UNIVERSIDADE FEDERAL DO CEARÁ
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JOSÉ WELLINGTON FRANCO DA SILVA

AQUISIÇÃO DE CONHECIMENTO DE MUNDO PARA SISTEMAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

FORTALEZA, CEARÁ

2013

JOSÉ WELLINGTON FRANCO DA SILVA

AQUISIÇÃO DE CONHECIMENTO DE MUNDO PARA SISTEMAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Marcelino Cavalcante Pequeno

Co-Orientador: Profa. Dra. Vládia Célia Monteiro Pinheiro

FORTALEZA, CEARÁ

2013

A000z .

Aquisição de Conhecimento de Mundo para Sistemas de Processamento de Linguagem Natural / José Wellington Franco da Silva. 2013.

88p.;il. color. enc.

Orientador: Prof. Dr. Marcelino Cavalcante Pequeno

Co-Orientador: Profa. Dra. Vládia Célia Monteiro Pinheiro

Dissertação(Ciência da Computação) - Universidade Federal do Ceará, Departamento de Computação, Fortaleza, 2013.

1. Aquisição de Conhecimento de Mundo 2. Entendimento de Linguagem Natural 3. Processamento de Linguagem Natural I. Prof. Dr. Marcelino Cavalcante Pequeno(Orient.) II. Universidade Federal do Ceará– Ciência da Computação(Mestrado) III. Mestre

~~CDD:000.0~~

A Deus, sempre. Aos meus pais, pelo exemplo de dedicação, trabalho e amor.

AGRADECIMENTOS

Agradeço em primeiro lugar a DEUS pelo o dom da vida, por todas as oportunidades que já tive em minha vida. Obrigado Senhor por me guiar e me iluminar em todos os momentos da minha vida.

Aos meus pais, Antônio Nelito e Lucia Helena que me educaram para vida, que me ensinaram as primeiras palavras, e que diversas vezes abriram mão de seus sonhos para que não faltassem nada para mim e para meu irmão. Ao meu irmão Rafael, por sempre iluminar a minha vida. Tenho certeza que você é um anjo que Deus enviou para nossa família.

A minha amiga e namorada Irinéia Raquel, pelo apoio e força. Com certeza foi bem mais fácil passar por tudo isso ao seu lado. Obrigado pela sua compreensão nos momentos que tive que está ausente.

À minha orientador Marcelino Pequeno, pela dedicação na elaboração deste trabalho e por ter aceitado o convite para me orientar.

À minha co-orientadora Vládia Pinheiro, por ter aceitado me co-orientar mesmo sendo de outra instituição. Obrigado também pela paciência comigo durante todo o desenvolvimento do trabalho e pela dedicação na elaboração do mesmo e por ter aceitado o convite para me orientar.

Aos amigos do grupo Logia, professores Carlos Brito, Ana Teresa e João Fernando pelas contribuições ao meu trabalho. E aos alunos Marcia Roberta, Carlos Filho, Henrique Viana, Thiago Alves, Gustavo Malkones, Luis Henrique e Arnaldo Junior pelo apoio durante o mestrado.

Aos amigos de mestrado Macedo Maia, David Araújo, Jeovane Regis, Anderson Boettge, Diego Sá, Fabiano Tavares, Juliano Efson, Régis Pires, Diego Victor e Manoel Siqueira pelos excelentes momentos de convívio e aprendizado que pudemos compartilhar ao longo dessa caminhada.

Ao Professor Vasco Furtado por ter cedido o Laboratorio de Engenharia do Conhecimento (LEC) na UNIFOR para a execução dos experimentos, além das importantes contribuições.

Aos amigos do LEC, Livio Freire, Janio Freire, Carlos Neto, Caio Cesar, Rodrigo Palheta, Rafael Bonfim e Marcelo Brito pelas inúmeras contribuições e incentivos durante o mestrado.

Aos amigos Willamy Fernandes e Regis Torquato pela ajuda na revisão do texto.

Aos meus gerentes Bruno Sabóia e Rute Castro, que sempre flexibilizaram meus horários para que eu resolvesse as coisas do mestrado.

Aos amigos do Great, Tiago Cunha, Katiuscia Morais, Leonel Alencar, Lenderson Abreu, Mardonio França, Francinaldo Madeira, Gracyane, Kevin Barros, Alexandre Arruda, Priscila Sampaio e Clayson Celes muito obrigado pelo incentivo e apoio.

Agradeço também aos diversos amigos que se preocuparam e me incentivaram durante esse processo. Em especial Italo Gois, Anderson Albuqueque, Aline Oliveira, Evenice Neta e todos os outros que não cito aqui.

Por fim, a todos que de alguma forma passaram na minha vida.

“A vida é como andar de bicicleta. Para manter-se em equilíbrio, você deve estar sempre em movimento.”

(A. Einstein)

RESUMO

Um dos desafios das pesquisas na área de Processamento de Linguagem Natural (PLN) é prover recursos semântico-linguísticos que expressem conhecimento de mundo para suportar tarefas como: extração de informação, recuperação de informação, sistemas de perguntas e respostas, sumarização de textos, anotação semântica de textos, dentre outras. Para esse desafio este trabalho propõe estratégias para aquisição de conhecimento de mundo. Propomos dois métodos. O primeiro é um método semiautomático que tem como ideia principal utilizar um processo de raciocínio semântico sobre o conhecimento pré-existente em uma base semântica. O segundo é um método de aquisição automática que utiliza a Wikipédia para a geração de conteúdo semântico. A Wikipédia foi utilizada como fonte de conhecimento devido à confiabilidade, dinamicidade e abrangência de seu conteúdo. Neste trabalho propomos um método para aquisição de relações semânticas entre conceitos a partir de textos de artigos da Wikipédia que faz uso de um conhecimento implícito existente na Wikipédia e em sistemas hipermídia: os links entre artigos. Ao longo do texto descritivo de um artigo da Wikipédia aparecem links para outros artigos que são evidências de que há uma relação entre o artigo corrente e o outro artigo referenciado pelo link. O método proposto objetiva capturar a relação semântica expressa no texto entre eles (artigo corrente e link para outro artigo), sem expressões regulares identificando relações similares através de uma medida de similaridade semântica.

Palavras-chave: Aquisição de Conhecimento de Mundo. Entendimento de Linguagem Natural. Processamento de Linguagem Natural.

ABSTRACT

One of the challenges of research in Natural Language Processing(NLP) is to provide semantic and linguistic resources to express knowledge of the world to support tasks such as Information Extraction, Information Retrieval systems, Questions & Answering, Text Summarization, Annotation Semantics of texts, etc. For this challenge this work proposes strategies for acquiring knowledge of the world. We propose two methods. The first is a semi-automatic method that has main idea of using a semantic reasoning process on pre-existing knowledge base semantics. The second is an acquisition method that utilizes automatic Wikipedia for generating semantical content. Wikipedia was used as a source of knowledge because of the reliability, dynamism and scope of its content. In this work we propose a method for acquiring semantic relations between concepts from the texts of Wikipedia articles that makes use of an implicit knowledge that exists in Wikipedia and in hypermedia systems: links between articles. Throughout the descriptive text of a Wikipedia article appear links to other articles that are evidence that there is a relationship between the current article and another article referenced by the link. The proposed method aims to capture the semantic relationship expressed in the text between them (current article and link to another article), no regular expressions identifying similar relationships through a semantic similarity measure.

Keywords: Acquisition of World Knowledge. Understanding Natural Language. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

Um dos desafios das pesquisas na área de Processamento de Linguagem Natural (PLN) é prover recursos semântico-linguísticos que expressem conhecimento de mundo para suportar o entendimento de linguagem natural por máquinas (ou, em inglês, *Natural Language Understanding - NLU*).

De acordo com [Ovchinnikova 2012], o principal objetivo de um sistema de NLU é interpretar um texto para que o conteúdo expresso sirva de base para realizar tarefas concretas, tais como extração de informação, recuperação de informação, sistemas de perguntas e respostas, sumarização de textos, análise de sentimentos. O processo de interpretação de um texto pode ser visto como uma tradução de uma língua natural para uma representação formal não ambígua, a qual expressa o conteúdo do texto.

[Ovchinnikova 2012] enumera três questões de fundo: Quais representações são consideradas adequadas para expressar o significado linguístico? E qual conhecimento é necessário para a compreensão da língua? E quais informações são consideradas uma parte do significado de língua natural? Para responder essas perguntas buscamos algumas referências na literatura.

Especialmente na área de Processamento de Linguagem Natural (PLN), há consenso de que a compreensão de textos por sistemas computacionais depende tanto de conhecimento de mundo como de conhecimento linguístico [Kay 2003]. Segundo [Ovchinnikova 2012], conhecimento de mundo associado ao conhecimento linguístico são importantes para sistemas de entendimento de linguagem natural. Considere os seguintes exemplos:

1. Se **SN** é um sintagma nominal e **V** é um verbo intransitivo, então a concatenação de **SN V** é uma oração.
2. A frase “x escreveu y” pode ser mapeada na proposição “escrever(x,y)”.
3. “Romeu e Julieta” foi escrito por *Shakespeare*.
4. Maria foi morta pelo marido.

Na sentença (1), temos um exemplo típico de regras sintáticas. Essas regras sintáticas estão incluídas em uma gramática e são dependentes da língua. Na sentença (2) temos o mapeamento do predicado escrever em lógica formal. Nas sentenças (3) e (4) temos fatos relativos ao mundo. O conhecimento de mundo que “*Romeu e Julieta*” é um livro e que *Shakespeare* é um escritor, os quais podem estar expressos em ontologias e bases semânticas, são importantes para o entendimento de (3). Para entender a sentença (4), *“Maria foi morta pelo seu marido”*, nos ajuda possuir conhecimento que *“Maria é uma pessoa”* e que *“maridos sentem ciúmes”*, para, por exemplo, inferir que aconteceu um crime passional. Em suma, as sentenças (1) e (2)

podem ser entendidas utilizando conhecimento linguístico e lógico. Já as sentenças (3) e (4) necessitam de um conhecimento extra-linguístico para serem entendidas.

Segundo [Ovchinnikova 2012], conhecimento de mundo é o conhecimento que é compartilhado por seres humanos pertencentes a uma mesma comunidade linguística e cultural não considerando aspectos individuais do discurso. Na definição de [Minsky 2007], o conhecimento de senso comum, que é um tipo de conhecimento de mundo, é um conjunto de fatos e conceitos que a maioria de nós sabemos. Por fim, segundo [Anacleto et al. 2007], o conhecimento de senso comum consiste em fatos e conhecimentos espaciais, físicos, sociais, temporais e psicológicos, possuídos pela maioria das pessoas, os quais são frutos da experiência da vida diária.

O conhecimento de mundo é fundamental para a solução de diversos problemas de PLN, dentre os quais podemos citar Ambiguidade, Metáforas, Resolução de Anáforas, Predicados Implícitos, etc.

Bases Semânticas que expressam conhecimento de mundo são amplamente utilizadas pelos sistemas de PLN. As mais conhecidas são WordNet [Miller 1995], FrameNet [Baker, Fillmore e Lowe 1998], ConceptNet [Liu e Singh 2004, Havasi, Speer e Alonso 2007, Speer e Havasi 2012] e InferenceNet [Pinheiro et al. 2010].

A WordNet é recurso léxico-semântico que tem como unidade básica de informação a palavra. Substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos chamados *synsets*. Esses *synsets* estão relacionados através de relações como hiponímia/hiperonímia, antonímia, etc. Por exemplo, a palavra *"house"* participa do seguintes *synsets*: *(house, firm),(house, theater),(house, domicile), etc.*

A FrameNet é uma base semântica baseada na teoria dos frames proposta por Minsky [Minsky 1974]. Nesta, cada frame representa um template de uma situação ou evento do mundo. Por exemplo, o evento de cometer um crime é representado, na FrameNet, por um frame "Commiting Crime" que relaciona um "criminoso" e um "crime" por meio do verbo "commit". De acordo com [Ruppenhofer et al. 2006], o principal objetivo da FrameNet é documentar o maior número de possibilidades sintáticas e semânticas de cada palavra em cada situação ou evento.

A ConceptNet é base de conhecimento de mundo que foi construída automaticamente a partir do *Open Mind Common-Sense Corpus (OMCS)* [Singh et al. 2002]. O OMCS é o corpus criado a partir do esforço colaborativo de pessoas. O conteúdo da ConceptNet é uma rede semântica que tem a seguinte estrutura: cada nó denota conceitos, tais como carro, casa, etc. e cada conceito se liga com outro conceito através de relações semânticas como por exemplo, *usadoPara(mesa da cozinha, comer café da manhã)*.

Na Universidade Federal do Ceará existe um grupo de pesquisadores que estudam Lógica e IA. Uma consequência do trabalho desse grupo foi o desenvolvimento de uma base conhecida como InferenceNet, o InferenceNet é um base de conhecimento construída por [Pinheiro et al. 2010] inspirada na teoria inferencialista de Brandom [Brandom 2001]. Para a abordagem inferencialista, o significado de uma sentença é o conjunto das inferências que podem

ser realizadas a partir de seu proferimento conjuntamente com todos os conjuntos de sentenças assumidos até então; ou seja, dado um conjunto de sentenças no qual acrescentamos uma nova sentença, o significado desta é o conjunto das inferências que podem ser realizadas com base no novo conjunto. O conteúdo da InferenceNet é uma rede semântica na qual os conceitos estão interligados por relações inferenciais que expressam as situações de uso dos conceitos - pré condições e pós-condições de uso. Por exemplo, capazDe(crime, ter vítima) é pré condição de uso do conceito crime e efeitoDe(crime, culpa, Pos) é uma pós-condição deste mesmo conceito.

A criação e manutenção de uma base de conhecimento não é uma tarefa fácil. Existe a dificuldade na geração de uma base completa, consistente e correta, visto que conhecimento de mundo tem ampla cobertura de domínios, grandes lacunas, imprecisões e ambiguidades. Este desafio é ainda maior quando consideramos a língua portuguesa, pois existem poucos recursos linguísticos que contemplam essa língua [Pardo, Caseli e Nunes 2009]. Dentre os recursos citados, existem versões para o português, por exemplo: a WordNet.Br [Silva, Felippo e Hasegawa 2006] contém atualmente na sua base 44.000 palavras, a FrameNet.Br contém um total de 32 frames e 38 unidades lexicais e, por fim, a ConceptNet.Br [Anacleto et al. 2006] está em andamento e conta, atualmente, com 160.000 afirmações de senso comum de seus colaboradores. Por fim, a InferenceNet.Br tem um total aproximado de 200.000 conceitos que se relacionam através de 800.000 relações. O principal problema dessas bases é que, em geral, são pouco representativas para a língua portuguesa.

1.2 Problemática

Com o intuito de termos recursos de conhecimento de mundo cada vez melhores, são necessários métodos de Aquisição de Conhecimento (AC) que garantam uma evolução contínua das bases de forma eficaz e com a agilidade que as aplicações exigem. Segundo [Scott, Clayton e Gibson 1991], os métodos de (AC) não envolvem apenas a identificação e a coleta de dados para bases de conhecimento, mas também engloba o armazenamento e organização da informação obtida, sempre com a preocupação na evolução dessas bases.

Segundo Boose (1984), os processos de AC podem ser divididos em dois tipos de métodos: métodos cognitivos e métodos automáticos. Os métodos cognitivos fazem a investigação e simulação computacional de processos de aprendizado humano. De acordo com [Gonçalves et al. 2012], esses métodos buscam simular o mecanismo de aprendizagem utilizado pelo cérebro humano na aquisição de conhecimento usando características como percepção, estimulo, adaptabilidade, dentre outras. Podemos citar como exemplos os sistemas multiagentes que aprendem de acordo com a percepção do meio em que estão inseridos.

Os métodos automáticos visam minimizar ou até eliminar a interferência humana no processo de AC e podem ser divididos em dois tipos: automáticos “puros” e semiautomáticos. No primeiro tipo, não há qualquer intervenção humana durante o processo de aquisição, enquanto que nos métodos semiautomáticos há necessidade de participação ou interferência humana, principalmente para validar ou direcionar o processo de aquisição de conhecimento.

Os métodos de AC automáticos são largamente usados em PLN para capturar o

conhecimento de mundo [Che et al. 2009, Baker, Ellsworth e Erk 2007]. As técnicas mais proeminentes são baseadas em aprendizado de máquina [Wu e Weld 2010] e expressões regulares [Fader, Soderland e Etzioni 2011, Xavier, Lima e Souza 2013]. O Wikipedia-based Open IE (WOE) [Wu e Weld 2010] propõe a extração automática de relações do tipo (arg1, relacao, arg2) utilizando técnicas de aprendizado supervisionado cujo o treinamento é realizado com um conjunto de exemplos etiquetados manualmente. Reverb [Fader, Soderland e Etzioni 2011] é outro ícone de sistemas que utilizam aprendizado supervisionado associado a um conjunto de expressões regulares e restrições lexicais. Em testes realizados, sua precisão alcançou 86%. Em [Xavier, Lima e Souza 2013], temos o *Lexical-Syntactic patterns based Open Extractor - (LSOE)*, um extrator de relações semânticas que utiliza um conjunto de padrões léxico-sintáticos para a extração de relações. A precisão desse método, segundo experimentos realizados pelos autores, é de 54%. Importante salientar que o conjunto de testes destes sistemas são diferentes e tais resultados não são comparáveis.

Esses sistemas são denominados de *Open IE Systems*, ou Sistemas Abertos de Extração de Informação, onde, em tese, não existe nenhuma restrição do tipo de informação a ser extraída [Li et al. 2011]. As principais desvantagens desses sistemas, em resumo, são a dependência de expressões regulares pré-definidas e a necessidade de um corpus anotado com informações semânticas para realização do treinamento. Uma oportunidade de melhoria desses sistemas é na identificação correta dos argumentos de uma relação semântica, que consiste em encontrar quais termos de fato estão relacionados. Por exemplo, da sentença *"I gave him 15 photographs"*, Reverb extrai a relação *(I, give, him)*, quando deveria extrair a relação *(I, give, 15 photographs)*. O problema na identificação dos argumentos é responsável por 52% dos erros dos sistemas Reverb [Fader, Soderland e Etzioni 2011] e WOE [Wu e Weld 2010].

Métodos de AC semiautomáticos tradicionais, por sua vez, enfrentam dificuldades em capturar conhecimento de mundo através de processos interativos. Essa dificuldade advém do fato de que as pessoas possuem esse tipo de conhecimento, mas não sabem explicitá-lo. Nós sabemos falar, assim como sabemos nadar ou andar de bicicleta, sabemos que uma consequência de cair de bicicleta é machucar-se, que quando alguém fala "Eu comprei doces" está implícito que gastou dinheiro, etc. No entanto, sentimos dificuldades em explicitar estes conhecimentos. O conhecimento está tão arraigado na mente das pessoas que é difícil lembrar e, mais ainda, externá-lo por meio de relações semânticas estruturadas. Outra questão é que, mesmo quando conseguimos explicitar conteúdo e relações conceituais, é difícil garantir a consistência com conteúdos já existentes na base de conhecimento, evitando a duplicação de conteúdo e fortalecendo a conexão entre os conceitos.

Desta classe de métodos, os mais relevantes são o projeto CYC [Lenat 1995] e o projeto OMCS [Singh et al. 2002]. O projeto Open CYC é uma ontologia produzida dentro do projeto Cyc, que começou em 1984, com o objetivo de construir uma base de conhecimento, incluindo tanto conhecimento científico quanto conhecimento de mundo em larga escala. Inicialmente foi criada uma base por um grupo de especialistas pagos. Nos primeiros anos do projeto, o CYC já tinha 1,6 milhão regras e 180.000 conceitos. O *Open Mind Common-Sense Corpus (OMCS)* é um projeto que tem como ideia principal criar uma base de conhecimento de mundo a partir da contribuição colaborativa de qualquer pessoa. As pessoas incluem as sen-

tenças respondendo a perguntas como “O efeito de comer é? ”. O projeto foi responsável por adquirir 300.000 conceitos e 1.6 milhões de relações para a base ConceptNet. É notório que o sucesso de projetos como estes, depende de esforço colaborativo de grande número de pessoas.

1.3 Solução Proposta

Como vimos, embora métodos de AC sejam propostos para garantir a evolução das bases de conhecimento de mundo, estes encontram problemas na realização dessa tarefa. Durante este trabalho de pesquisa foi possível delimitar os principais problemas na aquisição de conhecimento de mundo:

- Em métodos semiautomático de AC, o processo interativo utilizado implica em uma aquisição lenta e dispendiosa de conhecimento além de não garantir a consistências das bases.
- Métodos automáticos de AC que utilizam expressões regulares e padrões sintáticos são limitados por não contemplar as diversas expressões de relações semânticas em práticas linguísticas.
- Métodos automáticos de AC que utilizam técnicas de aprendizado de máquina são fortemente dependentes de um *corpus* para treinamento.

Com base nesses problemas, as nossas hipóteses de pesquisa foram as seguintes:

1. O conhecimento preexistente em uma base auxilia o usuário a explicitar e a validar relações semânticas para um novo conceito, otimizando o processo interativo de aquisição de conhecimento de mundo.
2. Textos de artigos da Wikipédia são uma importante e confiável fonte de conhecimento de mundo e sua semi-estrutura, provida pelos hyperlinks entre artigos, delimitam relações semânticas entre conceitos.

Resultados de experimentos realizados durante a pesquisa com adultos humanos suportam a primeira hipótese. Pedimos para usuários inserirem relações em uma base de conhecimento de mundo qualquer auxílio de uma base inicial com conteúdo relacionado. Percebeu-se, neste cenário, um baixo índice de inclusão de conceitos. O número de interações dobrou quando os usuários foram ajudados com conteúdo relacionado ao novo conceito.

Para fortalecer nossa crença na segunda hipótese, experimentos realizados evidenciaram que existem cerca de 300 mil relações entre links nos textos da Wikipedia, que podem ser usadas para uma evolução contínua de bases de conhecimento de mundo.

A partir das hipóteses, este trabalho de pesquisa resultou na proposta de dois métodos para aquisição de conhecimento de mundo.

1. Método semiautomático de Conhecimento de Mundo - este método apresenta como diferencial um módulo de raciocínio sobre conhecimento preexistente que visa oferecer ao usuário conteúdo inicial que o ajude a externar e a validar relações semânticas de novos conceitos. O processo de raciocínio se baseia em heurísticas e na análise sintática de sintagmas nominais. Por exemplo, o sintagma nominal "crime passional" é composto de um substantivo e de um adjetivo. Neste caso, o substantivo está sendo qualificado pelo adjetivo denotando uma especialização - "crime passional" é um tipo de "crime". Uma avaliação qualitativa do método proposto indicou 72% de acurácia nas relações semânticas adquiridas com o auxílio do método. Os resultados da avaliação também nos permitiu concluir que as interações realizadas foram mais produtivas e eficazes na medida em que o usuário era lembrado e instigado sobre relações semânticas acerca do novo conceito.

2. Método Automático para Extração de Relações Semânticas da Wikipédia - a Wikipédia tem se tornado uma importante fonte de conhecimento, principalmente devido a confiabilidade, dinamicidade e abrangência de seu conteúdo [Kittur e Kraut 2008]. O método, ora proposto, adquire automaticamente relações semânticas entre conceitos a partir do texto de documentos da Wikipédia e faz uso de um conhecimento implícito existente em sistemas hipermídia: os links entre artigos. O principal diferencial do método proposto é a independência de expressões regulares pré-definidas, o uso de links para definição dos argumentos das relações e a identificação de relações redundantes. Para além destas vantagens, o método proposto não aplica técnicas de aprendizagem supervisionada, cuja necessidade de anotação de corpus é sempre um gargalo para avanços nas pesquisas em PLN, principalmente para língua portuguesa. A avaliação realizada em 100 mil artigos da Wikipédia demonstrou uma precisão de 76%, resultado este que equivale o estado da arte em *Open IE Systems*.

