJ) *Treebank*).

A Tabela 2.3 apresenta as classes e alguns dos subtipos propostos por Ada Brunstein [Bru02]. A taxonomia totaliza 29 classes. A marcação das entidades é baseada naquela utilizada pelo MUC-6 [GS96].

As classes Pessoa, Instalação[2], Organização, *Geo-Political Entity* (GPE)[3] e Produto foram divididas em Nome e Descritor. Nomes são as entidades nomeadas como as conhecemos ("Luís Inácio Lula da Silva", por exemplo). Descritores são os qualificadores de nomes, como cargos destes ("presidente"). Quanto à classe NORP, é um acrônimo para *Nationality, other, religion, political* (que são justamente os seus subtipos).

2.1.3 Classificação de Sekine (2002)

Satoshi Sekine [Sek08] descreve um trabalho com resultados mais recentes, mas que também vem sendo trabalhado desde que foi apresentado em 2002 [SSN02]. O autor propõe uma taxonomia bastante completa, especificando várias classes de entidades que não receberam atenção em trabalhos anteriores, como produtos (dentre eles, leis e normas, que são as classes de interesse neste trabalho), medidas e fenômenos diversos.

Baseada naquela primeira taxonomia apresentada no MUC-6 [GS96], também divide as entidades em três grandes classes: nomes, tempo e expressões numéricas. A taxonomia foi elaborada com base em análise de textos e entidades candidatas extraídas destes.

A Figura 2.1, de Sekine, ilustra uma visão geral desta taxonomia. Segundo o autor, ela apresenta aproximadamente 150 classes. Devido ao grande número de classes, não será apresentada a taxonomia completa aqui, mas apenas alguns recortes de interesse.

[2] *Facility*, no original inglês.
[3] Entidade Geopolítica, em uma possível tradução

Tabela 2.3 – Classes detalhadas por Brunstein [Bru02] para categorização de entidades nomeadas.

Classe	Subtipos	Classe	Subtipos
Pessoa	-	**NORP**	Nacionalidade Outro Religião Política
Instalação	Edifício Ponte Aeroporto ...	**Organização**	Governo Corporação Educacional ...
GPE	País Cidade Estado ...	**Local**	Rio Região Latitude-Longitude ...
Produto	Arma Veículo Outro ...	**Data**	Data Duração Idade ...
Tempo	-	**Percentual**	-
Dinheiro	-	**Doença**	-
Quantidade	1D 2D 3D ...	**Evento**	Guerra Furacão Outro
Ordinal	-	**Cardinal**	-
Planta	-	**Animal**	-
Substância	Comida Droga Nuclear ...	**Obra de arte**	Livro Peça Música ...
Lei	-	**Língua**	-
Informação de contato	-	**Jogo/esporte**	-

Nesta proposta, as subclasses não são marcadas como subtipos, mas como classes que na taxonomia (e somente lá) estão atribuídas a uma superclasse. Isto fica claro no seguinte exemplo de marcação: `<PERSON>Edgar Allan Poe</PERSON>`. PERSON é subclasse de TOP, mas a marcação é feita diretamente com a classe à qual a entidade pertence. Desta forma, a marcação fica mais clara (sem atributos em demasia no arquivo anotado), mas a taxonomia apresenta um número consideravelmente maior de classes.

As Figuras 2.2 e 2.3 apresentam, respectivamente, um recorte das categorias de mais alto nível na hierarquia proposta e das categorias de PRODUCT. Esta última é a mais relevante no contexto deste trabalho, pois dela descende RULE (que refere-se a leis, normas e outros regulamentos).

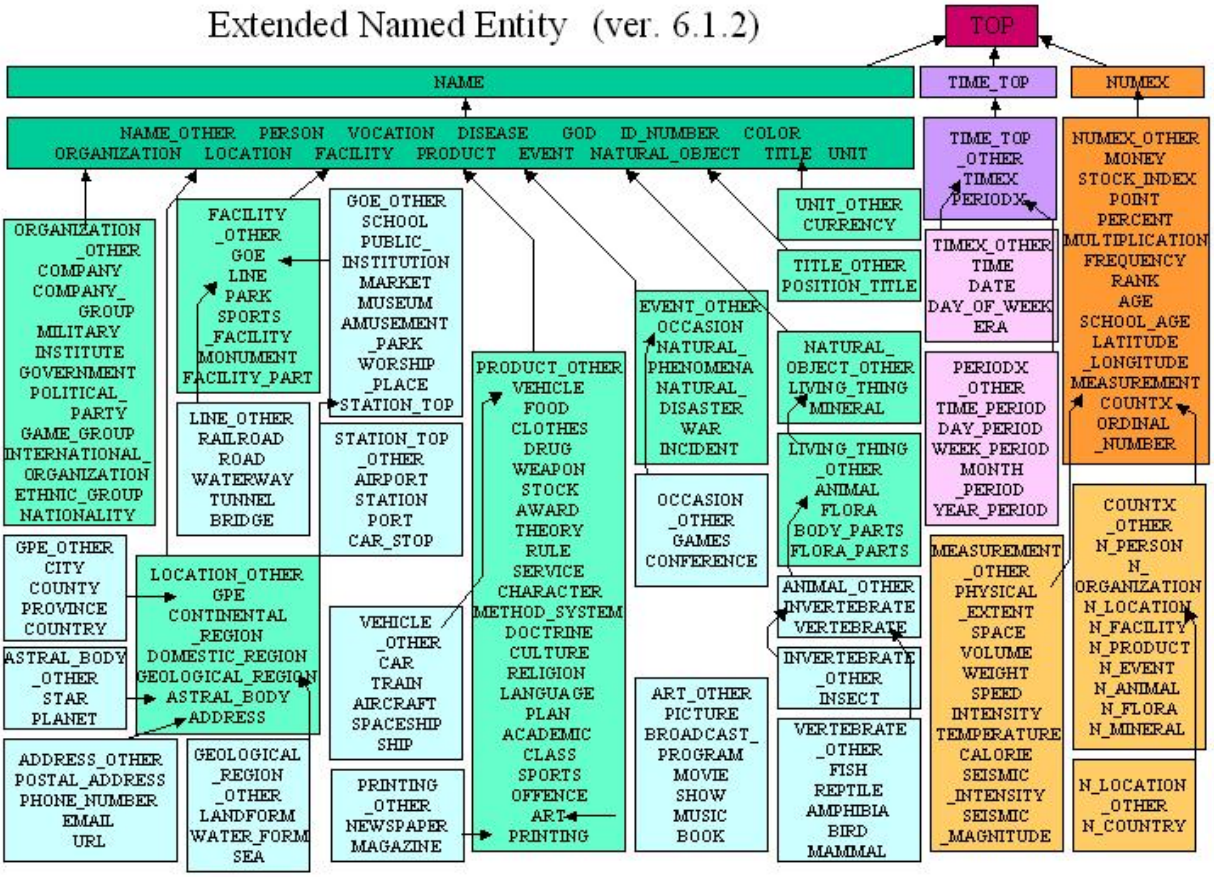


Figura 2.1 – Taxonomia para classificação de entidades nomeadas definida por Sekine [Sek08]. Figura retirada de `http://nlp.cs.nyu.edu/ene/`.

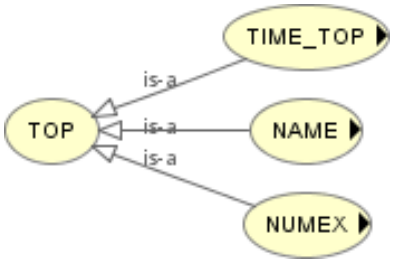


Figura 2.2 – Classes de mais alto nível da taxonomia de Sekine [Sek08].

2.1.4 Classificação do HAREM (2006)

Conforme previamente comentado, o HAREM é uma avaliação conjunta voltada exclusivamente para a língua portuguesa, e teve já duas edições (a última ocorrida em 2008) [SC07b, MS08].

A categorização proposta pelo HAREM para os participantes da avaliação é mais completa do que a de avaliações anteriores (como o MUC), apresentando um conjunto grande de possíveis classificações. Entretanto, há um número reduzido de classes e a sua separação em tipos e subtipos, com um nível a mais de detalhe que o MUC. A Tabela 2.4 apresenta esta taxonomia.

Nesta taxonomia, leis, decretos e outras normas seriam classificadas como OBRA do subtipo

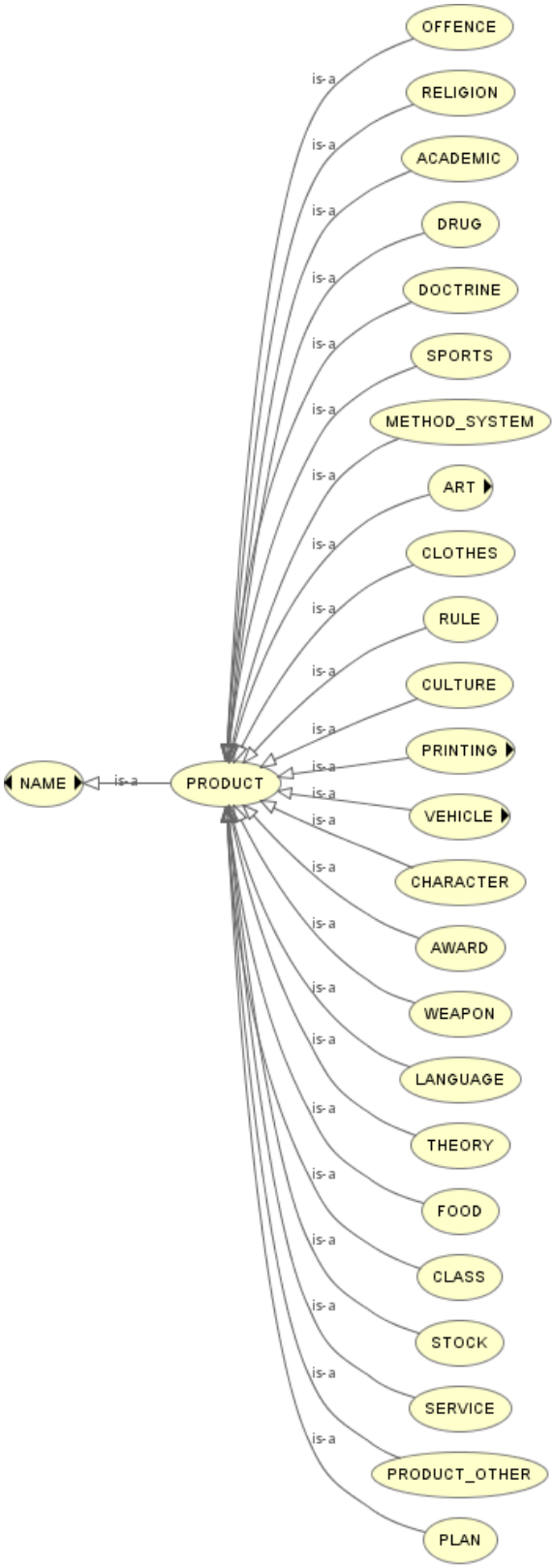


Figura 2.3 – Subclasses da categoria PRODUCT (por sua vez, subclasse de NAME) da taxonomia de Sekine [Sek08].

Tabela 2.4 – Classes propostas na avaliação conjunta do HAREM.

Classe	Tipos	Subtipos
ABSTRACCAO	DISCIPLINA ESTADO IDEIA NOME OUTRO	-
ACONTECIMENTO	EFEMERIDE EVENTO ORGANIZADO OUTRO	-
COISA	CLASSE MEMBROCLASSE OBJECTO SUBSTANCIA OUTRO	-
LOCAL	FISICO HUMANO VIRTUAL OUTRO	ILHA, AGUACURSO, PLANETA, ... RUA, PAIS, DIVISAO, ... COMSOCIAL, SITIO, OBRA, ...
OBRA	ARTE PLANO REPRODUZIDA OUTRO	-
ORGANIZACAO	ADMINISTRACAO EMPRESA INSTITUICAO OUTRO	-
PESSOA	CARGO GRUPOCARGO GRUPOIND GRUPOMEMBRO INDIVIDUAL MEMBRO POVO OUTRO	-
TEMPO	DURACAO FREQUENCIA GENERICO TEMPO_CALEND OUTRO	- HORA, INTERVALO, DATA, ...
VALOR	CLASSIFICACAO MOEDA QUANTIDADE OUTRO	-
OUTRO	-	-

PLANO. Os autores fornecem alguns exemplos em seu exemplário disponibilizado para os sistemas competidores:

- `Lei de Bases da Prevenção e da Reabilitação e Integração das Pessoas com Deficiência`
- `Lei n.º 67/98 de 26 de Outubro`
- `LEI DA PROTECÇÃO DE DADOS PESSOAIS`

2.1.5 Classificação do ACE (1999)

O ACE é uma avaliação conjunta que ocorre desde 1999 com o objetivo de desenvolver tecnologias e recursos para inferir o significado de conteúdo digital (em forma textual e multimídia) automaticamente. Nos últimos anos, a conferência tem crescido em tamanho e importância, e incluído diferentes idiomas e tarefas [DMP+04].

Na sua edição de 2008, as tarefas definidas para avaliação pelo ACE são de reconhecimento de entidades e relações entre estas, localmente e entre documentos diferentes [NIS08a].

A noção de entidade no contexto do ACE é mais abrangente do que as noções utilizadas em outras avaliações, incluindo também pronomes e descrições definidas. São considerados para marcação três níveis de menção:

- NAM (nome), por exemplo "*Joe Smith*";
- NOM (nominal), por exemplo "*the guy wearing a blue shirt*";
- PRO (pronominal), por exemplo "*he/him*" [Lin08a].

Ainda sobre as menções, estas podem ser classificadas como:

- SPC (referência a uma entidade específica, única no mundo),
- GEN (referência a uma classe de entidades),
- NEG (referência a um indivíduo que supostamente não existe, como em "*[No sensible lawyer] would take that case.*") e
- USP (uma referência pouco específica, quando não é possível mapear o referente; um exemplo seria "*[Many people] will participate in the parade.*") [NIS08a, Lin08a].

Quanto à tarefa de REN, ACE propõe a identificação de um subconjunto de entidades apenas, de acordo com sua classificação. A taxonomia completa para classificação das entidades é dada na Tabela 2.5.

Tabela 2.5 – Classes propostas na avaliação conjunta do ACE.

Tipo	Subtipos
FAC (Instalação)	Aeroporto, Prédio, Caminho, ...
GPE (Entidade Geopolítica)	Continente, Distrito, Nação, ...
LOC (Local)	Endereço, Fronteira, Massa de água, ...
ORG (Organização)	Comercial, Educacional, Entretenimento...
PER (Pessoa)	Grupo, Indeterminado, Indivíduo

2.2 Reconhecimento de Relações entre Entidades Nomeadas

O Reconhecimento de Relações entre Entidades Nomeadas trata de identificar relações entre entidades previamente reconhecidas e classificadas.

Dentre as relações comumente incluídas, estão as relações semânticas clássicas, como sinonímia, antonímia e outras ainda. Com o aumento da importância do correto e abrangente processamento de grandes volumes de texto para sumarizar, recomendar e traduzir este texto em informação real e valiosa para análise por sistemas e usuários finais, diferentes relações têm sido introduzidas em tarefas deste tipo, como relações entre pessoas, locais, organizações, ideias e comentários feitos na *Web*.

Por estes motivos, a tarefa de Reconhecimento de Relações entre Entidades Nomeadas tem tido muito mais atenção recentemente, inclusive em avaliações conjuntas, e tende a crescer ainda mais.

Na segunda edição do HAREM (2008), foi introduzida esta tarefa como uma adição ao Reconhecimento das Entidades, na trilha denominada Reconhecimento de Relações entre Entidades Mencionadas (ReRelEM). Nesta trilha, foram propostas quatro relações entre entidades: `ident`, `inclui`, `ocorre_em` e `outra`.

A relação `ident` é atribuída a entidades que refiram-se ao mesmo objeto no mundo. "Pequena Notável", "Carmen Miranda" e "Carmen", em diferentes posições do texto, são exemplos de referências a uma mesma entidade.

A relação `inclui` deve ser atribuída a entidades que incluam outras presentes no texto. Um exemplo seria a inclusão de um lugar em outro, como "Brasil" `inclui` "Rio Grande do Sul", que por sua vez `inclui` "Porto Alegre".

A relação `ocorre_em`, por sua vez, indica a ocorrência ou sede de entidades de EVENTO/ORGANIZAÇÃO em entidades de LOCAL. Alguns exemplos seriam: "Copa 2010" `ocorre_em` "África do Sul", "Exército Zapatista de Libertação Nacional" `ocorre_em` "México", e "São Leopoldo Fest" `ocorre_em` "São Leopoldo".

Finalmente, a relação `outra` seria atribuída a todas as entidades que possuíssem alguma relação diferente das três anteriores (como em "José" `outra` "Pedro", sendo José pai de Pedro, por exemplo).

Ainda no que se refere a avaliações conjuntas com tarefas relacionadas à descoberta e classificação de relações semânticas em textos, pode-se mencionar novamente o ACE [Lin08b]. Em sua

edição ocorrida em 2008, o ACE propunha o reconhecimento de inúmeras relações entre entidades no texto, sendo as relações categorizadas também por tipo e subtipo.

A tarefa de identificação das relações pressupunha a tarefa de REN executada. Além do número e variedade de relações, um diferencial nesta tarefa é a sua execução no contexto de múltiplos documentos. Uma transcrição da definição destas relações é ilustrada na Tabela 2.6.

Tabela 2.6 – Relações propostas na avaliação conjunta do ACE.

Tipo	Subtipos
ART (Artefato)	Usuário, Proprietário, Inventor...
GEN-AFF (Afiliação)	Cidadão, Residente, Religião, ...
METONYMY (Metonímia)	-
ORG-AFF (Afiliação a Organização)	Empregado, Fundador, Estudante...
PART-WHOLE (Parte-Todo)	Artefato, Geográfica, Subsidiária
PER-SOC (Pessoa-Social)	Negócio, Família, Duradouro-Pessoal
PHYS (Físico)	Localizado, Próximo

Fora do contexto de avaliações, a tarefa também tem sido bastante explorada. Trabalhos recentes desenvolvidos nesta área serão apresentados no próximo Capítulo, na seção referente a Reconhecimento de Relações entre Entidades Nomeadas.

2.3 Aprendizado e População de Ontologias

Ontologias têm feito parte e sido objeto de pesquisas desde os primórdios da Filosofia e Ciência, na Grécia Antiga, ou seja, desde muito antes de as áreas de Inteligência Artificial (IA) ou Representação do Conhecimento (RC) serem assim reconhecidas.

Recentemente, o assunto voltou a ser mais discutido com os avanços nas áreas associadas a IA, assim como pelo nascimento da *Web* Semântica. Existem diversas definições para ontologias, cada uma com seus próprios méritos.

Segundo a grande maioria dos trabalhos de Ciência da Computação, ontologias são *"uma especificação explícita de uma conceitualização"*. Esta definição é de Thomas Gruber [Gru95], que define conceitualização como uma visão abstrata e simplificada do mundo que deseja-se representar.

Apesar de correta, esta definição é vaga para os objetivos deste trabalho, já que não explica as potencialidades, expressividade ou limitações das ontologias. Embora este nível de abstração provavelmente fosse o objetivo do autor, em antecipação aos diferentes modelos que poderiam derivar-se daí e que seriam igualmente chamados ontologias, é necessário que definições mais precisas sejam exploradas.

Nicola Guarino [GG95], motivado pela mesma vagueza, discute os significados mais usuais do termo, assim como suas implicações, e apresenta um glossário simples, de acordo com sua avaliação. Neste glossário, o sentido mais comum de ontologias é o mesmo das conceitualizações, isto é, uma estrutura semântica que codifica regras de forma a modelar uma porção da realidade.

Outra definição mais específica que a de Gruber [Gru95], e portanto menos abrangente, mas já voltada para o domínio da Ciência da Computação e suas aplicações é dada por Neches e colegas [NFF+91]. Segundo ela, ontologias são modelos abstratos hierárquicos com informação de domínio, sendo esta composta de conceitos, regras e relações. Já é uma definição que pode ser trazida e aplicada em sistemas do mundo real.

Uma visão ainda mais pragmática é dada por Maedche e Staab [MS01a], que afirmam serem ontologias esquemas para representação de metadados, que apresentam um vocabulário controlado para especificação de conceitos. Este esquema seria representado de forma a ser processável por sistemas computacionais.

Já no contexto de ontologias para a *Web* Semântica, Asunción Gómez-Pérez e colegas [GPFLCG03] fazem um apanhado de várias destas definições e concluem que ontologias são objetos compartilhados e reutilizáveis que visam capturar conhecimento consensual. Afirmam também que estes objetos são construídos cooperativamente por grupos diferentes em locais diferentes.

Neste trabalho, será utilizada a noção de ontologia da *Web* Semântica, com todas as suas potencialidades e limitações associadas. Desta forma, ontologias referem-se a objetos para representação de conhecimento de forma declarativa, comumente por lógicas de descrição e linguagens baseadas nestas, através de conceitos e relacionamentos, e populados com instâncias.

A modelagem de ontologias pode ser definida como a aquisição de conhecimento e sua transferência para um modelo [Cim06]. Esta tarefa, executada de forma manual, ainda é uma tarefa difícil e custosa [MS01a], pois depende de conhecimento do domínio e de experiência de modelagem do engenheiro de ontologias.

Segundo Gómez-Pérez e colegas [GPFLCG03], até os anos 90 o processo de modelagem de ontologias podia ser descrito mais como arte do que como engenharia. Isto significa que não existiam processos bem definidos para a sua construção, e o sucesso dependia altamente das habilidades do indivíduo que criava o artefato. Nesta época, começaram a surgir metodologias para a modelagem de ontologias, mas estas só alcançaram o auge a partir do fortalecimento da percepção de necessidade de ontologias, motivado pela *Web* Semântica.

Gruber [Gru95] propõe que ontologias devem ser projetadas, como quaisquer outros artefatos de *software* normalmente são. Além disso, levanta cinco critérios que devem ser atendidos neste projeto. São eles: clareza, coerência, extensibilidade, viés mínimo de codificação (isto é, não devem ser usadas construções não-portáveis por conveniência, já que isto limita as possibilidades de reuso da ontologia em questão) e comprometimento ontológico mínimo (isto é, especificar apenas os termos necessários para a modelagem da visão desejada, não restringindo desnecessariamente o modelo).

Aprendizado de ontologias é a sua construção automática, evitando ou minimizando trabalho manual do especialista. Geralmente, esta automatização dá-se por meio de técnicas de Processamento de Linguagem Natural (PLN) ou Aprendizado de Máquina [GPFLCG03].

O termo foi cunhado por Maedche e Staab [MS01b], precisamente na época em que a *Web* Semântica ganhou *status* de tema de pesquisa e ontologias despontaram como um potencial formalismo para representação de conhecimento neste ambiente. Neste primeiro trabalho, os autores

ainda apresentam sua ferramenta *Text-To-Onto*, que se propõe a utilizar texto livre, dicionários e mesmo ontologias legadas para a construção de modelos ricos de forma a minimizar o trabalho do especialista do domínio.

Maedche e Staab [MS01b] apresentam a tarefa de aprendizado de ontologias como uma cooperação entre diversas áreas para gerar estes modelos, mas enfatizam a necessidade do especialista, afirmando que a execução totalmente automatizada desta tarefa deve ocorrer apenas em um futuro distante.

A definição dada por Gómez-Pérez e colegas [GPFLCG03] é mais detalhada, incluindo diferentes técnicas de manipulação e criação destes artefatos. Segundo eles, existem quatro principais formas de aprendizado de ontologias: Aprendizado de ontologias com o uso de *corpus*, na qual textos são lidos, processados e os conceitos, instâncias e relacionamentos extraídos dali; Aprendizado de ontologias a partir de instâncias; Aprendizado de ontologias a partir de modelos tais como Entidade-Relacionamento (ER); Aprendizado de ontologias visando interoperabilidade, que abrange mapeamento entre entidades (conceitos, instâncias ou relacionamentos) de duas ontologias.

Cimiano [Cim06] segue outro viés para classificação de técnicas para aprendizado de ontologias, apresentando a tarefa em sub-etapas diferentes, e não por fonte de dados do aprendizado como Gómez-Pérez [GPFLCG03]. Como fonte, considera fundamentalmente textos.