1.4 Objetivos e Contribuições

Este trabalho tem como principal objetivo a proposta e avaliação de dois métodos de aquisição de conhecimento de mundo. O primeiro, um método semiautomático, utiliza o próprio conhecimento existente em uma base para ajudar o usuário a incluir novas relações semânticas. O segundo método extrai automaticamente relações semânticas dos textos da Wikipédia utilizando o contexto existente entre os *links*.

Além dos métodos propostos, destacamos as seguintes contribuições deste trabalho de pesquisa:

- Elaboração de um algoritmo para aquisição de conhecimento a partir da estrutura gramatical de sentenças em língua portuguesa;
- Definição de estratégias para extração de conhecimento a partir dos textos e links da Wikipédia, independentes de expressões regulares e padrões sintáticos previamente definidos;
- Desenvolvimento de um algoritmo para aquisição automática de conhecimento a partir da Wikipédia;

- Definição de uma estratégia de clusterização de relações baseadas em uma medida de similaridade semântica.

1.4.1 Organização da Dissertação

Esta dissertação está organizada em sete capítulos descritos, resumidamente, na sequência:

- **Capitulo 1. Introdução**

  Este capítulo apresenta a motivação científica da pesquisa, a problemática atual das pesquisas para aquisição de conhecimento de mundo, uma visão geral da solução proposta e um resumo dos objetivos e contribuições deste trabalho.

- **Capitulo 2. Entendimento de Linguagem Natural**

  O capítulo 2 traz a revisão bibliográfica da área de conhecimento de mundo sobre dois enfoques: (i) A importância do conhecimento de mundo para os sistemas de entendimento de linguagem natural (ii) modelos de expressão de conhecimento semântico dos principais recursos linguístico-computacionais - WordNet, FrameNet, ConceptNet e InferenceNet. O capítulo termina com uma análise comparativa dos recursos semânticos descritos.

- **Capitulo 3. Aquisição de Conhecimento de Mundo: O Estado da Arte**

  O capítulo 3 traz a revisão bibliográfica da área de aquisição de conhecimento de mundo detalhando os principais projetos, métodos e ferramentas existentes na literatura. O capítulo termina relacionando os principais problemas encontrados nos métodos estudos.

- **Capitulo 4. Aquisição de Conhecimento de Mundo: Propostas de Solução**

  Este é o cerne do texto, apresentando os métodos para aquisição de conhecimento de mundo. Os fundamentos teóricos, conceitualização, arquitetura dos métodos Aquisição Semiautomática de Conhecimento de Mundo e Aquisição Automática de Relações Semânticas da Wikipédia são detalhadamente definidos e explanados. Ao final, são defendidas as vantagens e desvantagens, em relação aos outros métodos de aquisição de conhecimento de mundo.

- **Capitulo 5. Conclusão**

  Este é o remate da dissertação, aportando as contribuições científicas, resultados e produtos da pesquisa, restrições do método proposto, oportunidades de melhorias e indicações de trabalhos futuros.

2 ENTENDIMENTO DE LINGUAGEM NATURAL

Este capítulo versa sobre a fundamentação teórica deste trabalho. Inicialmente, discorremos sobre Processamento de Linguagem Natural e, especialmente, focamos na tarefa de entendimento de linguagem natural pelo computador. Por fim, apresentamos os principais recursos semânticos-linguísticos utilizados em sistemas de entendimento de linguagem natural, realizando uma análise comparativa.

2.1 Introdução

Com uma presença cada vez maior dos computadores no cotidiano das pessoas, cresce a necessidade de uma melhor comunicação entre homens e máquinas. A melhor forma de uma máquina se comunicar com um ser humano é através da língua natural, visto que estes utilizam esta linguagem para se comunicar. Quanto mais os computadores utilizarem a língua natural (falada e escrita) para se comunicar com os seres humanos, mais fácil será a interação com as máquinas.

Com a ideia de aproximar a língua natural, utilizada pelas pessoas, dos computadores, surgiu uma área de pesquisa chamada de Linguística Computacional. De acordo com [Vieira e Lima 2001], a "Linguística Computacional é uma área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas capazes de reconhecer e produzir informações em língua natural". Essa área utiliza as recursos tanto da linguística como da computação para a resolução de problemas como extrair e recuperar informações, sumariar textos, etc.

Segundo [Vieira e Lima 2001], a função do PLN é a implementação de programas capazes de entender e/ou gerar informação em língua natural. No que diz respeito ao entendimento de linguagem natural, esses programas devem ser capazes de manipular os signos linguísticos[1] para tomar decisões, extrair e recuperar informações, sumarizar textos, dentre outras tarefas que envolvem compreensão de sentenças e textos em língua natural.

Em PLN, os processadores linguísticos podem ser divididos nos níveis de processamento da língua: Morfologia, Sintaxe, Semântica e Pragmática. A figura 2.1 ilustra estes níveis, os recursos, semânticos e exemplifica alguns processadores usados comumente em sistemas de PLN. A seguir, detalhamos cada nível com base no trabalho de [Silva et al. 2007].

- Morfologia - neste nível as unidades mínimas dotadas de significado são isoladas para a compreensão do processo de formação e flexão das palavras. Este processo envolve a identificação e separação dos componentes significativos da sentença sob análise, comumente chamadas de *tokens*, tais como as palavras e os símbolos de pontuação. Além

[1] Segundo [Trask 2004], Signo linguístico - Um objeto linguístico dotado simultaneamente de forma e sentido. Por exemplo, a palavra portuguesa cachorro tem uma forma particular, que consiste numa sequência de seis fonemas destituídos de sentido, e também um significado particular - um tipo específico de animal. Os dois juntos formam um só e único signo, em português.

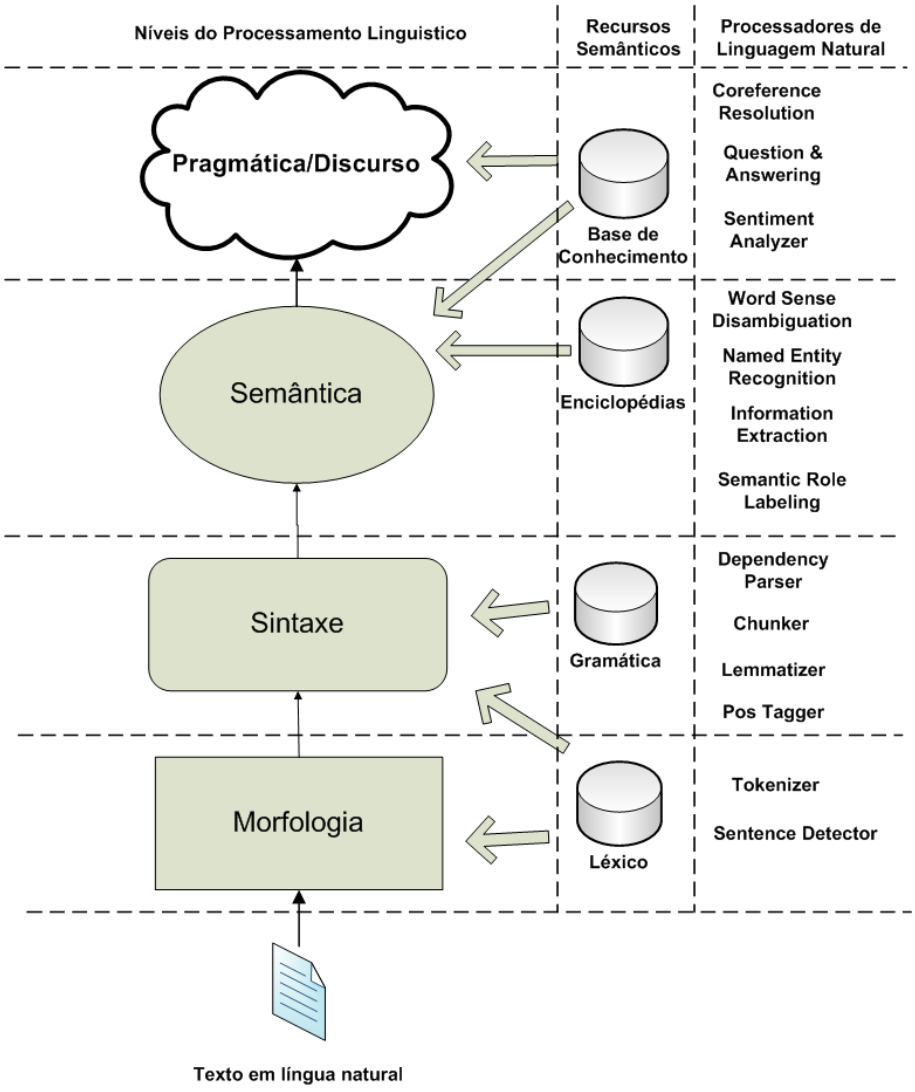


Figura 2.1: Visão geral dos diferentes níveis de processamento linguístico em PLN

disso, realiza a associação de atributos ou traços gramaticais e/ou semânticos a cada *token*, com base em consultas ao Léxico. Léxico é um recurso que guarda um conjunto de palavras e variações dessas palavras de acordo com a forma. Podemos citar nesse nível os seguintes processadores: Detector de Sentenças, *Tokenizador* e *POS-Tagger*. A primeira ferramenta é responsável por dividir o texto em sentenças. Já o *Tokenizador* tem como principal objetivo separar as sentenças em unidades menores que vamos chamar de *tokens*. No exemplo: "João gosta de futebol", temos os seguintes *tokens*: "João", "gosta", "de", "futebol" e ".". A ferramenta *POS-Tagger* classifica cada *token* de acordo com sua classe gramatical. Por exemplo: "João"-> Substantivo, "gosta"-> Verbo, "de"-> Preposição e "futebol"-> Substantivo.

- Sintaxe - neste nível considera-se que a organização das palavras e expressões resulta em determinadas funções que elas desempenham na sentença. Este nível é responsável por construir (ou recuperar) uma estrutura sintática válida para a sentença de entrada, também chamada de estrutura profunda. Para tanto, é guiado por uma representação da gramática da língua em questão. Em se tratando de uma língua natural, em geral adota-se uma gramática "parcial" da língua natural, que, embora não abranja todas as construções da língua, contempla aquelas construções válidas de interesse para a aplicação. Alguns exemplos de processadores são *Lemmatizer*, *Chunker* e *Dependence Parser*. O *Lemmatizer* tem como objetivo retornar o lemma de uma palavra. O lemma é a forma canônica de uma palavra. Por exemplo, na palavra "é" o lemma é "ser". O *Chunker* agrupa, identifica e classifica os sintagmas de uma sentença. Por exemplo, na sentença "Maria vai entregar os documentos a Pedro" temos os seguintes sintagmas:"Maria"-> Sintagma Nominal, "vai entregar"-> Sintagma Verbal, "os documentos"-> Sintagma Nominal e "a Pedro"-> Sintagma Preposicional. O *Dependence Parser* retorna a árvore de dependência sintática de uma sentença. O parser realiza a análise sintática automática de frases em termos de suas funções gramaticais, ou seja, se colocarmos uma frase como "o menino brinca", o parser é capaz de processá-la e dar como saída a sinalização do que é artigo, sujeito e complemento da frase. A saída então será uma árvore onde os nós folha são as palavras da sentença. No *Dependence Parser* cada palavra da frase tem uma relação com seus dependentes. Por exemplo, saber a identidade do verbo ajuda a determinar qual é o sujeito e qual é o objeto na frase.

- Semântica - o objetivo desse nível é identificar o conteúdo significativo da palavra que implica, por exemplo, em associar relações de natureza ontológica e referencial com objetos no mundo ou conceitos na mente. Existe um consenso que é necessário conhecimento de mundo para o processamento no nível semântico das línguas naturais. Este nível é responsável pela interpretação de componentes da sentença ou da sentença como um todo e está presente sempre que a aplicação exigir algum tipo de interpretação. Nesse caso, é necessário conhecimento mais específico do domínio, presente no Modelo do Domínio, p.ex., para distinguir a interpretação correta do termo manga (se parte de um vestuário ou objeto comestível). Podemos citar os seguintes processadores:

  - *Word Sense Disambiguation - WSD* [Stevenson e Wilks 2003] - tem como função identificar o sentido correto de uma palavra ambígua, quando usada em uma sen-

tença em particular. Por exemplo, na sentença “João gosta de manga” a palavra “manga” é usada no sentido de um objeto comestível.

- *Semantic Role Labeling - SRL* [Gildea e Jurafsky 2002] - Papéis semânticos representam as relações lógicas entre um evento e seus participantes. *Semantic Role Labeling - SRL* ou Anotação de papéis semânticos é o processo de extrair automaticamente estruturas de papéis semânticos que permitem a análise do significado das sentenças. Por exemplo, na sentença “João quebrou o vaso”, “João” assume o papel de Agente, ou seja, causador voluntário de uma ação, e o “vaso” o papel de Paciente, ou seja, quem ou o quê sofre a ação.
- *Named Entity Recognition - NER* [Nadeau e Sekine 2007] - esse processador tem como função identificar e classificar algumas entidades especiais do texto, tais como, localização, pessoas, tempo, organização, datas, quantidades e valores, etc. Por exemplo, na sentença “João Roberto trabalhou na UFC.”, o processador NER deve reconhecer “João Roberto” como pessoa e “UFC” como organização.
- *Information Extraction(IE)* [Cowie e Lehnert 1996] - consiste em localizar e extrair, de forma automática, informações relevantes em um documento ou coleção de documentos expressos em língua natural e estruturar tais informações para os padrões de saída, por exemplo, em banco de dados ou textos em língua natural, a fim de facilitar sua manipulação e análise.

• Pragmático/Discursivo - em textos, a força expressiva das palavras remete à identificação dos objetos do mundo em termos do seu contexto de enunciação e condições de produção discursiva. Analisadores discursivos permitem a obtenção de uma representação do significado da mensagem original, levando em conta aspectos pragmáticos da comunicação. Por exemplo, nem sempre o caráter interrogativo de uma sentença expressa exatamente o caráter de solicitação de uma resposta. A sentença “Você sabe que horas são?” pode ser interpretada como uma solicitação para que as horas sejam informadas ou como uma repreensão por um atraso ocorrido. No primeiro caso, a pergunta informa ao ouvinte que o falante deseja obter uma informação e, portanto, expressa exatamente o caráter interrogativo. Entretanto, no segundo caso, o falante utiliza o artifício interrogativo como forma de impor sua autoridade. Estas diferentes interpretações são claramente relacionadas com a prática da língua no dia-a-dia e com as relações sociais presentes no contexto discursivo. Para apreender essa prática, os processadores recorrem a bases de conhecimento de mundo, especificamente, a bases de senso comum. Alguns processadores desse nível são:

- *Coreference Resolution (CR)* [Soon, Ng e Lim 2001] - estes processadores tem como principal função identificar cadeias de correferência, ou seja, um grupo de palavras ou expressões que se referem a uma mesma entidade. Por exemplo: “João viajou. Ele estava na Jordania.”, a cadeia {“Ele”, “João”} refere-se a mesma entidade - uma pessoa chamada João.
- *Question & Answering* [Lee et al. 2001] - tem como principal função responder uma pergunta feita por um usuário. Por exemplo, “Onde nasceu Pelé?” o sistema deve procurar uma fonte confiável é responder:“Em Três Corações, Minas Gerais”.

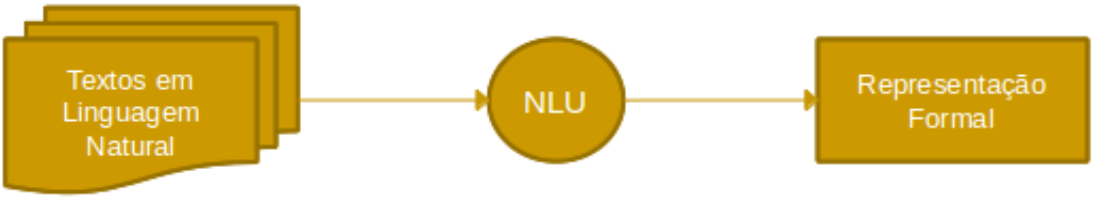


Figura 2.2: Visão geral de NLU

– *Sentiment Analyzer* [Taboada, Anthony e Voll 2006, Bontcheva et al. 2013] - refere-se ao uso de processamento de linguagem natural, análise de texto e linguística computacional para identificar e extrair informações subjetivas de textos como (tristeza, alegria, polaridade positiva ou negativa, etc.).

Vale a pena ressaltar que os processadores dos níveis semântico e pragmático se enquadram em sistemas/processadores de entendimento de linguagem natural e necessitam de conhecimento de mundo para realizar suas tarefas.

2.2 Entendimento de Linguagem Natural

De acordo com [Ovchinnikova 2012], entendimento de linguagem natural (ou, em inglês, *Natural Language Understanding - NLU*) é um sub-campo do Processamento de Linguagem Natural que lida com maquinas de leitura e compreensão de textos. O principal objetivo de um sistema de NLU é interpretar um texto para que o conteúdo expresso sirva de base para realizar tarefas concretas como mostrado na figura 2.2.

O entendimento de linguagem natural por sistemas computacionais é um dos problemas mais instigantes da história do PLN. A primeira tentativa de entendimento de linguagem natural vem da década de 60 com o programa STUDENT desenvolvido por Bobrow [Bobrow 1964]. Esse programa tinha como principal funcionalidade resolver problemas de álgebra do ensino médio. Sua estratégia era converter uma pergunta em linguagem natural em uma consulta para um banco de resolução de problemas.

Alguns anos depois, Weizenbaum [Weizenbaum 1966] propôs o Eliza, o primeiro *ChaterBot* conhecido. Os *ChatterBot* são programas de diálogos que tentam simular um ser humano durante uma conversa. Eliza funcionava com analisador morfossintático e tentava buscar padrões nas perguntas para achar uma resposta em seu banco de dados.

Esses dois programas foram de grande importância para o inicio das áreas de IA e PLN, mas não são considerados programas de entendimento de linguagem natural. De acordo com [Ovchinnikova 2012], existem dois critérios para julgar se realmente um sistema "entende" um fragmento de texto. O primeiro critério é classificado como *performance-based*.

Nesse critério, verificamos como o sistema atua em determinadas situações. A conferência TAC (*Text Analysis Conference*[2]*)* avalia sistemas de NLU de acordo com esse critério. Ela promove uma competição em tarefas específicas como sumarização, reconhecimento de padrões textuais. Nesta conferência os sistemas de entendimento de linguagem natural são comparados com outros sistemas existentes ou até mesmo com humanos na realização de tarefas.

O segundo critério leva em conta a representação interna do conteúdo gerado pelo sistema durante a interpretação de um texto. A avaliação é feita da seguinte forma: as representações geradas automaticamente são comparadas com anotações humanas no mesmo texto. A principal conferência que avalia os sistemas de NLU usando esse critério é *Semantic Evaluation (SemEval*[3] *)*, cujo foco muda a cada edição. Em 2010, uma das trilhas foi a tarefa de *Keyphrase Extraction*, ou extração de palavras-chave, que consiste na extração de um conjunto de palavras que expressam o conteúdo de um determinado texto. O sistema vencedor foi HUMB com 27,5% de acurácia nesta tarefa.

Um fator importante na construção de sistemas de entendimento de linguagem natural é a representação do significado. Segundo [Allen 1987, Schank 1975], para podermos representar o significado linguístico temos que obedecer três princípios.

- Sentenças com mesmo significado devem ser representadas da mesma forma;
- As representações devem ser precisas e sem ambiguidades;
- As informações implícitas devem ser explicitadas;

As teorias linguísticas propõem abordagens para representação do significado. A *Formal Semantics* tem como foco utilizar as propriedade da lógica para representação formal da linguagem natural, por exemplo, *"Shakespeare wrote a tragedy"* e *"A tragedy was written by Shakespeare"* pode ambas ser em representadas pela seguinte formula lógica: $\exists t,s,e(tragedy(t) \wedge Shakespeare(s) \wedge write(e,s,t))$. *Lexical Semantics* utiliza informações do/no léxico para expressar o significado de palavras ou expressões (itens lexicais). Por exemplo, *to drink* normalmente tem uma bebida como objeto direto. *Distributional Semantics* que utiliza técnicas estatísticas e de aprendizado de máquina para representar o significado de palavras. O princípio básico desta teoria é que palavras semelhantes ocorrem em contextos similares.

2.2.1 Conhecimento de Mundo Compartilhado para Entendimento de Linguagem Natural

Conhecimento de mundo é necessário para a resolução de alguns problemas, como ambiguidade, resolução de correferência, metáforas e metonímias e relações discursivas. OVCHINNIKOVA(2012) apresenta alguns exemplos contextualizados de problemas em que a expressão e raciocínio como conhecimento de mundo é imprescindível para uma solução.

[2]http://www.nist.gov/tac/

[3]http://aclweb.org/aclwiki/index.php?title=SemEvalPortal

Ambiguidade

Em uma língua é comum uma palavra ou expressão ter mais de um sentido, como na sentença "A mãe de Pedro entrou com seu carro na garagem" não sabemos dizer de quem é o carro. O fenômeno da ambiguidade deve-se a casos de polissemia ou homonímia. A polissemia é o caso em que uma dada palavra ou expressão adquire um novo sentido além de seu sentido original, guardando uma relação de sentido entre elas[4]. A homonímia é a propriedade onde duas ou mais formas, totalmente diversas pela significação ou função, terem a mesma estrutura fonológica, os mesmo fonemas, dispostos na mesma ordem e com o mesmo tipo de acentuação[5]. A ambiguidade afeta todos os níveis do processamento linguístico. Uma possível solução para esse problema é utilizar conhecimento de mundo para a desambiguação dos sentidos das palavras. No exemplo: "João foi ao banco abrir uma conta." não se sabe se a palavra "banco" refere-se a uma instituição financeira ou ao repositório de areia presente em rios e mares. Neste exemplo, o conhecimento de mundo que instituições financeiras permitem abertura de contas para depósito de valores e que é impossível abrir uma conta em um banco de areia nos auxilia a desambiguar a palavra "banco" para o sentido instituição financeira.

Resolução de Correferência

A identificação de cadeias de correferência, ou seja, um grupo de palavras ou expressões que se referem a uma mesma entidade pode ser auxiliada por conhecimento de mundo. Anáfora é um caso especial de correferência quando um pronome ou outro nome ou expressão faz referência a um sujeito anteriormente citado no texto. Vejamos o seguinte exemplo, na sentença "*We gave the bananas to the monkeys because they were hungry.*" não sabemos se "*they*" refere-se a "*monkeys*" ou a "*bananas*", mas o conhecimento que animais comem pode auxiliar nesta resolução.

Metáforas e Metonímias

A interpretação de metáforas e metonímias são problemas comuns em diversos sistemas de PLN. Para interpretação dessas figuras de linguagem o sistema de PLN deve ter um conhecimento profundo sobre determinado tema ou um bom conhecimento de mundo. Por exemplo, na sentença "O Planalto apoia o projeto da ficha limpa" o sistema deve ter conhecimento suficiente para inferir que a palavra "Planalto" se refere ao Palácio do Planalto que é o lugar onde fica localizado o gabinete do Presidente da República do Brasil.

Predicados Implícitos

Este fenômeno acontece em práticas linguísticas quando um predicado é tão previsível em um contexto que pode ser omitido. No exemplo, "José de Alencar finalizou o livro" apesar de não ter um sentido único, podemos inferir, com base no conhecimento de mundo que José de Alencar, é um escritor, que ele deve ter terminado "de escrever" o livro. Portanto, o predicado "escrever" estava implícito na sentença.

Relações Discursivas

[4] http://www.filologia.org.br/revista/36/10.htm

[5] http://www.dicionarioinformal.com.br/homon%C3%ADmia/

A definição de como dois segmentos de discurso diferentes estão logicamente relacionados também exige bom nível de conhecimento de mundo. No exemplo do texto “Joana acordou tarde hoje. Seu alarme quebrou.”, podemos inferir que há uma relação de causa e efeito entre as duas sentenças.

Estes exemplos de situações comuns no uso de linguagem natural denotam a importância de recursos semânticos que expressem conhecimento de mundo em suas bases e deem suporte a sistemas de PLN. Na seção a seguir descrevemos os principais recursos existentes.

2.3 Recursos Linguísticos

Em sistemas de PLN, comumente são necessários recursos linguístico como: gramáticas, léxicos, analisadores, etc., e bases com conhecimento semântico, as quais expressam o valor semântico de conceitos articulados em sentenças e textos.

Neste sentido, foram construídos diversos recursos semântico-linguísticos pela comunidade acadêmica que visam ajudar na construção de sistemas de PLN. Os recursos mais conhecidos são: WordNet [Miller 1995], FrameNet [Baker, Fillmore e Lowe 1998], ConceptNet [Singh et al. 2002] e InferenceNet [Pinheiro et al. 2010].

2.3.1 WordNet

A WordNet [Miller 1995] é uma larga base de dados léxico-semântica, inicialmente criada em inglês, desenvolvida na Universidade de Princeton. Nesta base semântica, a unidade básica de informação é a palavra. A WordNet está dividida em quatro redes semânticas, uma para cada classe aberta de palavras: substantivo, verbo, adjetivo e advérbio. As palavras são agrupadas em conjuntos de sinônimos - *synset*, cada um expressando um conceito particular. As palavras que pertencem a um mesmo *synset* estão relacionadas pela relação de sinonímia. As principais relações semânticas entre conceitos ou *synset* são: hiponímia/hiperonímia[6], meronímia/holonímia[7], similaridade, antonímia, *entailment (implicação)* e relação causal (para a classe dos verbos) etc. A figura 2.3, apresenta uma consulta ao portal da WordNet para a palavra “*house*”. Nesta figura, podemos identificar 12 sinônimos para a palavra “*house*”.

Ao longo do tempo, a WordNet evoluiu e incorporou alguns trabalhos que propunham melhorias. Um exemplo foi o trabalho de KOHL(1998), que apresentam os padrões sintáticos em que verbos podem ocorrer. A contribuição desse estudo foi a constatação, até certo ponto óbvia, de que verbos similares semanticamente nem sempre compartilham os mesmos comportamentos sintáticos. A WordNet foi construída de forma manual através de grande esforço de um grupo de linguistas.

[6]Hiponímia é a relação entre um membro e sua classe: X é hipônimo de Y se X é um (tipo de) Y. Sua relação inversa é a hiperonímia, que expressa relação entre a classe e seus membros: Y é hiperônimo de X se X é um (tipo de) Y

[7]Meronímia é a relação entre uma parte e o todo: X é merônimo de Y se X é parte de Y. Sua relação inversa é a holonímia, que expressa relação entre o todo e suas partes: Y é holônimo de X se X é parte de Y.

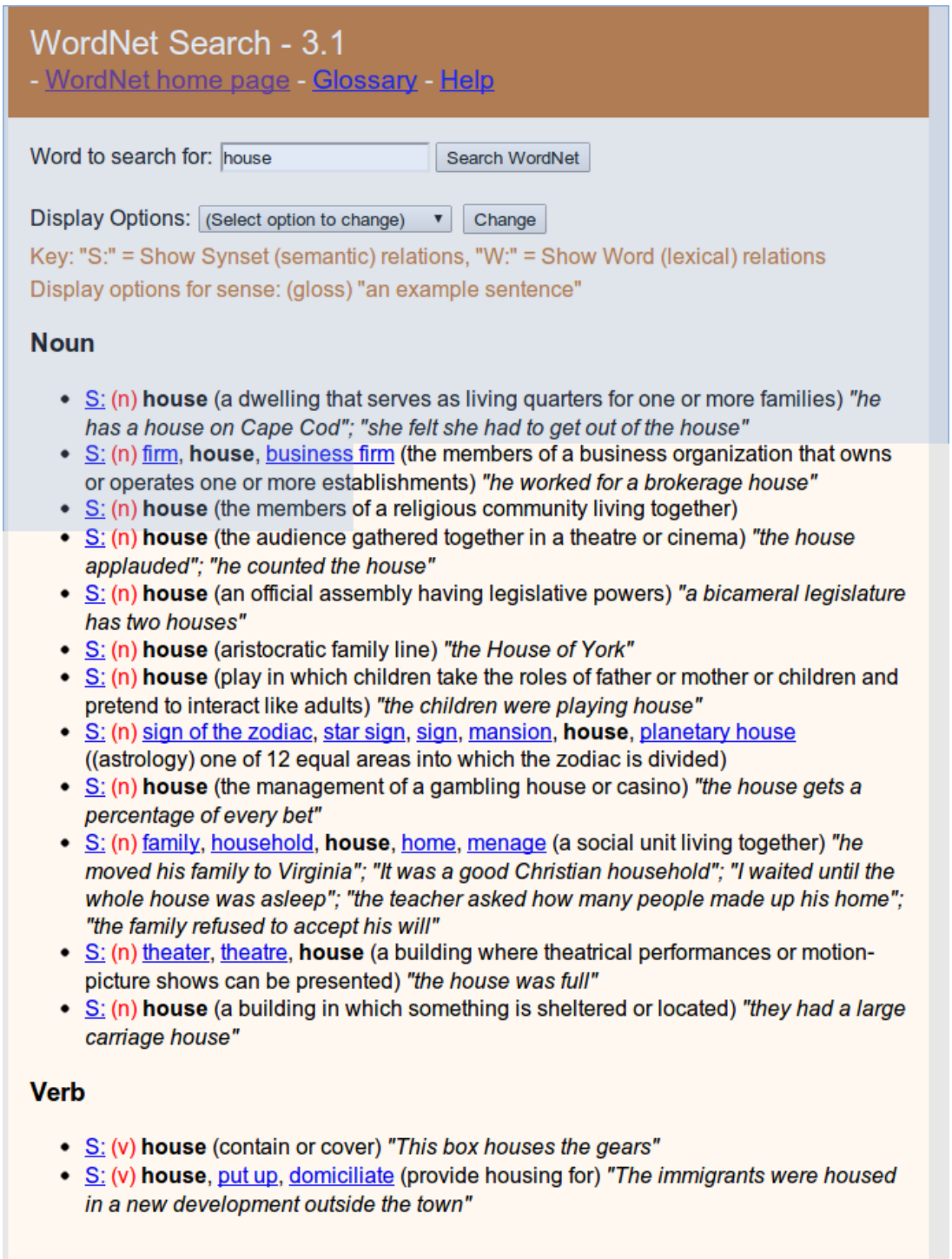


Figura 2.3: Screenshot do conteúdo do recurso WordNet para a palavra "*house*".

Categorias	Unidade Léxica	*synset*
Verbos	11.000	4.000
Substantivos	17.000	8.000
Adjetivos	15.000	6.000
Advérbios	1.000	500
Total	44.000	18.500

Tabela 2.1: Os números da WordNet.BR

Existe um esforço para a criação de uma versão da WordNet em português. Temos dois projetos que buscam esse objetivo. O primeiro é o WordNet.Pt [Marrafa et al. 2005] cujo foco é o português de Portugal. O outro projeto é o WordNet.Br [Silva, Felippo e Hasegawa 2006] cujo foco é o português do Brasil. A WordNet.Br contem atualmente na sua base 44.000 palavras que estão organizadas em 18.000 *synset*. A tabela 2.1, retirada de [Silva, Felippo e Hasegawa 2006] mostra como estão classificadas as palavras da WordNet.Br [Silva, Felippo e Hasegawa 2006, Paiva, Rademaker e Melo 2012].

2.3.2 FrameNet

A FrameNet [Baker, Fillmore e Lowe 1998] é uma base semântica desenvolvida na Universidade de *Berkeley* que foi baseada na teoria dos frames proposta por Minsky [Minsky 1974]. Um frame é uma estrutura hierárquica conceitual que define uma situação, objeto ou evento por meio de seus participantes e relacionamentos. Na figura 2.4 temos um *screenshot* do frame *Commiting_Crime* que apresenta uma parte do frame que relaciona um "criminoso" e um "crime" por meio dos verbos "cometer" ou "perpetrar".

O recurso FrameNet, fornece uma nova perspectiva para as bases léxico-semânticas. O significado de palavras ou unidades léxicas é dado no contexto das situações em que podem participar (frames) por meio dos papéis que podem assumir. A FrameNet foi construída a partir de um corpus e atualmente possui 11.836 frames.

Recentemente, foi iniciado no Brasil o projeto FrameNet Brasil [Salomão 2009]. Esse projeto tem como objetivo criar uma base semântica no idioma português brasileiro utilizando a teoria dos frames. Atualmente a versão da FrameNet.Br[8] contém um total de 32 frames e 38 unidades lexicais [Minghelli, Bertoldi e Chishman 2013].

Diferentemente da WordNet, ao invés de documentar os sentidos de palavras, a FrameNet abstrai aspectos gerais de entidades conceituais enquanto participantes de cenários ou frames situacionais. O ponto forte da WordNet são as relações entre *synsets*, enquanto que na FrameNet enfatiza a relação entre palavras e eventos.

[8] http://www.framenetbr.ufjf.br/, acessado em 31/10/2012.

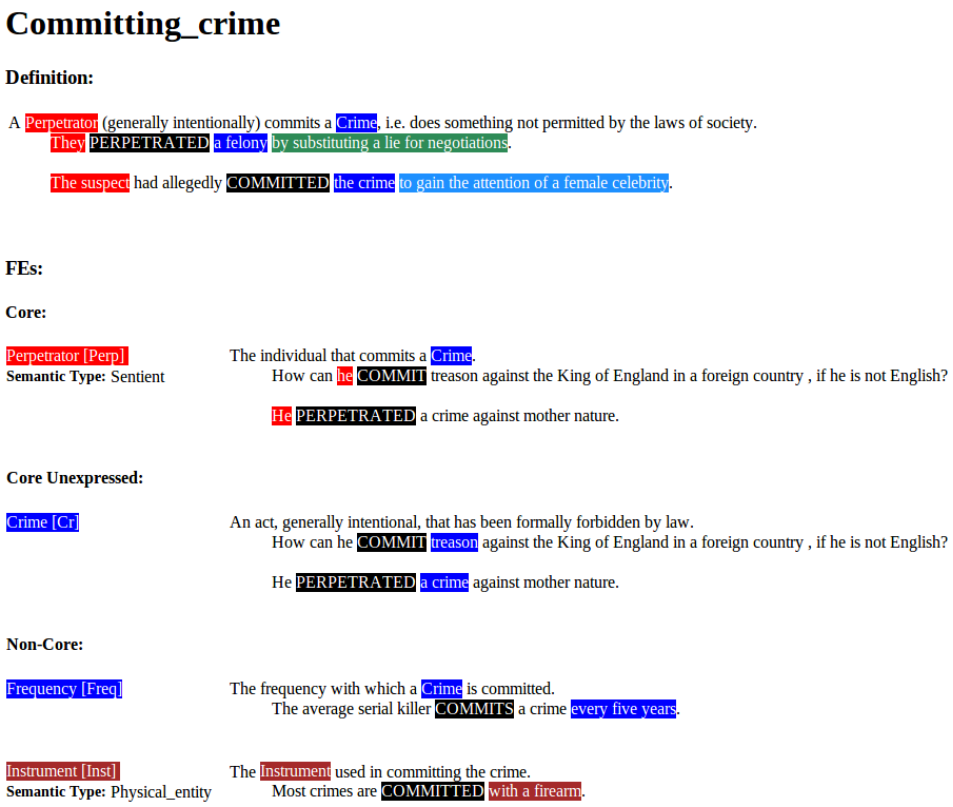


Figura 2.4: Screenshot do frame *Commiting_Crime* do site do recurso FrameNet.

2.3.3 ConceptNet

A ConceptNet [Liu e Singh 2004, Havasi, Speer e Alonso 2007, Speer e Havasi 2012] é uma rede semântica que representa o conhecimento de senso comum coletado através do projeto Open Mind Common Sense (OMCS) [Singh et al. 2002], um projeto com o objetivo de coletar de colaboradores voluntários, pela Internet, sentenças que expressam fatos da vida comum.

Conhecimento de senso comum consiste em fatos e conhecimentos espaciais, físicos, sociais, temporais e psicológicos, possuídos pela maioria das pessoas, os quais são frutos da experiência da vida diária [Anacleto et al. 2007]. Por exemplo quando alguém fala "*Eu comprei doces*", está implícito que usou dinheiro; que o efeito de cair de uma moto é você se machucar; que objetos rolam de superfícies inclinadas.

ConceptNet foi construída pela equipe do MediaLab do *Massachusetts Institute of Technology (MIT)*. A base tem como característica representar fatos que as pessoas sabem sobre o mundo. Na figura 2.5, temos um exemplo de consulta para o conceito *saxophone* e algumas relações de senso comum deste conceito *saxophone* como: *(saxophone, UsedFor, jazz)*, *(saxophone, IsA, wind instrument)*, etc.

A primeira versão da ConceptNet [Liu e Singh 2004], a ConceptNet 2, foi inicialmente construída na língua inglesa. Pouco tempo depois do lançamento da ConceptNet 2 surgiu no Brasil um projeto parceiro, o *Open Mind Common Sense* - Brasil (OMCS-Br) [Anacleto et

Figura 2.5: Screenshot do site do recurso ConceptNet.

al. 2006], que é um projeto do Laboratório de Interação Avançada (LIA) da UFSCar. Este projeto tem como objetivo a coleta de conhecimento de senso-comum em português, está em andamento e conta, atualmente, com 160.000 afirmações de senso comum de seus colaboradores.

A versão 3 da ConceptNet [Havasi, Speer e Alonso 2007] teve como principal diferença a construção da base em um banco de dados SQL e a expressão de senso comum de relações de senso comum de natureza negativa. Por exemplo, a relação *"Os cães não podem voar"* é uma relação de natureza negativa. Nessa versão, houve edições separadas da ConceptNet 3 para o inglês e para o português. A ConceptNet 4 é bastante semelhante ao ConceptNet 3, mas foi revisada e normatizada para que pudesse conter todas as línguas simultaneamente, de modo que, finalmente, poderia representar todo o conhecimento dos OMCS em um só lugar. A base incorporou contribuições de outros projetos, incluindo jogos on-line [Ahn, Kedia e Blum 2006], coletando conhecimentos em inglês, chinês e japonês. Para auxiliar a sua utilização como recurso em outros projetos, também foram adicionadas uma Web API para acessar e consultar os dados.

A ConceptNet 5 [Speer e Havasi 2012] veio sanar estes problemas através de diversas melhorias representacionais. O foco principal é fazer o armazenamento e consulta de conhecimento de senso comum de uma forma verdadeiramente distribuível, ou seja, crescer livremente e absorver o conhecimento de várias fontes, com contribuições de projetos diferentes. As novidades na ConceptNet 5 incluem a assimilação de conhecimento de senso comum a partir do conhecimento adquirido em outras fontes, por exemplo, Wikipédia e Wikidicionário, DBPedia [Auer et al. 2007], Freebase [Bollacker et al. 2008], e WordNet [Miller 1995]. Além de programas de extração de conhecimento a partir de textos como o Reverb [Fader, Soderland e Etzioni 2011], que extrai conhecimento relacional de páginas da Web.

2.3.4 InferenceNet

A base InferenceNet [Pinheiro et al. 2010, Pinheiro et al. 2012] foi construída a partir da ConceptNet, com a agregação de novas relações semânticas de senso comum e inferencialistas sobre conceitos e sentenças. InferenceNet é uma base bilíngue - em língua portuguesa e língua inglesa. As bases semânticas do recurso InferenceNet foram construídas de acordo

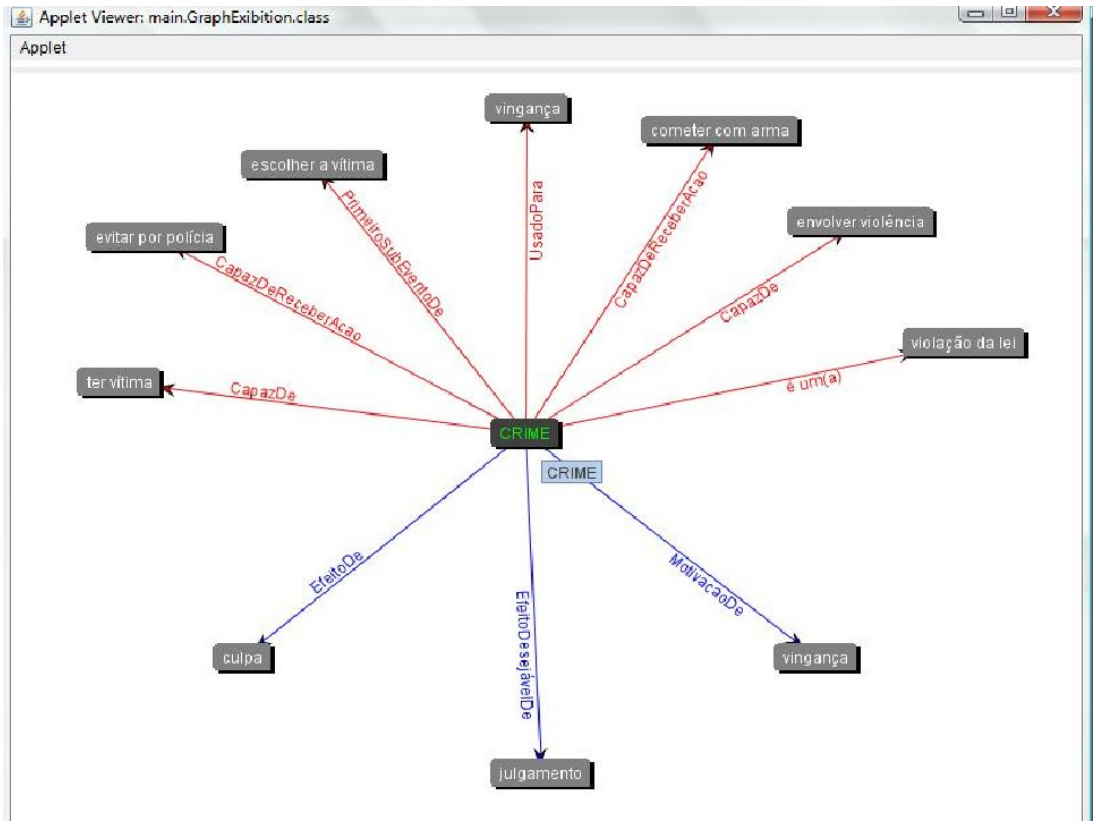


Figura 2.6: Visão parcial d grafo da rede inferencial do conceito “crime” no recurso InferenceNet.

com o Modelo Semântico Inferencialista (SIM) [Pinheiro] e expressam o caráter pragmático da língua natural através de pré condições e pós-condições do uso de conceitos e sentenças. As pre condições são relações que dão a alguém o direito de usar o conceito e o que poderia excluir tal direito, servindo de premissas para enunciados e raciocínios. As pós-condições são relações que permitem saber com o que alguém se compromete ao usar um conceito, servindo de conclusões do enunciados em si e de premissas para futuros enunciados e raciocínios. Na figura 2.6 temos o conceito “crime” tem as seguintes pré condições: *capazDe(crime, tervitima), éUma(crime, violação da lei), etc.* e as pós-condições: *efeitoDe(crime, culpa), motivaçãoDe(crime, vingança), etc.*. Em meados de 2012, Franco *et al*(2012), propôs uma *Description Logics (DLs)* ou Logica Descritiva [Baader et al. 2003] para o recurso.

Uma premissa do projeto InferenceNet era construir as bases de conceitos e de sentenças-padrão em larga escala, ou seja, com uma quantidade representativa de elementos que pudessem servir de insumo para aplicações reais. Por isso já na sua versão inicial, a base continha uma quantidade substancial de elementos nas suas bases. Na tabela 2.2 demostramos o atual quantitativo das bases do InferenceNet.

Na versão atual do InferenceNet, foram incluídas 37 relações. Dentre elas temos 9 relações negativas. Foram adquiridas cerca de 46535 conceitos e 77135 relações em português da ConceptNet 5.0.

Base Conceitual	**InferenceNet**
Conceitos	228764
Relações inferenciais entre conceitos	763121
Base de Sentenças-Padrão	
Sentenças-padrão	5.910
Relações inferenciais de sentenças-padrão	1.432

Tabela 2.2: Quantitativo do InferenceNet

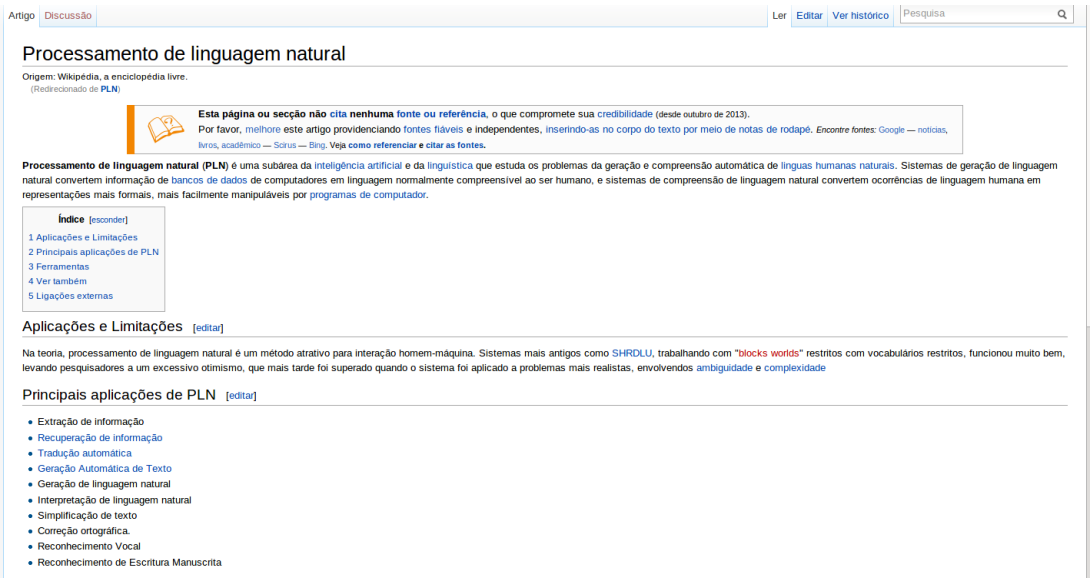


Figura 2.7: Screenshot do site da Wikipédia.

2.3.5 Wikipédia

Apesar de não ser uma base semântica propriamente dita, a Wikipédia, possui um conjunto de características, como definição de um grande números de artigos, organização dos artigos em categorias, etc. que faz um importante recurso léxico-semântico. A Wikipédia é uma enciclopédia multilíngue, colaborativa, sem fins lucrativos. Foi iniciada em 15 de janeiro de 2001. A versão em língua inglesa tem cerca de 4.069.555 artigos[9]. A versão em português foi disponibilizada no final do ano 2001 e tem cerca de 756.819 artigos. Na figura 2.7 tem um *screenshot* do site. A Wikipédia em português é considerada a décima maior edição em número de artigos[10]. Vale a pena ressaltar que a Wikipédia em português não é uma tradução da versão na língua inglesa [Xavier e Lima 2011]. Atualmente a Wikipédia tem cerca de 260 edições ativas[11].

O processo de edição da Wikipédia é colaborativo, único e aberto. Esse processo é todo desenvolvido em cima do conceito de Wiki [Leuf e Cunningham 2001], ou seja, qualquer pessoa que quiser tornar o conhecimento disponível para o público pode contribuir com um

[9] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[10] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[11] http://pt.wikipedia.org/wiki/Wikip\%C3\%A9dia:Wikip\%C3\%A9dia_em_outras_l\%C3\%ADnguas#Lista_das_edi.C3.A7.C3.B5es_da_Wikip.C3.A9dia

artigo além de ser capaz de corrigir os erros, aumentar o seu âmbito, ou compensar viés. Esse sistema é auto-organizado, mas controlado por políticas e guias de edição, que são algumas das estratégias criadas para evitar que sejam incluídas aberrações nos artigos.

Na Wikipédia, há um sistema de avaliação na qual o leitor do artigo pode avaliar a qualidade do artigo escrito. Já foi mostrado que a Wikipédia possui qualidade de conteúdo comparável a enciclopédias tradicionais e que atos de vandalismo são revertidos em minutos [Kittur e Kraut 2008].