São consideradas tarefas distintas o aprendizado de conceitos, sua hierarquia e relacionamentos, dentre outros. Nesta referência, a definição de aprendizado de ontologias segue aquela dada por Maedche e Staab [MS01b], ou seja, aquisição de um modelo a partir de dados.

A população de ontologias é apresentada como uma tarefa separada, descrita como sendo um processo através do qual extensões para conceitos e relações são aprendidos [Cim06]. É a tarefa de incluir no modelo as suas extensões. Estas extensões são comumente conhecidas por instâncias ou indivíduos. As relações, neste contexto, são comumente chamadas propriedades [Cim06].

Dentre as tarefas da área de Processamento de Linguagem Natural (PLN) relacionadas à população de ontologias, estão a EI e REN. Estas são grandes aliadas na população automática de ontologias através do processamento e análise de textos-fonte em busca de candidatos a indivíduos.

Esta não é uma tarefa cuja execução automática está totalmente resolvida, no entanto, e ainda existe trabalho em andamento na área. Na maior parte das vezes, este processo é executado manualmente com o suporte de ferramentas [GPFLCG03].

3. Trabalhos Relacionados

"A índole natural da ciência é a longanimidade;"
– Machado de Assis, em "O Alienista".

"Nothing shocks me–I'm a scientist."
– Indiana Jones, em *"Indiana Jones and the Temple of Doom"*.

Neste Capítulo, apresentamos trabalhos relacionados ao nosso em suas três principais tarefas, de forma similar ao Capítulo de Fundamentação Teórica: Reconhecimento de Entidades Nomeadas, Reconhecimento de Relações entre estas entidades e Aprendizado e População de Ontologias. Pelo fato de nosso trabalho ser um combinado destas três tarefas, trabalhos das três áreas podem ser relacionados. As seções a seguir apresentam estes trabalhos.

3.1 Reconhecimento de Entidades Nomeadas

Apesar de ser uma área com bastante tradição, o Reconhecimento de Entidades Nomeadas não encontra-se esgotado. Na verdade, muitos trabalhos e eventos com a finalidade específica de investigar o assunto têm sido apresentados nos últimos anos.

Recentemente, tem se observado uma tendência na utilização de recursos com informação semântica de forma a auxiliar na tarefa, aliada ao aprendizado de máquina e heurísticas com a utilização de padrões no texto.

Balasuriya e colegas [BRN+09] apresentam uma avaliação na área de Reconhecimento de Entidades Nomeadas (REN) utilizando artigos da Wikipedia como *corpus*. Os autores anotam um *corpus* composto de artigos desta enciclopédia e avaliam a dificuldade da tarefa sobre este *corpus*. Segundo eles, a enciclopédia é um recurso valioso para a área de Processamento de Linguagem Natural (PLN), mas também apresenta uma grande dificuldade para a composição de *corpus* para a tarefa, devido à heterogeneidade dos artigos. Afirmam ainda que *corpora* montados com matérias de jornal aparentam ser menos complexos para a utilização na tarefa de REN.

Nothman e colegas [NMC09a], por outro lado, apresentam um trabalho que usa a enciclopédia como material para treinamento para seu sistema. Os autores usam diferentes combinações de uso de recursos providos pela enciclopédia, e apresentam resultados muito bons, em torno de 80% de Medida-F, uma medida ponderada de precisão e abrangência em diferentes proporções, dependendo da fórmula utilizada. Os autores alertam para que seja utilizado material do mesmo domínio a que o sistema se destina a reconhecer, para obtenção de melhores resultados. Esta é uma verdade conhecida na área acerca do treino dos sistemas. A não ser em casos onde a aplicação será genérica, esta é uma abordagem conhecida.

Ainda utilizando este recurso para o desenvolvimento da tarefa de REN, o trabalho apresentado por Clark e Harrison [CH09] utiliza a Wikipedia para desambiguação de entidades nomeadas previamente identificadas, valendo-se do fato de a maioria dos artigos disponíveis na enciclopédia

referir-se a um conceito ou entidade, e de estes serem categorizados. O sistema apresentado afirma ser facilmente transposto para outras línguas que não o inglês, por ser apenas fracamente baseado em recursos dependentes de linguagem.

Sureka e colegas [SMI09] apresentam um *framework* que se propõe a induzir automaticamente a partir de aprendizado de máquina as regras para identificação e classificação de entidades nomeadas. Além disso, introduz uma plataforma na qual o usuário pode criar e manter suas próprias regras. O trabalho é apresentado como uma alternativa para facilitar o desenvolvimento de sistemas de REN.

Acerca da utilização de recursos para a bem-sucedida execução da tarefa, Ratinov e Roth [RR09] apontam a necessidade da utilização de informação não-local na identificação de entidades nomeadas e sua classificação, chamando a tarefa de *knowledge-intensive*[1]. Apesar de descartar a utilização de apenas abordagens baseadas neste tipo de informação e usar várias técnicas combinadas para seu sistema, afirma que técnicas baseadas em conhecimento adaptam-se muito bem a diferentes domínios.

Existem diversas abordagens para a resolução da tarefa de REN, e algumas estratégias adotadas atualmente utilizam várias delas simultaneamente. É o que propõe o trabalho desenvolvido por Kozareva e colegas [KFM$^+$07], voltado para língua espanhola. Combinando três abordagens diferentes (*Hidden Markov Models*, entropia máxima e aprendizado baseado em instâncias), os autores sugerem como hipótese que pode-se obter resultados melhores na tarefa de REN do que utilizando-as separadamente.

Os classificadores são treinados sobre o *corpus* e a abordagem utilizada para cada caso é decidida através de votação. O sistema desenvolvido identifica e classifica as entidades nomeadas em uma taxonomia simples de três grandes categorias. São elas: Pessoa, Organização e Local. Local inclui cidades e países (comumente classificadas como entidades geopolíticas em outras taxonomias) e acidentes geográficos, como rios e montanhas. Os resultados apresentados são excelentes para a identificação (98,5% de acurácia) e muito bons para a classificação das entidades (84,94% de acurácia).

Ainda sobre trabalhos em outras línguas que não o inglês, pode-se mencionar o sistema PALAVRAS-NER [Bic06]. Além de análise morfossintática, o analisador para o português PALAVRAS foi estendido para etiquetar com informação semântica nomes, adjetivos e outros. No que se refere a entidades nomeadas, o PALAVRAS categoriza-as segundo a classificação definida pelo HAREM, competição na qual o sistema foi o vencedor na tarefa de classificação das entidades (63,01%). O analisador utiliza Gramática de Restrições.

Recentemente, no NEWS'2010 (2010 *Named Entities Workshop*), evento promovido pela *Association for Computational Linguistics* (ACL), vários sistemas e tendências foram apresentados [KL10].

Ekbal, Sourjikova, Frank e Ponzetto [ESFP10] apresentam uma análise atual e relevante da tarefa, mencionando a dificuldade da classificação detalhada de entidades nomeadas. Os autores apresentam também um método não-supervisionado para recuperar entidades nomeadas, sua classe e contexto em que ocorrem em grandes bases de texto.

[1] "Muito dependente de conhecimento", em uma tradução possível

O método faz uso de padrões para identificação, heurísticas diversas e informação da WordNet. Os resultados apresentados são bastante razoáveis (59% de Medida-F no melhor caso), e deve-se considerar que este é um trabalho pioneiro na classificação detalhada destas entidades, o que torna o seu levantamento um *baseline* para outros sistemas que venham a trabalhar com o mesmo detalhe em sua pesquisa.

3.2 Reconhecimento de Relações entre Entidades Nomeadas

Conforme foi introduzido no Capítulo 2 (Fundamentação Teórica), o Segundo HAREM foi uma avaliação que ocorreu em 2008 e apresentou pela primeira vez na língua portuguesa a tarefa de Reconhecimento de Relações entre Entidades Mencionadas numa trilha que foi chamada ReRelEM.

Participaram desta trilha de avaliação três sistemas: o REMBRANDT [Car08], o SeRELeP [BVdS08] e o SEI-Geo [Cha08]. Cada um dos sistemas demonstrou objetivos diferentes, e propôs diferentes técnicas no reconhecimento de relações.

O sistema REMBRANDT enviou uma participação completa para avaliação no Segundo HAREM: com identificação e classificação de entidades, e posterior reconhecimento de relações entre elas. As técnicas deste sistema demonstraram resultados bastante interessantes, utilizando regras gramaticais em conjunto com a utilização de uma base de dados externa bastante ampla: a Wikipedia em português. O principal interesse do autor é no reconhecimento de entidades nomeadas associadas a locais geográficos.

Neste sistema, a utilização da Wikipedia vem a resolver a classificação das entidades e desambiguação, no caso de a entidade ter mais de um sentido; o autor cita "Varsóvia", que pode referir-se à cidade ou ao pacto, e "Cuba", nome de um país e de uma cidade portuguesa. Como subproduto do desenvolvimento do sistema REMBRANDT, também foi criada uma API de acesso e recuperação de informação da Wikipedia, a SASKIA.

O sistema SeRELeP propõe-se a demarcar as relações de identidade (*ident*), ocorrência (*ocorre_em*) e inclusão (*inclui*) entre as entidades. Partindo do reconhecimento e classificação de entidades efetuados pelo analisador PALAVRAS, o sistema processa a coleção de textos do HAREM e retorna a mesma coleção com a anotação das relações entre entidades.

Para a identificação e delimitação das entidades no texto, a marcação `prop` do PALAVRAS (nome próprio) é utilizada. Além disso, são usadas para a classificação as etiquetas semânticas deste sistema. O SeRELeP utiliza heurísticas baseadas em padrões sintáticos bastante simples para a extração das relações. Apesar disso, teve pontuação aproximada à melhor na maioria dos casos (sendo inclusive o sistema vencedor na extração da relação de co-ocorrência).

Quanto ao SEI-Geo, este sistema foca exclusivamente na relação de inclusão entre entidades relacionadas a LOCAL – domínio no qual obteve a maior precisão, dos três sistemas participantes. O sistema não anota a relação de identidade, que aparenta ser a mais básica, devido ao seu foco em inclusão entre entidades geográficas. O sistema baseia-se fortemente em ontologias geográficas, que são a motivação para o desenvolvimento deste trabalho.

Recentemente (2010), a avaliação SemEval [HKK+10][2], promovida pela ACL, propôs dentre as suas tarefas uma derivada do reconhecimento de relações simples: a classificação de relações semânticas entre pares de entidades já selecionadas.

As relações para classificação eram: Causa-Efeito, Instrumento-Agente, Produto-Produtor, Conteúdo-Contêiner, Entidade-Origem (como quando uma entidade é detalhada através de sua origem; o exemplo dado pelos proponentes da tarefa é *"letters from foreign countries"*), Entidade-Destino (que segue o mesmo princípio), Componente-Todo, Membro-Coleção, Comunicação-Tópico e, de forma similar ao Segundo HAREM, a relação Outra.

Participaram desta avaliação 10 sistemas. Dentre as abordagens mais comumente utilizadas pelos autores dos sistemas para resolver este problema, estavam combinações de aprendizagem de máquina com o uso de informação semântica, léxica e contextual.

Os três sistemas melhor classificados na tarefa foram: o UTD [RH10] em primeiro lugar, o FBK_NK [NK10] em segundo e o ISI [TH10] em terceiro.

O sistema UTD foi desenvolvido por Bryan Rink e Sanda Harabagiu [RH10]. O sistema faz uso de aprendizado de máquina e utiliza 45 atributos (*features*) para a classificação das relações. Dentre eles, vários com informação léxica e semântica, como hiperônimos da WordNet e outros da FrameNet e PropBank. O sistema realiza a tarefa objetivo em duas etapas distintas: 1) definição do tipo da relação semântica entre o par de entidades proposto, e 2) a direção da relação. O sistema UTD obteve 82.19% de Medida-F na avaliação.

Matteo Negri e Milen Kouylekov apresentam o sistema FBK_NK [NK10]. O sistema baseia-se no treinamento de um classificador bayesiano com um conjunto de atributos considerando principalmente o contexto em torno das entidades que formam o par da relação a ser classificada e informação semântica extraída da WordNet. O sistema FBK_NK obteve 77.62% de Medida-F na avaliação em sua melhor submissão.

Stephen Tratz and Eduard Hovy são os autores do terceiro sistema, o ISI [TH10]. Este utiliza um classificador de máxima entropia. Dentre os atributos utilizados, estão alguns de contexto das entidades, outros de identificação de padrões e busca em *gazetteers* e ainda outros baseados em informação provinda da WordNet e outros recursos similares. O sistema ISI obteve 77.57% de Medida-F na avaliação.

No que se refere ao desenvolvimento da tarefa fora de avaliações conjuntas, o tema também tem sido foco de bastante atenção recentemente.

Hasegawa, Sekine e Grishman [HSG04] apresentam um trabalho bem-sucedido nesta tarefa, usando um método baseado em aprendizado não-supervisionado para a detecção das relações. A principal ideia na qual o trabalho se baseia é o agrupamento de entidades duas a duas de acordo com seu contexto, e a posterior procura de outros pares de entidades que encaixem-se no mesmo padrão. Isto permite a obtenção de relações não estabelecidas previamente. Os autores reportam uma faixa de Medida-F entre 75 e 82%, a depender da relação.

Carlson e colegas [CBHM09] descrevem o uso de aprendizado semi-supervisionado para catego-

[2]Informações, participantes e resultados disponíveis em `http://semeval2.fbk.eu/`.

rizar nomes quaisquer (não somente aqueles que conhecemos por entidades nomeadas) e relações entre estes. O domínio das entidades e relações é predominantemente esportivo, apresentando relações como *"PlaysFor(Athlete, Sports Team)"*. O trabalho apresenta como desafio e principal objetivo a população de uma ontologia com instâncias ditas "de alta confiança", que servirão de modelo de treino para o aprendizado das demais instâncias e relações.

Estas instâncias iniciais são obtidas a partir de texto anotado com informação de POS (*Part-Of-Speech*) através da identificação de múltiplas ocorrências do mesmo padrão. É necessário que haja significância estatística da entidade ou relação no *corpus*.

Foram realizados experimentos preliminares, e a precisão apresentada mostra-se muito boa para a maioria das classes (algumas chegando a 100%). As relações apresentam resultados inferiores mas ainda competitivos para a tarefa, numa média de precisão de 42% entre as relações identificadas.

3.3 Aprendizado e População de Ontologias

Desde que ontologias despontaram como o principal mecanismo para representação de conhecimento no contexto da *Web* Semântica, pesquisadores começaram a investigar o seu aprendizado de forma automática.

Na pesquisa apresentada por Dario Bianchi e Rodolfo Delmonte [BD05], os autores apresentam um método para o aprendizado de uma ontologia com classes e instâncias com foco voltado para uma aplicação de perguntas e respostas. O método utiliza um modelo de discurso gerado por uma outra ferramenta para as etapas de definição das classes e, posteriormente, das instâncias.

Nesta última etapa, é usada a estrutura de predicado-argumento do verbo, uma técnica utilizada em outros trabalhos na área de Extração de Informação (EI) [WSC04]. Esta técnica consiste em identificar no predicado o papel do sujeito do qual se fala. O que foi feito por Bianchi e Delmonte, de forma resumida, é uma espécie de mapeamento destes papeis para as classes previamente identificadas na ontologia.

O trabalho de Amardeilh, Laublet e Minel [ALM05] segue um viés diferente, apresentando uma plataforma para anotação e população semiautomática de ontologias a partir de textos. Esta plataforma usa uma ferramenta para anotação semântica dos textos, e a partir dos textos anotados são mapeados os conceitos, atributos e relações entre conceitos. Este processo é realizado por meio de regras para aquisição de conhecimento que são ativadas quando etiquetas semânticas específicas são identificadas.

O trabalho foi experimentado em um *corpus* do domínio de Direito em francês composto por processos e decisões judiciais, e é formado por um conjunto de ferramentas. Segundo os autores, o processo dá-se em três grandes etapas (seguintes ao pré-processamento pelo analisador semântico): 1) Verificação da árvore conceitual gerada a partir da análise linguística; 2) Definição das regras de mapeamento de etiqueta semântica para conceitos na ontologia; e 3) Processamento automático destas regras no *corpus* e população da ontologia.

A plataforma apresenta resultados positivos (55% de abrangência e 79% de precisão geral), mas

ainda possui a limitação de depender fortemente da criação manual das regras para aquisição.

Por fim, outro trabalho relacionado ao aqui apresentado é chamado Sem@ntica [MFWJ08], de McMichael e colegas. O sistema apresentado se propõe a extrair conhecimento de grandes bases textuais através de REN e do uso de ontologias de domínio. Compreende seis subsistemas (Captura de Documentos, Análise de Estrutura, Extração de Informação - REN, Análise Semântica, Pesquisas e Suporte no Portal).

Os dois últimos subsistemas referem-se a uma funcionalidade especial da ferramenta, que é permitir pesquisas sobre a base de conhecimento gerada e interação com esta através de um portal na Internet. Os subsistemas de interesse nesta revisão bibliográfica são os de Análise de Estrutura do Documento, Extração de Informação e Análise Semântica. O sistema Sem@ntica faz uso da plataforma GATE (*General Architecture for Text Engineering*) e seus resultados preliminares são excelentes (entre 90 e 95%).

Clarissa Xavier e Vera Strube de Lima [XL09] apresentam um estudo sobre a extração de uma estrutura ontológica contendo relações de hiponímia e localização a partir da Wikipedia em língua portuguesa. O principal objeto deste recurso que foi utilizado pelas autoras foi a estrutura de categorias da enciclopédia, que provê um rico material semântico, que pode ser utilizado em diversas aplicações na área de Extração de Informação (EI) e PLN.

O experimento desenvolvido pelas autoras não extrai somente instâncias, mas também toda a estrutura taxonômica da ontologia-alvo, além de relações de localização entre as instâncias identificadas. O estudo de caso apresentado é voltado para o domínio de Turismo, e portanto utiliza este recorte das categorias da Wikipedia. O artigo apresenta os resultados da atribuição das relações de localização, que é de 74,09% de Medida-F.

Outro trabalho recente na área é apresentado por Damljanovic, Amardeilh e Bontcheva [DAB09]. Este não apresenta um foco específico, ao contrário do anterior. As autoras propõem um *framework* para a criação de processos de população de ontologias e anotação semântica baseados em recursos da *Web*. O sistema baseia-se na *Unstructured Information Management Architecture* (UIMA), um recurso bastante poderoso para a criação de aplicações de PLN e EI.

As autoras ilustram a flexibilidade e escalabilidade de seu *framework* introduzindo quatro diferentes ontologias que foram criadas ou atualizadas através do sistema, utilizando repositórios e subsistemas diferentes. Elas afirmam que o sistema apresenta 100% de abrangência na recuperação de informação, porém não explicitam a sua precisão.

O trabalho apresentado por Lucas Drumond e Rosario Girardi [DG10] extrai estruturas taxonômicas a partir de texto, usando abordagens estatísticas (especialmente redes de Markov). Os autores apresentam a técnica motivadora do trabalho, denominada *Probabilistic Relational Hierarchy Extraction* (PREHE), que faz a extração das estruturas através de reconhecimento de relações e outras técnicas de PLN.

Este sistema também apresenta um estudo de caso no domínio do turismo. Os resultados apresentados e comparados com sistemas similares, e ficam em torno de 50% de Medida-F, com abrangência e precisão semelhantes.

Hoifung Poon e Pedro Domingos [PD10] apresentam um trabalho bastante similar, usando basicamente as mesmas técnicas. O sistema desenvolvido pelos autores tem o nome de OntoUSP. Ele aprende relações taxonômicas usando como matéria-prima grupos de expressões lógicas.

O estudo de caso apresentado é composto de resumos de artigos de Biomedicina, e a acurácia resultante é bastante notável: 91%, superando vários trabalhos relacionados apresentados pelos autores.

4. Recursos e Ferramentas

"Never trust a computer you can't throw out a window."
– Steve Wozniak.

"– Open the pod bay doors, HAL.
– I'm sorry, Dave. I'm afraid I can't do that."
– Diálogo entre Dave e HAL 9000 em "2001: A Space Odyssey".

Este Capítulo descreve os recursos e ferramentas utilizados no desenvolvimento da parte prática deste trabalho. Alguns destes recursos foram produzidos em preparação a este trabalho pela autora e seus colegas no projeto *Privacy*/APAO (como a ontologia *Legal* e o *corpus Privacy*, seções 4.6 e 4.7 respectivamente), enquanto outros (NLTK, WordNet, Wikipedia, OWL-API e Protégé, seções 4.1 a 4.5) são de autoria de terceiros e utilizados neste trabalho de diferentes formas. O Capítulo apresenta estas ferramentas, seu histórico e desenvolvedores ou criadores.

4.1 NLTK

NLTK (*Natural Language Tool Kit*) é uma suíte composta por diversos algoritmos de Processamento de Linguagem Natural (PLN) e *corpora* de vários tipos e idiomas com diferentes anotações. Ela foi desenvolvida em 2001 na Universidade da Pennsylvania, em conjunto com um curso de Linguística Computacional. O seu foco é principalmente educacional, e tem se tornado bastante popular no ensino e pesquisa [LB02, BKL09].

O projeto do NLTK foi feito de forma a arranjar um grande número de módulos minimamente dependentes entre si. A ideia principal é facilitar o reuso de pequenas partes da suíte, nas mais diversas aplicações onde isso se fizer necessário. Existe, entretanto, um conjunto de módulos centrais, que definem os principais tipos de dados utilizados nas tarefas de PLN e que são usados em toda a suíte [BL04].

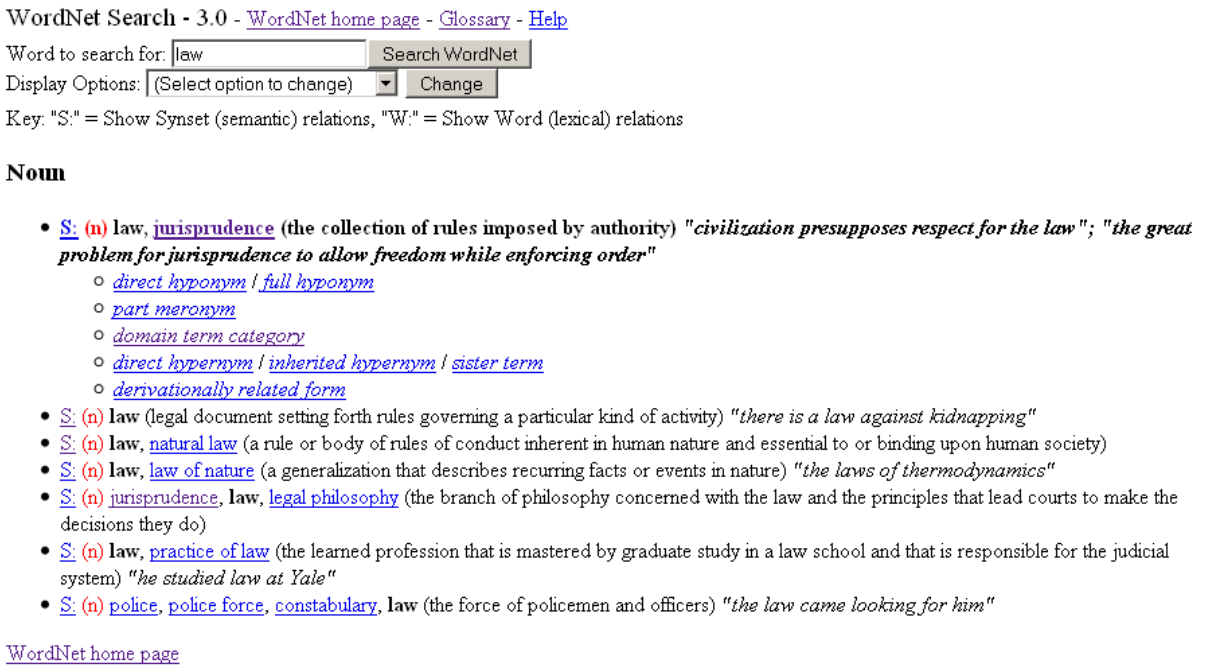
A suíte é livre para uso e distribuição, e foi desenvolvida em Python[1]. A escolha da linguagem deve-se a sua facilidade de aprendizado, para incentivar o uso dos algoritmos e outros módulos por estudantes de PLN. É também a linguagem escolhida para a maior parte do protótipo deste trabalho.