Um artigo na Wikipédia é uma matéria sobre determinado assunto. Uma característica peculiar dos artigos da Wikipédia é que na primeira frase já é definido o seu tipo e sua entidade. Por exemplo, no artigo "Brasil" a primeira frase é "*oficialmente República Federativa do Brasil, é o maior país da América do Sul e o quinto maior do mundo em área territorial (equivalente a 47% do território sul-americano) e população, com mais de 192 milhões de habitantes.*". Nesta frase já é definido que o artigo se refere a um país (categoria), que ele pertence à América do Sul (categoria), dentre outras informações referentes a seu tipo e categoria.Neste trabalho de pesquisa, consideramos que um artigo é equivalente a um conceito ou instância de conceitos (no caso de nomes próprios).

A Wikipédia está estruturada com os seguintes elementos: artigos, página de redirecionamento, páginas de desambiguação, hyperlinks, estrutura de categorias, infoboxes e interwikis.

As páginas de redirecionamento são páginas contendo apenas texto em forma de diretiva, ou seja, o texto será exibido sem flexão de gênero, número e grau. O principal objetivo é encontrar um único artigo, para os termos equivalentes. Por exemplo, se o usuário pesquisar "casas", a página de redirecionamento vai remetê-lo para o artigo "casa".

As páginas de desambiguação contêm os possíveis significados (artigos) de um mesmo termo. Apresentando uma breve descrição do termo. Por exemplo, para a palavra "manga" a página de desambiguação mostra links para artigos de "manga", a fruta e "manga", a parte da roupa. Assim o usuário pode escolher qual artigo quer ler.

O hyperlink é uma ligação entre artigos através de links contidos no texto. Os links apresentam indícios de relações semânticas entre conceitos e demostram uma semiestrutura semântica entre eles.

As páginas de categorias dos artigos da Wikipédia apresentam os artigos organizados em uma estrutura hierárquica. Os artigos estão categorizados em super categorias ou subcategorias. A estrutura de categorias não é uma árvore portanto algumas categorias apresentam múltiplas supercategorias e um artigo pode pertencer a várias categorias. Assim como os links, as categorias demostram uma estrutura semântica entre os artigos da Wikipédia.

O infobox é uma tabela com informações básicas sobre a entidade descrita no artigo. Por exemplo, o infobox de uma cidade deve trazer informações como: população, extensão geográfica, localização, etc.

O último elemento da Wikipédia do qual falaremos são o interwikis. Os interwikis

são links que permitem a navegação entre artigos de línguas diferentes. Por exemplo na url do artigo temos `http://xx.wikipedia.org/wiki/Neymar` o xx pode representar o artigo sobre Neymar em outras línguas como es, en e pt.

Com o tempo, dentro da área de PLN, a Wikipédia se mostrou bem mais do que uma enciclopédia. Viu-se que ela poderia dar suporte a outras tarefas como: ser um corpus linguístico (por ter um grande volume de texto e parcialmente anotado através de tags da própria Wikipédia), ser um grande banco de dados (tem um grande volume de informações estruturadas e a existência de projetos que tentam disponibilizar os seus dados como, por exemplo, a DBPedia), ser uma grande rede semântica de conceitos (a estrutura de hyperlinks permite saber como os artigos estão ligados), entre outras tarefas.

Nesse sentido existe um esforço da própria comunidade em disponibilizar a informação estruturada na Wikipédia. A DBPedia tem 3.64 milhões de conceitos que estão relacionados em 1 bilhão de tuplas em RDF[12] em 97 línguas.

O principal objetivo da DBPedia é facilitar o acesso a incrível quantidade e variedade de informações na Wikipédia seja usada de maneiras novas e interessantes, e que inspirem novos mecanismos de navegação, interligação e melhoramento da própria enciclopédia.

2.3.6 Outras bases semânticas

Em meados de 2007, *Suchanek et al.* [Suchanek, Kasneci e Weikum 2007] propôs a ontologia conhecida como YAGO. YAGO foi construída utilizando a WordNet e a Wikipédia. A WordNet fornece grande quantidade de entidades e Wikipédia fornece uma ótima taxonomia. YAGO atualmente tem cerca de 1 milhão de conceitos e 5 milhões de relações. A acurácia de suas relações, que foi medida através de uma avaliação manual, mostrou 95% de acurácia.

Um exemplo menos conhecido é a FreeBase [Bollacker et al. 2008]. Essa base é um banco de dados em forma de grafo e escalável usado para estruturar o conhecimento humano. Os seus dados estão divididos por domínios como musica, livros, filmes, etc. que podem ser criados de forma colaborativa, estruturado e mantido por pessoas e softwares. Atualmente a FreeBase[13] contém cerca de 1.8 milhões de conceitos e 39 milhões de relações

Outro exemplo de base léxico-semântica é a VerbNet [Kipper, Dang e Palmer 2000]. Essa base consiste em um conjunto de verbos separados em classes. A sua principal motivação é fornecer a ligação entre sintaxe e semântica dos verbos. A versão inglesa da VerbNet está na versão 3.0 tem atualmente 3769 verbos divididos em 274 classes.

2.3.7 Análise Comparativa

A tabela 2.3 mostra uma comparação entre os principais recursos semântico-linguísticos estudados nesse trabalho. A tabela também evidencia que tipo de conteúdo semântico é ex-

[12] `http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/`
[13] http://www.freebase.com/

Base Semântica	**Conhecimento**	**Tipos de relações semânticas**	**Como foram construídas?**	**Tamanho das bases**	**Língua Principal**
WordNet	Taxonômico	Sinonímia, antonímia, hiponímia/hiperonímia, meronímia/holonímia, similaridade, *entailment* , causal	Elemento por elemento	180.000 conceitos	EN
FrameNet	Relações entre frames	Específicas por frame	Forte dependência de corpus	11.836 conceitos	EN
ConceptNet	Senso comum	Coisas, espacial, causal, afetiva, funcional, agente, eventual	De forma colaborativa na Web	11.401.322 de conceitos	EN
Wikipédia	Senso comum não estruturado	Categorias e links entre os artigos	De forma colaborativa na Web	4.069.555 de artigos	EN
InferenceNet	Senso comum e Inferencialista	Coisas, espacial, causal, afetiva, funcional, agente, eventual	De forma colaborativa na Web	228.764 de conceitos	PT

Tabela 2.3: Comparação entre os recursos linguísticos.

presso em cada base, como elas foram construídas e qual é a língua principal. Com esta visão, percebe-se que a maior parte do conteúdo que alimenta as bases é adquirido de forma colaborativa e língua inglesa domina grande parte dos recursos. Daí a importância de métodos de aquisição de conhecimento na língua portuguesa.

3 AQUISIÇÃO DE CONHECIMENTO DE MUNDO: O ESTADO DA ARTE

Este capítulo visa apresentar o estado da arte em abordagens, métodos e ferramentas para aquisição de conhecimento de mundo, às quais são a base para sistemas de NLU. Segundo [Ovchinnikova 2012], conhecimento de mundo é o conhecimento que é compartilhado por seres humanos pertencentes a uma mesma comunidade linguística e cultural, não considerando aspectos situacionais e individuais de discurso. Dentro do conhecimento de mundo temos vários tipos de conhecimento, um bem popular é conhecimento de senso comum. Minsky, em seu trabalho "*The Emotion Machine*" [Minsky 2007], define este conhecimento como um conjunto de fatos, conceitos e habilidades comuns à maioria das pessoas e aplicações em tarefas cotidianas como, por exemplo, levar o cachorro para passear, ação que não necessita de nenhum conhecimento científico para sua realização. De acordo com [Anacleto et al. 2007], senso comum consiste em fatos e conhecimentos espaciais, físicos, sociais, temporais e psicológicos possuídos pela maioria das pessoas, os quais são frutos da experiência da vida diária. Como este trabalho está focado na aquisição de conhecimento, escolhemos na literatura o que melhor se aplica a esta tarefa e domínio. Para efeito didático, o capítulo está dividido em 1) Métodos de AC semiautomáticos e 2) Métodos de AC automáticos.

3.1 Métodos Semiautomáticos de Aquisição de Conhecimento de Mundo

Os métodos semiautomáticos têm como principal característica a necessidade de interferência humana no processo de aquisição. A consequência principal dessa interferência é a diminuição da taxa de erros no processo de AC, já que o conteúdo é validado antes de ser adquirido. Os métodos semiautomáticos são bastante utilizados na aquisição de conhecimento de mundo [Lenat 1995, Witbrock et al. 2003, Singh et al. 2002, Speer et al. 2009].

O projeto CYC [Lenat 1995] foi o primeiro esforço para construção de um banco de dados de conhecimento de senso comum. Inicialmente foi criada uma base de dados por um grupo de especialistas pagos. Nos primeiros anos do projeto, o CYC já tinha 1,6 milhão de regras e 180.000 conceitos. Esse foi o passo inicial, porém mais conceitos e relações foram sendo coletados através da combinação de conceitos/relações já existentes. Um problema dessa abordagem é que o número de conceitos/relações depende de colaboradores especialistas. De acordo com [Zang et al. 2013], atualmente a base tem cerca de 500.000 termos e 7.000.000 de relações entre esses termos. Para melhoria do processo de AC, [Witbrock et al. 2003] propôs um sistema de AC que se baseia na construção de um diálogo entre usuário e sistema. Antes da inclusão de um novo conceito, o usuário escolhe um conceito similar que pertence à base CYC e, de forma interativa, pode aceitar ou rejeitar um conjunto de afirmações do conceito similar, essas respostas serão copiadas para o novo conceito. Por exemplo, se o conceito que o usuário deseja incluir é "computador" e se há o conceito "notebook" na base CYC, o usuário pode selecionar "notebook" e ser guiado através de um processo interativo de perguntas e respostas que visam a aquisição de relações de senso comum para "computador " com base no que já é conhecido sobre "notebook". [Witbrock et al. 2003] não informa como este processo de AC foi avaliado.

Em meados de 2000, foi lançado o projeto *Open Mind Common Sense* (OCMS) [Singh et al. 2002] com o objetivo de coletar, a partir da Internet e de colaboradores voluntários, relações que expressam fatos da vida comum. O OCMS dá suporte a várias línguas incluindo Inglês, Chinês, Português, Alemão, etc. O corpus OMCS deu origem à base de conhecimento de senso comum conhecida como ConceptNet [Havasi, Speer e Alonso 2007]. O processo de inclusão de relações semânticas do OCMS passou por algumas melhorias para deixar o processo de aquisição de relações menos oneroso. Em 2007, o OMCS [Speer 2007] já fornecia funcionalidades que ajudavam o usuário a refinar e validar o conhecimento coletado. Essa versão trazia como novidade a expansão do projeto para outras línguas e relações de senso comum de natureza negativa como, por exemplo, a relação "os cães não podem voar". Em [Singh et al. 2002] foi descrita a avaliação das relações adquiridas da seguinte forma: sete especialistas avaliaram cerca de 3.000 relações que foram incluídas na Web seguindo os critérios "verdade", "generalidade", "sentido" e "neutralidade. Os resultados observados foram os seguintes: 75% de corretude das sentenças, 82% foram consideras neutras e 85% faziam sentido.

Estratégias utilizadas para aquisição de conhecimento com a utilização de jogos estão cada vez mais sendo propostas. A principal vantagem de usar um jogo no processo de aquisição de conhecimento é tornar o processo lúdico e menos oneroso. Nessa perspectiva, [Speer et al. 2009] propuseram um jogo interativo chamado "20 Questions", cujo objetivo principal é motivar contribuições voluntárias para o projeto OMCS, aumentando a taxa de aquisição de novos conceitos. O jogo utiliza um modelo de *cluster* hierárquico para definir um conjunto de 20 perguntas que serão utilizadas para motivar o usuário a incluir relações sobre um conceito. Métodos de clusterização visam agrupar objetos com alguma similaridade [Tryon e Bailey 1970, Fraley e Raftery 1998]. Exemplificamos a seguir o método para a aquisição de conteúdo para o conceito "maçã".

- É um exemplo de lugar? Resposta: Não
- É um exemplo de comida? R: Sim
- Pode ser encontrado em uma loja? R: Sim ...

Com base nas respostas a estas perguntas, o algoritmo de agrupamento pode definir se o novo conceito "maçã" pertence ao mesmo grupo de conceitos como "queijo", "pão","carne", etc. A avaliação desse método levou em conta apenas aspectos como ludicidade, intuitividade e performance. Em média, 80% dos usuários avaliaram que o jogo "20 Questions" é mais divertido do que o método tradicional, porém 56% dos usuários não o considerou intuitivo. Além disso, foi observado que usuários que utilizam o jogo realizam a tarefa em metade do tempo de usuários que não o utilizam. Não houve qualquer avaliação sobre a qualidade do conteúdo adquirido através do método proposto.

O projeto Verbosity [Ahn, Kedia e Blum 2006] baseia-se em jogo interativo para AC de senso comum, cuja ideia principal é transformar o processo de AC de senso comum em algo divertido e interessante. Verbosity consiste em um jogo de adivinhação, onde duas pessoas desempenham os papéis de narrador e adivinhador. O narrador escolhe uma palavra e dá dicas para o adivinhador descobrir o conceito relacionado. Essas dicas são formuladas utilizando um

modelo com um conjunto de relações pré-determinadas, por exemplo: *é um*, *tipo de*, *é sobre*, *é o oposto de*, *é utilizada*, *está dentro*, etc. No final do processo, se o adivinhador for capaz de descobrir o conceito que o narrador escolheu, o conjunto das relações sobre o conceito é adquirido para uma base de senso comum. Experimentos indicaram que a média de inclusões de relações foi de 29,58 relações por pessoa e o tempo médio de uso foi de 23,58 minutos. Outro experimento realizado foi para avaliar a qualidade do conteúdo inserido através do jogo. A metodologia utilizada foi a seguinte: foram escolhidas aleatoriamente 200 relações incluídas através do jogo, e depois essas relações foram avaliadas por especialistas. O resultado foi uma acurácia de 85%.

Na tabela 3.1 temos uma comparação entre as principais estratégias de aquisição de conhecimento existentes. O CYC tem como principal vantagem a geração de conteúdo a partir de interações de especialistas com o sistema, porém um grande número de interações pode deixar o processo cansativo e oneroso. O OMCS tem como principal característica utilizar um conjunto de questionários para fazer aquisição. Sua principal vantagem é a flexibilidade - rapidamente o método pode ser importado para outras línguas - e a desvantagem é deixar o processo lento e oneroso. Utilizar um jogo para AC de senso comum pode deixar a aquisição mais divertida e lúdica, porém a necessidade de duas pessoas para utilizar o Verbosity pode atrapalhar a aquisição. Outro ferramenta mencionada foi o jogo "20 Questions", que traz como principal vantagem a utilização de um método de clusterização no processo de inclusão de novas sentenças. No entanto, não se mostrou intuitivo para os usuários que o utilizaram.

Método de AC semiautomático	**Estratégia utilizada para AC**	**Base Semântica ou Corpus Gerado**	**Avaliação Realizada**
Projeto CYC	Interação com usuário	CYC/Open CYC	Avaliação extrínseca.
OMCS	Questionários	ConceptNet	75% de acurácia
20 Questions	Perguntas e Respostas	ConceptNet	80% mais divertido
Verbosity	Jogo	OMCS	85% de acurácia

Tabela 3.1: Comparação entre as principais estratégias de aquisição semiautomática de conhecimento.

3.2 Métodos Automáticos de Aquisição de Conhecimento de Mundo

Inicialmente, explanamos os sistemas e abordagens que se aplicam a qualquer fonte de conhecimento textual.

Sistemas Abertos de Extração de Informação (*Open IE Systems*), é um paradigma de extração de informação em que não existe nenhuma restrição do tipo de informação a ser extraída [Li et al. 2011], ou seja, os sistemas *Open IE* utilizam padrões gerais para extrair todas as possíveis relações entre entidades. Esse paradigma de sistema apresenta algumas vantagens como: a sua boa precisão (na literatura temos cerca de 86% [Fader, Soderland e Etzioni 2011]) e à generalidade de relações que podem ser extraídas.

Reverb [Fader, Soderland e Etzioni 2011] - é um programa que identifica e extrai relações binárias de sentenças em inglês automaticamente de textos e documentos na Web. Reverb foi utilizado para extrair relações semânticas para a base ConceptNet. [Speer e Havasi 2012] mostraram que apenas de 20% das relações extraídas com Reverb foram incorporadas a base devido a má qualidade das extrações. Para realizar a extração de relações binárias, o sistema implementa duas restrições: a primeira é a restrição sintática, que consiste em conjunto de expressões regulares - por exemplo, $V^1 \mid V\ P^2 \mid V\ W^3\ P$ - para reconhecer relações e modificações morfológicas, tais como a conjugação de verbos no infinitivo; a segunda restrição é a léxica, que destina-se a descartar relações mal formadas ou complexas através da contagem de ocorrências da relação no corpus e se a relação não tiver um número mínimo de ocorrências no corpus será descartada. Por exemplo, na frase "O governo Obama está oferecendo metas modestas para a redução dos gases de efeito estufa na conferência", o Reverb extrairia a relação "X está oferecendo metas modestas para a redução de gases de efeito estufa em Y" com os argumentos X = "Obama" e Y = "conferência". Essa relação não atende às restrições lexicais porque a relação é muito específica.

O Reverb funciona da seguinte forma: a partir de uma sentença de entrada como *Hudson was born in Hampstead, which is a suburb of London*, o primeiro passo é verificar se as relações satisfazem as restrições sintáticas e léxicas (nesse exemplo *was, born in e is a suburb of* satisfazem ás restrições); o próximo passo é encontrar os argumentos. Usando o mesmo exemplo, os argumentos encontrados são *(Hudson, Hampstead)*) e, por fim, temos as seguintes relações *(Hudson, was born in, Hampstead)* e *(Hampstead, is a suburb of, London)*.

A aplicação livre de padrões sintáticos pelo Reverb faz com que seu nível de precisão seja baixo. Para melhorar este resultado, o Reverb adota nove *features ad hoc* para definir uma função de confiança aprendida por um classificador linear. As *features ad hoc* são número de palavras nos argumentos, se os nomes próprios são argumentos, "para" como última preposição na relação, "sobre" como última preposição na relação, "de" como última preposição na relação, a relação tem no máximo 10 palavras, não existir verbos nos argumentos, a relação não pode ser uma pergunta e ter verbos nas relações.

O modelo de AC do Reverb é específico para o idioma inglês. Para avaliar este sistema, foram escolhidas aleatoriamente 500 relações extraídas pelo Reverb a partir de textos na Web para revisão por dois avaliadores humanos. Como resultado, 86% das relações extraídas pelo Reverb foram corroboradas pelos avaliadores [Fader, Soderland e Etzioni 2011].

Argumentamos que as desvantagens do Reverb são a dependência de expressões regulares pré-definidas, a necessidade de análise da aplicação de *features* adequadas para outras línguas (como o Português) e a não identificação de tipos de relações redundantes. Além disso, [Speer e Havasi 2012] relata que 65% das extrações incorretas feitas pelo ReVerb foram decorrentes da incapacidade de identificar corretamente os argumentos da relação, por exemplo na sentença *"I gave him 15 photographs,"* a relação extraída é *(I,gave, him)*. Isso ocorre devido ao fato de o Reverb sempre tentar extrair relações binárias das sentenças.

[1] V = Verbo/Adverbio

[2] Preposição

[3] Nome, Adjetivo, Determinante, Pronome

DepOE [Gamallo, Garcia e Fernández-Lanza 2012] - sistema que propõe a extração de relações em outras línguas além do inglês para melhoria dos métodos de *Open IE*. Esse sistema faz a extração não supervisionada utilizando um analisador baseado em regras. Inicialmente, o texto é processado por *dependency parser*, depois são extraídas da árvore de dependência, as cláusulas para a identificação e extração das relações. A avaliação do método mostrou cerca de 68% de acurácia contra 52% do Reverb para o mesmo conjunto de textos.

LSOE [Xavier, Lima e Souza 2013] - um método de extração de informação que utiliza padrões léxico-sintáticos aplicados a um texto analisado por um *POS-Tagger*. Os padrões utilizados compõem a estrutura de Qualia. Essa estrutura foi apresentada por [Pustejovsky 1991] em 1991 e tem com a ideia representar os atributos que compõem um objeto, suas partes e sua função. Com isso, foi gerada um conjunto de expressões regulares. Por exemplo, a expressão regular **A NP OF NP IS NP** aplicada na sentença *A central branch of metaphysics is ontology* extrai a relação *(ontology, is a central branch, metaphysics)*. O método teve precisão de 64% contra 49% do Reverb para o mesmo conjunto de textos analisados.

[Cankaya e Moldovan 2009]: os autores propõem um método automático para gerar novas relações de conhecimento de mundo baseadas em regras de Lógica de Primeira Ordem (LPO). O algoritmo proposto procura automaticamente na WordNet[4] conceitos que têm uma determinada propriedade e gera novos axiomas usando regras de conhecimento de mundo. Por exemplo, se "vidro" tem a propriedade de "transparente", e "ver através" é uma característica de "transparente", então podemos concluir que "ver através" também é característica de "vidro". Cerca de 50 axiomas gerados pelo método foram escolhidos aleatoriamente e avaliadores humanos foram convidados a separar os grupos dos axiomas corretos dos que não faziam sentido. Esse experimento obteve 98% de precisão.

Em um segundo grupo de abordagens, destacamos e detalhamos aquelas que exploram o conteúdo textual dos artigos da Wikipédia e os links existentes entre eles, como fonte de informação.

Semantic Wiki [Völkel et al. 2006] - é uma extensão da Wikipédia que explicita semanticamente as ligações entre artigos e seus links. Essa explicitação é feita pelo usuário no momento da edição de um artigo. Por exemplo, no artigo sobre a cidade de Londres temos o seguinte texto: **Londres** é a capital da [[*Inglaterra*]] e do [[*Reido Unido*]][5] e o *Semantic Wiki* anota a relação semântica ao lado do link: **Londres** é a capital da [[*capital of::Inglaterra*]] e [[*capital of::Reido Unido*]].

[Stoutenburg, Kalita e Hawthorne 2009] é proposta a extração de relações entre artigos da Wikipédia utilizando expressões regulares para detectar padrões linguísticos e inferir relações entre links de artigos. São utilizadas seis expressões regulares, como **ARTICLE-NAME.* VBZ DD. *[[(NN)+]]**, para extrair as seguintes sentenças: *isA*, *partOf*, *bornOnDate*, *diedOnDate*, *bornIn* e *locatedIn*, além de uma expressão para extração de relações genéricas. No artigo **Automóvel**, por exemplo, temos o texto "Um *automóvel* é um *veículo* de passageiro que leva o seu próprio *motor*". Serão extraídas relações como *Automóvel Éum veiculo*, *Automó-*

[4] `http://xwn.hlt.utdallas.edu`, acessado em 10 de fevereiro de 2011.

[5] Os termos em negrito são links no artigo da Wikipédia.

vel Tem motor, etc. A definição de expressões regulares para um conjunto restrito de relações obviamente é adequada para finalidades específicas e, embora seja uma estratégia que apresente boa precisão (em torno de 80%), possui baixa cobertura de tipos de relações semânticas.

(WOE) Wikipedia-based Open IE [Wu e Weld 2010] - propõe a extração automática de relações do tipo *(arg1, relação, arg2)* utilizando técnicas de aprendizado supervisionado em um conjunto de exemplos etiquetados manualmente. O algoritmo de extração automática é treinado com exemplos que são construídos através de uma heurística que utiliza os valores contidos nos infoboxes da Wikipédia. Embora não restrinja o conjunto de relações semânticas a serem extraídas, apresenta em torno de 36% de relações pouco informativas ou incoerentes, como na sentença "Se ela é nomeada pelo presidente Bush": a sentença extraída é *(ela,eNomedaPor,Bush)*. Na avaliação realizada, foram extraídas 300 sentenças utilizando o WOE. As sentenças foram avaliadas por dois avaliadores humanos tendo como resultado uma precisão em torno de 73%.

Por fim, detalhamos a seguir as abordagens para extração de conhecimento que usam a estrutura da Wikipédia (infoboxes e a árvore de categorias), como fonte de informação.