4.2 WordNet

A WordNet é uma base de dados que relaciona unidades léxicas diversas (nomes, verbos, adjetivos e advérbios) através de relações semânticas predefinidas, como sinonímia, antonímia, hipo e hipernímia, dentre outras [Mil95].

Para cada unidade pesquisada, a WordNet recupera um conjunto de definições para esta palavra, associando a seus sinônimos. A Figura 4.1 apresenta o resultado da pesquisa pela unidade *"law"*.

[1] Especificação e *software* disponíveis em `http://www.python.org/`.


WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: law Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) law, jurisprudence (the collection of rules imposed by authority) *"civilization presupposes respect for the law"; "the great problem for jurisprudence to allow freedom while enforcing order"*
  - direct hyponym / full hyponym
  - part meronym
  - domain term category
  - direct hypernym / inherited hypernym / sister term
  - derivationally related form
- S: (n) law (legal document setting forth rules governing a particular kind of activity) *"there is a law against kidnapping"*
- S: (n) law, natural law (a rule or body of rules of conduct inherent in human nature and essential to or binding upon human society)
- S: (n) law, law of nature (a generalization that describes recurring facts or events in nature) *"the laws of thermodynamics"*
- S: (n) jurisprudence, law, legal philosophy (the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do)
- S: (n) law, practice of law (the learned profession that is mastered by graduate study in a law school and that is responsible for the judicial system) *"he studied law at Yale"*
- S: (n) police, police force, constabulary, law (the force of policemen and officers) *"the law came looking for him"*

WordNet home page

Figura 4.1 – Resultado da pesquisa na WordNet pela palavra *"law"*. A pesquisa neste recurso retorna todas as definições disponíveis para a palavra pesquisada, permitindo a navegação por seus sinônimos e outras unidades relacionadas. No caso de *"law"*, uma das unidades com relação de sinonímia recuperadas foi *"jurisprudence"*.

A WordNet é um recurso muito popular entre os pesquisadores de PLN desde seu lançamento, em 1995. Desde então, vários outros projetos associados foram criados, propondo versões da WordNet em outras línguas e com focos diversos. Um exemplo é a Jur-WordNet [STB04], que estende a ItalWordNet (em língua italiana), especializando-a com conteúdo do domínio do Direito.

4.3 Wikipedia

A Wikipedia é uma enciclopédia *online*, gratuita e de uso livre. Foi fundada em 2001, e desde então cresceu em número de artigos e línguas. Em julho de 2010, a Wikipedia em língua inglesa contava com mais de 3 milhões de artigos. Além do inglês, existem versões da Wikipedia para outras 272 línguas, inclusive Português[2].

É um recurso bastante rico, particularmente em conteúdo textual, útil para tarefas de Processamento de Linguagem Natural (PLN), e que vem sendo bastante usado para tal. É o caso de áreas como Aprendizado de Ontologias [SKW07, XL09] e Reconhecimento de Entidades Nomeadas [NMC09b], dentre muitas outras.

A Wikipedia também disponibiliza para *download* diversas coleções para processamento *offline*. Dentre estas coleções, estão: conjunto de artigos de determinada Wikipedia (em língua inglesa, por exemplo) com todos os textos e imagens, conjunto de categorias de artigos, e mesmo conjunto de nomes de artigos.

[2] Conteúdo disponível em `http://www.wikipedia.org/`.

Estas e outras coleções são disponibilizadas pela equipe oficial da Wikipedia em `http://download.wikimedia.org/`, e por um projeto independente em `http://wiki.dbpedia.org/`.

4.4 OWL-API

A OWL-API é uma biblioteca para programação na linguagem Java[3] que permite a fácil interpretação, consulta, criação e modificação de ontologias nos formatos OWL (*Web Ontology Language*)[4] e RDF (*Resource Description Framework*)[5], definidos pelo *World Wide Web Consortium* (W3C).

É a biblioteca atualmente usada no desenvolvimento do editor de ontologias Protégé, que é referência na área de Aprendizado de Ontologias e Engenharia do Conhecimento.

Desde sua primeira versão em 2003 [BVL03], os projetistas e desenvolvedores da OWL-API sempre tiveram como principal objetivo construir um componente que fosse reusável em diferentes aplicações, de editores de ontologias a ferramentas de anotação e agentes de pesquisa.

O lançamento da OWL2 ocasionou mudanças na arquitetura da biblioteca, mas a filosofia de desenvolvimento continua a de prover um componente altamente reusável. Entretanto, seu projeto agora é mais próximo da especificação da linguagem. Seus módulos centrais consistem em interfaces para carregamento, inspeção, manipulação e raciocínio com ontologias OWL2 [HB09].

A biblioteca é extremamente poderosa, e apresenta muito mais recursos do que os atualmente usados nesta pesquisa. A forma de uso e sua instalação são tão simples quanto a de qualquer biblioteca externa a um projeto em Java.

4.5 Protégé

Dentre as ferramentas atualmente disponíveis para criação e manutenção de ontologias *Web Ontology Language* (OWL), provavelmente a mais conhecida e utilizada é o Protégé [GMF+03, KFNM04]. O Protégé é um ambiente de código aberto para criação de ontologias, desenvolvido em Java e extensível através da instalação de *plugins*.

Nesta ferramenta, pode-se facilmente criar conceitos, instâncias e relacionamentos, assim como realizar inferências sobre estes. Dentre as verificações possíveis, e facilitadas pela ferramenta, podemos mencionar a verificação de consistência entre os conceitos e relações (se estes não se contradizem em algum ponto). Na criação de modelos para representação de domínios, esta é uma verificação muito importante, uma vez que pode indicar prováveis erros de modelagem e que levariam a inferências incorretas no futuro.

O Protégé está atualmente em sua versão 4.0, com uma versão de testes por usuários finais (*beta*) 4.1 já disponibilizada[6]. A Figura 4.2 mostra a interface do editor na versão 4.0, com o

[3] Especificação e *software* disponíveis em `http://www.oracle.com/technetwork/java/`.
[4] Especificação disponível em `http://www.w3.org/standards/techs/owl`.
[5] Especificação disponível em `http://www.w3.org/standards/techs/rdf`.
[6] Informações disponíveis em `http://protege.stanford.edu/` em outubro/2010.

suporte de *plugins* adicionais para visualização (OWL Viz[7], ACE View[8]).

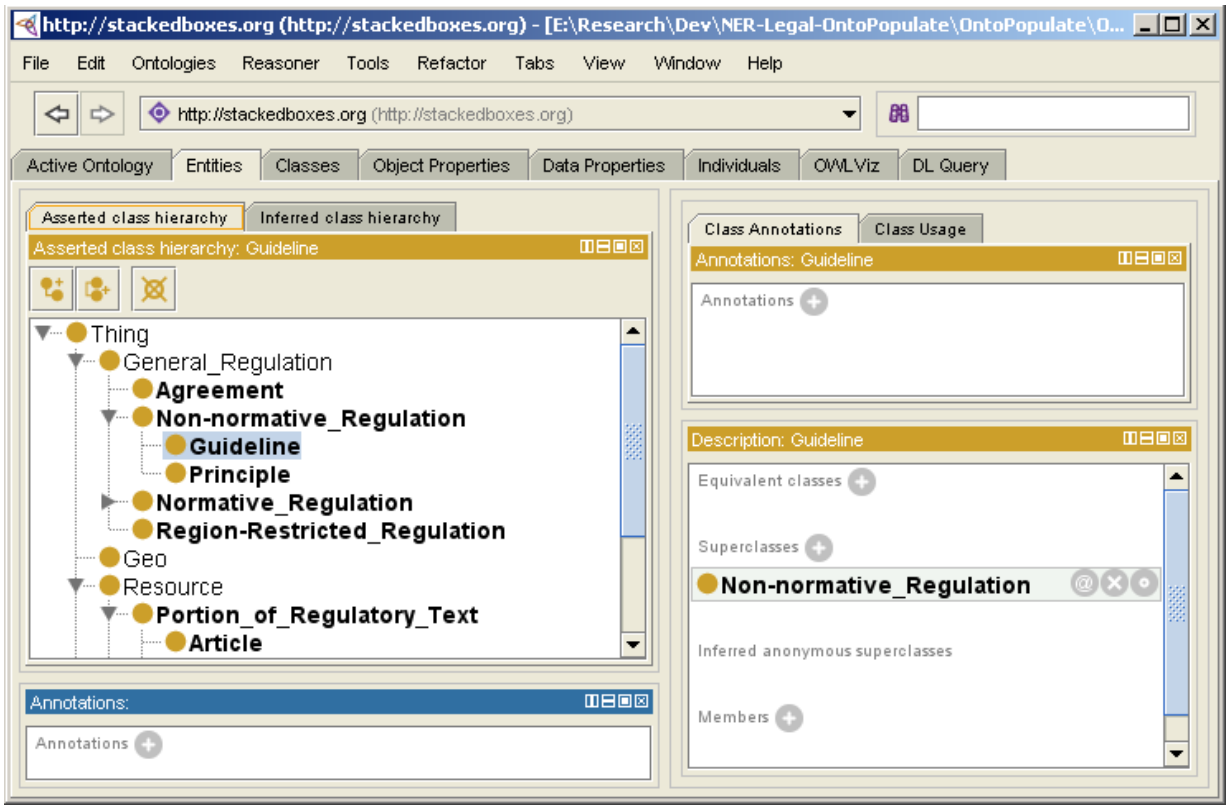


Figura 4.2 – Interface do editor de ontologias Protégé. A ferramenta oferece visões diferentes da ontologia sendo editada, como visão somente de classes, propriedades e instâncias (ou indivíduos). A ontologia visível nesta figura é a *Legal*, que foi desenvolvida para este trabalho e será detalhada na próxima seção.

4.6 Ontologia *Legal*

No contexto do projeto *Privacy*, realizado no Centro de Pesquisa em Computação Aplicada (CPCA) da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), foi desenvolvida uma ontologia que propunha-se a modelar conceitos e relações referentes a normas e regulações, particularmente no domínio de privacidade e responsabilização. Esta ontologia, inicialmente chamada de *Transfer* porque se propunha a modelar especificamente os aspectos legais de ações de transferência de dados de organizações, foi posteriormente chamada de *Legal* e remodelada para contemplar todo o domínio legal e abstrair as ações executadas. A ontologia foi modelada desde o seu princípio pela autora desta dissertação como parte da pesquisa do mestrado.

Desta ontologia foi derivada a taxonomia para a tarefa de Reconhecimento de Entidades Nomeadas (REN) deste trabalho. Desta forma, a ontologia *Legal* propõe-se a ser um modelo genérico para representação de normas e regulamentações, e relações entre estas. Atualmente, focamos principal-

[7] *Software* disponível em `http://www.co-ode.org/downloads/owlviz/`.
[8] *Software* disponível em `http://attempto.ifi.uzh.ch/aceview/`.

mente nas classes de interesse direto para este trabalho, ou seja, normas e regulamentações e suas especializações.

A Tabela 4.1 traz algumas métricas da ontologia em sua versão 4.2. Posteriormente, esta ontologia foi estendida automaticamente com novas classes, como será detalhado mais adiante no Capítulo 6.

Tabela 4.1 – Métricas da ontologia *Legal*.

Número de classes	21
Número de propriedades	3
Número de atributos	2
Número de instâncias	0

É importante observar que, na tabela, este é o número de instâncias antes da população automática da ontologia, isto é, a ontologia inicial sem quaisquer instâncias. Além disso, chamam-se aqui propriedades apenas aquelas que ligam conceitos a outros conceitos, não a literais, isto é, *object properties*. Quanto aos atributos, estes são as propriedades chamadas *data properties*.

A Figura 4.3 apresenta a ontologia *Legal*. As classes de topo principais são: `Regulation`, `Resource`, `Subject` e `Geo`, sendo as duas primeiras as de maior importância.

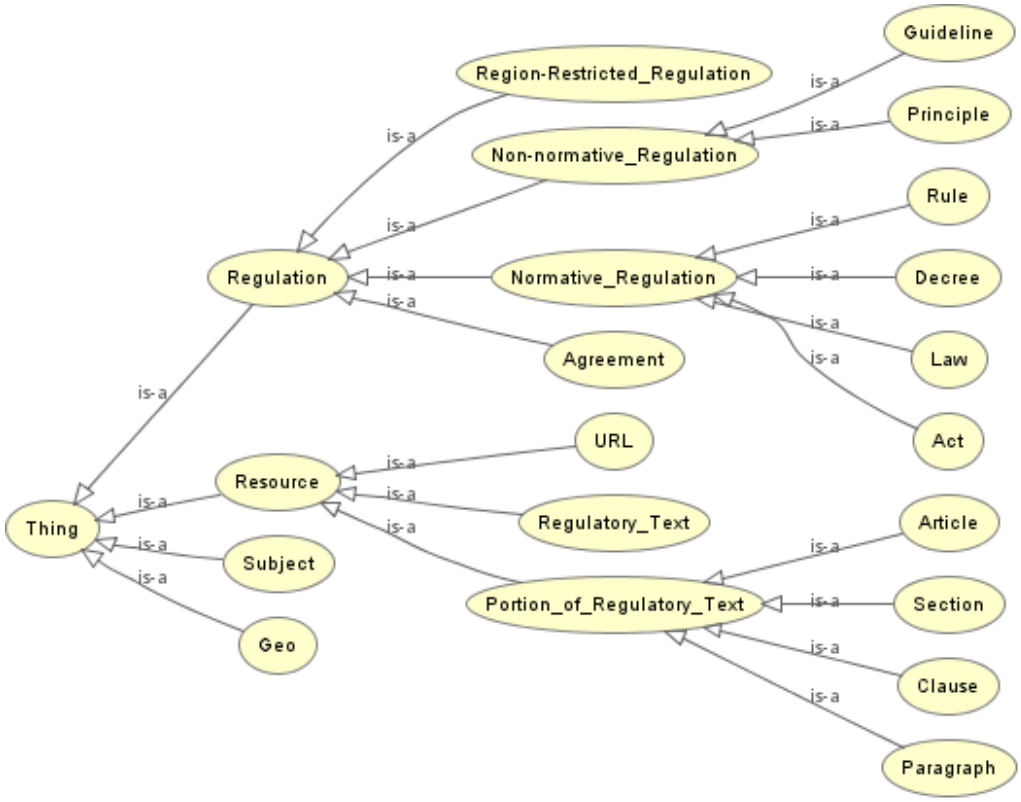


Figura 4.3 – Classes e relações taxonômicas da ontologia *Legal*. A ontologia visa mapear conceitos de interesse do domínio de privacidade e responsabilização, sob um viés legal.

Instâncias de `Regulation` são qualquer entidade abstrata que dite regras para serem seguidas

sob certas circunstâncias. Esta classe já figurou de classificações anteriores para entidades nomeadas, como `OBRA/PLANO` [SC07a], `NAME/PRODUCT/RULE` [Sek08] e `Norm` [HBBB07]. A classe `Resource` foi projetada para recursos que documentem regulamentações, como `URL` por exemplo. `Subject` é o assunto abordado por uma instância de `Regulation` (alguns exemplos são *"transborder data flow"* e *"health information"*). `Geo`, por fim, refere-se a qualquer entidade geopolítica; países, estados, cidades e distritos, por exemplo.

4.7 *Corpus Privacy*

O *corpus Privacy* foi montado pela equipe do projeto de pesquisa no qual este trabalho está inserido. Este recurso é atualmente composto por 100 textos em língua inglesa que versam sobre privacidade e proteção de dados. Em número de palavras, o *corpus* é composto por 1.122.836 palavras. A relação destes textos é apresentada no Apêndice A.

Os textos que compõem o *corpus* são em sua maioria de caráter normativo, mas não unicamente. Também estão incluídos no *corpus* guias de boas práticas de empresas, assim como materiais informativos de governos acerca do tema de privacidade. A Figura 4.4 ilustra a proporção com que os documentos que compõem o *corpus* estão distribuídos de acordo com o seu tipo.

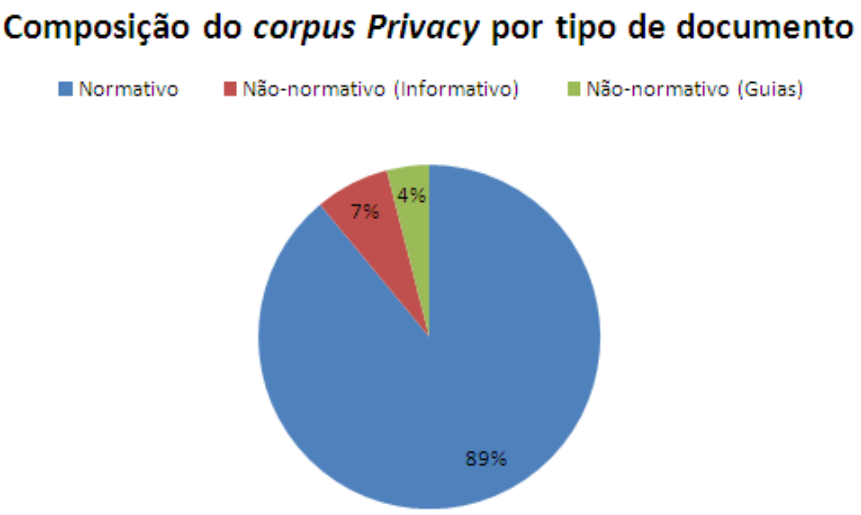


Figura 4.4 – Composição do *corpus Privacy* por tipo de documento. Compõem o *corpus* documentos normativos e não-normativos, sendo destacados dentre estes últimos guias de boas práticas de privacidade e material informativo desenvolvido por governos de países onde existe legislação de privacidade recente. A maior parte deste material é voltado para empresas, para que estas possam cumprir os requisitos legais no que se refere ao assunto.

Documentos de diversos países de origem foram incluídos, mas a maioria destes (aproximadamente metade) são de legislação dos Estados Unidos. A Figura 4.5 ilustra esta proporção. É importante lembrar que muitos países (como o Brasil) ainda não possuem legislação especificamente voltada a privacidade, enquanto muitos outros ainda caminham na direção da elaboração e implementação de tais leis.

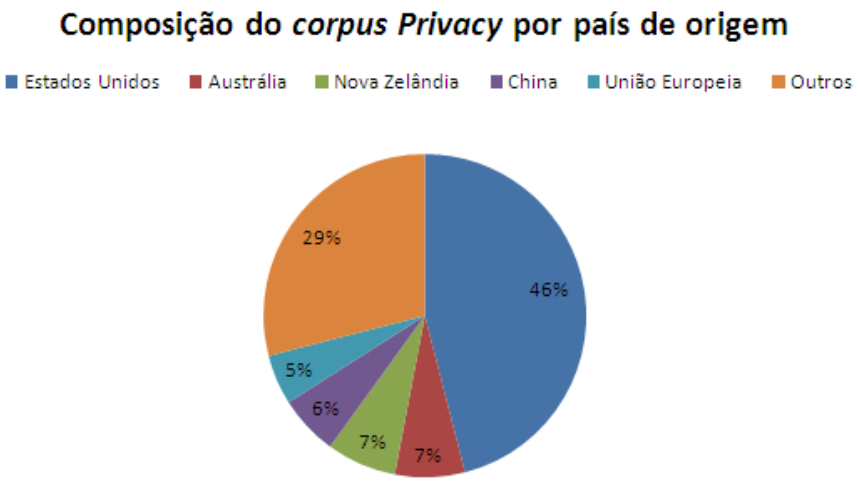


Figura 4.5 – Composição do *corpus Privacy* por país de origem. A maioria dos documentos do *corpus* é dos Estados Unidos por ser um dos primeiros países a instituir leis para a preservação de privacidade.

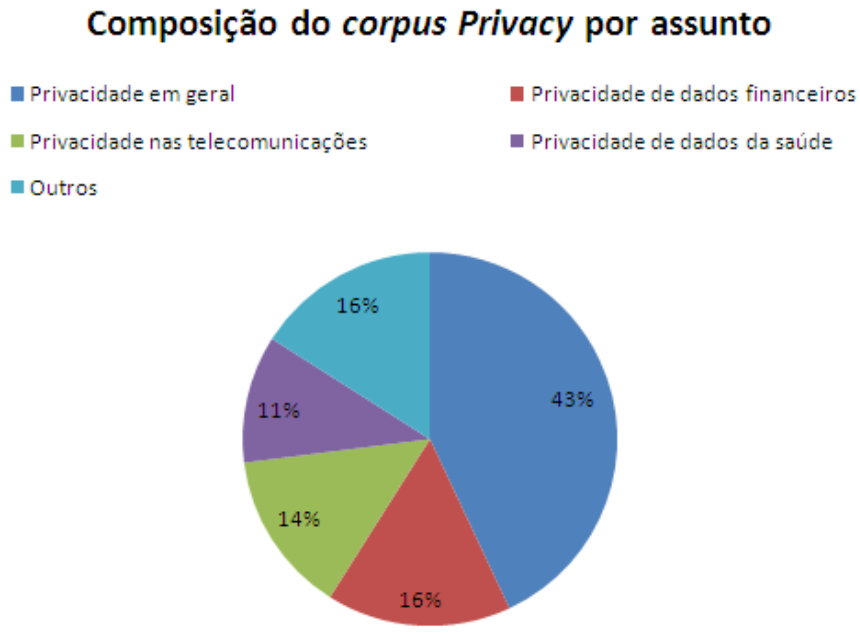


Figura 4.6 – Composição do *corpus Privacy* por assunto principal do documento. A maioria do *corpus* é composta por documentos que tratam de privacidade de forma geral, enquanto outros abordam privacidade específica de algumas áreas, como saúde e finanças.

Esta é uma das razões pelas quais a maior parte do *corpus* é de documentos dos Estados Unidos, onde há bastante tempo já existe legislação sobre o assunto, assim como a cultura de preservação da privacidade dentro das organizações. Outra razão é a disponibilidade destas leis em língua inglesa, requisito para a utilização no projeto no qual este trabalho de mestrado está enquadrado.

No que se refere a assunto, foram coletados documentos que abordavam privacidade, proteção a dados e ações necessárias para a sua garantia de forma genérica, assim como outros em que privacidade em assuntos específicos (privacidade de dados financeiros e bancários e privacidade de dados médicos, dentre outros). A Figura 4.6 ilustra a distribuição do *corpus* no que se refere aos assuntos dos documentos que o compõem.

5. Método para População de Ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações

"(...) et chaque vérité que je trouvois étant une règle qui me servoit après à en trouver d'autres,"
– René Descartes, em *"Discours de la méthode"*.

"Normal people don't understand this concept; they believe that if it ain't broke, don't fix it. Engineers believe that if it ain't broke, it doesn't have enough features yet."
– Scott Adams, em *"The Dilbert Principle"*.

Em busca de automatizar a população de ontologias a partir de texto, focando no domínio estudado, que é de privacidade e responsabilização, foi elaborado um método para a utilização de técnicas de Processamento de Linguagem Natural (PLN) e Extração de Informação (EI) no auxílio a esta tarefa. Estas técnicas são o Reconhecimento de Entidades Nomeadas (REN) e o Reconhecimento de Relações entre estas. O método é detalhado neste Capítulo.

5.1 Visão Geral

O método desenvolvido para a população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações neste trabalho consiste em 5 etapas bem definidas, conforme é ilustrado na Figura 5.1. O método pressupõe um conjunto de textos do domínio estudado e uma ontologia com conceitos do mesmo domínio. Estes recursos são apresentados no Capítulo 4.