DBPedia surgiu de um esforço da comunidade científica para construção de uma base semântica a partir da extração de informação estruturada da Wikipédia [Lehmann et al. 2013]. O seu conteúdo é disponibilizado na Web através do formato *Resource Description Framework (RDF)* [Beckett e McBride 2004, Lehmann et al. 2013]. O método de aquisição de conhecimento utilizado neste projeto realiza a extração a partir do infobox da Wikipédia. Por exemplo, o infobox de um país traz informações como população, extensão geográfica, localização, etc. Na DBPedia o conceito **Brasil** tem a propriedade *PopulatedPlace/areaTotal*, cujo valor é extraído do respectivo infobox.

YAGO [Suchanek, Kasneci e Weikum 2007, Suchanek, Kasneci e Weikum 2008] é uma ontologia que foi construída de forma automática pela combinação das categorias e infoboxes da Wikipédia com as relações taxonômicas da WordNet [Miller 1995, Kohl et al. 1998]. Cada *synset* da WordNet é uma *classe* na base Yago, e a hierarquia entre categorias da Wikipédia corresponde à relação de hiperonímia na WordNet. Por exemplo, na Wikipédia tem-se a categoria "Povo norte-americano no Japão", a qual é associada como *subClassOf* da classe "Pessoa" da WordNet. Atualmente, a base YAGO contém mais de 10 milhões de entidades e 120 milhões de relações[6]. Segundo [Zang et al. 2013], foi realizada uma avaliação empírica na base para avaliar a qualidade do conteúdo que mostrou uma acurácia de 95%.

Wikipedia Thesaurus [Nakayama, Hara e Nishio 2007, Pei et al. 2008] - é um dicionário extraído da Wikipedia através da exploração de URLs e estrutura de links. A construção do thesaurus baseia-se na medida de similaridade entre os artigos calculada pelos links que o artigo referencia. Por exemplo, UFO tem 65% de similaridade com Objeto Voador Não-identificado.

Em [Syed e Finin 2010], temos a construção de um método de aquisição de conhecimento a partir dos hyperlinks. A extração é feita utilizando aprendizado não supervisionado

[6] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

sobre os links da Wikipédia. O método extrai conceitos, atributos de conceitos, tipos de conceitos e a hierarquia entre os conceitos. A base da WordNet é utilizada para encontrar sinônimos e assim aumentar a eficácia do algoritmo proposto. Em [Fogarolli 2011], mostra-se que a Wikipédia pode ser usada como uma fonte de conhecimento para a criação de diversas aplicações semânticas e descreve uma aplicação automática que extrai conceitos e relações através dos links contidos nos artigos, páginas de desambiguação e páginas de artigos. Outros trabalhos, como os apresentados em [Nastase et al. 2010, Suchanek, Kasneci e Weikum 2008], também fazem a extração de conhecimento utilizando as categorias, infoboxes, artigos e interwikis contidos na Wikipédia. [Xavier e Lima 2012] utiliza as categorias da Wikipédia para extração de novos conceitos e os links entre conceitos que fazem parte da mesma categoria ou subcategoria para extração de relações. Vale ressaltar que as relações extraídas nesse trabalho são genéricas do tipo "*relacionado com a*". Para avaliar esse método foram selecionados aleatoriamente 50 relações, e um juiz teve que avaliar a qualidade dessas relações. O experimento obteve 77% de precisão.

3.3 Análise Comparativa

A tabela 3.2, resume as principais características das abordagens automáticas para AC de mundo. O primeiro grupo são os trabalhos que utilizam infobox e arvores de categorias no processo de AC, a principal vantagem desses métodos é a qualidade do conteúdo extraído e sua desvantagem é a grande dependência da estrutura da Wikipédia, com a utilização dos infoboxes e das categorias, para a extração de relações. Nos métodos que utilizam links ou texto da Wikipédia e nos *Open IE Systems* temos como característica principal a utilização de expressões regulares e aprendizado de máquina no processo de extração. A principal vantagem destes métodos é grande número e diversidade de relações extraídas. A principal desvantagem de utilizar expressões regulares é a dificuldade de externar todas as relações de uma língua através de um conjunto de expressões regulares. No caso do aprendizado da máquina, existe uma dificuldade na extração dos argumentos de uma relação. Isso acontece devido à dificuldade de encontrar um corpus grande o suficiente para treinar os algoritmos.

Método de AC automático	**Estratégia utilizada para AC**	**Base Semântica ou Corpus Gerado**	**Avaliação Realizada**
Métodos que utilizam infobox e arvores de categorias da Wikipédia			
DBPedia	Infobox da Wikipédia	DBPedia	Avaliação extrínseca.
YAGO	Combinação das categorias e infoboxes da Wikipédia com as relações taxonômicas da WordNet.	YAGO	95% de acurácia
[Xavier and de Lima 2012]	Categorias da Wikipédia para a extração de novos conceitos; Relações genéricas do tipo "relacionado a".	Corpus próprio	77% de relações corretas
Métodos que utilizam links ou texto da Wikipédia			
Semantic Wiki	Extensão da Wikipédia que descreve semanticamente as ligações entre artigos e seus links.	Wikipédia	Avaliação extrínseca
[Stoutenburget al. 2009]	Extração de relações entre links de artigos utilizando expressões regulares.	Corpus Próprio	70% de precisão
Wikipedia-based Open IE (WOE)	Aprendizado supervisionado em um conjunto de exemplos etiquetados manualmente.	Wikipédia	73% de precisão
Open IE Systems			
ReVerb	Expressões regulares pré-definidas e anotação de corpus.	ConceptNet	86% de precisão
LSOE	Padrões léxico-sintáticos em textos analisados sintaticamente.	Corpus Próprio	54% de precisão
DepOE	Regras	Corpus Próprio	68% de precisão

Tabela 3.2: Comparação entre as principais estratégias de aquisição automática.

4 AQUISIÇÃO DE CONHECIMENTO DE MUNDO: PROPOSTAS DE SOLUÇÃO

4.1 Introdução

O maior desafio na aquisição de conhecimento de mundo é conseguir um processo evolutivo, sistemático e com a dinamicidade que as aplicações exigem, além de favorecer uma aquisição de conhecimento consistente, correta e completa. Os sistemas de AC vistos na literatura apresentam diversos problemas, conforme destacado no Capítulo 3:

- Semiautomáticos - utilizam questionários ou se baseiam nas interações dos usuários com o sistema, o que pode deixar o processo de AC oneroso e gerar inconsistências nas bases;
- Automáticos - dependência de corpus abrangente para treinar os modelos de aprendizado de máquina, dificuldade de encontrar expressões regulares que cobriam as relações semânticas encontradas em uma língua, e dificuldade em descobrir os argumentos de uma relação.

Neste trabalho, propomos dois métodos para aquisição de conhecimento de mundo.

1. Um método semiautomático, cujo diferencial é um módulo de raciocínio sobre conhecimento preexistente que visa oferecer ao usuário conteúdo inicial que o ajude a externar e a validar relações semânticas de novos conceitos [Pinheiro et al. 2013, Pinheiro et al. 2011].
2. Um método automático para aquisição de relações semânticas entre conceitos, a partir de artigos da Wikipédia, que faz uso de um conhecimento implícito existente em sistemas hipermídia: os links entre artigos. O principal diferencial do método proposto é a independência de expressões regulares pré-definidas, o uso de links para definição dos argumentos das relações e a identificação de relações redundantes [Franco et al. 2013].

Neste capítulo apresentamos cada um destes métodos, os protótipos desenvolvidos e a avaliação realizada.

4.2 Aquisição Semiautomática de Conhecimento de Mundo

A nossa primeira hipótese de pesquisa afirma que o conhecimento preexistente em uma base auxilia o usuário a explicitar e a validar relações semânticas para um novo conceito, otimizando o processo interativo de aquisição de conhecimento de mundo. Baseado nessa hipótese, propomos um método semiautomático de aquisição de conhecimento de mundo que consiste em um algoritmo heurístico para geração de relações conceituais a partir da análise sintática de sintagmas nominais. Essa seção está dividida da seguinte forma: inicialmente discorremos sobre sintagma nominais e sua importância para o processo de aquisição, ora proposto.

Em seguida apresentamos o método heurístico para aquisição de conceitos em língua natural. Por fim, apresentamos os resultados de uma avaliação qualitativa com usuários que interagiram com um protótipo de sitema, construído com base no método proposto.

4.2.1 Sintagmas Nominais

De acordo com [Silva e Koch 1989], em qualquer enunciado, os signos linguísticos ligam-se uns aos outros formando grupos. Esses grupos são chamados sintagmas. O sintagma tem um elemento principal, chamado de núcleo, que define a natureza do sintagma. Na língua portuguesa, são definidos os seguintes sintagmas [Lemle 1984]:

- Sintagma Nominal (SN), quando o núcleo do sintagma é um nome ou substantivo;
- Sintagma Adjetival (SAdj), quando o núcleo do sintagma é um adjetivo;
- Sintagma Verbal (SV), quando o núcleo do sintagma é um verbo;
- Sintagma Preposicional (SP), quando o núcleo do sintagma é uma preposição;
- Sintagma Adverbial (SAdv), quando o núcleo do sintagma é um advérbio.

Na análise sintática de uma sentença os sintagmas são identificados e devidamente qualificados. Por exemplo, a figura 4.1 apresenta o resultado da análise sintática da sentença "*Os ladrões oportunistas agiram impunemente durante a greve da Polícia Militar do Ceará*", realizada pelo parser PALAVRAS [Bick 2000]. Foram identificados os seguintes sintagmas, cujos núcleos estão destacados: "*os ladrões oportunistas*" (SN); "*agiram*" (SV); "*impunemente*" (SAdv); "*durante*" (SP); "*a greve*" (SN); "*de*" (PRP); "*a Polícia Militar do Ceará*" (SN).

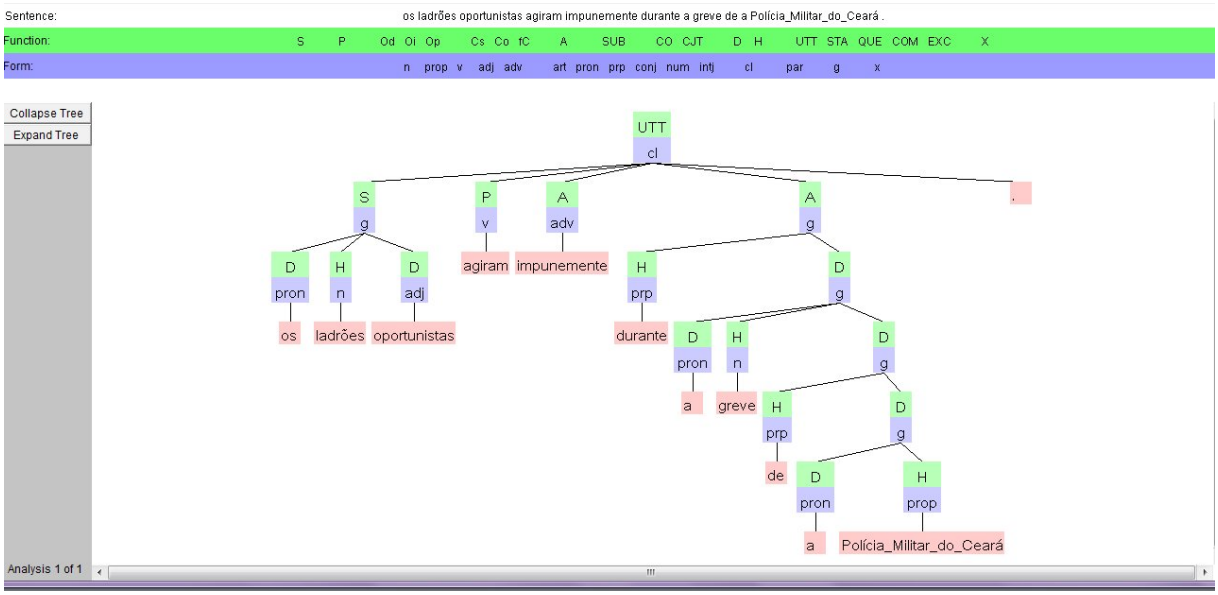


Figura 4.1: Análise sintática da sentença "Os ladrões oportunistas agiram impunemente durante a greve da Polícia Militar do Ceará".

Sintagmas Nominais (SN) são normalmente usados para denominar as "coisas" do mundo e, por isso, são primordialmente usados para nomear conceitos em língua natural. O núcleo do sintagma nominal (SN) pode ser constituído de um substantivo (próprio ou comum) ou um pronome substantivo (pessoal, demonstrativo, indefinido, interrogativo, possessivo ou relativo). Quando o núcleo do SN é um pronome substantivo, este, por si só, representa o SN.

Além do núcleo (Nome-N), o SN pode apresentar determinante(s) (DET) e/ou modificadore(s) (MOD). Os determinantes antecedem o núcleo. Os determinantes do SN são representados pelos artigos, pelos pronomes e pelos numerais. Podemos citar como exemplo de SN com DET + N: ***a luz, o sol, um jornal, certas tardes***.

Os modificadores (MOD) são consequentes ao núcleo e são representados por adjetivos e locuções adjetivas. Eles têm como função caracterizar ou expressar uma avaliação sobre os substantivos. Para caracterizar o substantivo utilizam-se, com maior frequência, as locuções adjetivas. Os seguintes SN, da forma N + MOD, são exemplos nos quais uma locução adjetiva (sublinhada) está caracterizando o núcleo: **bola de futebol**, **panela de arroz**, **pista de corrida**, **amor de mãe**, etc. Os modificadores são usados também para expressar uma avaliação sobre o substantivo. Neste caso, utilizam-se, com maior frequência, os adjetivos simples. Os SN a seguir, da forma N + MOD, são exemplos nos quais um adjetivo simples (sublinhado) expressa uma avaliação sobre o núcleo: **bola estragada**, **juiz ladrão**, **criminoso cruel**, **crime passional**, **amor materno**, etc.

A tabela 4.1 apresenta as principais estruturas de sintagmas nominais com os respectivos exemplos.

Estrutura de SN	**Exemplos**
DET + N + MOD	*os aguaceiros de verão*
N + MOD	*chuva grossa*
DET + DET + N	*uma certa crença*
DET + N + DET + N + MOD	*a terra e a areia assentadas*
MOD + N + MOD	*grande movimentação de bichos*
DET + DET + N + MOD	*uma certa alegria despropositada*

Tabela 4.1: Sintagmas Nominais utilizados e nominais.

4.2.2 Método Semiautomático de Conhecimento de Mundo

A figura 4.2 apresenta as fases do nosso método semiautomático de AC. Inicialmente, o usuário informa com uma expressão $<$ EXP $>$ ao sistema de AC. Se a expressão é um conceito contido na base de conhecimento, as relações semânticas que expressam o conteúdo do conceito serão apresentadas ao usuário para validação. Caso contrário, um novo conceito deve ser adquirido e, então, o método é executado de acordo com os seguintes passos:

1. Análise sintática de EXP, a fim de definir a estrutura gramatical do SN;
2. Execução das heurísticas para adquirir o conteúdo do novo conceito pelo raciocinador;
3. Apresentação para o usuário da lista de relações semânticas de mundo inferidas pelo raciocinador. Essa lista servirá como arcabouço inicial para o novo conceito a ser adquirido;
4. Validação, pelo usuário, da lista de relações semânticas e definição do conteúdo do novo conceito.

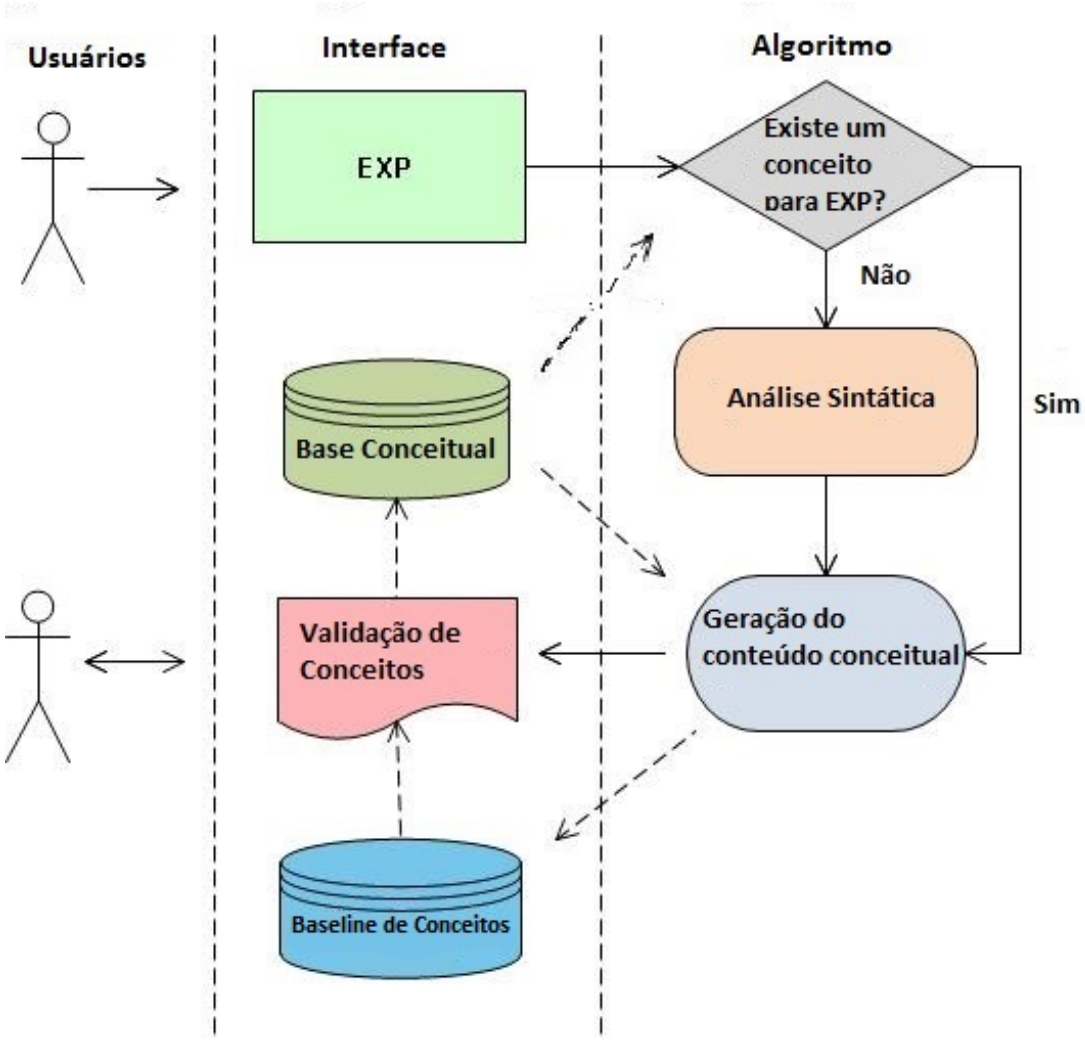


Figura 4.2: O método semiautomático de aquisição de conhecimento de mundo para conceitos em língua natural.

O método proposto consiste em um raciocínio heurístico aplicado sobre conteúdo conceitual pré-existente e semanticamente relacionado, o qual gera uma base inicial de conhecimento para o novo conceito. O método possibilita, adicionalmente, um processo interativo com o usuário, o qual pode incluir novas relações semânticas bem como excluir relações propostas, encerrando um mecanismo de validação do conteúdo semântico do novo conceito. A seguir detalhamos as heurísticas aplicadas no passo 2.

4.2.3 Heurísticas para Aquisição de Conhecimento de Mundo

As heurísticas são responsáveis pela geração e proposição do conteúdo semântico para a expressão linguística de entrada, EXP, a qual nomeia o novo conceito a ser adquirido. A tabela 4.2 apresenta as estruturas gramaticais de sintagmas nominais, contempladas pelas heurísticas.

Estrutura de SN - EXP	**Exemplos**
$< nome >$	*vingança, pistolagem*
$< nome >< adjetivo >$	*crime passional, impunidade penal*
$< adjetivo >< nome >$	*má urbanização*
$< adjetivo_1 > < nome >< adjetivo_2 >$	*má iluminação pública*
$< nome_1 >$ <"DE"> $< nome_2 >$	*aula de português, bola de plástico*

Tabela 4.2: Principais estruturas de SN.

1. <*nome*> ou <*adjetivo*> - Quando EXP não é encontrada na base conceitual, é apresentado para o usuário um conjunto de conceitos pré-existentes na base, os quais são: (i) semanticamente relacionados (p.ex. sinônimos); (ii) nomeados com o mesmo radical de <*nome*> ou <*adjetivo*>; (iii) nomeados com a forma primitiva de <*nome*>; (iv) substantivos relacionados semanticamente a <*adjetivo*>. Por exemplo, para a expressão linguística "*torcedor*", a heurística apresenta os conceitos "*fã*", "*torcida*" e "*torcer*". Para a palavra "*passional*", a heurística apresenta o conceito "*paixão*". Em seguida, o usuário seleciona qual, dentre os conceitos apresentados, pode ser usado como base para aquisição do novo conceito. A heurística retorna, em seguida, uma lista de relações semânticas do conceito selecionado, previamente contidas na base.

2. <*nome*><*adjetivo*> ou <*adjetivo*><*nome*> - Nesses casos, <*adjetivo*> caracteriza <*nome*> , indicando-lhe atributo, propriedade, estado, modo de ser ou aspecto. Percebe-se, portanto, um caso de especialização no qual "<*nome*><*adjetivo*>" ou "<*adjetivo*><*nome*>" expressa uma situação particular ou um tipo de <*nome*>. Por exemplo, no caso da expressão "*crime passional*", o adjetivo "*passional*" está caracterizando o nome "*crime*", atribuindo-lhe propriedades relativas à "*paixão*" e especificando um tipo de "*crime*". A heurística está definida nos seguintes passos:

   (a) Chamada recursiva à heurística (1) para EXP_1 =<*nome*> e EXP_2 = <*adjetivo*>, retornando uma lista de relações semânticas dos conceitos associados a EXP_1 e EXP_2;

   (b) Herança do conteúdo de <*nome*> para o novo conceito "<*nome*><*adjetivo*>" ou "<*adjetivo*><*nome*>", pois em ambos tem-se a expressão de um caso particular ou um tipo de <*nome*> e, assim sendo, todo o conteúdo de <*nome*> pode ser transcrito (ou herdado) para "<nome><adjetivo>" ou "<adjetivo><nome>". Por exemplo, a relação inferencial "<***crime***><*capazDe*><***ter vítima***>" é transcrita para uma nova relação semântica "<***crime passional***><*capazDe*> <***ter vítima***>";

(c) Transcrição parcial do conteúdo de <adjetivo> para o novo conceito "<*nome*> <*adjetivo*>" ou "<*adjetivo*><*nome*>". Neste caso, <*adjetivo*> está caracterizando <*nome*> e algumas relações semânticas de <*adjetivo*> devem ser sugeridas para <*nome*> de forma a atribuir-lhe características ou qualidades. A seguinte metarregra é usada neste passo:

$$\frac{\text{<A>e caracterizado por <B>,<B> <nome\_rel> C}}{\rightarrow \text{<A caracterizado por B> <nome\_rel> C}}$$

Para definir quais $<$ *nome_rel* $>$ tornam esta inferência válida, cada $<$ *nome_rel* $>$ da base semântica deve ser analisada conforme a natureza da relação semântica. Relações semânticas estruturais (por exemplo: eUm, efeitoDe, parteDe) comumente não devem ser herdadas, pois expressam conteúdo restrito a <*adjetivo*>. Por exemplo, o fato de que "$<$***paixão***$><eUm><$***sentimento***$>$" não implica que "$<$***crime passional***$><eUm><$***sentimento***$>$". As relações semânticas pragmáticas como relações funcionais, causais, eventuais, motivacionais, comumente suscitam características que são atribuídas de $<$ *adjetivo* $>$ para $<$ *nome* $>$. Por exemplo, "$<$***paixão***$><efeitoDe><$***ciúme***$>$" autoriza a geração do conteúdo "$<$***crime passional***$><efeitoDe><$***ciúme***$>$". A tabela 4.3 apresenta os tipos de relações semânticas da InferenceNet definidas para aplicação da metarregra acima. Ao final do processo, a heurística retorna a lista de relações semânticas geradas, as quais foram associadas a "<*nome*><*adjetivo*>" ou "<*adjetivo*><*nome*>". Na figura 4.3, temos um exemplo do conjunto de relações apresentadas aos usuários após a execução do algoritmo.