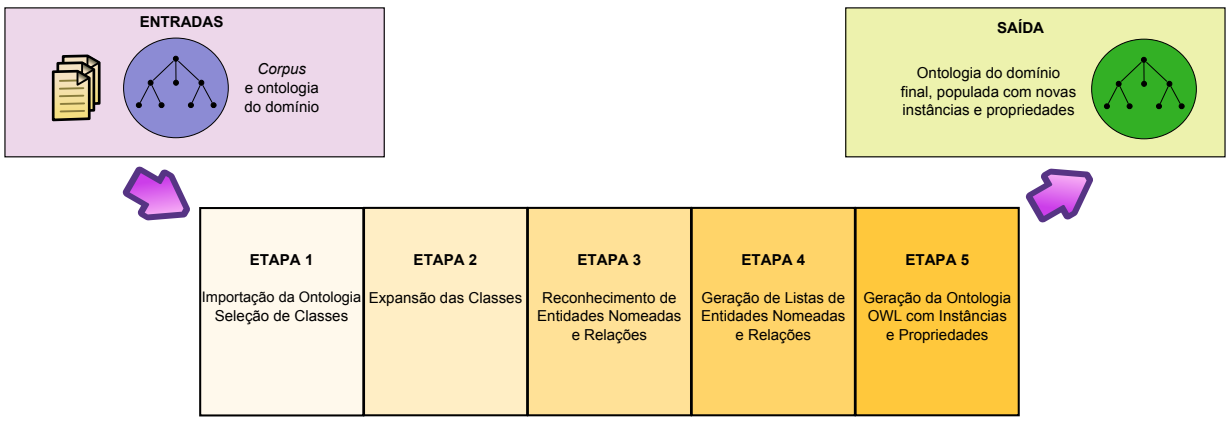


Figura 5.1 – Visão geral do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho. O método é composto de cinco etapas, desde a definição manual de classes da ontologia original a serem instanciadas até a geração da ontologia final em formato OWL contendo as classes originais e novas, instâncias e relações entre estas.

São entradas para o processo a ontologia e o *corpus* desenvolvidos neste trabalho e já mencionados. A partir da ontologia, dá-se na Etapa 1 a seleção de classes a serem focadas nas etapas posteriores.

A Etapa 2 envolve a expansão destas classes a partir do acesso à WordNet. Estas classes serão populadas e terão relações atribuídas entre si na Etapa 3, a partir do Reconhecimento de Entidades Nomeadas e Relações no *corpus* de entrada. Entrada adicional para a Etapa 3 são as heurísticas utilizadas para o Reconhecimento de Entidades Nomeadas e Relações, assim como outros recursos eventuais de conhecimento específico do domínio.

No restante da dissertação, serão abordados os recursos utilizados para a instanciação do método apresentado neste trabalho. A Etapa 4 gera as listas com entidades nomeadas, para posterior população na ontologia OWL final (Etapa 5).

As seções que se seguem apresentam estas etapas detalhadamente.

5.2 Etapa 1: Importação da Ontologia OWL e Definição de Classes para População

Nesta etapa, a ontologia previamente modelada em formato OWL deve ser importada e suas classes apresentadas para o usuário, para que este possa selecionar as classes para expansão (Etapa 2) e população pelo método de Reconhecimento de Entidades Nomeadas e Relações (Etapa 3). A Figura 5.2 ilustra esta etapa.

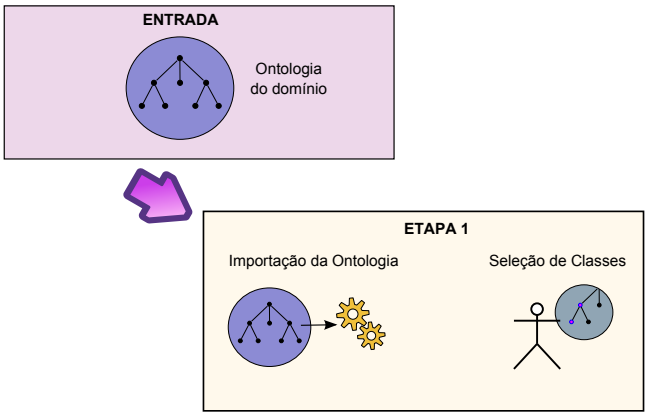


Figura 5.2 – Etapa 1 do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho: importação da ontologia OWL e definição manual de classes para expansão (Etapa 2) e população (Etapa 3).

Esta é uma etapa preparatória e operacional apenas. Nela, há entrada de parâmetros de forma manual e processamento de um recurso já pronto e validado. A inteligência do método concentra-se nas próximas etapas.

Como alternativa à interferência direta do usuário nesta fase, também pode-se entrar direto com as classes de interesse juntamente com a ontologia, a depender do funcionamento da aplicação em questão.

5.3 Etapa 2: Expansão de Classes Definidas

O objetivo desta etapa é a expansão das classes definidas na etapa anterior, para aumentar a abrangência da etapa de descoberta de possíveis instâncias e propriedades, através do Reconhecimento de Entidades Nomeadas e Relações (Etapa 3). Estas classes virão a complementar a ontologia final, na Etapa 5.

A expansão da lista de classes para população se dá através de consulta aos conjuntos de sinônimos destas classes, via WordNet. A Figura 5.3 ilustra esta etapa.

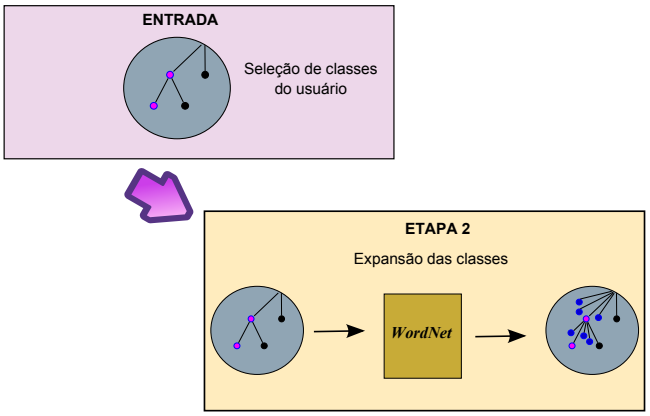


Figura 5.3 – Etapa 2 do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho: expansão de classes definidas na etapa anterior, para posterior população e atribuição de relações (Etapa 3).

5.4 Etapa 3: Reconhecimento de Entidades Nomeadas e Relações

As entradas esperadas para esta etapa são: as classes definidas e expandidas das etapas anteriores, o *corpus* do domínio escolhido e as heurísticas para o Reconhecimento de Entidades Nomeadas e Relações.

Com base nas classes definidas e expandidas nas etapas anteriores, esta etapa analisa o *corpus* de entrada e executa a tarefa de REN, classificando as entidades nomeadas encontradas de acordo com as classes e as heurísticas predefinidas. Cada domínio envolve diferentes tipos de classes, entidades e relações, o que significa diferentes mecanismos para a sua detecção também. O Capítulo 6 apresenta como se deu a aplicação deste método para o domínio de escolha deste trabalho.

É importante observar que, embora as tarefas de Reconhecimento de Entidades Nomeadas e de Reconhecimento de Relações sejam depedentes entre si, frequentemente associadas neste trabalho, e até pertençam à mesma etapa no método proposto, elas são diferentes e exigem técnicas e heurísticas diversas.

Parte da tarefa de Reconhecimento de Relações é inclusive a definição das relações a serem identificadas, que neste método não é restrita a uma relação ou um conjunto delas. Esta definição é entrada para esta etapa, juntamente com as heurísticas para o reconhecimento das relações. A

tarefa de Reconhecimento de Relações segue a de Reconhecimento de Entidades Nomeadas, uma vez que as relações são atribuídas a pares destas entidades.

A Figura 5.4 apresenta os passos desta etapa.

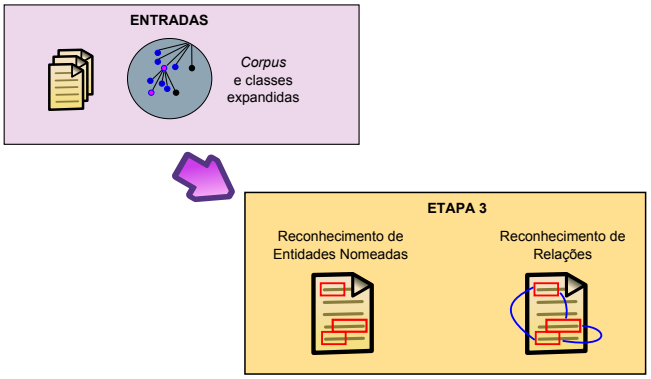


Figura 5.4 – Etapa 3 do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho: Reconhecimento de Entidades Nomeadas e Relações a partir das classes definidas e expandidas, e do *corpus* do domínio escolhido.

5.5 Etapa 4: Geração de Listas de Entidades Nomeadas (EN) e Relações

A Etapa 4 é a geração dos resultados da Etapa 3 de forma a ser possível o processamento pela próxima etapa, que deve ler as entidades nomeadas e relações e associar estas informações à ontologia. A geração de diferentes listas com entidades nomeadas e relações convém a este propósito.

As listas de entidades nomeadas devem conter, acerca de cada futura instância: o seu nome e a classe a que pertence (definida quando da sua identificação e classificação dentre as classes definidas e expandidas nas primeiras etapas do método). As listas de relações devem conter, acerca de cada atribuição de relação: o nome da relação, a instância do lado esquerdo da atribuição e a instância do lado direito da atribuição.

A Figura 5.5 ilustra esta etapa.

5.6 Etapa 5: Geração da Ontologia OWL Final

Esta etapa é a finalização do processo e geração da ontologia OWL final a partir da ontologia original, classes expandidas, e entidades nomeadas (que se tornam instâncias e ajudam a popular a ontologia) e relações (que se tornam propriedades entre as instâncias).

Primeiramente, as novas classes, geradas a partir da expansão das classes selecionadas para população, devem ser incorporadas à ontologia. Estas classes podem ter instâncias associadas a si ou não, a depender do resultado da etapa de REN.

Em um segundo momento, devem ser incluídas na ontologia as instâncias identificadas, classificadas e relacionadas na lista de entidades nomeadas. Esta inclusão deve seguir a classificação feita

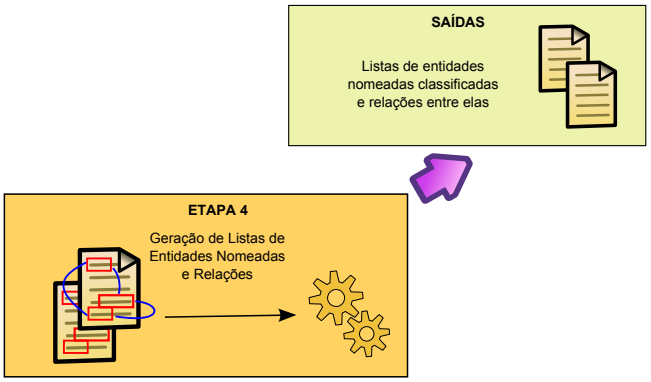


Figura 5.5 – Etapa 4 do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho: geração de listas de Entidades Nomeadas e Relações para posterior população da ontologia final OWL (Etapa 5).

na Etapa 3, isto é, deve ser feita de acordo com as classes informadas no começo do processo para população da ontologia e suas expansões.

Uma vez tendo a ontologia populada, deve ocorrer a atribuição das propriedades entre instâncias (também chamadas *object properties*) a partir das listas de relações entre entidades nomeadas. O processo deve associar duas instâncias existentes na ontologia (incluídas no passo anterior desta mesma etapa) pela relação explicitada na lista de relações.

A Figura 5.6 ilustra esta etapa de finalização do processo e geração da ontologia em formato OWL.

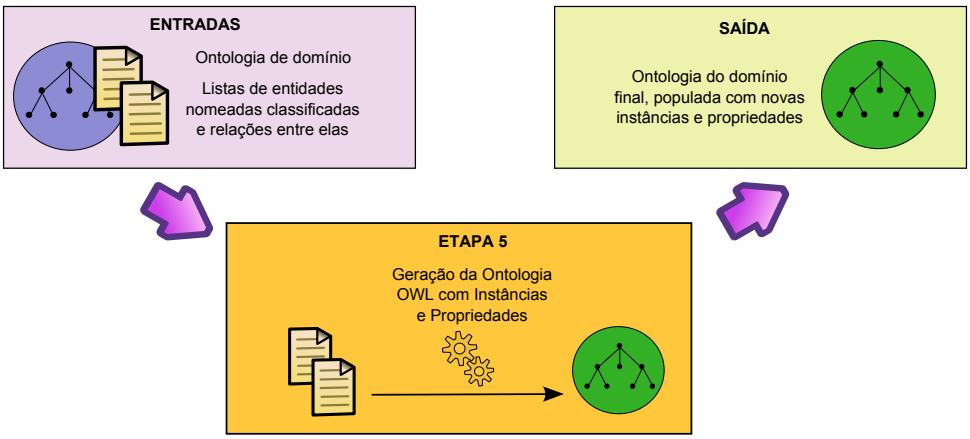


Figura 5.6 – Etapa 5 do método de população de ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações proposto neste trabalho: geração da ontologia final em formato OWL com instâncias e propriedades, com base nas listas geradas na Etapa 4.

A utilização do formato OWL como entrada e saída final deste método de população de ontologias é devida ao seu caráter de padrão para representação de conhecimento no contexto da *Web* Semântica.

OWL é uma linguagem criada com base em *Resource Description Framework* (RDF), provendo vocabulário e semântica adicionais a este, sendo portanto mais adequada a um grande número de aplicações, particularmente aquelas voltadas à descrição de domínios na *Web* Semântica [Hor08].

6. Sistema para Reconhecimento de Entidades Nomeadas e População de Ontologias

"Do, or do not. There is no try."
– Yoda, em *Star Wars: Episode V – The Empire Strikes Back*.

"That's the secret to life... replace one worry with another..."
– Charlie Brown, de *Peanuts*.

Para a experimentação do método genérico para população de ontologias proposto, foi desenvolvido um sistema dividido em módulos ou subsistemas. Estes subsistemas processam a ontologia e o *corpus* do domínio estudado, que é Privacidade e Responsabilização, executam as tarefas de Reconhecimento de Entidades Nomeadas e Relações e transpõem para a ontologia as instâncias e relações resultantes deste processamento. O sistema completo, sua arquitetura e funcionamento são detalhados neste Capítulo.

6.1 Visão Geral

O sistema desenvolvido para a validação do método proposto segue as etapas apresentadas no Capítulo 5, porém reagrupando estas etapas em módulos com funções específicas. Estes módulos são: Módulo 1, de Pré-processamento e Parametrização; Módulo 2, de Reconhecimento de Entidades Nomeadas (REN) e Relações (chamado *NER-Legal*); e Módulo 3, de População de Ontologias (chamado *OntoPopulate*). A Figura 6.1 ilustra como foram distribuídas as etapas em módulos no Sistema apresentado neste Capítulo.

A linguagem de programação Python foi usada para a prototipação da maior parte do sistema. O módulo 1 faz uso também da linguagem de programação Java e da API OWL-API, além da linguagem Python e da suíte NLTK, utilizadas para a integração com a WordNet para a expansão das classes da ontologia, além da separação de sentenças e unidades léxicas (*tokens*), e etiquetação de *Part-Of-Speech* (POS), pré-processando o *corpus* que será usado adiante. O sistema utiliza a linguagem Python nos módulos 2 e 3, de REN e População de Ontologias. A Figura 6.2 apresenta a arquitetura geral do sistema com todos os seus módulos, que serão detalhados nas próximas seções.

6.2 Módulo 1: Pré-Processamento e Parametrização

Este módulo envolve duas etapas do método de População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações: Etapa 1, de Importação da Ontologia e Seleção de Classes, e Etapa 2, de Expansão de Classes. Além disso, ele pré-processa o *corpus* de domínio para a Etapa 3, de Reconhecimento de Entidades e Relações, que é executada pelo módulo seguinte.

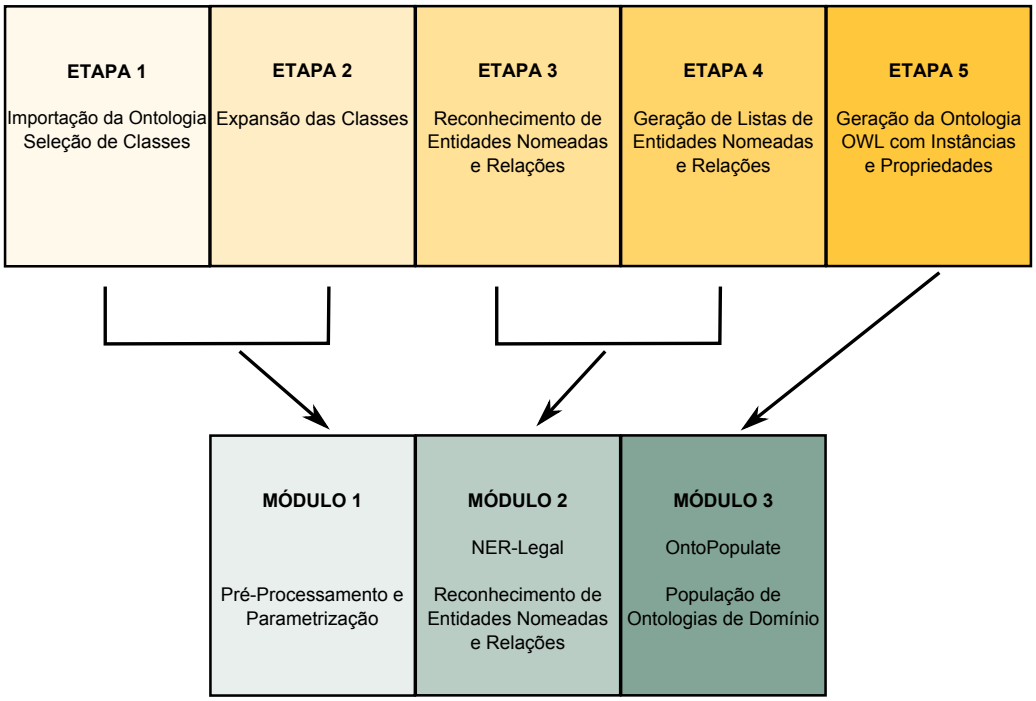


Figura 6.1 – Distribuição das cinco etapas propostas no Método de População de Ontologias a partir do Reconhecimento de Entidades Nomeadas e Relações nos três módulos do sistema desenvolvido para validação do método.

A importação da ontologia *Web Ontology Language* (OWL) original para população é feita usando a linguagem Java e a biblioteca OWL-API, já que não foi encontrada uma biblioteca robusta para a manipulação de ontologias OWL em Python, a linguagem principal de desenvolvimento deste sistema. A seleção de classes é feita através do arquivo de configuração `ner-legal.conf`, que obedece a uma estrutura simples e flexível: linhas começadas com # representam comentários, e linhas começadas com `class:` representam classes que foram selecionadas para população pela aplicação. Pode-se observar este arquivo na versão utilizada nesta dissertação no Exemplo 6.1. Outras configurações podem ser incluídas no mesmo arquivo para a inclusão no sistema obedecendo a mesma estrutura.

Exemplo 6.1 – Arquivo de configuração do Sistema de População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações.

```
# Classes to populate
class: Law
class: Act
class: Rule
```

Para a execução da Etapa 2, de Expansão das Classes selecionadas pelo usuário, é utilizada a linguagem Python com o módulo de ligação do NLTK com a WordNet. Esta etapa é executada para ampliar a abrangência de entidades a serem buscadas na próxima Etapa, quando da Identificação de Entidades Nomeadas no *corpus*. Assim, quando o sistema procurar entidades para a classe `Law`, ele automaticamente utilizará a expansão feita para esta classe e passa a buscar também `Jurisprudence`. De igual forma, quando procurar entidades para a classe `Rule`, ele automatica-

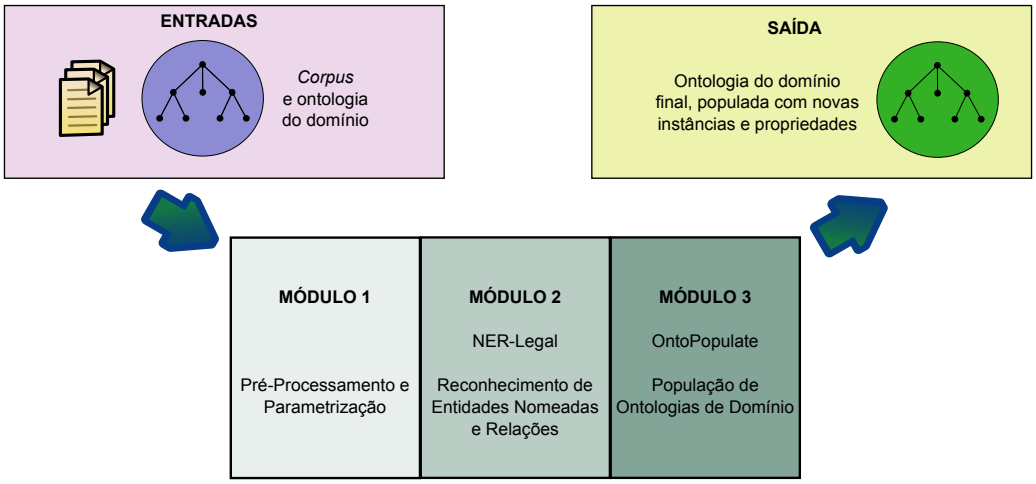


Figura 6.2 – Visão geral da arquitetura do Sistema de População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações, dividida em três grandes módulos com funções distintas.

mente utilizará a expansão desta classe e passará a usar também `Convention`. É um mecanismo bastante simples de expansão da abrangência de busca, mas se provou bastante efetivo, geralmente retornando entidades que relacionam-se com o domínio pretendido.

O pré-processamento do *corpus* para a execução do módulo de Reconhecimento de Entidades Nomeadas e Relações dá-se através de três processos executados na ordem enumerada a seguir, cada um retornando a entrada para o próximo. Os Exemplos 6.2 e 6.3 ilustram trechos de um mesmo texto do *corpus* em suas versões original e após a execução destes três processos.

1. `nltk-annotate`: processo responsável por separar as unidades léxicas (*tokens*) de cada texto do *corpus*, incluindo a sua anotação de POS. Esta separação de unidades léxicas e anotação de POS é aquela provida pela suíte NLTK.

2. `sentence-splitter`: processo responsável por separar as frases de cada texto do *corpus*. Esta separação é aquela provida pela suíte NLTK.

3. `sent-check`: processo responsável por corrigir alguns aspectos da separação de frases providas pelo NLTK. Como a separação de frases provida por esta suíte é baseada fortemente em pontuação, expressões como *"Pub.L."* ficam divididas em duas frases distintas, dificultando a tarefa de Reconhecimento de Entidades Nomeadas, executada a seguir. Neste processo, o módulo 1 rearranja as frases corrigindo alguns destes problemas.

Exemplo 6.2 – Trecho de texto original do *corpus* do domínio utilizado.

```
This part implements the Children's Online Privacy Protection Act of 1998, (15 U.S.C. 6501, et
 seq.,) which prohibits unfair or deceptive acts or practices in connection with the
collection, use, and/or disclosure of personal information from and about children on the
Internet. The effective date of this part is April 21, 2000.
```

Exemplo 6.3 – Trecho de texto do *corpus* do domínio utilizado, anotado e corrigido pelo módulo 1 de Pré-Processamento e Parametrização.

```
('This', 'DT'), ('part', 'NN'), ('implements', 'VBZ'), ('the', 'DT'), ('Children', 'NNP'), ("'
s", 'POS'), ('Online', 'NNP'), ('Privacy', 'NNP'), ('Protection', 'NNP'), ('Act', 'NNP'), ('of
', 'IN'), ('1998', 'CD'), (',', ','), ('(', ':'), ('15', 'CD'), ('U.S.C', 'JJ'), ('.', '.')
('6501', 'CD'), (',', ','), ('et', 'NN'), ('seq.', 'NNP'), (',', ','), (')', 'NNP'), ('which',
 'WDT'), ('prohibits', 'VBZ'), ('unfair', 'JJ'), ('or', 'CC'), ('deceptive', 'JJ'), ('acts', '
NNS'), ('or', 'CC'), ('practices', 'NNS'), ('in', 'IN'), ('connection', 'NN'), ('with', 'IN'),
 ('the', 'DT'), ('collection', 'NN'), (',', ','), ('use', 'VBP'), (',', ','), ('and/or', 'NNP'
), ('disclosure', 'NN'), ('of', 'IN'), ('personal', 'JJ'), ('information', 'NN'), ('from', 'IN
'), ('and', 'CC'), ('about', 'IN'), ('children', 'NNS'), ('on', 'IN'), ('the', 'DT'), ('
Internet', 'NNP'), ('.', '.')
('The', 'DT'), ('effective', 'JJ'), ('date', 'NN'), ('of', 'IN'), ('this', 'DT'), ('part', 'NN
'), ('is', 'VBZ'), ('April', 'NNP'), ('21', 'CD'), (',', ','), ('2000', 'CD'), ('.', '.')
```

Todos os três processos de pré-processamento foram desenvolvidos em Python, sendo apenas os dois primeiros suportados pela suíte NLTK.

A Figura 6.3 apresenta a arquitetura geral do módulo 1 do Sistema de População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações.

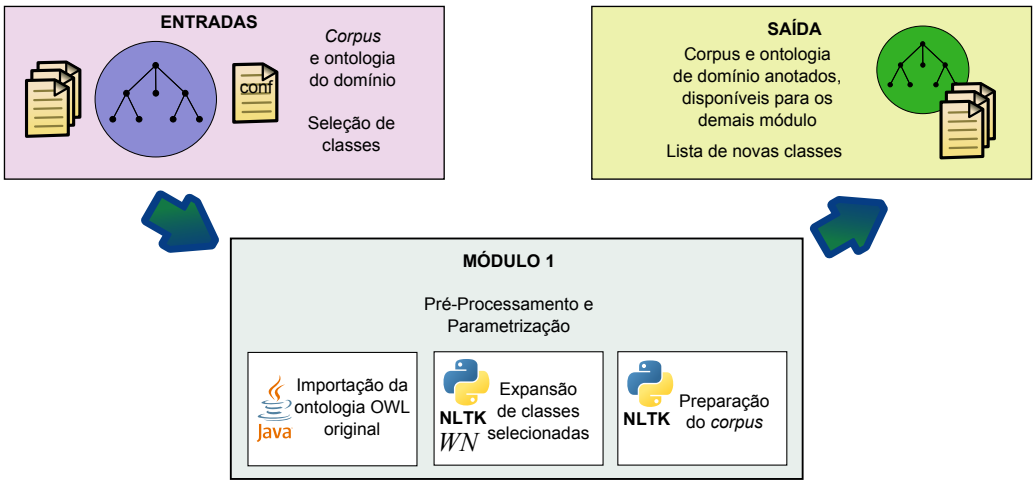


Figura 6.3 – Arquitetura do módulo 1, de Pré-Processamento e Parametrização, do Sistema de População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações.

6.3 Módulo 2: Reconhecimento de Entidades Nomeadas e Relações (*NER-Legal*)

Este módulo envolve duas etapas do método de População de Ontologias: Etapa 3, de Reconhecimento de Entidades Nomeadas e Relações, e Etapa 4, de Geração de Listas de Entidades Nomeadas e Relações. A este módulo dá-se o nome de *NER-Legal*. A maior parte da inteligência do sistema encontra-se neste módulo, por concentrar a forma com que entidades e relações, os principais objetos desta pesquisa, são identificados no *corpus*. Foi desenvolvido totalmente com a linguagem de programação Python.

A Figura 6.4 apresenta a arquitetura geral do módulo 2, NER-Legal.

O módulo envolve dois processos distintos, embora interdependentes: o Reconhecimento de Entidades Nomeadas e o Reconhecimento de Relações. Da execução bem-sucedida destes dois

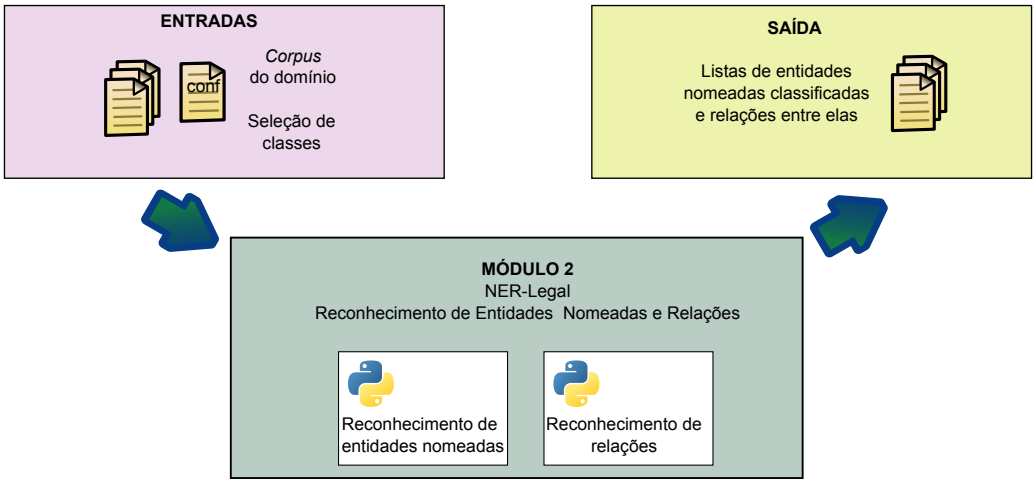


Figura 6.4 – Arquitetura do módulo de REN do Sistema de Reconhecimento de Entidades Nomeadas. Este módulo recebe o *corpus* pré-processado e as seleções de classes para população feitas pelo usuário via parametrização no módulo 1 e concentra a maior parte da inteligência do sistema como um todo. As tarefas de Reconhecimento de Entidades Nomeadas e Reconhecimento de Relações são centrais para o método.

processos depende o resultado final do sistema, daí a importância deste módulo. Nas próximas duas subseções, estes processos serão abordados. A subseção final aborda o processo de geração das listas de saída deste módulo, que serão entrada para o módulo 3, de População de Ontologias.

6.3.1 Reconhecimento de Entidades Nomeadas

No módulo NER-Legal, o processo de Reconhecimento de Entidades Nomeadas identifica dentro do *corpus* as possíveis instâncias (neste momento, ainda como entidades nomeadas) das classes pré-selecionadas e expandidas da ontologia e as classifica. No experimento realizado nesta dissertação, as classes pré-selecionadas foram: `Law`, `Act` e `Rule`.

Para a execução deste processo, foi necessário definir uma heurística para o Reconhecimento das Entidades em si. Esta heurística faz uso intensivo dos nomes das classes pré-selecionadas e aquelas originadas de sua expansão. Nos passos de execução da heurística, os nomes destas classes são chamados de "padrões", e compõem os mecanismos de busca das entidades nomeadas. Os passos desta heurística são enumerados abaixo:

1. Busca de fragmentos do texto onde os padrões apareçam como nomes (e não verbos);

2. Delimitação (à direita) do fragmento de texto pelo próprio padrão;

3. Delimitação (à esquerda) do fragmento de texto por pronome definido ou demonstrativo (*"the"*, *"this"*);

    - Exceção: existe pontuação no intervalo delimitado entre pronome definido e o padrão (neste caso, a delimitação se dará pela unidade léxica posterior à pontuação)

- Exceção: existe uma conjunção no intervalo delimitado entre pronome definido e o padrão (neste caso, a delimitação se dará pela unidade léxica posterior à conjunção)

4. Se houver número ou identificador posterior ao padrão, deve inclui-lo também, estendendo o tamanho da entidade;

5. Se houver preposição qualificadora (*"of"*) posterior ao padrão e que seja seguida de número, deve inclui-los também, estendendo o tamanho da entidade;

6. Normalização

   - Entidades de menos de duas ou mais de dez unidades léxicas são consideradas inválidas
   - Remoção de apóstrofo, aspas, parênteses, espaços e outros caracteres, para fins de unificação com o gabarito (anotação manual do *corpus*)

Exemplos de entidades encontradas e que vêm a popular a ontologia são: "Do Not Call Register Act 2003", "COPPA", "Children's Online Privacy Protection Act" e "Pub.L. 104-____".

6.3.2 Reconhecimento de Relações

O processo de Reconhecimento de Relações baseia-se fortemente no Reconhecimento de Entidades Nomeadas previamente executado. São reconhecidas as seguintes relações, cada uma com seu conjunto específico de requisitos de identificação. Estas relações são descritas a seguir.

Relação 1: same_as

Tendo como domínio e imagem a classe `Thing`, esta relação é uma revisão da relação de "identidade" proposta no Segundo HAREM, e recorrente na área de Processamento de Linguagem Natural como uma subparte da tarefa de Resolução de Correferência. Esta relação contém pares de instâncias que se refiram ao mesmo objeto no mundo. Esta relação é baseada na busca de acrônimos.

Para cada entidade nomeada, é feita uma busca em todo o *corpus* por acrônimos desta entidade que tenham entre 3 e 10 caracteres (caso tenha menos ou mais do que este intervalo, o acrônimo é considerado inválido), que estejam escritos em maiúsculas e que não contenham caracteres numéricos. Além disso, o possível acrônimo deve estar marcado no *corpus* como nome, não verbo ou numeral. Uma vez atribuída a relação, o acrônimo é incluído na lista de entidades nomeadas, sendo atribuída a ele a mesma classe da entidade que deu origem a si.

Exemplos de instâncias desta relação são:

- (*Employee Retirement Income Security Act of 1974*, *ERISA*)
- (*Health Insurance Portability and Accountability Act*, *HIPAA*)
- (*International Telecommunications Convention*, *International Telecommunications Convention of 1973*).

Quando duas entidades originais diferentes (e não acrônimos) têm atribuída a si uma relação `same_as` com o mesmo acrônimo, estas entidades também passam a ter atribuída a relação entre si. Desta forma, quando a relação é atribuída aos pares:

- (*Health Insurance Portability and Accountability Act*, *HIPAA*) e
- (*Health Insurance Portability and Accountability Act of 1996*, *HIPAA*),

então ela também passa a valer para o par:

- (*Health Insurance Portability and Accountability Act*, *Health Insurance Portability and Accountability Act of 1996*).

Relação 2: `references`

Tendo como domínio e imagem a classe `Regulation`, esta relação deve ser atribuída a pares de instâncias de `Regulation` em que uma referencie a outra; a relação não é simétrica, mas pode acontecer de em algum caso um par de instâncias referenciar uma à outra. A identificação desta relação é proposta como um passo adiante na criação de uma rede de recursos normativos e não-normativos, ligados por referências dentro do texto.

A identificação desta relação baseia-se na estrutura de cada texto do *corpus*, assim como no Reconhecimento das Entidades Nomeadas deste. O reconhecimento desta relação assume que cada texto, individualmente, possui um título como primeiro elemento na estrutura do texto, e que contido neste título está o nome do texto. Esta entidade nomeada que dá nome ao texto é associada a todas as demais entidades que figuram no texto através da relação `references`.

Exemplos de instâncias encontradas para esta relação são:

- (*Data Protection Law*, *National Liability Law*)
- (*Law 3471/2006*, *Law 2703/1999*)
- (*Telecommunications Business Act*, *Youth Protection Act*).

Relação 3: `applies_to_geo`

Esta relação contém pares de instância na forma (`Region-Restricted_Regulation`, `Geo`), e é atribuída a instâncias de `Regulation` que se apliquem a um espaço físico-geográfico do mundo.

O reconhecimento desta relação é feito com o uso de um recurso simples, que é uma lista de países e seus gentílicos. Para cada entidade nomeada previamente identificada, é verificado se seu nome contém uma expressão que designa um país ou gentílico contido nesta lista, e em caso positivo, esta entidade é associada com a entidade de país (ou do país do gentílico, se for o caso) através da relação `applies_to_geo`. Caso negativo, um segundo teste ainda é executado: caso na

mesma sentença da entidade nomeada investigada haja uma menção a país ou gentílico do recurso mencionado, a relação também é atribuída, da mesma forma.

A forma com que este processo foi implementado tornou o reconhecimento destas relações bastante flexível: caso um recurso mais completo seja disponibilizado (com regiões ou outras entidades geopolíticas e seus adjetivos pátrios, por exemplo), o sistema não demanda quaisquer alterações.

Todos os nomes de países associados a outra entidade através desta relação são incluídos na lista de entidades nomeadas reconhecidas pelo sistema, e virão a compor a ontologia final populada na classe `Geo`.

Exemplos de instâncias encontradas para esta relação são:

- (*Austrian Federal Law*, *Austria*)
- (*Medicare Australia Act 1973*, *Australia*)
- (*Do Not Call Register Act 2006*, *Australia*).

6.3.3 Geração de Listas com Entidades Nomeadas e Relações

Este módulo retorna diversas saídas, para diferentes finalidades. O total de listas de saída deste processo são:

- Quatro listas com entidades nomeadas, a saber:
    1. Lista com todas as menções a entidades nomeadas, estejam elas duplicadas ou não;
    2. Lista com entidades únicas; entidades repetidas que figuram na lista anterior repetidamente, nesta aparecem apenas uma vez – esta é a lista utilizada para população da ontologia pelo módulo 3, *OntoPopulate*.
    3. Lista com entidades únicas que estão contidas na relação de títulos de artigos na Wikipedia. Quanto à utilização da Wikipedia, estamos no momento fazendo uso da lista de nomes de artigos. Considerando que muitas das leis e atos de legislação, principalmente norte-americanos e europeus, já figuram desta enciclopédia, a lista com os nomes dos artigos da Wikipedia é um mecanismo simples e eficiente para o controle da precisão.
    4. Lista com entidades de `Geo` criadas a partir da associação da relação `applies_to_geo`. O reconhecimento destas entidades não é foco neste trabalho, mas estas entidades eram necessárias para a correta atribuição desta propriedade na ontologia.
- Três listas com relações, a saber:
    5. Atribuições da relação `same_as`;
    6. Atribuições da relação `references`;
    7. Atribuições da relação `applies_to_geo`;

- Uma lista com classes, a saber:

    8. Uma lista com as classes geradas a partir da expansão de classes previamente selecionadas; estas classes virão a complementar a ontologia no módulo 3, de População de Ontologias.

Os formatos das listas de entidades, relações e classes são ilustrados pelos Exemplos 6.4, 6.5 e 6.6, respectivamente.

Exemplo 6.4 – Trecho do arquivo de saída do módulo NER-Legal com entidades nomeadas.

```
[Class=Act]35 and 36 Privacy Act
[Class=Principle]Openness Principle 6
[Class=Act]Local Government Act 2002
[Class=Act]Crown Entities Act 2004
[Class=Act]Coal Industry Act 1994
[Class=Act]Local Government (Wales) Act 1994
[Class=Principle]Data Quality Principle 53
[Class=Law]2001 USA PATRIOT Act Public Law 107-56
[Class=Act]Criminal Procedure (Scotland) Act 1995
(...)
```

Exemplo 6.5 – Trecho do arquivo de saída do módulo NER-Legal com relações.

```
Employee Retirement Income Security Act of 1974 same_as => ERISA
TCPA same_as => Telephone Consumer Protection Act of 1991
Medical Practitioners Act same_as => MPA
Official Information Act same_as => OIA
Official Information Act same_as => Official Information Act 1982
Courts Constitution Act same_as => CCA
Farm Credit Act of 1971 same_as => FCA
Federal Educational Rights and Privacy Act same_as => FERPA
ERA same_as => Employment Rights Act 1996
(...)
```

Exemplo 6.6 – Trecho do arquivo de saída do módulo NER-Legal com novas classes.

```
Normative_Regulation=>Police
Normative_Regulation=>Jurisprudence
Normative_Regulation=>Constabulary
Normative_Regulation=>Ruler
Normative_Regulation=>Normal
Normative_Regulation=>Pattern
Normative_Regulation=>Prescript
Normative_Regulation=>Regulation
Normative_Regulation=>Principle
(...)
```

Os formatos dos arquivos são bastante simples e autoexplicativos. Os arquivos com entidades, além dos nomes das entidades em si (que serão instâncias na ontologia), também incluem suas classes. Os arquivos com relações, incluem a relação e as duas entidades relacionadas em cada uma das atribuições. Quanto ao arquivo de classes, é incluída em cada linha, além da classe, a sua superclasse.

6.4 Módulo 3: População de Ontologias de Domínio (*OntoPopulate*)

Este módulo é responsável pelo cumprimento da Etapa 5 e final do Método de População de Ontologias baseado no Reconhecimento de Entidades Nomeadas e Relações. Ele é o fechamento do sistema apresentado neste Capítulo, populando a ontologia com as instâncias e propriedades identificadas e classificadas nos módulos anteriores. Foi desenvolvido totalmente com a linguagem de programação Python. Não foi usada nenhuma biblioteca específica para a criação do arquivo OWL final.

São três os processos envolvidos neste módulo:

1. Inclusão de Classes resultantes da expansão das classes selecionadas no módulo 1;
2. População da Ontologia com as entidades identificadas e classificadas no módulo 2;
3. População da Ontologia com as relações entre entidades identificadas no módulo 2.

A arquitetura do módulo é ilustrada na Figura 6.5, e cada um dos processos é detalhado a seguir. O resultado final do sistema é a ontologia OWL original com classes adicionais, instâncias e propriedades, identificadas a partir dos módulos anteriores.

A avaliação e resultados do sistema apresentado neste Capítulo são apresentados no Capítulo seguinte, juntamente com considerações acerca destes resultados.

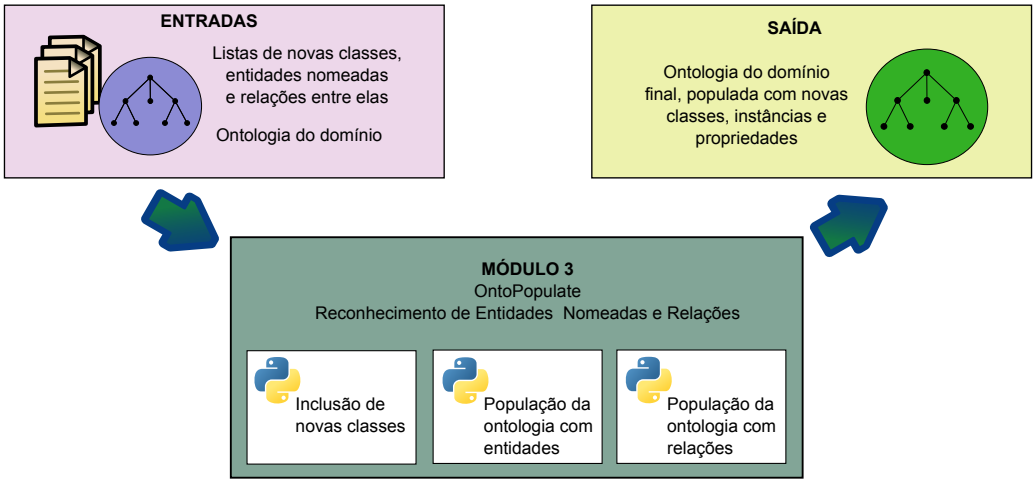


Figura 6.5 – Arquitetura do módulo *OntoPopulate* para População de Ontologias de Domínio.

6.4.1 Inclusão de Classes

Todas as classes resultantes da expansão são incluídas na ontologia, tendo sido identificadas entidades nomeadas para elas ou não. Esta decisão é decorrente da percepção que a ontologia gerada ao final deste módulo não é um artefato final, e sim um recurso que será alterado e complementado de forma manual ou por outros sistemas, e que as novas classes são uma inclusão importante para este fim.

Cada uma das novas classes é incluída como classe irmã daquela que a originou através da expansão. Um trecho de código OWL gerado nesta etapa é dado pela inclusão da classe `Constabulary`, subclasse de `Normative_Regulation`, no Exemplo 6.7.

Exemplo 6.7 – Trecho do arquivo de saída do módulo *OntoPopulate* com novas classes.

```
<owl:Class rdf:ID="Constabulary">
<rdfs:subClassOf rdf:resource="#Normative_Regulation"/>
</owl:Class>
```

6.4.2 População da Ontologia

As entidades nomeadas únicas resultantes do processo de Reconhecimento de Entidades Nomeadas do módulo NER-Legal são incluídas como instâncias neste processo. A classificação da entidade dá-se juntamente com a sua identificação, e no momento atual (de População da Ontologia), esta informação já é dada, não sendo necessário nenhum raciocínio adicional.

Um trecho de código OWL gerado nesta etapa é dado pela inclusão da instância *Civil Rights Act of 1968* à classe `Act`, no Exemplo 6.8.

Exemplo 6.8 – Trecho do arquivo de saída do módulo *OntoPopulate* com instâncias.

```
<!-- http://stackedboxes.org#Civil_Rights_Act_of_1968 -->

<owl:Thing rdf:about="#Civil_Rights_Act_of_1968">
<rdf:type rdf:resource="#Act"/>
</owl:Thing>
```

6.4.3 Atribuição das Propriedades

A atribuição de propriedades entre instâncias na ontologia é resultado da listagem de relações entre entidades nomeadas, que é entrada para este processo. A atribuição das propriedades dá-se de forma semelhante aos dois processos anteriores, isto é, apenas incluindo um novo trecho de código no arquivo OWL.

Um trecho de código OWL gerado nesta etapa é dado pela inclusão da relação `applies_to_geo` no par (*Austrian Federal Law*, *Austria*), no Exemplo 6.9.

Exemplo 6.9 – Trecho do arquivo de saída do módulo *OntoPopulate* com relações.

```
<Law rdf:ID="Austrian_Federal_Law">
<applies_to_geo>
<Geo rdf:ID="Austria"/>
</applies_to_geo>
</Law>
```

7. Avaliação e Resultados Obtidos

"I endeavour to be accurate."
– Spock, em *"Star Trek: The Original Series"*, episódio *"Errand of Mercy"*.

"You know how people are. They only recognize greatness when some authority confirms it."
– Calvin, em *"Homicidal Psycho Jungle Cat"*.

Objetivando calcular e analisar os resultados obtidos com o sistema (Capítulo 6) que implementa o método proposto (Capítulo 5), neste Capítulo apresentamos as avaliações feitas sobre este trabalho. Foram avaliadas as diversas etapas e construções deste trabalho, cada uma da forma julgada mais pertinente ao seu caso.

Também foram produzidos recursos voltados à etapa de avaliação: a anotação do *corpus Privacy*, já apresentado, com entidades nomeadas e sua classificação; e a ferramenta de avaliação do reconhecimento e classificação destas entidades.

Além destes recursos, as seguintes avaliações são detalhadas neste capítulo: das tarefas de Reconhecimento de Entidades Nomeadas e de Reconhecimento de Relações, e da anotação manual do *corpus*.

7.1 Visão Geral

Por ser este um trabalho que estuda tópicos diversos e desempenha tarefas em diferentes áreas, a avaliação também se deu de forma fragmentada. Parte da avaliação foi feita quantitativamente, através de gabarito anotado pela própria autora em boa parte após o desenvolvimento da ferramenta. Este gabarito é a anotação de entidades sobre o *corpus Privacy*, descrito na próxima seção.

Outra parte da avaliação foi feita de forma qualitativa: especialistas do domínio de privacidade avaliaram determinados aspectos do trabalho.

A avaliação quantitativa restringiu-se à tarefa de Reconhecimento de Entidades Nomeadas (REN). A avaliação qualitativa deu-se sobre o reconhecimento de relações e população da ontologia. Os especialistas participantes desta avaliação são pesquisadores do *Privacy Office* da empresa *Hewlett-Packard* (HP).

7.2 Recursos Produzidos

Durante a avaliação ou em preparação a ela, alguns recursos foram produzidos neste trabalho. São eles: a anotação do *corpus Privacy* com entidades nomeadas e sua classificação, e uma ferramenta para avaliação automática da tarefa de REN. Estes recursos são apresentados nesta seção.

7.2.1 *Corpus Privacy* Anotado

A anotação do *corpus Privacy* era necessária para possibilitar a avaliação quantitativa automática da tarefa de REN. Como o *corpus* foi montado pela equipe do projeto a que este trabalho de mestrado pertence, não havia anotação manual previamente existente, e esta teve que ser realizada.

A anotação do *corpus* segue o mesmo padrão adotado para as listas de entidades nomeadas resultantes do Módulo de Reconhecimento de Entidades Nomeadas descrito como parte do Sistema de População de Ontologias no Capítulo Anterior, conforme ilustrado no Exemplo 7.1.