Natureza da Relação	Tipo de Relação Semântica	Tipo de Relação Inferencial
RELATIVA A PROPRIEDADE	PropriedadeDe	Pré condição
RELATIVA A EVENTO	EventoPreRequisitoDe, PrimeiroSubEventoDe, SubEventoDe, UltimoSubEventoDe	Pré condição
CAUSAL	EfeitoDe, EfeitoDesejavelDe	Pós-condições
MOTIVACIONAL	MotivacaoDe, DesejoDe	Pré condição
FUNCIONAL	UsadoPara;	Pré condição

Tabela 4.3: Tipos de relações semânticas de InferenceNet que serão herdadas de <*adjetivo*> para <*nome*><*adjetivo*> ou <adjetivo><nome>

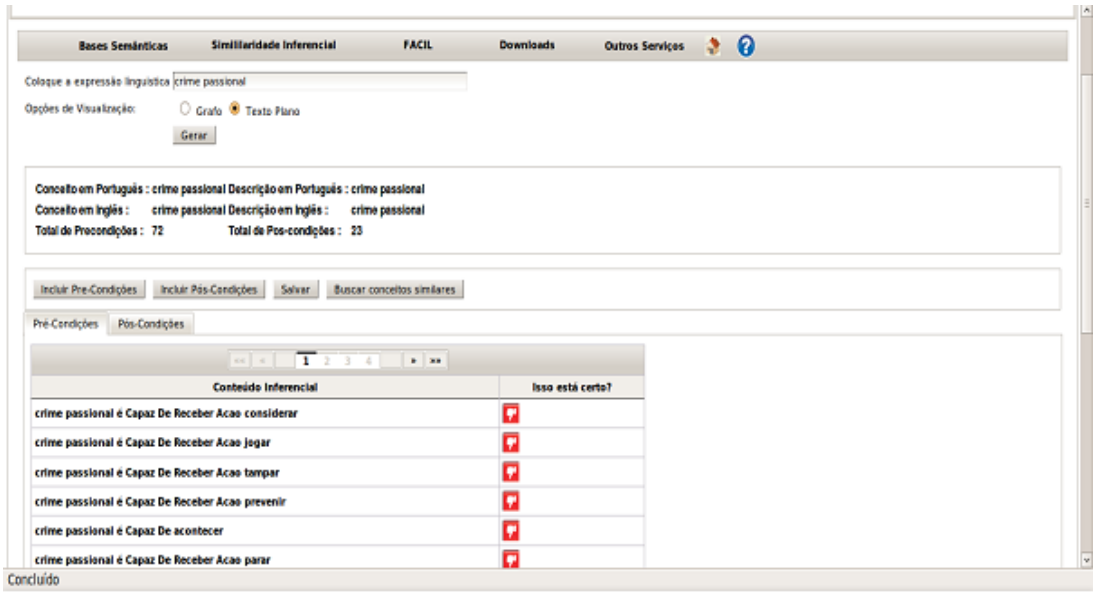


Figura 4.3: Captura de tela do protótipo com o exemplo “crime passional”

3. $< adjetivo_1 >< nome >< adjetivo_2 >$ - Neste caso, o usuário questionado se $< adjetivo_1 >$ está qualificando “$< nome >< adjetivo_2 >$”, por exemplo, como acontece em “*má iluminação pública*”. Caso o usuário confirme, a heurística (2) é chamada para EXP=“$< nome >< adjetivo_2 >$” e, em seguida, para EXP=“$< adjetivo_1 >< np_2 >$” com $< np_2 >$ = “$< nome >< adjetivo_2 >$”. Caso contrário, a heurística (2) é chamada para EXP= “$< adjetivo_1 >< nome >$” e EXP= “$< nome >< adjetivo_2 >$”. Ao final, a heurística retorna a lista de relações semânticas selecionadas de forma recursiva.

4. $< nome_1 >< DE >< nome_2 >$ - Neste caso, o usuário é questionado se $< nome_2 >$ está caracterizando “$< nome_1 >$”, por exemplo, como acontece em “*aula de português*”. Caso o usuário confirme, a heurística (2) é chamada para EXP=“$< nome_1 >< nome_2 >$”

com $< nome_2 >$ sendo uma locução adjetiva que está expressando uma caracterização de $< nome_1 >$. Caso contrário, a heurística (1) é chamada para EXP= "$< nome_1 >$" e para EXP= "$< nome_2 >$". Ao final, a heurística retorna a lista de relações semânticas selecionadas de forma recursiva.

A figura 4.4 apresenta o algoritmo que implementa as heurísticas propostas, vamos exemplificar o funcionamento do algoritmo com o sintagma de entrada *crime passional*. Na primeira interação o algoritmo recebe a expressão *crime passional*, inicialmente verifica se o conceito existe na base semântica, caso contrário, é feita a análise sintática e verificado em qual dos casos o sintagma se encaixa. O nosso exemplo, *crime passional*, se encaixa na heurística (2), que se refere a <*nome*><*adjetivo*>, em seguida são realizadas chamadas recursivas ao algoritmo para EXP =<*nome*> e EXP = <*adjetivo*> e, por fim, o conteúdo gerado é apresentado ao usuário para validação.

```
SemanticRelations [ ] generateContent (exp)

// 1ª iteration:  exp=" crime passional"
// 2nd  iteration: exp = "crime"
// 3rd  iteration: exp = "passional"

If knowledgeBaseExists (exp) then
      return retrieveContent(exp);
      // 2nd iteration: for exp = "crime"
      // retrieved semantic relations of "crime"
      // Examples: (capableOf, "crime","envolver violência", Pre); (effectOf, "crime","sofrimento", Pos)
else {
      if structure(exp) = "<noun>" or "<adjective>":    // HEURISTIC (1)
            relatedConcepts [ ] = retrieveRelatedConcepts (exp);
            relatedConcept = selectionUser(relatedConcepts [ ]);
            return retrieveContent(relatedConcept);

            // 3rd iteration: for exp = "passional"
            // retrieved semantic relations of "paixão" – relatedConcept selected by user
            // Examples: (eventPreRequisitOf,"paixão","amante", Pre);
            //                (effectOf,"paixão","sofrimento", Pos)
            //                (usedFor ,"paixão","romance", Pre); (isA, "paixão","sentimento", Pos)

      If structure(exp) = "<noun><adjective>" or "<adjective><noun>": // HEURISTIC (2)
                        // 1ª iteration:  exp="crime passional"

            exp1 = firstTerm(exp);                                  // exp1 = "crime"
            exp2 = secondTerm(exp);                                 // exp2 = "passional"
                  content1 = generateContent(exp1)      ;     // recursive  call for exp1= "crime"
                  content2 = generateContent(exp2);         // recursive  call for exp2= "passional"
            if structure(exp1) = "<adjective>" then {
                  content1 = selectContent(content1);
            } else {
                  content2 = selectContent(content2):
                  // selects relations of content2 (referring to exp2="passional") per step 2.c
                  //  Examples: (eventPreRequisitOf,"paixão", "amante", Pre);
                  //                (effectOf, "paixão", "sofrimento", Pos)
                  //                (usedFor, "paixão", "romance", Pre);
                  //               Note: the relation "isA("paixão", "sentimento", Pos)" was not selected
            }
            return content1+content2;
      if structure(exp) = "<adjective1><noun><adjective2>" :  // HEURISTIC (3)
            if <adjective1> qualifies "<noun><adjective2>" then {
                  exp1 = <adjective1>;
                  exp2 = <noun> <adjective2>;
                  content1 = generateContent(exp1);
                  content2 = generateContent(exp2);
                  content1 = selectContent(content1);
            } else {
                  exp1 = <adjective1><noun>;
                  exp2 = <noun><adjective2>;
                  content1 = generateContent(exp1);
                  content2 = generateContent(exp2);
            }
            return content1+content2;
      if structure(exp) = "<noun1><DE><noun2>" :  // HEURISTIC (4)
            if <noun2> characterizes "<noun1>" then {
                  exp1 = <noun1>;
                  exp2 = <noun2>;
                  content1 = generateContent(exp1);
                  content2 = generateContent(exp2);
                  content2 = selectContent(content2);
            } else {
                  exp1 = <noun1>;
                  exp2 = <noun2>;
                  content1 = generateContent(exp1);
                  content2 = generateContent(exp2);
            }
            return content1+content2;

}
```

Figura 4.4: Algoritmo para geração de conteúdo de conceitos.

4.2.4 Avaliação do Método Semi-Automático de Aquisição de Conhecimento de Mundo

A avaliação realizada visou analisar dois aspectos: (1) o quão as heurísticas facilitam a aquisição de conhecimento de mundo para a língua portuguesa; (2) a qualidade do conteúdo gerado pelas heurísticas, ou seja, se o conteúdo proposto realmente expressa o valor semântico do conceito desejado pelo usuário. Nesta avaliação, o algoritmo foi implementado para AC de conceitos para a base InferenceNet.BR e utilizou o *parser* PALAVRAS [Bick 2000]. No entanto, o método pode ser aplicado em outras bases de conhecimento de mundo e outros *parsers* para língua portuguesa também podem ser utilizados.

4.2.4.1 Metodologia de Avaliação

A metodologia de avaliação seguiu os passos delineados na sequência.

1. Seleção de 20 pessoas adultas com experiência em sistemas interativos da Internet e que não tinham conhecimento sobre o método de AC semiautomático, proposto neste trabalho. As pessoas foram distribuídas aleatoriamente em 2 (dois) grupos de 10 pessoas, um grupo para cada cenário de teste;

2. Seleção de conceitos usados na língua portuguesa que não existiam previamente na base InferenceNet: "*crime passional*", "*violência policial*", "*má iluminação pública*", "*bom juiz honesto*", "*aula de português*", "*bola de plástico*". Esses conceitos foram selecionados de forma que todas as heurísticas propostas nesse trabalho fossem avaliadas.

3. Definição de cenários de testes:

   - **Cenário 1** - Usuários incluem, sem limite de tempo, relações semânticas para os conceitos escolhidos, através do portal `http://www.inferencenet.org`, o qual possui uma interface interativa que permite a entrada de relações de senso comum e inferencialista na base InferenceNet. Na figura 4.5 temos uma captura de tela do protótipo no primeiro cenário.

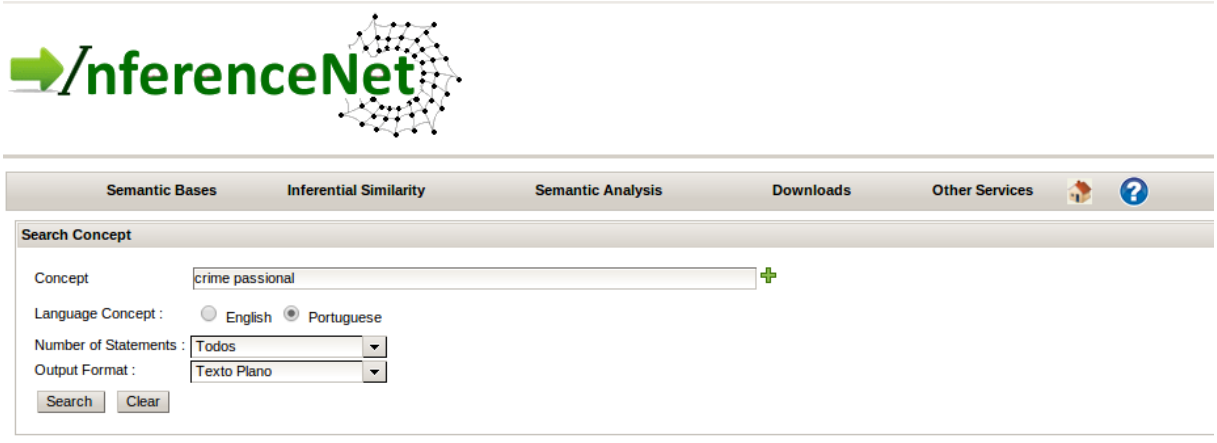


Figura 4.5: Captura de tela do protótipo com o exemplo "crime passional" no Cenário 1

- **Cenário 2** - No portal `http://www.inferencenet.org`, o usuário informa a expressão (EXP) correspondente ao conceito e interage com o portal para validar o conteúdo conceitual gerado pelo algoritmo implementado. Os usuários foram orientados a alterar e a excluir relações semânticas caso não concordassem com elas, além de incluir novas relações caso julgassem ainda necessárias, sem limite de tempo. Na figura 4.6 temos a captura de tela do protótipo no segundo cenário.

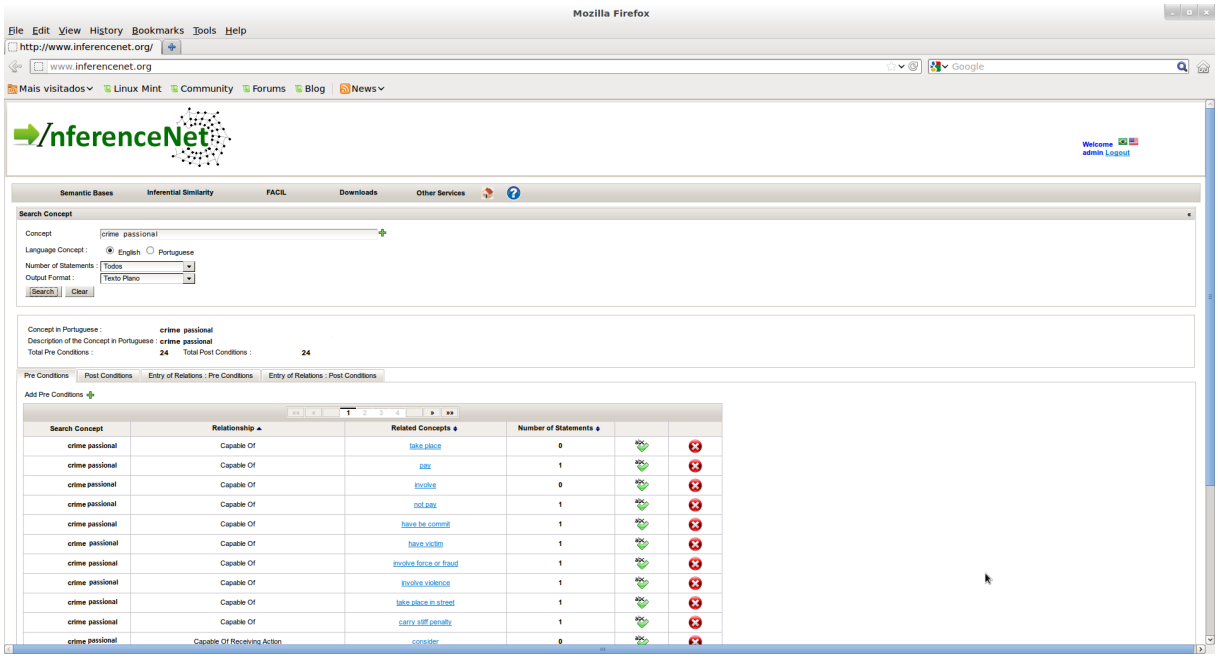


Figura 4.6: Captura de tela do protótipo com o exemplo “crime passional” no Cenário 2

4. Geração do conteúdo dos conceitos, onde um avaliador humano validou as relações semânticas geradas pelo algoritmo e definiu uma lista para os conceitos desta avaliação. Essa lista serviu como Coleção Dourada(CD) para análise qualitativa do conteúdo semântico ao final do processo de AC experimentado pelos 10 usuários no Cenário 2. No anexo A temos alguns exemplos de relações validadas pelo avaliador.

Em cada cenário, foi medido o tempo para realização da atividade e quantas relações semânticas foram incluídas e excluídas para cada conceito selecionado. As tabelas 4.4, 4.5 e 4.6 apresentam os resultados médios coletados. No Cenário 2, o algoritmo implementado gerou as seguintes quantidades de relações para os conceitos em questão: ***crime passional*** - 45 pré condições e 17 pós-condições; ***violência policial*** - 67 pré condições e 1 pós-condição; ***má iluminação publica*** - 13 pré condições; ***bom juiz honesto*** - 69 pré condições e 7 pós-condições; ***aula de português*** - 53 pré condições e 1 pós-condição; ***bola de plástico*** - 808 pré condições e 1 pós-condição. Como resultado principal, tem-se que 72% das relações geradas foram validadas por humanos (considerando a média de exclusões da lista de conceitos dos 10 usuários que participaram do Cenário 2).

Cenários	**“crime passional”**					**“violência policial”**				
	RG[1]	**Incl**	**Excl**	**RF**[2]	**Tempo**	**RG**	**Incl**	**Excl**	**RF**	**Tempo**
Cenário 1	0	2.9 pre 1.6 pos	- -	5	00:02:55	0	4.1 pre 2 pos	- -	6	00:02:31
Cenário 2	62	0 pre 0 pos	8.2 pre 0.7 pos	53	00:03:46	68	0 pre 0.1 pos	17.9 pre 0.4 pos	50	00:03:52
CD	62	4 pre 1 pos	10 pre 5 pos	52	n/a	68	5 pre 1 pos	23 pre 1 pos	50	n/a

Tabela 4.4: Resultados coletados nos dois cenários de avaliação e da lista de conceitos.

Cenários	**“bom juiz honesto”**					**“má iluminação pública”**				
	RG	**Incl**	**Excl**	**RF**	**Tempo**	**RG**	**Incl**	**Excl**	**RF**	**Tempo**
Cenário 1	0	5.7 pre 2,2 pos	- -	8	00:03:49	0	2,9 pre 1,2 pos	- -	5	00:02:25
Cenário 2	76	0.9 pre 0.1 pos	5,8 pre 0,6 pos	70	00:02:16	13	0 pre 0,1 pos	6,3 pre 0 pos	6	00:01:42
CD	76	- -	15 pre 2 pos	59	n/a	13	4 pre 3 pos	7 pre 0 pos	13	n/a

Tabela 4.5: Resultados coletados nos dois cenários de avaliação e da lista de conceitos.

Cenários	**“aula de português”**					**“bola de plástico”**				
	RG	**Incl**	**Excl**	**RF**	**Tempo**	**RG**	**Incl**	**Excl**	**RF**	**Tempo**
Cenário 1	0	6,1 pre 2 pos	- -	8	00:04:54	0	6,9 pre 1,9 pos	- -	9	00:04:22
Cenário 2	54	1 pre 0 pos	9.2 pre 0 pos	55	00:04:20	809	0.7 pre 0.2 pos	15.7 pre 0.7 pos	793	00:07:49
CD	54	- -	12 pre -	42	n/a	809	-	240 pre 0 pos	569	n/a

Tabela 4.6: Resultados coletados nos dois cenários de avaliação e da lista de conceitos.

4.2.4.2 Análise dos Resultados

A partir dos resultados coletados, tem-se que o método proposto possibilita interações mais produtivas para AC: no Cenário 1, os usuários levaram em média, 2min31s para inclusão de 4,9 relações semânticas (média de pré condições e pós-condições incluídas para os seis conceitos), enquanto que, no Cenário 2 os usuários realizaram 11,2 exclusões e inclusões de relações semânticas em 3min6s (tempo médio). Observamos que, no Cenário 1, os usuários encontraram dificuldade em externar relações semântica de senso comum sobre o conceito e, em alguns casos, até mesmo em se lembrar o que caracterizaria semanticamente aquele conceito. No Cenário 2, o usuário é instigado a interagir com as relações semântica geradas, resultando em uma melhor relação inclusões/exclusões por minuto (3,61 no Cenário 2 contra

1,96 no Cenário 1). Outro fato interessante é que a quantidade de relações semânticas geradas pelo método no Cenário 2 é bem maior do que as inclusões feitas pelos usuários no Cenário 1, mesmo considerando as exclusões realizadas (Ver coluna RF - Relações finais, nas tabelas 4.4, 4.5, e 4.6).

Em relação à qualidade do conteúdo conceitual gerado pelas heurísticas, comparamos os grafos conceituais dos seis conceitos, após as inclusões e exclusões realizadas pelos usuários, e os grafos conceituais da Coleção Dourada(CD). Como resultado, o método proposto possibilitou 72% de acurácia, em média, para os seis conceitos analisados, considerando as relações mais excluídas pelos usuários. Importante salientar que, no Cenário 2, os usuários se limitaram a excluir as relações que lhes pareceram inválidas para o conceito, e praticamente não incluíram novas relações.

Vale a pena ressaltar que o estudo inicial sobre esse método começou em [Silva 2010] e foi concluído durante a dissertação de mestrado.

Por fim, podemos concluir que as heurísticas facilitam a aquisição de conhecimento já que o número de relações incluídas na base semântica é maior. Sobre a qualidade do conteúdo semântico gerado com o auxilio das heurísticas temos 72% de acurácia.

4.3 Aquisição Automática de Relações Semânticas da Wikipédia

O método automático aqui proposto consiste em um algoritmo para extrair relações semânticas a partir do texto contido entre *links* de artigos da Wikipédia. O método sustenta-se em nossa segunda hipótese de pesquisa, o texto contido entre *links*, esses *links* foram editados por usuários, pode fornecer relações semânticas interessantes. Esta seção está dividida da seguinte forma: inicialmente mostramos como a Wikipédia é cada vez mais usada como fonte de conhecimento de mundo. Em seguida, mostramos uma comparação entre o conhecimento semântico expresso nas bases InferenceNet e Wikipédia. O objetivo deste experimento foi explicitar as semelhanças e diferenças entre o conteúdo presente nestas bases e descobrir se vale a pena adquirir relações dos textos da Wikipédia. Por fim, apresentamos o método proposto, a avaliação realizada e os resultados que subsidiaram uma discussão sobre nossa investigação.

4.3.1 Wikipédia como fonte para extração de conhecimento de Mundo

Dentre as características da Wikipédia, destacamos os *links* presentes no corpo dos artigos, os quais são indícios de relações semânticas entre os conceitos (representados pelos artigos). Esta característica suscitou-nos uma ideia de como resolver o problema presente em vários *Open IE Systems* que é a identificação dos argumentos de uma relação. Além disso, utilizar a Wikipédia como fonte para extração de conhecimento de mundo trás algumas vantagens como: alta qualidade e dinamicidade do conteúdo, um abrangente corpus multilíngue, a existência de uma semi-estrutura que facilita o processo de extração de conhecimento. Porém, existem algumas desvantagens: dependência da estrutura para realização da extração, e um volume menor de informação em comparação com toda a informação presente na Web.

4.3.2 Comparação entre Wikipédia e InferenceNet

Existia, no inicio deste trabalho de pesquisa, a crença na utilidade de um método automático para aquisição de conhecimento de mundo a partir da Wikipédia. Porém, algumas questões surgiram e precisavam de respostas: Será que existe conteúdo semântico a ser adquirido na Wikipédia? Qual a espécie de conhecimento expresso na Wikipédia? Existe conhecimento compartilhado entre a Wikipédia e bases proeminentes de senso comum?

Para responder estas perguntas, realizamos uma análise minuciosa e comparativa do conteúdo expresso na Wikipédia em português e na base InferenceNet, por esta ser a maior base de senso comum para a língua portuguesa [Pinheiro et al. 2010].

Visando simplicidade e um maior valor agregado ao experimento, o escopo foi reduzido para os sintagmas nominais, especificamente para os substantivos comuns, presentes em ambas as bases.

Como na Wikipédia um conceito equivale a um artigo e não temos qualquer indicação do tipo de substantivo que está representando o conceito, propomos uma heurística para definir se um artigo refere-se a um substantivo próprio ou a um substantivo comum. Esse algoritmo tem como ideia principal verificar como o artigo é citado em textos da própria Wikipédia. Se, na maioria das vezes, o artigo for citado com letra inicial minúscula, este refere-se a um substantivo comum, caso contrário, se na maioria das vezes é citado com inicial maiúscula, então o artigo refere-se a um substantivo próprio. A figura 4.7 ilustra o algoritmo que implementa esta estratégia.