Exemplo 7.1 – Exemplo de anotação do *corpus Privacy*, utilizada na avaliação automática da tarefa de Reconhecimento de Entidades Nomeadas.

```
[Class=Act]Computer Fraud and Abuse Act
[Class=Act]Atomic Energy Act of 1954
[Class=Act]Fair Credit Reporting Act
[Class=Act]Farm Credit Act of 1971
[Class=Act]Securities Exchange Act of 1934
[Class=Act]International Banking Act of 1978
[Class=Act]Federal Reserve Act
[Class=Law]Pub.L. 98-473
```

A anotação totalizou um número de 4863 menções a entidades nomeadas e um total de 1191 entidades únicas do domínio estudado dentro do *corpus Privacy*. Este gabarito é o utilizado pela ferramenta de avaliação, apresentada logo a seguir, para o cálculo de Precisão, Abrangência e Medida-F.

7.2.2 Ferramenta de Avaliação

A ferramenta desenvolvida para avaliação automática da tarefa de REN chama-se *NE-comparacao*.

Desenvolvida em Python, esta ferramenta sempre analisa comparativamente dois arquivos, que são informados como entrada para execução do programa: um de gabarito e um a ser avaliado. A comparação dá-se segundo os seguintes critérios: nome da entidade e classificação da entidade.

No caso da avaliação ser referente a menções de entidades, isto é, considerando menções repetidas a uma mesma entidade, também a contagem de cada menção influencia no resultado. Um exemplo seria o caso de alguma entidade que segundo a anotação manual figura 20 vezes no *corpus* e nos resultados do sistema tem apenas 10 menções contabilizadas. A abrangência no caso desta entidade é de 10/20, afetando também a Medida-F.

A fórmula de Medida-F utilizada é aquela comumente chamada F1, e representa a média harmônica entre precisão e abrangência. A fórmula para F1 é ilustrada na equação 7.1.

$$\mathrm{F1} = \frac{(2 \times \mathrm{Precisão} \times \mathrm{Abrangência})}{\mathrm{Precisão} + \mathrm{Abrangência}} \tag{7.1}$$

Na avaliação deste trabalho, a execução da ferramenta deu-se em três instâncias, com resultados a serem detalhados mais adiante na Seção 7.3:

- avaliação de entidades únicas, comparada contra o gabarito de entidades únicas encontradas na anotação manual no *corpus*, que desconsidera menções repetidas a uma mesma entidade;

- avaliação de entidades únicas com artigo na Wikipedia em língua inglesa, comparada contra o gabarito de entidades únicas encontradas na anotação manual no *corpus*, que desconsidera menções repetidas a uma mesma entidade;

- avaliação de menções a entidades, comparada contra o gabarito de menções a entidades encontradas na anotação manual do *corpus*, que considera igualmente menções repetidas à mesma entidade.

Como saída após o processamento e avaliação das entidades, a ferramenta devolve as medidas Abrangência, Precisão e Medida-F. O Exemplo 7.2 apresenta a saída desta ferramenta para a avaliação de menções a entidades.

Exemplo 7.2 – Saída da ferramenta de avaliação de Reconhecimento de Entidades Nomeadas para a lista de menções a entidades.

```
Abrangencia:  0.30022619782 (1460/4863)
Precisao...:  0.604805302403 (1460/2414)
Medida F...:  0.401264257249
```

7.3 Avaliação do Reconhecimento de Entidades Nomeadas

A avaliação foi executada sobre o *corpus Privacy* estendido, contendo 100 textos e mais de 1 milhão de palavras. O experimento resultou em 2414 menções a entidades (considerando todas as menções feitas a elas), 971 entidades únicas (desconsiderando menções repetidas a estas entidades) e 177 entidades únicas com artigo próprio na Wikipedia em língua inglesa (desconsiderando menções repetidas a estas entidades).

A Tabela 7.1 apresenta, separadamente por classe, a quantidade de menções a entidades, entidades únicas e entidades únicas com artigo na Wikipedia. As novas classes que foram geradas a partir da expansão das originais através da consulta de sinônimos da WordNet estão marcadas com *.

A precisão e abrangência foram computadas pela ferramenta de avaliação em comparação com uma anotação manual do mesmo *corpus* e são apresentadas na Tabela 7.2.

Resultados comparáveis podem ser obtidos da avaliação conjunta ACE 2008, na tarefa de detecção e reconhecimento de entidades localmente (*"Local Entity Detection and Recognition (EDR)"*). Os melhores resultados apresentados nesta tarefa no ACE são de 52,6% [NIS08b], considerando apenas as classes mais comuns para identificação e classificação de entidades: Pessoa, Organização e Local, dentre alguns outros.

Os números que podem ser comparados a esta avaliação conjunta são os relacionados em "Menções a entidades", já que o ACE considera todas as menções a entidades em seus resultados.

Tabela 7.1 – Quantidade de entidades reconhecidas por classe.

	Menções a entidades	Entidades únicas	Entidades únicas (Wikipedia)
Act	2015	737	161
Law	133	78	3
Rule	111	61	5
Constabulary*	2	1	1
Convention*	5	4	1
Enactment*	1	1	0
Number*	10	9	1
Police*	52	17	4
Principle*	52	40	1
Regulation*	33	23	0

Tabela 7.2 – Resultados obtidos na tarefa de REN, integrante do Sistema para População de Ontologias baseado em Reconhecimento de Entidades Nomeadas e Relações.

	Precisão	Abrangência	Medida-F
Menções a entidades	60,48% (1460/2414)	30,02% (1460/4863)	40,13
Entidades únicas	40,06% (389/971)	32,66% (389/1191)	35,99
Entidades únicas (Wikipedia)	74,01% (131/177)	11,00% (131/1191)	19,15

Entretanto, esta avaliação (nem nenhuma outra que tenhamos conhecimento) não provê um *corpus* no domínio que trabalhamos. Assim, para uma comparação alternativa, a Tabela 7.3 apresenta os resultados obtidos pelo *baseline* desenvolvido e apresentado no Plano de Estudo e Pesquisa (PEP) após uma execução completa sobre o *corpus* novo. Este *baseline* foi desenvolvido como um experimento preliminar para esta dissertação, e utilizou-se também da busca de padrões no *corpus*, porém sem a expansão de padrões (classes) e utilizando apenas uma heurística posicional simples a partir deste padrão. Pode-se observar que todas as medidas foram melhoradas.

Neste cenário, consideramos os nossos resultados atuais como muito positivos, especialmente considerando o foco e a natureza das entidades de interesse (leis e outras).

7.4 Avaliação do Reconhecimento de Relações entre Entidades Nomeadas

Esta avaliação foi executada pelos dois especialistas do domínio previamente mencionados. Os especialistas avaliaram duas das três relações reconhecidas pelo sistema: `same_as` e `applies_to_geo`.

Tabela 7.3 – Resultados obtidos na execução do *baseline* do protótipo de REN, desenvolvido pela autora a partir de uma heurística simples de REN.

	Precisão	Abrangência	Medida-F
Menções a entidades	59,09% (923/1562)	18,98% (923/4863)	28,73
Entidades únicas	40,38% (252/624)	21,16% (252/1191)	27,77

A avaliação da relação `references` é apresentada posteriormente.

Não é apresentada a abrangência da relação uma vez que não foi desenvolvido gabarito para esta tarefa, e sim apenas sua precisão na visão dos especialistas.

7.4.1 Relação `same_as`

O sistema identificou 185 instâncias desta relação no *corpus*, e os avaliadores classificaram as instâncias em Correta ou Incorreta. A Tabela 7.4 apresenta os resultados numéricos desta avaliação, feita pelos especialistas do domínio. A precisão da identificação desta relação segundo a sua avaliação varia entre 53 e 67%.

Tabela 7.4 – Resultados da avaliação dos especialistas da relação `same_as`.

	Avaliador 1	Avaliador 2
Corretas	52,97% (98/185)	67,03% (124/185)
Incorretas	47,03% (86/185)	32,97% (60/185)

A Tabela 7.5 apresenta o grau de concordância entre os avaliadores quanto a esta relação.

Tabela 7.5 – Matriz de confusão representando o grau de concordância dos avaliadores quanto à relação `same_as`.

		Avaliador 1	
		Corretas	Incorretas
Avaliador 2	Corretas	89	35
	Incorretas	9	51

Ainda, a relação `same_as` foi questionada por um dos avaliadores por agrupar, como se fossem uma mesma, entidades distintas unidas pela mesma sigla. É o caso de *"False Claims Act"* e *"Federal Courts Act"*, por exemplo, incorretamente associadas pela relação `same_as`.

7.4.2 Relação `applies_to_geo`

O sistema identificou 119 instâncias desta relação no *corpus*, e os avaliadores classificaram as instâncias em Correta, Incorreta, Vaga e ainda Não soube informar. As duas últimas opções foram inseridas pelos próprios avaliadores durante o processo de avaliação, já que nem todas as relações eram tão claras, assim como a abrangência das atuais legislações acerca do assunto não é. Muitas das legislações atualmente em vigor sobre o assunto abrangem regiões ou entidades geopolíticas menores do que um país, e em outros casos a abrangência é a nível nacional de fato. Para certificar-se da precisão de cada uma das relações, seria necessário conhecer legislações muito específicas de locais isolados, além daquelas de maior abrangência e mais conhecidas.

Os comentários dos avaliadores voltaram-se principalmente para a falta de especificidade de muitas relações (atribuídas a pares de entidades legais e entidades geopolíticas como (United States

of) America e United Kingdom) e do fato de entidades com mesmo nome não serem necessariamente as mesmas – diferentes países têm sua própria legislação, e muitas vezes o nome é similar ou até o mesmo. Praticamente todos os países de língua inglesa que tiveram a sua legislação estudada durante a montagem do *corpus* têm o seu *"Privacy Act"*, por exemplo.

Por outro lado, algumas recomendações e diretivas (como as da União Europeia) se aplicam a diversos países ao mesmo tempo, que devem aplicar estas a sua própria legislação. Embora não tenham sido reconhecidas instâncias desta relação para entidades deste tipo, este é claramente um aspecto que o sistema deve abordar em um aprimoramento futuro.

A Tabela 7.6 apresenta os resultados numéricos desta avaliação, feita pelos especialistas. A precisão da identificação desta relação segundo a avaliação dos dois especialistas de domínio varia entre 43 e 50%, incluindo apenas aquelas relações marcadas como Corretas. Dentre aquelas marcadas como Vagas e Não soube informar, estão 24 a 29% das ocorrências das relações.

Tabela 7.6 – Resultados da avaliação dos especialistas da relação `applies_to_geo`.

	Avaliador 1	Avaliador 2
Corretas	50,42% (60/119)	42,86% (51/119)
Incorretas	26,05% (31/119)	27,73% (33/119)
Vagas	-	24,37% (29/119)
Não soube informar	23,53% (28/119)	5,04% (6/119)

A Tabela 7.7 apresenta o grau de concordância entre os avaliadores quanto a esta relação. Foram desconsideradas na montagem desta tabela aquelas atribuições de relação que foram avaliadas como Não soube informar ou Vaga por um dos avaliadores (ou ambos). Do total de 119 instâncias desta relação, apenas 60 constam na matriz. Pode-se inferir daí a dificuldade da tarefa mesmo para avaliadores humanos, especialistas do domínio que se investiga.

Tabela 7.7 – Matriz de confusão representando o grau de concordância dos avaliadores quanto à relação `applies_to_geo`.

		Avaliador 1	
		Corretas	Incorretas
Avaliador 2	Corretas	37	8
	Incorretas	11	4

7.4.3 Relação `references`

A relação `references` teve uma análise prévia pela equipe do projeto a que pertence este trabalho e não foi incluída na avaliação pelos especialistas, por apresentar muito ruído na avaliação geral. Nesta análise, julgamos que seu desempenho foi bastante prejudicado pela forma com que foi projetada e desenvolvida. A motivação para não chegarmos a submetê-la aos especialistas foi para que estes pudessem focar naquelas relações que tinham resultados mais promissores.

O sistema apresentou 1062 relações deste tipo. A dificuldade desta da relação estava na identificação da entidade que nomeava o texto processado, e foi onde o reconhecimento desta relação falhou. Das 1062 relações apresentadas, apenas 44 apresentavam o lado esquerdo da relação correto, isto é, identificaram corretamente o documento sendo processado. Como já comentado, não há como calcular a abrangência do reconhecimento de relações. Entretanto, sua precisão é bastante baixa, ficando em 4%.

7.5 Avaliação da Anotação Manual do *Corpus*

A avaliação da tarefa de REN deu-se com base na anotação manual realizada pela autora. Uma amostragem desta anotação foi submetida à avaliação, uma vez que a autora não é especialista no domínio investigado. O procedimento foi executado por um dos especialistas do domínio participantes desta avaliação.

Foi selecionada aleatoriamente uma relação de 100 entidades identificadas e classificadas para fins de gabarito, que foram avaliadas como corretas ou não pelo especialista. O resultado é apontado na Tabela 7.8.

Tabela 7.8 – Resultados sobre a avaliação da anotação do *corpus*. A avaliação foi feita pelo especialista de domínio sobre uma amostragem das entidades nomeadas previamente identificadas e classificadas que vieram a compor o gabarito.

Resultado	Quantidade
Corretas	97
Incorretas	3

Embora tenha sido executada apenas sobre uma pequena amostra do conjunto total de entidades que vieram a compor o gabarito, o resultado da avaliação do especialista do domínio dá confiabilidade à anotação, por apresentar uma concordância bastante elevada com a autora da anotação.

7.6 Análise dos Resultados e Considerações

A nossa pesquisa nesta área é motivada como um mecanismo de suporte à verificação automática de conformidade de ações em projetos com legislação e melhores práticas na área. A difícil tarefa de identificar possíveis falhas nos processos pode ser automatizada através de técnicas de Processamento de Linguagem Natural (PLN) e recursos adequados de Representação do Conhecimento, como ontologias. Aplicações que utilizem-se de tal suporte poderiam ser utilizadas tanto na indústria como em órgãos do governo.

Adicionalmente, como mecanismo de auxílio a especialistas de domínio na população de ontologias para posterior refinamento manual, nosso método e sistema mostra-se eficaz e rápido, sendo ainda passível de fácil modificação e adaptação.

Na tarefa de REN, os resultados demonstrados pelo experimento são bastante positivos, e apesar da especificidade da tarefa, apresenta resultados comparáveis aos apresentados em uma avaliação

conjunta respeitada. A utilização de parte de uma ontologia como taxonomia para classificação das entidades e a expansão destas classes mostrou-se uma tática bem-sucedida para a execução da tarefa, com resultados bastante promissores.

Exemplos de entidades reconhecidas corretamente são: *"Civil Rights Act"*, *"PIPEDA"* (*"Personal Information Protection and Electronic Documents Act"*) e *"TSR"* (*"Telemarketing Sales Rule"*). Exemplos de entidades falsamente reconhecidas, ou seja, falsos positivos, são: *"Royal Canadian Mounted Police"*, *"P-21 An Act"* e *"US Social Security Number"*.

A identificação das entidades que foram detectadas como possuindo um artigo com seu nome na Wikipedia em língua inglesa alcançaram uma precisão muito boa, embora a baixa abrangência seja um limitador importante quanto ao uso de tal técnica para aumento de precisão. Quanto à precisão, que poderia ser esperado que fosse ainda mais alta, convém lembrar que o fato de ser encontrado na enciclopédia um artigo com o nome da entidade não assegura que a entidade seja de fato da classe pretendida, como no caso de *"Royal Canadian Mounted Police"* e *"US Social Security Number"*, por exemplo.

A tarefa de Reconhecimento de Relações entre Entidades Nomeadas, por outro lado, demonstrou resultados mais baixos. As heurísticas trabalhadas até então demonstraram que devem ser refinadas o suficiente até poderem passar por uma nova avaliação.

A relação `same_as` tem potencial a ser investigado no uso de informação específica da área legal: entidades com nomes totalmente diferentes (e não apenas acrônimos) podem ser referentes a uma mesma entidade no mundo. Um exemplo seria *"Title XIII of the Code of Federal Regulations (CFR)"*, que é comumente designado por *"Children's Online Privacy Protection Act of 1998"*.

Por outro lado, o uso do acrônimo como marcador da relação deve ser limitado, uma vez que entidades diversas no universo podem remeter a um mesmo acrônimo, o que causa erros graves nos resultados finais. Um exemplo é o já mencionado par (*"False Claims Act"*, *"Federal Courts Act"*), incorretamente associado através da relação `same_as`.

Quanto à relação `applies_to_geo`, esta também pode fazer uso de recursos geográficos mais completos, mas também o seu uso pode ser bastante aprimorado com alguns refinamentos; um exemplo claro é o caso de *"British Columbia Act"*, que foi associada ao Reino Unido por esta relação, enquanto British Columbia é uma província do Canadá, e não parte do Reino Unido. Esta associação deveu-se ao gentílico de Great Britain, que ocorre também no nome desta província. Além disso, esta relação pode ser grandemente aprimorada com o aumento de sua especificidade.

Um dos avaliadores comentou em sua avaliação sobre esta relação que frequentemente a legislação é diferente em diferentes partes do Reino Unido, e o mesmo se dá em diversos outros casos. Para alcançar a especificidade desejada, deve-se fazer um trabalho mais aprofundado de entendimento do texto processado, e possivelmente consultas a bases de dados jurídicas mais completas. Como estes são recursos escassos, o que é inclusive parte da motivação deste trabalho, ainda existe um extenso caminho a ser percorrido neste sentido.

Por fim, a relação `references` mostrou os resultados mais insatisfatórios das três relações, devido à confiança não-correspondida da sua heurística na estrutura dos textos do *corpus*. Esta

relação deve ser reformulada para fazer uso de bases de dados de leis, e possivelmente confiar em um conjunto de atributos para definir qual é o texto de que se fala. Quanto à atribuição da relação entre todas as ocorrências do par (texto processado, entidade referida neste texto), esta ainda parece estar de acordo com a realidade observada nos textos e nas instâncias resultantes desta relação.

A população da ontologia dá-se com base em um combinado destas tarefas todas realizadas em conjunto. Desta forma, o resultado da população da ontologia mostrou-se bastante adequado para facilitar o trabalho de especialistas de domínio, que hoje vêem-se com muito e custoso trabalho manual na construção, população e associação de relações em ontologias de domínio. Mesmo que os resultados no que tange as relações não seja tão animador quanto no que se refere às entidades, eles ainda podem ser bastante melhorados com a adição de heurísticas e atributos para a associação das entidades através das relações, além das abordagens adicionais sugeridas nesta seção, como o uso de recursos e o aprimoramento das heurísticas já existentes.

8. Conclusão

"O todo é mais que a mera soma das partes."
– Aristóteles.

"Todos os problemas são insolúveis. A essência de haver um problema é não haver uma solução."
– Fernando Pessoa, em "O Livro do Desassossego".

Na conclusão do presente trabalho, apresentamos nossas considerações finais e percepções acerca do trabalho apresentado nesta dissertação e de seus resultados. Além disso, também relacionamos as contribuições científicas deste trabalho e propostas de trabalhos futuros.

8.1 Considerações Finais

Este trabalho apresentou um conjunto de mecanismos visando contribuir com o avanço da pesquisa nas áreas de Extração de Informação (EI), Processamento de Linguagem Natural (PLN) e Engenharia de Ontologias, adicionalmente às áreas que investigam a utilização de ontologias em aplicações práticas no domínio de privacidade de dados e responsabilização. A principal construção do trabalho é um método para população de ontologias a partir das seguintes tarefas das áreas mencionadas: Reconhecimento de Entidades Nomeadas (REN) e Reconhecimento de Relações entre Entidades Nomeadas (EN).

Este método parte da percepção que estas tarefas podem sugerir instâncias e relações para uma ontologia do domínio investigado, enriquecendo a ontologia de forma a que esta possa ser revisada por especialistas e aplicada neste contexto.

Os resultados das avaliações quantitativas e qualitativas apresentados no Capítulo anterior são bastante satisfatórios e demonstram a viabilidade do método proposto. Entretanto, eles ainda podem ser significativamente melhorados explorando possibilidades como as propostas na Seção de Trabalhos Futuros neste Capítulo.

A avaliação qualitativa realizada em conjunto com os especialistas do domínio investigado mostrou-se extremamente proveitosa, principalmente devido aos comentários feitos por estes e que podem vir a contribuir com melhorias significativas para trabalhos futuros baseados neste.

Adicionalmente, esta avaliação enriquece o trabalho incluindo uma visão que é subjetiva, mas é justamente a visão do público-alvo a que se destinam os recursos e mecanismos desenvolvidos por pesquisas na área que investigamos.

Os resultados do trabalho como um todo são muito completos, tanto pela integração das diferentes tarefas empregadas na resolução do problema apresentado quanto pela qualidade dos recursos finais produzidos.

Um grande escopo das áreas de PLN e Engenharia de Ontologias foi investigada durante o desenvolvimento deste trabalho: REN, Reconhecimento de Relações entre EN e Aprendizado de

Ontologias. Consideramos que as contribuições resultantes deste são um acréscimo significativo tanto a cada uma delas individualmente e em particular ao todo. Entendemos que da integração entre estas áreas é possível fazer a importante consideração de que este tipo de aplicação de diferentes técnicas é possível, útil e traz resultados positivos como os aqui demonstrados.

8.2 Contribuições

Nesta seção, relacionamos algumas das contribuições deste trabalho nos contextos acadêmico e industrial para o conhecimento produzido nas áreas de PLN, EI e Engenharia de Ontologias. São elas:

- Contribuições principais
  - Método de população de ontologias de domínio a partir da execução das tarefas de REN e Reconhecimento de Relações entre EN em um *corpus* do domínio em questão;
  - Sistema para validação do método mencionado no item anterior, com suporte a todas as etapas previstas no método e a efetiva população de uma ontologia *Web Ontology Language* (OWL) com instâncias e propriedades, obtidas a partir das EN e relações;
  - Construção de heurísticas para REN no domínio legal;
  - Avaliação dos resultados obtidos na experimentação do sistema em duas formas: quantitativa e qualitativa, esta última com o apoio de especialistas do domínio de privacidade e responsabilização.
- Recursos
  - Ontologia *Legal*, que em sua versão anterior à execução do sistema acima provê uma taxonomia para categorização de EN do domínio investigado, e que posteriormente à população provê ainda um recurso adicional para o refinamento e utilização por pesquisadores e profissionais que trabalhem com privacidade de dados na academia e indústria de *software*;
  - *Corpus Privacy* e anotação deste com EN, desenvolvidos visando este trabalho em particular, embora não existam quaisquer restrições para a sua utilização em outros projetos e pesquisas;
  - Sistema para avaliação automática da tarefa de REN, que utiliza um gabarito para a comparação das EN reconhecidas e classificadas pelo sistema para o cálculo das medidas de Abrangência, Precisão e Medida-F.
- Artigo
  - *"Named entity recognition in the legal domain for ontology population"*, aceito no *workshop "Semantic Processing of Legal Texts (SPLeT)"* do LREC'2010, com resultados preliminares do trabalho até o primeiro semestre de 2010 [BNdS+10].

8.3 Trabalhos Futuros

No decorrer deste trabalho, algumas ideias de trabalhos futuros baseados e em continuação a este foram elaboradas. Algumas destas ideias são detalhadas nesta seção. São elas:

- Desenvolvimento de heurísticas adicionais para a tarefa de REN;
  Foi possível observar no decorrer deste trabalho que com um conjunto reduzido de heurísticas simples, pode-se atingir abrangência e precisão bastante satisfatórias para a tarefa. Julgamos que o aprimoramento das atuais heurísticas, para uma melhoria da precisão, somado à adição de novas, utilizando diferentes atributos além dos atualmente usados (posicionais, semânticos e sintáticos) podem trazer um grande ganho nos resultados.

- Desenvolvimento de heurísticas adicionais para a tarefa de Reconhecimento de Relações entre EN;
  As heurísticas atualmente utilizadas para a tarefa demonstram resultados competitivos com trabalhos relacionados, mas ainda muito aquém do que pode ser atingido com um refinamento das heurísticas atuais em todas as relações apresentadas neste trabalho. No que se refere à relação `references`, em particular, uma adaptação importante deve ser experimentada, no sentido de identificar o texto processado mais efetivamente. Uma sugestão é a pesquisa na *Web* por uma porção do texto e o Reconhecimento de Entidades Nomeadas nos resultados, procurando por entidades que se refiram ao domínio legal.