Algoritmo 1 isNomeComum(*Artigoa*)

```
1: S ← a.ondeArtigoEcitado()
2: while S ≥ 0 do
3:     if S == comecaLetraMinuscula then
4:         N_c++
5:     else
6:         N_p++
7:     end if
8:     S ← S − 1
9: end while
10: retorna N_c > N_p
```

Figura 4.7: Algoritmo de identificação de substantivos comuns

Para analisar a Wikipédia utilizamos a ferramenta WikipediaMiner[3], um *toolkit* para exploração da Wikipédia. As principais vantagens de utilizar o WikipediaMiner são as seguintes: fornece um acesso simplificado aos dados da Wikipédia através de uma estrutura orientada á objetos ligada à estrutura existente na Wikipédia; faz a detecção e eliminação de ambiguidades em tópicos da Wikipédia e, por fim, fornece uma medida de relacionamento semântico entre os artigos [Milne e Witten 2008]. Devido a uma instabilidade da versão mais atual do WikipediaMiner utilizamos uma versão 1.0 para realização das análises.

[3]http://wikipedia-miner.cms.waikato.ac.nz/

A primeira análise foi em relação a quantidade e conteúdo de substantivos comuns presentes nas bases. Verificamos que cerca de 25% da base InferenceNet são substantivos comuns, enquanto que substantivos comuns estão presentes em 8% da Wikipédia. No que concerne as relações entre conceitos, temos que cerca de 71% das relações presentes na base do InferenceNet são entre nomes comuns. Esse número cai para 13% quando analisamos a Wikipédia. A figura 4.8 apresenta diagramas destes conjuntos.

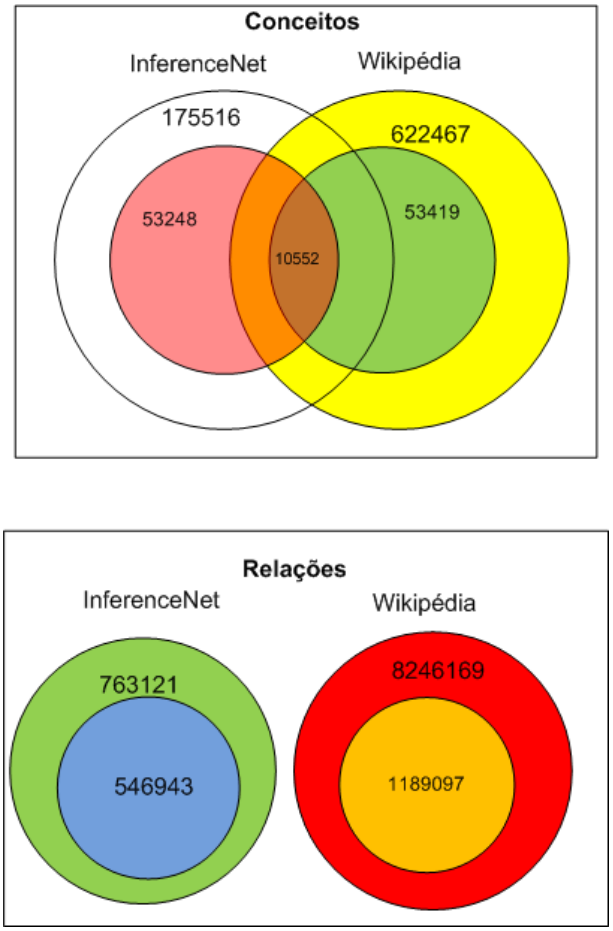


Figura 4.8: Distribuição de conceitos/relações na Wikipédia e InferenceNet

Após delimitar quais conceitos e relações são objeto de estudo em nosso trabalho, precisamos saber se existe algum conteúdo semântico para ser adquirido da Wikipédia pela base InferenceNet e qual é a parcela de conhecimento partilhado entre as bases. Verificamos que as duas bases têm aproximadamente 10500 conceitos em comum. Para esses conceitos, existem cerca de 3500 relacionamentos em comum entre a Wikipédia e a base InferenceNet. Ou seja, existe ainda conteúdo de 42.919 conceitos da Wikipédia a serem adqueridos. A figura 4.9 apresenta um diagrama com estes resultados.

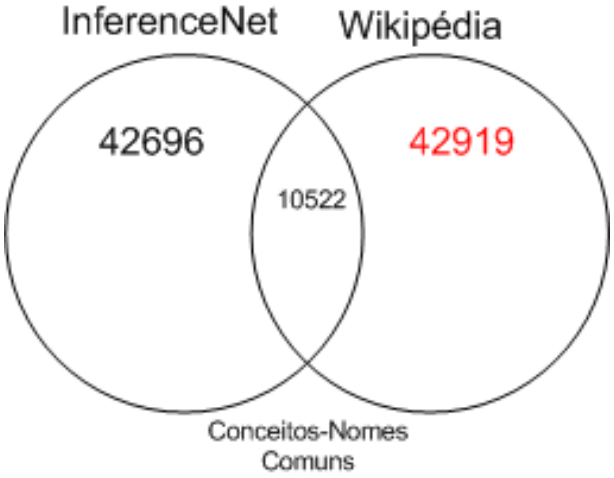


Figura 4.9: Diagrama de conceitos em comum entre InferenceNet e Wikipédia

Dentre esses 10552 conceitos em comum, descobrimos que os conceitos em comum tem 419554 relações que pertencem somente a InferenceNet e 350787 relações que pertencem somente a Wikipédia. Podemos concluir que existe ainda muito conteúdo semântico para ser extraído da Wikipédia para a base InferenceNet. A figura 4.10 apresenta um diagrama com estes resultados.

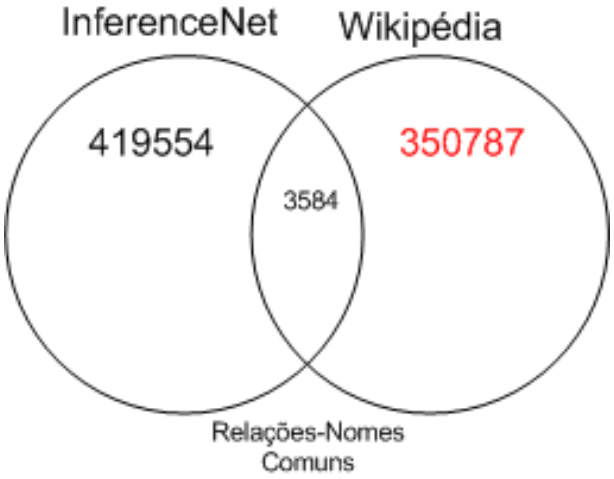


Figura 4.10: Diagrama de relações em comum entre InferenceNet e Wikipédia.

A questão de pesquisa suscitada foi: como adquirir e se apropriar do conteúdo semântico entre os artigos da Wikipédia?

4.3.3 Método Automático de Extração de Relações Semânticas da Wikipédia

Nesta seção detalhamos um método automático de extração de relações semânticas a partir de textos em linguagem natural presentes nos artigos da Wikipédia. As relações extraídas expressam um relacionamento semântico entre dois conceitos c_1 e c_2 e podem ser representadas na forma *(c_1, relacao, c_2)*, por exemplo, *(agricultura, ser conjunto de, técnica)*.

O principal diferencial do método proposto é a independência de expressões regulares pré-definidas, a identificação de relações redundantes e identificação dos argumentos das relações semânticas através dos *links*. Além destas vantagens, o método proposto não aplica técnicas de aprendizagem supervisionada, cuja necessidade de anotação de corpus é sempre um gargalo. A figura 4.11 apresenta o método proposto com as seguintes etapas: Mineração e Seleção de Sentenças, Clusterização de Sentenças e Aquisição de Relações Semânticas.

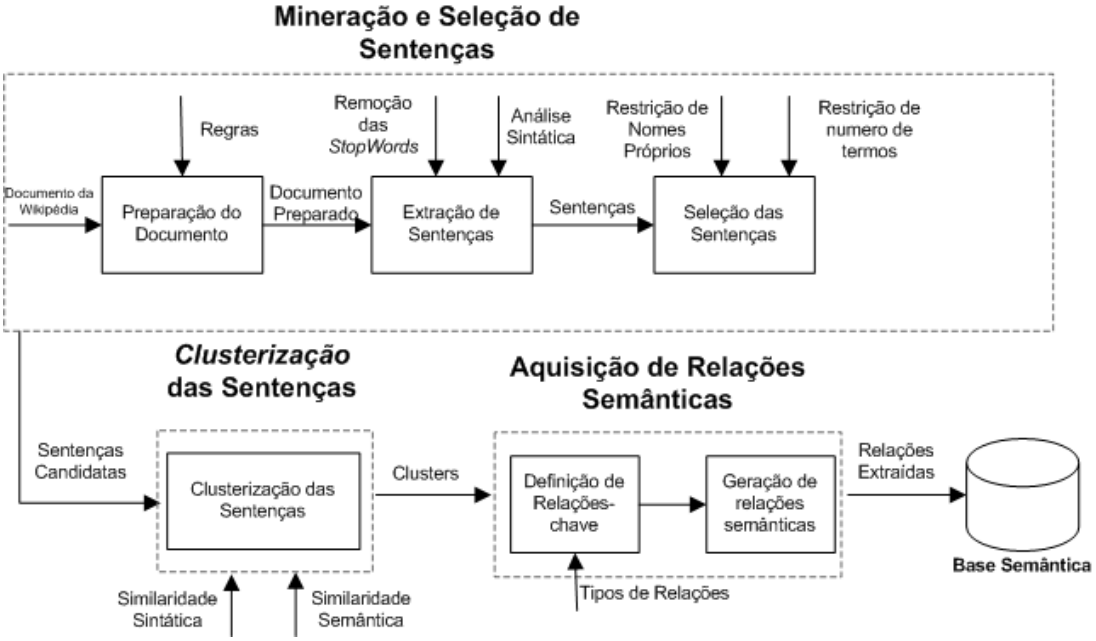


Figura 4.11: Método de Aquisição Automática de Relações Semânticas.

A seguir, cada etapa do método é detalhada com o auxilio de dois exemplos de artigos da Wikipédia: “Agricultura” e “Casa”.

4.3.3.1 Mineração e Seleção de Sentenças

O objetivo deste módulo é separar e analisar as sentenças do texto do artigo de entrada (documento da Wikipédia), gerando um conjunto de sentenças candidatas. Para isso, inicialmente, deve ser realizada a preparação do documento de entrada com a aplicação de regras e parâmetros específicos, tais como: qual segmento do texto do artigo será processado, quais trechos de textos podem ser descartados, escolha de um tema específico dos artigos analisados, etc. Essas regras são definidas pelo usuário do método e podem ser parametrizadas de acordo com a finalidade de aplicação do método.

Como exemplos de regras e parâmetros citamos os seguintes:

- Apenas o primeiro parágrafo do texto deve ser processado, pois este, em geral, apresenta uma descrição sucinta do conceito referenciado no artigo, contendo as informações mais relevantes;
- Trechos de textos entre parênteses devem ser descartados. Na Wikipédia esses trechos entre parenteses trazem informações sobre a fonética do nome do artigo.

As figuras 4.12 e 4.13 trazem exemplos de artigos que ilustram a aplicação dessas regras. O artigo “Agricultura” apresenta em destaque o seu primeiro parágrafo sem nenhum descarte de texto. No artigo “Casa” temos alguns termos entre parenteses que serão descartados durante a aplicação das regras.

Figura 4.12: Screenshot do artigo **Agricultura** da Wikipédia em português com seu primeiro parágrafo em destaque.

Figura 4.13: Screenshot do artigo **Casa** da Wikipédia em português com seu primeiro parágrafo em destaque.

Após a preparação do documento, a próxima atividade deste módulo é a extração de sentenças. Nesta, é realizada a análise, separação e extração de sentenças entre a expressão que nomeia o artigo (o nome do artigo) e as expressões que são *links* presentes no documento preparado. Na Wikipédia, diversos *links* são inseridos ao longo do texto de um artigo, os quais indicam uma relação entre o artigo corrente e outro artigo referenciado pelo link. Por exemplo, na Figura 4.12, no primeiro parágrafo do artigo **Agricultura** existem os *links* em destaque: técnicas, plantas, alimentos, fibras, energia, matéria-prima, roupas, construções, medicamentos, ferramentas e estética.

Neste caso, as seguintes sentenças foram extraídas para o artigo **Agricultura**:

- *Agricultura é um conjunto de técnicas*
- *Agricultura utilizadas para cultivar plantas*
- *Agricultura com o objetivo de obter alimentos*
- *Agricultura com o objetivo de obter fibras*
- *Agricultura com o objetivo de obter energia*
- *Agricultura com o objetivo de obter matéria-prima*
- *Agricultura para roupas*
- *Agricultura para construções*
- *Agricultura para medicamentos*
- *Agricultura para ferramentas*
- *Agricultura ou apenas para contemplação estética*

O próximo passo dessa etapa é fazer a análise morfossintática usando um *POS tagger* que classifica as palavras e expressões (*Part Of Speech*) do texto. A separação em *tokens* consiste em separar todos os elementos de uma sentença. Por exemplo, a sentença "Agricultura é o conjunto de técnicas utilizadas para cultivar plantas." é separada da seguinte forma: "Agricultura", "é", "o", "conjunto", "de", "técnicas", "utilizadas", "para", "cultivar", "plantas" e ".". A literatura apresenta processadores *POS tagger* com acurácia em torno de 97% para a língua portuguesa [Padró e Stanilovsky 2012] e, nesse trabalho, utilizamos o FreeLing, um *parser* multilíngue e *open-source* desenvolvido pela Universidade Politécnica da Catalunha com cerca de 97% de acurácia. A figura 4.14 apresenta a sentença "Agricultura é o conjunto de técnicas utilizadas para cultivar plantas." anotada pelo parser FreeLing com as classes gramaticais e a forma primitiva das palavras. Por exemplo, "agricultura" é classificado como NP (substantivo comum), "é" é classificado como VMIP3S0 (Verbo no infinitivo, terceira pessoa do singular), etc.

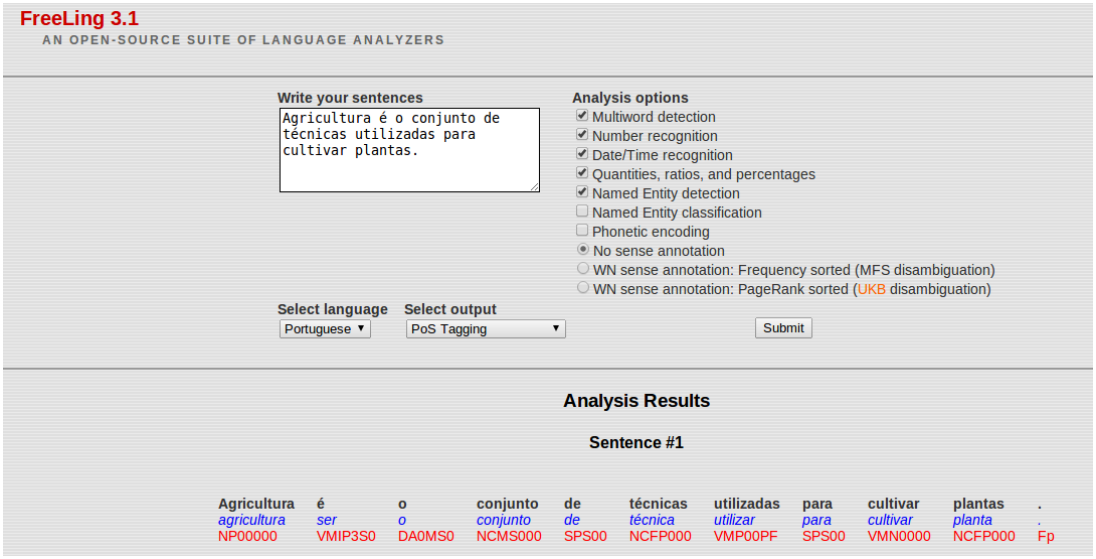


Figura 4.14: Screenshot da sentença “Agricultura é o conjunto de técnicas utilizadas para cultivar plantas.” analisada pelo parser FreeLing.

Em seguida, são removidas as *stop words* do texto entre os *links*. *Stop words* são palavras muito frequentes e que não possuem valor semântico, servindo apenas como elementos estruturadores de uma sentença (tais como artigos, conjunções, etc.) [Silva e Ribeiro 2003]. Podemos citar como exemplos de *stop words* as seguintes palavras: “a”, “agora”, “ainda”, “alguém”, etc. Normalmente as proposições são consideradas *stop words* porém no nosso método não vamos descarta-las. No exemplo, as sentenças extraídas do artigo “Agricultura” após a remoção das *stop words* ficaram:

- *Agricultura é conjunto de técnicas*
- *Agricultura utilizadas para cultivar plantas*
- *Agricultura com objetivo de obter alimentos*
- *Agricultura com objetivo de obter fibras*
- *Agricultura com objetivo de obter energia*
- *Agricultura com objetivo de obter matéria-prima*
- *Agricultura para roupas*
- *Agricultura para construções*
- *Agricultura para medicamentos*
- *Agricultura para ferramentas*
- *Agricultura para contemplação estética*

Por fim, são extraídas sentenças do texto de entrada no formato <*artigo*><*segmento_texto*><*link*>, onde:

- <*artigo*> é a expressão que nomeia o artigo da Wikipédia, geralmente ocorre no início do texto;
- <*segmento_texto*> é o texto entre o artigo e um link, que representa a relação entre <*artigo*> e <*link*>. As palavras são representadas em sua forma canônica, por exemplo, a palavra *"é"* trocado pela forma canônica do verbo *"ser"*;
- <*link*> é a expressão que é marcada como um link para outro artigo da Wikipédia;

Seguindo o mesmo exemplo do artigo "*Agricultura*", ao final tem-se as seguintes sentenças:

- <*agricultura*> <*ser conjunto de*> <*técnica*>
- <*agricultura*> <*utilizar para cultivar*> <*planta*>
- <*agricultura*> <*objetivo de obter*> <*alimentos*>
- <*agricultura*> <*objetivo de obter*> <*fibra*>
- <*agricultura*> <*objetivo de obter*> <*energia*>
- <*agricultura*> <*objetivo de obter*> <*matéria-prima*>
- <*agricultura*> <*para*> <*roupas*>
- <*agricultura*> <*para*> <*construção*>
- <*agricultura*> <*para*> <*medicamentos*>
- <*agricultura*> <*para*> <*ferramentas*>
- <*agricultura*> <*para contemplação*> <*estética*>

Na última atividade desse modulo, realiza-se a seleção de sentenças candidatas para extração de relações semânticas. Para a realização dessa seleção, são aplicados os seguintes critérios :

- Existência de verbos no <segmento_texto> - a obrigatoriedade de verbos em <*segmento_texto*> deve-se ao argumento que um verbo denota uma relação significativa entre os conceitos. Por exemplo, *(assassinato, éUm(a), crime)*.
- Existência de substantivos comuns, advérbios, adjetivos, ou preposições em <*segmento_texto*> - além do verbo o <*segmento_texto*> pode conter outros elementos para enriquecer a relação.

- Número limite de termos em <*segmento_texto*> - a restrição do número limite de termos advém do fato de que <*segmento_texto*> com um número excessivo de termos, por exemplo 10 (dez), indica uma complexidade estrutural na sentença, indicando uma relação muito específica entre conceitos. Por exemplo, a sentença extraída "<*Arqueologia*> <incluir em campo de estudo intervenção fazer por homem em> <*meio ambiente*>" é complexa demais para ser classificada.

- <*artigo*> e/ou <*link*> não sejam nomes próprios - a restrição de nomes próprios é devido ao fato de que conceitos são, em geral, expressos por nomes comuns, e nomes próprios ou entidades nomeadas são instâncias de conceitos. Para identificação de nomes próprios, adotamos a heurística apresentada na figura 4.7, segundo a qual um artigo ou link é um nome próprio se, na maioria das vezes em que o mesmo é citado na Wikipédia, usa-se letra inicial maiúscula.

Vale a pena salientar que este conjunto de restrições pode variar de acordo com o objetivo da extração de informação. Por exemplo, caso o usuário queira extrair conteúdo restrito a o domínio da política, a aplicação do último critério não é adequada.

No final dessa etapa, temos um conjunto de sentenças candidatas, extraídas de um conjunto de documentos de entrada. No exemplo do artigo "Agricultura", as sentenças candidatas são:

- <*agricultura*> <*ser conjunto de*> <*técnica*>
- <*agricultura*> <*utilizar para cultivar*> <*planta*>
- <*agricultura*> <*objetivo de obter*> <*alimentos*>
- <*agricultura*> <*objetivo de obter*> <*fibra*>
- <*agricultura*> <*objetivo de obter*> <*energia*>
- <*agricultura*> <*objetivo de obter*> <*matéria-prima*>

4.3.3.2 Clusterização das Sentenças

O objetivo deste módulo é identificar grupos de relações similares usando algoritmos de clusterização [Veyssieres e Plant 1998]. Nosso argumento é que muitos tipos de relações semânticas são expressas por <*segmento_texto*> similares e devem ser adquiridas como relações semânticas de mesmo tipo. Por exemplo, nas sentenças abaixo, apesar dos <*segmento_texto*> serem diferentes eles expressam uma mesma relação semântica.

- "<*mastigação*> <*obter*> <*bolo alimentar*>"
- "<*célula*> <*adquirir*> <*herança genética*>"

Para identificar relações similares utilizamos duas medidas de similaridade entre dois <*segmento_texto*> s_1 e s_2.

A primeira é uma medida sintática (1) $Sin(s_1, s_2)$, que calcula a porcentagem de termos idênticos em s_1 e s_2, descartando-se as preposições.

$$(1)\ Sin(s_1, s_2) = \frac{\sum_{i=1}^{n}(\mu t_j)}{n}$$

Onde:

- μt_j é o valor de similaridade sintática. Esse é calculado a partir da verificação do conjunto de termos da sentença. Caso dois termos sejam iguais, μt_j será 1, senão μt_j será 0. Os termos são representados em T_1 e T_2, onde:
  - T_1 é o conjunto de termos de s_1, descartando-se as preposições;
  - T_2 é o conjunto de termos de s_2, descartando-se as preposições;
  - $t_j = (t_1, t_2)$, tal que $t_1 \in T_1$ e $t_2 \in T_2$;
- n é a quantidade total de elementos em $T_1 x T_2$.

Por exemplo, para s_1=<*transmitir por*> e s_2=<*transmitir*>, como a preposição "por" é descartada de s_1, temos que $Sin(s_1, s_2) = 1$ (ou 100%).

A segunda medida utilizada é uma medida de similaridade semântica $Sem(s_1, s_2)$ (2), que define o quão dois segmentos de textos são semanticamente similares entre textos [Franco et al. 2013]. Esta medida é calculada pela média ponderada do somatório das similaridades semânticas entre conceitos de s_1 e s_2. A medida proposta é independente de qualquer medida de similaridade semântica entre conceitos. Esta independência é interessante devido ao método proposto poder ser aplicado a qualquer base de conhecimento.

$$(2)\ Sem(s_1, s_2) = \frac{\sum_{i=1}^{n}(\sum_{j=1}^{qi} \theta t_j) * P_i}{\sum_{n}^{i=1} q_i * P_i}$$

Onde:

- θt_j é o valor de similaridade semântica entre os conceitos representados pelos termos de T_1 e T_2, onde:
  - T_1 é o conjunto de termos de s_1;
  - T_2 é o conjunto de termos de s_2;
  - $t_j = (t_1, t_2)$, tal que $t_1 \in T_1$ e $t_2 \in T_2$;
  - $T_1 x T_2$: produto cartesiano entre termos da mesma classe gramatical (substantivo x substantivo, verbo x verbo) de T_1 e T_2;

- q_i é a quantidade de elementos de cada classe gramatical em $T_1 x T_2$. Este valor significa a quantidade de comparações a realizar, conforme a abordagem definida para a criação do conjunto T_1 x T_2.
- P_i é o peso da i-ésima classe gramatical.
- n é a quantidade de classes gramaticais em $T_1 x T_2$.