- Experimentação de diferentes recursos semânticos para o aprimoramento das relações propostas neste trabalho;
  Entendemos que nas duas tarefas que constituem o cerne deste trabalho, ou seja, Reconhecimento de Entidades Nomeadas e Reconhecimento de Relações, a utilização de mais recursos com informação semântica seria um grande adendo à acurácia do sistema. Alguns destes recursos já estão disponíveis e não foram utilizados (ou não foram utilizados em toda a sua potencialidade) por questão de foco, porém em outros casos os recursos ainda não estão disponíveis para pesquisadores, o que constitui parte da motivação para esta pesquisa. Dentre os recursos já disponíveis, podemos mencionar o caso das ontologias e recursos similares com informação geográfica e geopolítica (Geonames e OpenStreetMap, por exemplo), e os diversos recursos enciclopédicos disponibilizados pela Wikipedia e pelo projeto DBPedia, além da WordNet e a FrameNet. Na outra categoria de recursos, não foram encontrados recursos abrangentes referentes aos domínios legal e de privacidade de dados.

- Proposição de relações adicionais na tarefa de Reconhecimento de Relações;
  Muitas relações foram pensadas para o desenvolvimento deste trabalho, mas novamente por questão de foco, foram descartadas. Dentre estas relações, muitas são úteis e seu uso seria interessante em diversas aplicações. Um trabalho futuro poderia explorar suas possibilidades de desenvolvimento e posterior uso. São elas: `refers_to`, aplicável a pares de

instâncias na forma (Regulation, Subject); documents, aplicável a pares de instâncias de (Resource, Regulation); implements, aplicável a pares de (Normative_Regulation, Non-Normative_Regulation); e includes, aplicável a pares de instâncias na forma (Thing, Thing). Esta última, embora por sua definição possa ser aplicada a instâncias de Geo, poderia incluir principalmente o reconhecimento desta relação para pares de Resource e Regulation, aproveitando trabalhos já extensos e com resultados consolidados na área de ontologias geográficas para pares do tipo Geo.

- Extensão do método para ampliar a sua abrangência, sugerindo um maior leque de novas classes a partir do *corpus* processado além daquelas originadas pela expansão das classes selecionadas;
  Neste trabalho, focamos na população da ontologia com instâncias e relações, apenas incluindo novas classes na ontologia como uma tarefa acessória e decorrente das tarefas anteriormente mencionadas. Entretanto, há espaço para este aprimoramento, que complementaria o trabalho de maneira bastante completa, interessante e aplicável em sistemas de Engenharia e Aprendizado de Ontologias, dos quais há uma grande necessidade atualmente.

- Experimentação do método proposto para um domínio diferente;
  O método proposto pode ser aplicado a outros domínios. A aplicação deste trabalho considerou o domínio de privacidade de dados e responsabilização na indústria de software, o que particulariza o conjunto de heurísticas desenvolvidas para as tarefas de REN e Reconhecimento de Relações entre EN. Não houve oportunidade ou tempo hábil para explorar outros domínios e fazer um estudo sobre adaptação das heurísticas e verificação do comportamento do sistema para outros casos. Entretanto, entendemos que a experimentação deste em um novo domínio, não relacionado ao atualmente investigado, é um trabalho extremamente importante e que merece atenção em uma oportunidade futura.

Bibliografia

[ALM05] Florence Amardeilh, Philippe Laublet, e Jean-Luc Minel. Document annotation and ontology population from linguistic extractions. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pp. 161–168, New York, NY, USA, 2005. ACM.

[BD05] Dario Bianchi e Rodolfo Delmonte. Learning domain ontologies from text analysis: an application for question answering. In *Proceedings of the 2nd Meaning Workshop (Meaning-2005)*, Trento, Italy, 2005.

[Bic06] Eckhard Bick. Functional aspects in portuguese ner. In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira, e Maria Carmelita Dias, editores, *PROPOR*, volume 3960 of *Lecture Notes in Computer Science*, pages 80–89. Springer, 2006.

[BKL09] Steven Bird, Ewan Klein, e Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

[BL04] Steven Bird e Edward Loper. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[BNdS+10] Mírian Bruckschen, Caio Northfleet, Douglas Michaelsen da Silva, Paulo Bridi, Roger Granada, Renata Vieira, Prasad Rao, e Tomas Sander. Named entity recognition in the legal domain for ontology population. In *3rd Workshop on Semantic Processing of Legal Texts*, Valletta, Malta, 2010.

[BRN+09] Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, e James R. Curran. Named entity recognition in wikipedia. In *People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP*, pp. 10–18, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Bru02] Ada Brunstein. Annotation guidelines for answer types, 2002.

[BVdS08] Mírian Bruckschen, Renata Vieira, e José Guilherme Camargo de Souza. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, chapter Sistema SeRELeP para o reconhecimento de relações. Linguateca, 2008.

[BVL03] Sean Bechhofer, Raphael Volz, e Phillip Lord. Cooking the semantic web with the owl api. In Dieter Fensel, Katia Sycara, e John Mylopoulos, editores, *The Semantic Web –*

ISWC 2003: The Second International Semantic Web Conference, volume 2870/2003, pp. 659–675, Sanibel Island, Florida, USA, 2003. Springer.

[Car08] Nuno Cardoso. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, chapter REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Linguateca, 2008.

[CBFR99] Nancy Chinchor, Erica Brown, Lisa Ferro, e Patty Robinson. 1999 named entity recognition task definition. Technical report, MITRE, Corp., Agosto 1999. Technical Report Version 1.4.

[CBHM09] Andrew Carlson, Justin Betteridge, Estevam R. Hruschka, e Tom M. Mitchell. Coupling semi-supervised learning of categories e relations. In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pp. 1–9, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[CH09] Peter Clark e Phil Harrison. Large-scale extraction and use of knowledge from text. In *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*, pp. 153–160, New York, NY, USA, 2009. ACM.

[Cha08] Marcírio Chaves. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, chapter Geo-ontologias para reconhecimento de relações entre locais: a participação do SEI-Geo no Segundo HAREM. Linguateca, 2008.

[Cim06] Philipp Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[DAB09] Danica Damljanovic, Florence Amardeilh, e Kalina Bontcheva. Ca manager framework: creating customised workflows for ontology population and semantic annotation. In *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*, pp. 177–178, New York, NY, USA, 2009. ACM.

[DG10] Lucas Drumond e Rosario Girardi. Extracting ontology concept hierarchies from text using markov logic. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1354–1358, New York, NY, USA, 2010. ACM.

[DMP+04] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, pp. 837–840, 2004.

[ESFP10] Asif Ekbal, Eva Sourjikova, Anette Frank, e Simone Paolo Ponzetto. Assessing the challenge of fine-grained named entity recognition and classification. In *Proceedings of the 2010 Named Entities Workshop*, pages 93–101, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[GG95] N. Guarino e P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pp. 25–32, 1995.

[GMF+03] John H. Gennari, Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, e Samson W. Tu. The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.

[GPFLCG03] Asunción Gómez-Pérez, Mariano Fernández-López, e Oscar Corcho-Garcia. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[Gru95] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[GS96] Ralph Grishman e Beth Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pp. 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[HB09] Matthew Horridge e Sean Bechhofer. The OWL API: A Java API for working with OWL 2 ontologies. In *OWLED*, 2009.

[HBBB07] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, e Alexander Boer. The LKIF Core ontology of basic legal concepts. In Pompeu Casanovas, Maria Angela Biasiotti, Enrico Francesconi, and Maria Teresa Sagri, editores, *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007)*, June 2007.

[HKK+10] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[Hor08] Ian Horrocks. Ontologies and the Semantic Web. *Commun. ACM*, 51(12):58–67, 2008.

[HSG04] Takaaki Hasegawa, Satoshi Sekine, e Ralph Grishman. Discovering relations among named entities from large corpora. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 415, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[KFM+07] Z. Kozareva, O. Ferrández, A. Montoyo, R. Mu A. Suárez, and J. Gómez. Combining data-driven systems for improving named entity recognition. *Data Knowl. Eng.*, 61(3):449–466, 2007.

[KFNM04] Holger Knublauch, Ray W. Fergerson, Natalya F. Noy, e Mark A. Musen. The protégé owl plugin: An open development environment for semantic web applications. In *The Semantic Web–ISWC 2004*, pp. 229–243. Springer, 2004.

[KL10] A Kumaran e Haizhou Li, editores. *Proceedings of the 2010 Named Entities Workshop*. Association for Computational Linguistics, Uppsala, Sweden, July 2010.

[Kri81] Saul A. Kripke. *Naming and necessity*. Wiley Blackwell, 1981.

[LB02] Edward Loper e Steven Bird. NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pp. 63–70, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Lin08a] Linguistic Data Consortium. ACE english annotation guidelines for entities. Technical report, Linguistic Data Consortium, 2008.

[Lin08b] Linguistic Data Consortium. ACE english annotation guidelines for relations. Technical report, Linguistic Data Consortium, 2008.

[MFWJ08] Daniel W. McMichael, R. Fu, Simon Williams, e Geoff A. Jarrad. Sem@ntica: A system for semantic extraction and logical querying of text corpora. In *ISI*, pp. 277–278. IEEE, 2008.

[Mil95] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[MS01a] Alexander Maedche e Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[MS01b] Alexander Maedche e Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[MS08] Cristina Mota e Diana Santos, editores. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. 2008.

[NFF+91] Robert Neches, Richard Fikes, Tim Finin, Tom Gruber, Ramesh Patil, Ted Senator, e William R. Swartout. Enabling technology for knowledge sharing. *AI Mag.*, 12(3):36–56, 1991.

[NIS08a] NIST. Automatic content extraction 2008 evaluation plan (ace08). Technical report, NIST, April 2008.

[NIS08b] Linguistic Data Consortium NIST/LDC. Automatic content extraction evaluation (ace08) official results, 2008.

[NK10] Matteo Negri e Milen Kouylekov. Fbk_nk: A wordnet-based system for multi-way classification of semantic relations. In *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 202–205, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[NMC09a] Joel Nothman, Tara Murphy, e James R. Curran. Analysing wikipedia and gold-standard corpora for ner training. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 612–620, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[NMC09b] Joel Nothman, Tara Murphy, e James R. Curran. Analysing wikipedia and gold-standard corpora for ner training. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 612–620, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[NS07] David Nadeau e Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007.

[PD10] Hoifung Poon e Pedro Domingos. Unsupervised ontology induction from text. In *ACL '10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 296–305, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[RH10] Bryan Rink e Sanda Harabagiu. Utd: Classifying semantic relations by combining lexical and semantic resources. In *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 256–259, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[RR09] Lev Ratinov e Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 147–155, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[SC07a] Diana Santos e Nuno Cardoso. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, November 2007. ISBN: 978-989-20-0731-1.

[SC07b] Diana Santos e Nuno Cardoso. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, November 2007.

[Sek08] Satoshi Sekine. Extended named entity ontology with attribute information. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, e Daniel Tapias, editores, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Marrocos, Maio 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[SKW07] Fabian M. Suchanek, Gjergji Kasneci, e Gerhard Weikum. Yago : a large ontology from wikipedia and wordnet. Research Report MPI-I-2007-5-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, December 2007.

[SMI09] Ashish Sureka, Pranav Prabhakar Mirajkar, e Kishore Varma Indukuri. A rapid application development framework for rule-based named-entity extraction. In *Bangalore Compute Conf.*, page 25, 2009.

[SSCV06] Diana Santos, Nuno Seco, Nuno Cardoso, e Rui Vilela. Harem: An advanced ner evaluation contest for portuguese. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik, e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pp. 1986–1991, Genoa, Italy, 22–28 May 2006. ELRA.

[SSN02] S. Sekine, K. Sudo, e C. Nobata. Extended named entity hierarchy. In M. Gonzáles Rodríguez e C. Paz Suárez Araujo, editores, *Proceedings of 3^{rd} International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1818–1824, Canary Islands, Spain, May 2002.

[STB04] Maria Teresa Sagri, Daniela Tiscornia, e Francesca Bertagna. Jur-wordnet. Technical report, Institute for Theory and Techniques for Legal Information (ITTIG), 2004.

[TH10] Stephen Tratz e Eduard Hovy. Isi: Automatic classification of relations between nominals using a maximum entropy classifier. In *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 222–225, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[WSC04] Tuangthong Wattarujeekrit, Parantu Shah, e Nigel Collier. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155, 2004.

[XL09] Clarissa Castellã Xavier e Vera Lúcia Strube de Lima. Construção de uma estrutura ontológica de domínio a partir da wikipédia. In *The 7th Brazilian Symposium in Information and Human Language Technology*, São Carlos, SP, Brasil, 2009.

Apêndice A. Relação de Textos do *Corpus Privacy*

- Argentina
  - Personal Data Protection Act
  - Security Measures
- Asia-Pacific Economic Cooperation (APEC)
  - APEC Privacy Framework
- Áustria
  - Federal Act concerning the Protection of Personal Data
- Austrália
  - Do Not Call Register Act 2006
  - Information Privacy Principles
  - National Privacy Principles
  - Spam Act 2003
  - National Health Act 1953 - Section 13 5AA
  - National Health Act 1953 - Section 13 5AB
  - National Health Act 1953 - Section 13 5AC
- Bahamas
  - Managing a Data Security Breach
  - Privacy Act
- Canadá
  - Key Steps for Organizations in Responding to Privacy Breaches
  - Personal Information Protection and Electronic Documents Act (PIPEDA)
  - Privacy Act 1985
- China
  - China's Anti-Spam Law
  - Computer-Processed Personal Data Protection Law

  - Measures for the Administration of Internet E-mail Services
  - Procedures of Shanghai Municipality on the Administration of the Collecting of Information Relating to the Credit Standing of Individual Persons (for Trial Implementation)
  - Provisions of Shanghai Municipality on Open Government Information
  - Rules of Shenzhen Municipality on the Publication of Government Information
- Alemanha
  - Telecommunications Act
- Espanha
  - Royal Decree 1720-2007
- União Europeia
  - Directive 2002-58-EC
  - Directive 2006-24-EC
  - Directive 95-46-EC
  - Directive 97-66-EC
  - Regulation (EC) No 45-2001
- França
  - Act N. 78-17 of 6 january 1978 on data processing, data files and individual liberties
- Grécia
  - Law 2472-1997-on the Protection of Individuals with regard to the Processing of Personal Data
  - Law 3471-Protection of personal data and privacy in Protection of personal data and privacy
- Hong Kong
  - Personal Data (Privacy) Ordinance
- Índia
  - The Information Technology Act
- Japão
  - Act on the Protection of Personal Information (Japan)

- Coréia do Sul
  - Act on Promotion of Information and Communications Network Utilization and Data Protection, etc. (Republic of Korea)
  - RFID Privacy Protection Guideline (Republic of Korea)
  - South Korean Act
- Nova Zelândia
  - Credit Reporting Privacy Code 2004
  - Health Information Privacy Code 1994
  - Privacy Act 1993
  - Privacy Amendment Act 2009
  - Privacy At Work Book
  - Telecommunications Information Privacy Code 2003
  - Unsolicited Electronic Messages Act 2007
- Organisation for Economic Co-operation and Development (OECD)
  - Guidelines on the Protection of Privacy and Transborder Flows of Personal Data
- IBM
  - IBM Privacy Pratices on the Web
- Microsoft
  - Microsoft Privacy Guidelines
- Portugal
  - Act on the Protection of Personal Data
- Rússia
  - Federal Law of 27 July 2006 N 152-FZ on Personal Data
  - Federal Law on Information, Informatization, and the Protection of Information
- Reino Unido
  - Data Protection Act
  - Freedom Of Information Act 2000
  - International transfers legal guidance

- Estados Unidos
  - 15 USC, Subchapter I, Sec. 6801-6809-Gramm-Leach-Bliley Act
  - Authentication in an Internet Banking Environment Description
  - California Civil Code Section 1798.29
  - California SB 1436
  - Coppa
  - ElectroicFundTransfer
  - FACTA-2003
  - FairHousing
  - FCRA
  - FOIA
  - HIPAA
  - Interagency Guidance on Authentication in an Internet Banking Environment
  - pci audit procedures v1-1
  - Telecommunication's Rule 1996
  - Telemarketing sales rule - 15 U.S.C Sec.6101 et seq
  - Telemarketing sales rule - 16 CFR Part 31
  - The Computer Fraud and Abuse Act
  - The Federal Trade Commission's Privacy Initiatives Unfairness & Deception Enforcement
  - TITLE 12–BANKS AND BANKING-225.28-List of permissible nonbanking activities
  - Title 12-PART 332-PRIVACY OF CONSUMER FINANCIAL INFORMATION
  - Title 12-RFPA
  - Title 16-PART 1014-POLICIES AND PROCEDURES IMPLEMENTING THE PRIVACY ACT OF 1974
  - Title 16-PART 308-TRADE REGULATION RULE PURSUANT TO THE TELEPHONE DISCLOSURE AND DISPUTE RESOLUTION ACT OF 1992
  - Title 16-PART 310-TELEMARKETING SALES RULE
  - Title 16-PART 312-CHILDREN'S ONLINE PRIVACY PROTECTION RULE
  - Title 16-PART 313-PRIVACY OF CONSUMER FINANCIAL INFORMATION
  - Title 16-PART 314-STANDARDS FOR SAFEGUARDING CUSTOMER INFORMATION
  - Title 16-PART 316-CAN-SPAM RULE

- Title 16-PART 318-HEALTH BREACH NOTIFICAT
- Title 16-Part 682-Disposal
- Title 18-18 USC CHAPTER 121-18C121
- Title 18-18 USC CHAPTER 206-18C206
- Title 18-18 USC-CHAPTER 119-18C119
- Title 18-Section 2721-Drivers Privacy Protection Act
- Title 18-Video Privacy Protection Act
- Title 20-CHAPTER 31–GENERAL PROVISIONS CONCERNING EDUCATION
- Title 42-Chapter 21a-THE PUBLIC HEALTH AND WELFARE-CHAPTER 21A - PRIVACY PROTECTION
- Title 42-PART 2-CONFIDENTIALITY OF ALCOHOL AND DRUG ABUSE PATIENT RECORDS
- Title 42-PART 2a-PROTECTION OF IDENTITY-RESEARCH SUBJECTS-
- Title 42-PART 3-PATIENT SAFETY ORGANIZATIONS AND PATIENT SAFETY WORK PRODUCT
- Title 42-PART 93-PUBLIC HEALTH SERVICE POLICIES ON RESEARCH MISCONDUCT
- Title 45-PART 164-SECURITY AND PRIVACY
- TITLE 47–Chap5-SUBCHAPTER II-TELEGRAPHS, TELEPHONES, AND RADIOTELEGRAPHS
- TITLE 47–Chap5-SUBCHAPTER Va-TELEGRAPHS, TELEPHONES, AND RADIOTELEGRAPHS
- Title 47-PART 22-PUBLIC MOBILE SERVICES
- Title 47-PART 42-PRESERVATION OF RECORDSOF COMMUNICATION COMMON CARRIERS

Apêndice B. Planilha de Avaliação da Tarefa de Reconhecimento de Relações entre Entidades Nomeadas

The purpose in this evaluation is to assess the NE recognizer system.

First there is a list of instantiations for normative documents. Please indicate whether the named entities are correctly identified.

After that there are two relations sets. Please indicate whether the relations identified are correct

The relations are: **same_as**, and **applies_to_geo**.
- same_as means that two (or more) entities refer to the same object in the real worlc
- applies_to_geo means that one legal entity is applied to a geopolitical entity, such as country, province or cit

An example is given for same_as relation below.

Class	Named entity	Relation	Class	Named entity	Correct?
Act	Employee Retirement Income Security Act of 1974	same_as	Act	ERISA	

First column (Class) represents the class of the entity, named in the second column (Named entity)
Third column (Relation) brings the name of the relation being evaluated.
Fourth and fifth columns represent class and named entity recognized as the same.

Sixth column (Correct?) is for the evaluators to say **Y** if the relation was correctly attributed or **N** if not.

The example above means that both "Employee Retirement Income Security Act of 1974" and "ERISA" were found as legal named entities in a set of texts, and the system identified them as refering to the same object in the world. The evaluator should say if this relation is considered correct or not.

Thanks for your time and availability in answering this evaluation,
your cooperation is much appreciated.
-- Mírian Bruckschen, PUCRS, Brazil.

Original classes (only subclasses of Normative_Regulation):
- Act, Law, Rule

Derived classes:
- original: Law
- generated: Police Jurisprudence Constabulary

- original: Rule
- generated: Ruler, Normal, Pattern, Prescript, Regulation, Principle, Convention, Formula, Dominion

- original: Act
- generated: Enactment, Number, Turn, Routine, Deed, Bit

Non-instantiated derived classes:
- Bit, Decree, Deed, Dominion, Formula, Jurisprudence, Normal, Pattern, Prescript, Routine, Ruler, Turn

Instances
[Class=Act]Access to Personal Files Act 1987
[Class=Act]Act 25,236
[Class=Act]ACT No. 78-17 OF 6 JANUARY 1978
[Class=Act]Act No. 78-758 of 17 July 1978
[Class=Act]Act of August 9, 1989
[Class=Act]Act of November 12, 1999
[Class=Act]Act on the Promotion, etc. of Utilization of Information System
[Class=Act]Acts and Regulations Publication Act 1989
[Class=Act]Agricultural Marketing Act 1958
[Class=Act]Anti-Car Theft Act of1992
[Class=Act]Australian Communications and Media Authority Act 2005
[Class=Act]Bankruptcy Act
[Class=Act]Building Societies Act 1965
[Class=Act]Canadian Security Intelligence Service Act

[Class=Act]Children, Young Persons and Their Families Act 1989
[Class=Act]Civil Rights Act of 1968
[Class=Act]Commonwealth Electoral Act 1918
[Class=Act]Competition Act
[Class=Act]Controlled Substances Act
[Class=Act]Cree-Naskapi (of Quebec) Act
[Class=Act]Criminal Records (Clean Slate) Act 2004
[Class=Act]Data Protection Act 1998
[Class=Act]Designs Act 1953
[Class=Act]DSG 2000
[Class=Act]Electoral Act 1993
[Class=Act]Electronic Signature Act
[Class=Act]Enforcement Proceedings Act
[Class=Act]FACT Act
[Class=Act]Fair Housing Amendments Act of 1988
[Class=Act]Federal Business Records Act
[Class=Act]Federal Educational Rights and Privacy Act
[Class=Act]Federal Trade Commission Act (''FTC Act'')
[Class=Act]Financial Administration Act
[Class=Act]Foreign Service Act
[Class=Act]Gambling Act 2003
[Class=Act]Gramm-Leach-Bliley Act
[Class=Act]Health Insurance Act 1973
[Class=Act]Home Owners' Act
[Class=Act]Housing Restructuring and Tenancy Matters (Information Matching) Amendment Act 2006
[Class=Act]Immigration and Refugee Protection Act
[Class=Act]Indian Trusts Act, 1882
[Class=Act]Insolvency Act 1967
[Class=Act]International Business Companies Act, 2000
[Class=Act]Judicial Pensions and Retirement Act 1993
[Class=Act]Law Enforcement Technology Advertisement Clarification Act of 1997
[Class=Act]Local Government (Rating) Act 2002
[Class=Act]Local Government Official Information and Meetings Act 1987

[Class=Act]Medicines Act
[Class=Act]Misuse of Drugs Amendment Act 1978
[Class=Act]National Housing Act
[Class=Act]New Zealand Security Intelligence Service Act 1969
[Class=Act]Northern Ireland Assembly Disqualification Act 1975
[Class=Act]Optometrists and Dispensing Opticians Act 1976
[Class=Act]Patient Safety Act
[Class=Act]Pharmacy Act 1970
[Class=Act]Prison Act (Northern Ireland) 1953
[Class=Act]Privacy Act 1993Medicines Act 1981
[Class=Act]Privacy Amendment Act 2005
[Class=Act]Psychologists Act 1981
[Class=Act]Public Records Act (Northern Ireland) 1923
[Class=Act]Radio New Zealand Act
[Class=Act]Registered Architects Act 2005
[Class=Act]resolution A/RES/51/162
[Class=Act]Sarbanes-Oxley Act of 2002
[Class=Act]Self-Governing Schools etc. (Scotland) Act 1989
[Class=Act]Social Security Administration (Northern Ireland) Act 1992
[Class=Act]Social Workers Registration Act 2003
[Class=Act]Summary Proceedings Act 1957
[Class=Act]Taxation (Business Taxation and Remedial Matters) Act 2007
[Class=Act]Telecommunications Act of 1996
[Class=Act]Television Signals Transmission Act of 14 November 1997
[Class=Act]Trade Marks Act 2002
[Class=Act]Truth in Lending Act
[Class=Act]US PATRIOT Act
[Class=Act]Video Privacy Protection Act
[Class=Act]Westbank First Nation Self-Government Act
[Class=Code]Administrative Penal Code 1991
[Class=Code]Code of Criminal Procedure
[Class=Code]Constitution
[Class=Code]Heritage Code

Class	Named entity	Relation	Class	Named entity	Correct?
Act	Employee Retirement Income Security Act of 1974	same_as	Act	ERISA	
Act	TCPA	same_as	Act	Town and Country Planning Act 1990	
Act	TCPA	same_as	Act	Telephone Consumer Protection Act of 1991	
Act	Medical Practitioners Act	same_as	Act	MPA	
Act	ECOA	same_as	Act	Equal Credit Opportunity Act	
Act	Official Information Act	same_as	Act	OIA	
Act	Official Information Act	same_as	Act	Official Information Act 1982	
Act	Courts Constitution Act	same_as	Act	CCA	
Act	Courts Constitution Act	same_as	Act	Crime Control Act of 1990	
Act	Courts Constitution Act	same_as	Act	Court Costs Act	
Act	Courts Constitution Act	same_as	Act	Consumer Credit Act 1974	
Act	Farm Credit Act of 1971	same_as	Act	FCA	
Act	Farm Credit Act of 1971	same_as	Act	False Claims Act	
Act	Farm Credit Act of 1971	same_as	Act	Federal Courts Act	
Act	False Claims Act	same_as	Act	FCA	
Act	False Claims Act	same_as	Act	Farm Credit Act of 1971	
Act	False Claims Act	same_as	Act	Federal Courts Act	
Act	Small Business Act	same_as	Act	SBA	
Act	Federal Educational Rights and Privacy Act	same_as	Act	FERPA	
Act	ERA	same_as	Act	Employment Rights Act 1996	
Act	ERA	same_as	Act	Education Reform Act 1988	
Act	ERA	same_as	Act	Employment Relations Act 2000	
Act	ERA	same_as	Act	Employment Relations Act	
Act	Public Records Act 2005	same_as	Act	PRA	
Act	Public Records Act 2005	same_as	Act	Paperwork Reduction Act	
Act	Public Records Act 2005	same_as	Act	Public Records Act	
Act	Public Records Act 2005	same_as	Act	Paperwork Reduction Act of 1995	
Act	Public Records Act 2005	same_as	Act	Public Records Act 1958	
Act	Crime Control Act of 1990	same_as	Act	CCA	
Act	Crime Control Act of 1990	same_as	Act	Courts Constitution Act	
Act	Crime Control Act of 1990	same_as	Act	Court Costs Act	
Act	Crime Control Act of 1990	same_as	Act	Consumer Credit Act 1974	
Act	DEA	same_as	Act	307D Electoral Act	
Act	Health Insurance Portability and Accountability Act	same_as	Act	HIPAA	
Act	Health Insurance Portability and Accountability Act	same_as	Act	Health Insurance Portability and Accountability Act of 1996	
Act	Fair Housing Act	same_as	Act	FHA	
Act	Public Records Act 1958	same_as	Act	PRA	
Act	Public Records Act 1958	same_as	Act	Paperwork Reduction Act	

Act	Public Records Act 1958	same_as	Act	Public Records Act 2005	
Act	Public Records Act 1958	same_as	Act	Public Records Act	
Act	Public Records Act 1958	same_as	Act	Paperwork Reduction Act of 1995	
Act	Gramm-Leach- Bliley Act	same_as	Act	GLBA	
Act	Gramm-Leach- Bliley Act	same_as	Act	Gramm Leach Bliley Act	
Act	GLBA	same_as	Act	Gramm-Leach- Bliley Act	
Act	GLBA	same_as	Act	Gramm Leach Bliley Act	
Convention	International Telecommunications Convention	same_as	Convention	ITC	
Convention	International Telecommunications Convention	same_as	Convention	International Telecommunications Convention of 1973	
Police	IAP	same_as	Police	Italian American Police	
Act	Consumer Credit Act 1974	same_as	Act	CCA	
Act	Consumer Credit Act 1974	same_as	Act	Crime Control Act of 1990	
Act	Consumer Credit Act 1974	same_as	Act	Courts Constitution Act	
Act	Consumer Credit Act 1974	same_as	Act	Court Costs Act	
Principle	Individual Participation Principle	same_as	Principle	IPP	
Principle	Individual Participation Principle	same_as	Principle	Individual Participation Principle 58	
Act	Fair Credit Reporting Act	same_as	Act	FCRA	
Act	Fair Credit Reporting Act	same_as	Act	Federal Credit Reform Act of 1990	
Act	62C Building Act	same_as	Act	CBA	
Act	Court Costs Act	same_as	Act	CCA	
Act	Court Costs Act	same_as	Act	Crime Control Act of 1990	
Act	Court Costs Act	same_as	Act	Courts Constitution Act	
Act	Court Costs Act	same_as	Act	Consumer Credit Act 1974	
Rule	TSR	same_as	Rule	Telemarketing Sales Rule	
Act	Employment Relations Act 2000	same_as	Act	ERA	
Act	Employment Relations Act 2000	same_as	Act	Employment Rights Act 1996	
Act	Employment Relations Act 2000	same_as	Act	Education Reform Act 1988	
Act	Employment Relations Act 2000	same_as	Act	Employment Relations Act	
Act	Federal Credit Reform Act of 1990	same_as	Act	FCRA	
Act	Federal Credit Reform Act of 1990	same_as	Act	Fair Credit Reporting Act	
Act	Domestic Violence Act 1995	same_as	Act	DVA	
Act	Domestic Violence Act 1995	same_as	Act	Domestic Violence Act	
Act	FCA	same_as	Act	Farm Credit Act of 1971	
Act	FCA	same_as	Act	False Claims Act	
Act	FCA	same_as	Act	Federal Courts Act	
Act	Domestic Violence Act	same_as	Act	DVA	
Act	Domestic Violence Act	same_as	Act	Domestic Violence Act 1995	
Act	SBA	same_as	Act	Small Business Act	
Principle	Individual Participation Principle 58	same_as	Principle	IPP	

Principle	Individual Participation Principle 58	same_as	Principle	Individual Participation Principle	
Act	Town and Country Planning Act 1990	same_as	Act	TCPA	
Act	Town and Country Planning Act 1990	same_as	Act	Telephone Consumer Protection Act of 1991	
Police	Italian American Police	same_as	Police	IAP	
Rule	Telemarketing Sales Rule	same_as	Rule	TSR	
Act	FERPA	same_as	Act	Federal Educational Rights and Privacy Act	
Act	Age Discrimination Act of 1975	same_as	Act	ADA	
Act	Official Information Act 1982	same_as	Act	OIA	
Act	Official Information Act 1982	same_as	Act	Official Information Act	
Act	CBA	same_as	Act	62C Building Act	
Act	Tax Reform Act of 1986	same_as	Act	TRA	
Act	Health Insurance Portability and Accountability Act of 1996	same_as	Act	HIPAA	
Act	Health Insurance Portability and Accountability Act of 1996	same_as	Act	Health Insurance Portability and Accountability Act	
Act	DVA	same_as	Act	Domestic Violence Act 1995	
Act	DVA	same_as	Act	Domestic Violence Act	
Act	Equal Credit Opportunity Act	same_as	Act	ECOA	
Act	FHA	same_as	Act	Fair Housing Act	
Act	Federal Courts Act	same_as	Act	FCA	
Act	Federal Courts Act	same_as	Act	Farm Credit Act of 1971	
Act	Federal Courts Act	same_as	Act	False Claims Act	
Act	Employment Rights Act 1996	same_as	Act	ERA	
Act	Employment Rights Act 1996	same_as	Act	Education Reform Act 1988	
Act	Employment Rights Act 1996	same_as	Act	Employment Relations Act 2000	
Act	Employment Rights Act 1996	same_as	Act	Employment Relations Act	
Act	Financial Administration Act	same_as	Act	FAA	
Act	Financial Administration Act	same_as	Act	Final Agreement Act	
Act	Financial Administration Act	same_as	Act	Federal Aviation Act of 1958	
Act	Gramm Leach Bliley Act	same_as	Act	GLBA	
Act	Gramm Leach Bliley Act	same_as	Act	Gramm-Leach- Bliley Act	
Act	RFA	same_as	Act	Regulatory Flexibility Act	
Act	IPA	same_as	Act	Investigatory Powers Act 2000	
Act	Investigatory Powers Act 2000	same_as	Act	IPA	
Law	National Consumer Law	same_as	Law	NCL	
Act	ADA	same_as	Act	Age Discrimination Act of 1975	
Act	Paperwork Reduction Act of 1995	same_as	Act	PRA	
Act	Paperwork Reduction Act of 1995	same_as	Act	Paperwork Reduction Act	
Act	Paperwork Reduction Act of 1995	same_as	Act	Public Records Act 2005	
Act	Paperwork Reduction Act of 1995	same_as	Act	Public Records Act	
Act	Paperwork Reduction Act of 1995	same_as	Act	Public Records Act 1958	

Act	Federal Aviation Act of 1958	same_as	Act	FAA	
Act	Federal Aviation Act of 1958	same_as	Act	Final Agreement Act	
Act	Federal Aviation Act of 1958	same_as	Act	Financial Administration Act	
Act	PIPEDA	same_as	Act	Personal Information Protection and Electronic Documents Act	
Act	PIPEDA	same_as	Act	PIPED Act	
Act	Telecommunication Installations Act	same_as	Act	TIA	
Act	Telecommunication Installations Act	same_as	Act	Tribunals and Inquiries Act 1992	
Convention	International Telecommunications Convention of 1973	same_as	Convention	ITC	
Convention	International Telecommunications Convention of 1973	same_as	Convention	International Telecommunications Convention	
Act	Tribunals and Inquiries Act 1992	same_as	Act	TIA	
Act	Tribunals and Inquiries Act 1992	same_as	Act	Telecommunication Installations Act	
Act	Employment Relations Act	same_as	Act	ERA	
Act	Employment Relations Act	same_as	Act	Employment Rights Act 1996	
Act	Employment Relations Act	same_as	Act	Education Reform Act 1988	
Act	Employment Relations Act	same_as	Act	Employment Relations Act 2000	
Act	FAA	same_as	Act	Final Agreement Act	
Act	FAA	same_as	Act	Federal Aviation Act of 1958	
Act	FAA	same_as	Act	Financial Administration Act	
Act	Personal Information Protection and Electronic Documents Act	same_as	Act	PIPEDA	
Act	Personal Information Protection and Electronic Documents Act	same_as	Act	PIPED Act	
Act	PIPED Act	same_as	Act	PIPEDA	
Act	PIPED Act	same_as	Act	Personal Information Protection and Electronic Documents Act	
Act	FCRA	same_as	Act	Federal Credit Reform Act of 1990	
Act	FCRA	same_as	Act	Fair Credit Reporting Act	
Act	Public Records Act	same_as	Act	PRA	
Act	Public Records Act	same_as	Act	Paperwork Reduction Act	
Act	Public Records Act	same_as	Act	Public Records Act 2005	
Act	Public Records Act	same_as	Act	Paperwork Reduction Act of 1995	
Act	Public Records Act	same_as	Act	Public Records Act 1958	
Act	Telephone Consumer Protection Act of 1991	same_as	Act	TCPA	
Act	Telephone Consumer Protection Act of 1991	same_as	Act	Town and Country Planning Act 1990	
Act	PRA	same_as	Act	Paperwork Reduction Act	
Act	PRA	same_as	Act	Public Records Act 2005	
Act	PRA	same_as	Act	Public Records Act	
Act	PRA	same_as	Act	Paperwork Reduction Act of 1995	
Act	PRA	same_as	Act	Public Records Act 1958	
Act	MPA	same_as	Act	Medical Practitioners Act	
Act	TRA	same_as	Act	Tax Reform Act of 1986	
Convention	ITC	same_as	Convention	International Telecommunications Convention of 1973	

Convention	ITC	same_as	Convention	International Telecommunications Convention	
Act	OIA	same_as	Act	Official Information Act	
Act	OIA	same_as	Act	Official Information Act 1982	
Act	Paperwork Reduction Act	same_as	Act	PRA	
Act	Paperwork Reduction Act	same_as	Act	Public Records Act 2005	
Act	Paperwork Reduction Act	same_as	Act	Public Records Act	
Act	Paperwork Reduction Act	same_as	Act	Paperwork Reduction Act of 1995	
Act	Paperwork Reduction Act	same_as	Act	Public Records Act 1958	
Principle	IPP	same_as	Principle	Individual Participation Principle 58	
Principle	IPP	same_as	Principle	Individual Participation Principle	
Act	307D Electoral Act	same_as	Act	DEA	
Law	NCL	same_as	Law	National Consumer Law	
Act	TIA	same_as	Act	Telecommunication Installations Act	
Act	TIA	same_as	Act	Tribunals and Inquiries Act 1992	
Act	ERISA	same_as	Act	Employee Retirement Income Security Act of 1974	
Act	Regulatory Flexibility Act	same_as	Act	RFA	
Act	Education Reform Act 1988	same_as	Act	ERA	
Act	Education Reform Act 1988	same_as	Act	Employment Rights Act 1996	
Act	Education Reform Act 1988	same_as	Act	Employment Relations Act 2000	
Act	Education Reform Act 1988	same_as	Act	Employment Relations Act	
Act	Final Agreement Act	same_as	Act	FAA	
Act	Final Agreement Act	same_as	Act	Federal Aviation Act of 1958	
Act	Final Agreement Act	same_as	Act	Financial Administration Act	
Act	HIPAA	same_as	Act	Health Insurance Portability and Accountability Act of 1996	
Act	HIPAA	same_as	Act	Health Insurance Portability and Accountability Act	
Act	CCA	same_as	Act	Crime Control Act of 1990	
Act	CCA	same_as	Act	Courts Constitution Act	
Act	CCA	same_as	Act	Court Costs Act	
Act	CCA	same_as	Act	Consumer Credit Act 1974	

Class	Named entity	Relation	Class	Named entity	Correct?
Law	1996 Public Law 104-208, the Omnibus Consolidated Appropriations Act for Fiscal Year 1997, Title II, Subtitle D, Chapter 1	applies_to_geo	Geo	America	
Law	1998 Public Law 105-347	applies_to_geo	Geo	America	
Law	2001 USA PATRIOT Act Public Law 107-56	applies_to_geo	Geo	America	
Act	26 October Act	applies_to_geo	Geo	Portugal	
Regulation	Access Services and Facilities 30 Rates Regulation 31	applies_to_geo	Geo	Germany	
Act	Act 35	applies_to_geo	Geo	Australia	
Police	Alabama State Police	applies_to_geo	Geo	America	
Act	American Recovery and Reinvestment Act of 2009	applies_to_geo	Geo	America	
Act	American Reinvestment and Recovery Act of 2009	applies_to_geo	Geo	America	
Regulation	Approved Rates 38 Ex Post Rates Regulation	applies_to_geo	Geo	Germany	
Act	Article 10 Act	applies_to_geo	Geo	Germany	
Act	Australian Communications and Media Authority (Consequential and Transitional Provisions) Act 2005	applies_to_geo	Geo	Australia	
Act	Australian Communications and Media Authority Act 2005	applies_to_geo	Geo	Australia	
Law	Austrian Federal Law	applies_to_geo	Geo	Austria	
Act	Bankers' Books Evidence Act	applies_to_geo	Geo	India	
Act	British Columbia Act	applies_to_geo	Geo	United Kingdom	
Police	British Transport Police	applies_to_geo	Geo	United Kingdom	
Police	Cable and Telecommunications Committee NJ Police	applies_to_geo	Geo	America	
Act	Canada Evidence Act	applies_to_geo	Geo	Canada	
Act	Canada Evidence Act 41	applies_to_geo	Geo	Canada	
Act	Canada Evidence Act 701	applies_to_geo	Geo	Canada	
Police	Canada Royal Canadian Mounted Police	applies_to_geo	Geo	Canada	
Act	Canadian Personal Information Protection and Electronic Documents Act	applies_to_geo	Geo	Canada	
Act	Canadian Security Intelligence Service Act	applies_to_geo	Geo	Canada	
Act	Children and Young Persons Act	applies_to_geo	Geo	Ireland	
Act	Children and Young Persons Act 1933	applies_to_geo	Geo	Ireland	
Rule	China Telecommunication Rule	applies_to_geo	Geo	China	
Act	Civil Rights Act of 1968	applies_to_geo	Geo	America	
Act	Consumer Credit Reporting Reform Act of 1996	applies_to_geo	Geo	Georgia	
Act	Consumer Reporting Employment Clarification Act of 1998	applies_to_geo	Geo	America	
Act	Contents Privacy Act	applies_to_geo	Geo	Canada	
Act	Courts and Legal Services Act 1990	applies_to_geo	Geo	Ireland	
Act	Crime (Money Laundering) Act	applies_to_geo	Geo	Canada	
Act	Crime (Money Laundering) and Terrorist Financing Act	applies_to_geo	Geo	Canada	
Act	Criminal Procedure (Scotland) Act 1995	applies_to_geo	Geo	Ireland	
Number	Damages and Injunctive Relief 45 Customer Protection Ordinance 46 Number	applies_to_geo	Geo	Germany	
Act	Data Protection Act	applies_to_geo	Geo	Ireland	
Act	Data Protection Act 1984	applies_to_geo	Geo	Ireland	

Act	Data Protection Act 1998	applies_to_geo	Geo	Ireland	
Act	Data Protection Act 1998	applies_to_geo	Geo	Wales	
Law	Data Protection Law	applies_to_geo	Geo	Germany	
Act	Do Not Call Register Act 2006	applies_to_geo	Geo	Australia	
Act	Education (Scotland) Act 1980	applies_to_geo	Geo	Scotland	
Act	Fair and Accurate Credit Transactions Act of 2003	applies_to_geo	Geo	Georgia	
Act	Fatal Accidents and Sudden Deaths Inquiries (Scotland) Act 1976	applies_to_geo	Geo	Scotland	
Act	Final Agreement Act	applies_to_geo	Geo	United Kingdom	
Act	Financial Administration Act	applies_to_geo	Geo	Canada	
Law	Fiscal Year 1998 Public Law 105-107	applies_to_geo	Geo	America	
Act	Further and Higher Education (Scotland) Act 1992	applies_to_geo	Geo	Scotland	
Act	Government Employees Compensation Act	applies_to_geo	Geo	Canada	
Act	Gramm-Leach- Bliley Act	applies_to_geo	Geo	America	
Law	Gramm-Leach- Bliley Act Public Law 106-102	applies_to_geo	Geo	America	
Rule	Health Breach Notification Rule	applies_to_geo	Geo	America	
Act	Health Insurance Portability and Accountability Act of 1996	applies_to_geo	Geo	America	
Act	Health Records Act 1990	applies_to_geo	Geo	Ireland	
Act	Health Service Commissioners Act 1993	applies_to_geo	Geo	Wales	
Act	Home Owners' Act	applies_to_geo	Geo	Samoa	
Act	Immigration and Nationality Act	applies_to_geo	Geo	America	
Act	Immigration and Refugee Protection Act	applies_to_geo	Geo	Canada	
Police	Inc Illinois Police	applies_to_geo	Geo	America	
Police	Inc Michigan Police	applies_to_geo	Geo	America	
Law	Inc Southern Poverty Southern Poverty Law	applies_to_geo	Geo	Mexico	
Act	Indian Evidence Act	applies_to_geo	Geo	India	
Act	Indian Health Care Improvement Act	applies_to_geo	Geo	India	
Act	Indian Succession Act	applies_to_geo	Geo	India	
Act	Indian Trusts Act	applies_to_geo	Geo	India	
Act	Information Technology Act	applies_to_geo	Geo	India	
Act	Intercept and Obstruct Terrorism Act of 2001	applies_to_geo	Geo	America	
Police	Ireland 60 The Police	applies_to_geo	Geo	Ireland	
Act	Ireland Assembly Disqualification Act 1975	applies_to_geo	Geo	Ireland	
Police	Italian American Police	applies_to_geo	Geo	America	
Act	Labrador Inuit Land Claims Agreement Act	applies_to_geo	Geo	United Kingdom	
Regulation	Levying and Agreeing Rates 29 Rates Regulation	applies_to_geo	Geo	Germany	
Act	Local Government (Scotland) Act 1975	applies_to_geo	Geo	Scotland	
Act	Local Government (Wales) Act 1994	applies_to_geo	Geo	Wales	
Act	Local Government Act	applies_to_geo	Geo	Ireland	
Act	Local Government Act 1972	applies_to_geo	Geo	Wales	

Act	Medicare Australia Act 1973	applies_to_geo	Geo	Australia	
Act	Mental Health (Compulsory Assessment and Treatment) Act 1992	applies_to_geo	Geo	England	
Law	National Consumer Law	applies_to_geo	Geo	America	
Act	National Health Service (Scotland) Act 1978	applies_to_geo	Geo	Scotland	
Act	National Health Service and Community Care Act 1990	applies_to_geo	Geo	Scotland	
Act	Negotiable Instruments Act	applies_to_geo	Geo	India	
Police	New Jersey Police	applies_to_geo	Geo	America	
Act	New Tax System (Australian Business Number) Act 1999	applies_to_geo	Geo	Australia	
Number	New Tax System Australian Business Number	applies_to_geo	Geo	Australia	
Act	Northern Ireland (Emergency Provisions) Act	applies_to_geo	Geo	Ireland	
Act	Northern Ireland (Sentences) Act 1998	applies_to_geo	Geo	Ireland	
Act	Northern Ireland Act 1998	applies_to_geo	Geo	Ireland	
Act	Northern Ireland Assembly Disqualification Act 1975	applies_to_geo	Geo	Ireland	
Law	Office (Carol Beyers) McIntyre-Supp McIntyre Law	applies_to_geo	Geo	America	
Act	Omnibus Consolidated Appropriations Act	applies_to_geo	Geo	America	
Regulation	Orders Subchapter 2 Regulation	applies_to_geo	Geo	Germany	
Act	P-21 An Act	applies_to_geo	Geo	Canada	
Act	Personal Data Federal Act	applies_to_geo	Geo	England	
Act	Personal Files Act 1987	applies_to_geo	Geo	Ireland	
Act	Prisons (Scotland) Act 1989	applies_to_geo	Geo	Ireland	
Act	Provisions Act	applies_to_geo	Geo	Australia	
Act	Public Records Act	applies_to_geo	Geo	Ireland	
Act	Public Records Act 1958	applies_to_geo	Geo	Ireland	
Act	Quebec Act	applies_to_geo	Geo	Canada	
Regulation	Regulation 99	applies_to_geo	Geo	America	
Regulation	Retail Services 39 Rates Regulation	applies_to_geo	Geo	Germany	
Police	Royal Canadian Mounted Police	applies_to_geo	Geo	Canada	
Act	Royal Canadian Mounted Police Act	applies_to_geo	Geo	Canada	
Law	Russian Federation Law	applies_to_geo	Geo	Russia	
Act	School Standards and Framework Act 1998	applies_to_geo	Geo	Wales	
Act	Scotland Act 1989	applies_to_geo	Geo	Scotland	
Act	Scotland Act 1994	applies_to_geo	Geo	Scotland	
Act	Sechelt Indian Band Self-Government Act	applies_to_geo	Geo	Canada	
Act	Sechelt Indian Band Self-Government Act	applies_to_geo	Geo	India	
Act	Social Security Administration (Northern Ireland) Act 1992	applies_to_geo	Geo	Ireland	
Act	Social Security Contributions and Benefits (Northern Ireland) Act 1992	applies_to_geo	Geo	Ireland	
Act	Social Work (Scotland) Act 1968	applies_to_geo	Geo	Scotland	
Regulation	Subchapter 3 Regulation	applies_to_geo	Geo	Germany	
Regulation	Telecommunication Regulation	applies_to_geo	Geo	China	