Para exemplificar o cálculo de similaridade semântica , seja o seguinte par de sentença candidata "<*agricultura*> <*ser conjunto de*> <*tecnica*>" e "<*agricultura*> <*utilizar para cultivar*> <*planta*>", com os segmentos de texto s_1 = <ser conjunto de> e s_2 = <utilizar para cultivar>.

As preposições são desconsideradas Na definição do conjunto T_1 e T_2. Assumimos que o peso para a classe dos substantivo é 1 e para a classe dos verbos é 3 já que verbos importam maior significado para as relações semânticas. Portanto, para o caso do exemplo, temos os seguintes conjuntos e valores:

- T_1 = ser(V), conjunto(S)[4]
- T_2 = utilizar(V), cultivar(V)[5]
- T_1 x T_2 = (ser-utilizar), (ser-cultivar)
- $P_{verbo} = 3$
- $P_{substantivo} = 1$
- $q_{verbo} = 2$
- $q_{substantivo} = 0$
- $n = 1$
- $\theta(ser, utilizar) = 40\%$ e $\theta(ser, cultivar) = 30\%$[6].

Aplicando os valores acima em (2), tem-se que

(2) $Sem(s_1, s_2) = \frac{\sum_{i=1}^{1}(\sum_{j=1}^{2}\theta t_j)*3}{3}$

$Sem(s_1, s_2) = \theta(ser, utilizar) + \theta(ser, cultivar) = 70\%$

O algoritmo de clusterização inicialmente tenta agrupar <*segmento_texto*> utilizando a medida de similaridade semântica. Para isto é necessário como paramêtro um valor de

[4] S = Substantivo
[5] V = Verbo
[6] Valores hipotéticos são utilizados nesse exemplo

corte (V_C) que representa a porcentagem de similaridade desejada no agrupamento. Caso não exista similaridade semântica, que atenda a restrição, seja maior que V_C o algoritmo utiliza a medida de similaridade sintática. Por fim, o algoritmo, agrupa pares de conceitos c_1 e c_2 relacionados por <*segmento_texto*>. Na figura 4.15 é apresentado o algoritmo de clusterização proposto neste trabalho.

```
cluster [] retornaSentencas(sentencas[]) {
cluster[] = new cluster ()
para n de 0 ate sentenças[].tamanho faca
    para n+1 de 0 ate sentenças[].tamanho -1 faca
        //caso a similaridade semântica seja maior que Valor_Corte
        if(Sem(sentencas[n].relacao,sentencas[n+1].relacao) > V_corte)
            cluster (sentencas[n],sentencas[n+1])
        //caso exista similaridade sintática
        else if(simSin(sentencas[n].relacao,sentencas[n+1] .relacao))
            cluster (sentencas[n],sentencas[n+1])
retorne relações }
```

Figura 4.15: Algoritmo de clusterização de sentenças.

Com essa estratégia de clusterização, conseguimos aumentar o numero de relações adquiridas por que aumentamos o número de relações em cada grupo de relações, tornando-os mais significativos. Um exemplo hipotético e simples ajuda a entender nosso argumento. Na figura 4.16, temos duas relações *obter* e *adquirir* que, apesar de similares, possuem pares de conceitos distintos. Com a clusterização conseguimos agrupar as relações *obter* e *adquirir*.

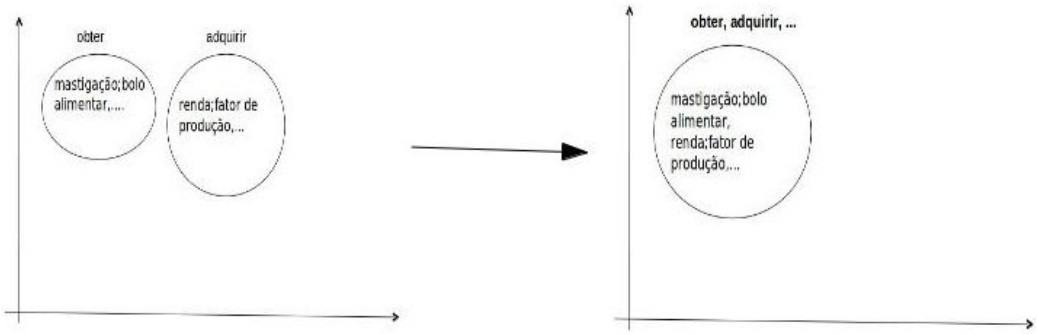


Figura 4.16: Clusterização com as relações “obter” e “adquirir”

4.3.3.3 Aquisição de Relações Semânticas

No último módulo do método proposto, o objetivo é a aquisição final de relações semânticas da forma *(c_1, relacao, c_2)* para uma base de conhecimento *B*. A entrada desta etapa são um conjunto de *clusters* Cl_i, cada um sendo identificado com um conjunto S_i de <*segmento_texto*> similares e contendo *n* pares de conceitos (c_k, c_l) relacionados. Por exemplo, em $cl_1 = (obter; para\_obter; adquirir)$ temos os seguintes pares de conceitos *{ (mastigação;bolo alimentar), (cartel;lucro), (renda;fator de produção),* $\cdots$ *}*. Inicialmente, somente *clusters* com mais de *k* pares de conceitos são pré-selecionados, pois representam tipos de relações relevantes no corpus analisado. Em [Fader, Soderland e Etzioni 2011], experimentos mostraram que $k = 20$ é um valor ótimo para eliminar *clusters* pouco relevantes.

O próximo passo é definir a $relacao_i$ que melhor representa o conjunto S_i de cada cluster $cl_i \in Cl_i$. Como dito, o método proposto é independente da base de conhecimento *B*, podendo, por exemplo, ser aplicado para bases como ConceptNet [Speer e Havasi 2012], WordNet [Miller 1995] e InferenceNet [Pinheiro et al. 2010]. Todas estas propõem um conjunto finito e bem controlado de tipos de relações semânticas. Por exemplo, a ConceptNet 5.0 [Speer e Havasi 2012] contém 54 tipos de relações, tais como: *partOf, locationOf, motivationOf, etc.*

O conjunto de tipos de relações pré-definidas de *B* é utilizado como parâmetro desta atividade. Para o conjunto S_i de <*segmento_texto*> similares, é selecionado aquele $s_j \in S_i$ com maior valor de similaridade semântica com os demais s_k e com os tipos de relações pré-definidas de B. Ao final, para cada cluster $cl_i \in Cl_i$, é definido $relacao_i = s_j$.

Exemplificando, seja considerado o seguinte cluster <*localizar entre;localizar em; localizar a;localizar;estar localizar em*>. O valor que tem uma maior similaridade comparado com outros elementos é *localizar*. Logo esse elemento é escolhido para representar esse grupo, o qual equivale a relação *locationOf*. Essa relação está presente nas bases InferenceNet e na ConceptNet.

Por último, são geradas as relações semânticas *(c_k, $relacao_i$, c_l)* para cada par de conceitos (c_k,c_l), contido em cada $cl_i \in Cl_i$ que tenha numero de pares de conceitos $\geq k$ (definido pelo usuário). Podemos citar como exemplos de relações extraídas: *(pólen, locationOf, gâmeta),(delta, partOf, rio), (banheira, locationOf, banheiro)* etc.

4.3.4 Avaliação do Método

A avaliação realizada visa analisar dois aspectos fundamentais do método. O primeiro foi a qualidade do conteúdo extraído pelo método proposto, ou seja, se as relações semânticas extraídas de fato contribuirão para evolução de bases de conhecimento. O segundo aspecto é se a identificação de relações redundantes otimiza a quantidade de relações extraídas.

Para esta avaliação, o método foi implementado em um protótipo na linguagem de programação Java[7] e aplicado para aquisição de relações para a base InferenceNet [Pinheiro et al. 2010]. Foi utilizado o parser FreeLing 3.0 [Padró e Stanilovsky 2012]. Escolhemos o

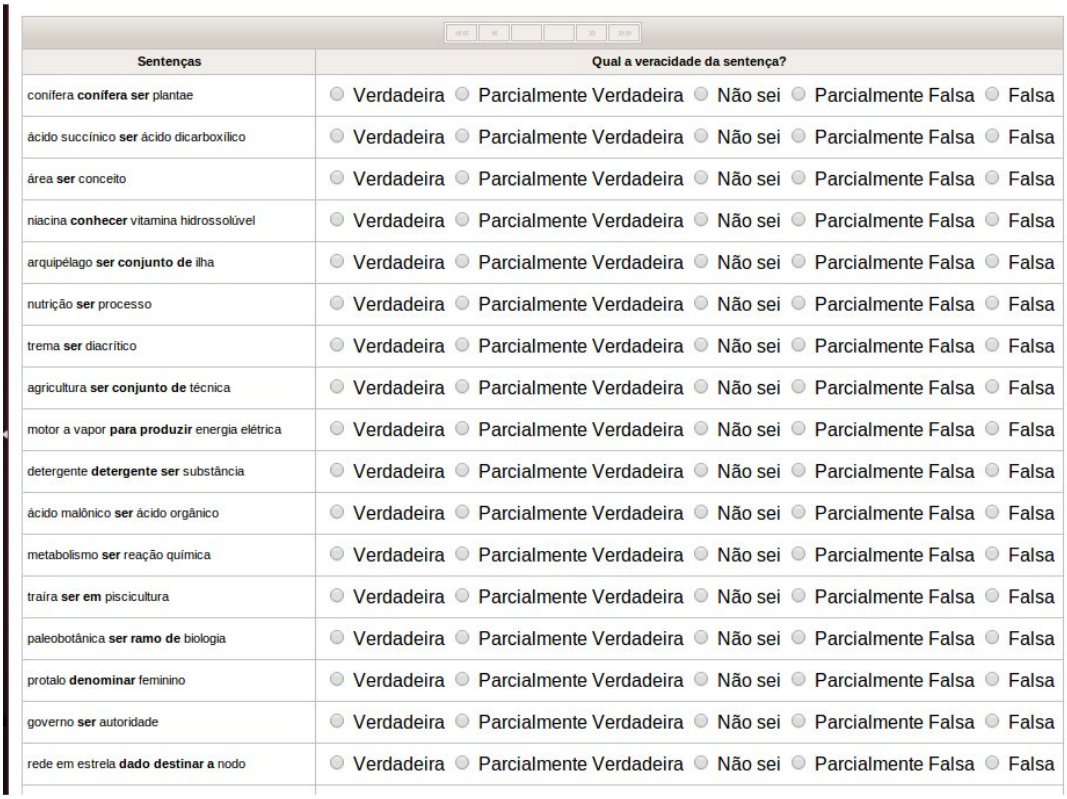
[7] http://www.oracle.com/technetwork/pt/java/javase/downloads/index.html

FreeLing para os nossos experimentos devido a sua alta acurácia e a facilidade de utilização em nossos experimentos. No entanto, o método pode ser aplicado para outras bases de conhecimento e pode ser utilizado outro analisador morfossintático para língua portuguesa. Para o processamento da Wikipédia em português utilizou-se a ferramenta WikipediaMiner [Milne e Witten 2008].

4.3.4.1 Metodologia de Avaliação

A metodologia de avaliação contemplou os seguintes passos:

1. Seleção aleatória de 100 mil artigos da Wikipédia;
2. Execução do método proposto para aquisição de relações para base InferenceNet. Os dados gerados, parâmetros, regras e ferramentas utilizadas são apresentados na tabela 4.7. Por exemplo, somente o primeiro parágrafo é analisado, são selecionadas sentenças com V, N e PREP e com no máximo $\leq$ três termos.
3. Desenvolvimento de uma aplicação Web que seleciona aleatoriamente 20 relações semânticas dentre as relações adquiridas neste experimento. As relações eram apresentadas em linguagem natural para facilitar a leitura de avaliadores.
4. Avaliação humana das relações semânticas extraídas. Avaliadores humanos são convidados a participar via e-mail ou redes sociais. Os avaliadores tinham de 20 a 50 anos entre homens e mulheres e foi explicado a eles do que se tratava a avaliação. Para completar a avaliação eles deveriam classificar as sentenças quanto a veracidade das mesmas: “Verdadeira”, “Parcialmente Verdadeira”, “Eu não sei”, “Vaga ou Parcialmente Falsa”, “Falsa”. Na Figura 4.17 temos *screenshot* de um exemplo de questionário aplicado aos avaliadores.



Sentenças	Qual a veracidade da sentença?
conífera **conífera ser** plantae	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
ácido succínico **ser** ácido dicarboxílico	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
área **ser** conceito	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
niacina **conhecer** vitamina hidrossolúvel	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
arquipélago **ser conjunto de** ilha	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
nutrição **ser** processo	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
trema **ser** diacrítico	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
agricultura **ser conjunto de** técnica	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
motor a vapor **para produzir** energia elétrica	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
detergente **detergente ser** substância	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
ácido malônico **ser** ácido orgânico	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
metabolismo **ser** reação química	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
traíra **ser em** piscicultura	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
paleobotânica **ser ramo de** biologia	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
protalo **denominar** feminino	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
governo **ser** autoridade	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa
rede em estrela **dado destinar a** nodo	Verdadeira Parcialmente Verdadeira Não sei Parcialmente Falsa Falsa

Figura 4.17: Screenshot da aplicação Web com questionário aplicado aos avaliadores.

5. Resultados foram coletados e analisados, conforme apresentado na próxima seção.

A tabela 4.7 apresenta o resultado de cada etapa do método. Na primeira etapa, para um conjunto de 100 mil artigos e utilizando regras, como: seleção do primeiro paragrafo, exclusão de sentenças com argumentos maiores de 3 palavras, etc., foram selecionadas um total de 21.802 sentenças. Na segunda etapa aplicamos o algoritmo de clusterização que agrupou as sentenças em 287 *clusters*. Por fim, aplicamos de *clusters* com 20 elementos, foram adqueridas 12.362 relações.

Etapa do método	Entrada	Saída	Parâmetros
Mineração e Seleção de Sentenças	100.000 artigos da Wikipédia	143.395 sentenças 33.985 sentenças 21.802 sentenças	- Somente 1º.parágrafo analisado - Parser FreeLing - Seleção de sentenças com V,N e PREP - Artigos/links de nomes comuns - Sentenças com $\leq$ 3 termos
Clusterização de Sentenças	21.802 sentenças candidatas	287 *clusters*	- SimSem da InferenceNet [Franco et al. 2013]
Aquisição de Relações Semânticas	287 *clusters*	64 *clusters* 12.362 relações	- Seleção de *clusters* com $\geq$20 elementos

Tabela 4.7: Dados gerados, parâmetros, regras e ferramentas utilizadas em cada etapa do método.

4.3.4.2 Análise dos Resultados

Após 48h com a aplicação Web disponível, 60 avaliadores julgaram 1.395 relações semânticas distintas e 295 relações em duplicidade. Foram computados o somatório de relações avaliadas em cada classe e 76% das relações foram avaliadas como "verdadeiras" ou "parcialmente verdadeiras". Neste cômputo foram desconsideradas as respostas da classe "Não sei". As 295 relações que foram avaliadas por mais de um humano foram contabilizadas com o pior valor de avaliação. Apesar de não ter sido aplicado em *corpus* equivalentes, consideramos este resultado promissor, pois é equiparável ao estado da arte. Os resultados do Reverb são diversos. Em [Fader, Soderland e Etzioni 2011] mostra 86% de aprovação. Nos trabalhos de [Gamallo, Garcia e Fernández-Lanza 2012] e [Xavier, Lima e Souza 2013] o Reverb teve 52% e 49% de precisão respectivamente. Creditamos esta diferença a diferenças na Coleção Dourada de avaliação. Outro resultado importante foi que apenas 20% das relações extraídas pelo o Reverb para a ConceptNet foram aprovadas por avaliadores [Speer e Havasi 2012].

Além disso, o método aqui proposto eliminou uma das principais causas de incorretude do ReVerb - a identificação incorreta dos argumentos de uma relação. Isto se deve ao uso da estrutura de links ao longo do texto de um artigo da Wikipédia, pois estes indiciam quais conceitos estão sendo, de fato, relacionados pelo texto.

4.3.4.3 Análise da Etapa de Clusterização de Sentenças

A nossa principal motivação para a utilização de *clusters* no processo de aquisição de relações semânticas é que esse mecanismo pode aumentar o número de relações adquiridas. Para verificar a veracidade desse argumento realizamos o seguinte experimento.

Foram gerados *clusters* somente pela igualdade sintática dos <*segmento_texto*>, ou seja, quando os <*segmento_texto*> eram idênticos. Esta abordagem é a mesma utilizada pelo Reverb. Neste caso, foram gerados 881 *clusters* e 94 destes com mais de 20 pares de conceitos

(argumentos) relacionados. Neste cenário, apenas 5.312 relações semânticas seriam extraídas e geradas para a base de conhecimento InferenceNet. A estratégia proposta neste trabalho, que utiliza a similaridade semântica para identificar tipos de relações redundantes, possibilita que um número menor de *clusters* seja formado e com mais elementos, adquirindo, assim, um número maior de relações semânticas (12.362, conforme tabela 4.7.)

4.4 Conclusão

Neste capitulo, foram apresentados dois métodos para aquisição de conhecimento de mundo para bases semânticas. O primeiro trata-se de uma abordagem semiautomática que utiliza o conhecimento existente na base para ajudar o usuário a externar novas relações. A principal vantagem desse método é deixar o processo menos oneroso, além evitar problemas de inconsistência na inclusão de novas relações. A principal desvantagem desse método é a necessidade de existir conteúdo pré-existente para ser utilizado.

Na segunda abordagem, a automática, utilizamos a estrutura de links da Wikipédia para fazer a extração de relações semânticas. As principais vantagens desse método são: a redução de relações semânticas redundantes; consistência nos argumentos das relações já que são formados a partir dos links nos artigos da Wikipédia; a não utilização de expressões regulares; independência de língua e a não da utilização de um corpus para treinamento do método. Algumas desvantagens são a dependência da Wikipédia e utilização de links para definição dos argumentos das relações, pois alguns artigos não tem links no corpo do seu texto como o mostrado na figura 4.18.

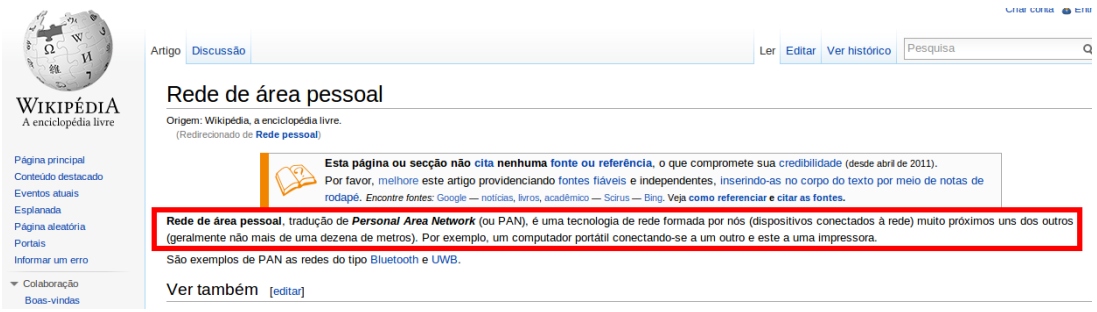


Figura 4.18: Screenshot do artigo “Rede Pessoal”.

5 CONCLUSÃO

Um dos desafios das pesquisas na área de Processamento de Linguagem Natural (PLN) é prover recursos semântico-linguísticos que expressem conhecimento de mundo para suportar o entendimento de linguagem natural por máquinas. O principal objetivo de um sistema de NLU é interpretar um texto para que o conteúdo expresso sirva de base para realizar tarefas concretas, tais como extração de informação, recuperação de informação, sistemas de perguntas e respostas, sumarização de textos, análise de sentimentos.

Existem atualmente diversos recursos semânticos-linguísticos que expressam conhecimento de mundo, como por exemplo: WordNet, FrameNet, ConceptNet e InferenceNet. No entanto, existe a necessidade da evolução do conteúdo desses recursos de maneira rápida e consistente. O conhecimento de mundo é dinâmico e métodos de aquisição de conhecimento devem suportar a evolução de tais recursos.

Após investigação dos métodos de aquisição de conhecimento automáticos comumente usados em PLN identificamos que:

- Semiautomáticos - utilizam questionários ou se baseiam nas interações dos usuários com o sistema, o que pode deixar o processo de AC oneroso e gerar inconsistências nas bases;
- Automáticos - dependência de corpus abrangente para treinar os modelos de aprendizado de máquina, dificuldade de encontrar expressões regulares que recuperem as relações semânticas encontradas em uma língua, e dificuldade em descobrir os argumentos de uma relação.

Com base nestes problemas, nossas hipóteses de pesquisas são:

- O conhecimento preexistente em uma base auxilia o usuário a explicitar e a validar relações semânticas para um novo conceito, otimizando o processo interativo de aquisição de conhecimento de mundo.
- Textos de artigos da Wikipédia são uma importante e confiável fonte de conhecimento de mundo e sua semi-estrutura, provida pelos hyperlinks entre artigos, delimitam relações semânticas entre conceitos.

Neste trabalho, propomos dois métodos para aquisição de conhecimento de mundo. O diferencial do primeiro método em relação ao estado da arte é um processo automático de raciocínio que gera novos fatos de senso comum e pragmáticos para conceitos da língua portuguesa, a partir de conteúdo de outros conceitos similares e conforme a estrutura gramatical de sintagmas nominais. Um conjunto de heurísticas foi proposta e a interação com o usuário final permite uma validação das relações semânticas geradas e, consequentemente, melhora a qualidade na aquisição de conhecimento dessa natureza. O método foi implementado e avaliado para a base de conhecimento de mundo da língua portuguesa - InferenceNet, e obteve 72% de acurácia. Além disso, o método proposto possibilitou interações mais produtivas para AC, pois,

com uma base inicial para validação, o usuário é instigado sobre relações semânticas de senso comum acerca do novo conceito.

O segundo método propõe a aquisição de relações semânticas entre conceitos a partir do texto de documentos da Wikipédia, aproveitando-se dos links entre artigos para identificar os argumentos da relação. Adicionalmente, prescinde da definição prévia de expressões regulares ou de um processo oneroso de anotação de corpus. Visando a otimização da aquisição de conhecimento, identifica tipos de relações similares através de uma medida de similaridade semântica. O método foi utilizado em um corpus da Wikipédia em português de 100 mil artigos e 12.632 relações semânticas foram geradas para a base de conhecimento InferenceNet. Um grupo de 80 avaliadores humanos analisou a veracidade de 1395 relações semânticas, selecionadas aleatoriamente. Os resultados obtidos indicaram que a acurácia do método é de 76%, superior ao estado da arte. Além disso, a etapa de clusterização de tipos de relações por similaridade semântica permitiu a aquisição de um maior número de relações.

5.1 Contribuições da Pesquisa

A seguir listamos algumas contribuições da nossa pesquisa:

- Foi elaborado um algoritmo para aquisição de conhecimento de mundo a partir da estrutura gramatical de sentenças em língua portuguesa. Nosso algoritmo utiliza estratégias baseadas em sintagmas nominais para geração de conteúdo para um novo conceito;
- Definição de estratégias para extração de conhecimento a partir dos textos e links da Wikipédia, independentes de expressões regulares e padrões sintáticos previamente definidos;
- Desenvolvimento de um algoritmo para aquisição automática de conhecimento a partir da Wikipédia;
- Definição de uma estratégia de clusterização de relações baseadas em uma medida de similaridade semântica. Esta estratégia tem como objetivo aumentar o número de relações semânticas extraídas.

5.2 Trabalhos Futuros

Uma pesquisa não deve somente ser avaliada pelos seus resultados imediatos, mas pelas questões que propõe para a comunidade científica. Esta investigação lança vários desafios que detalhamos a seguir, os quais são bons indicativos de pesquisas futuras.

Neste sentido, apresentamos a seguir as motivações para investigações futuras suscitadas com base neste trabalho de pesquisa:

- Foram realizadas análises qualitativas de partes amostrais do conhecimento adquirido para a base InferenceNet, com a execução dos métodos propostos. No entanto, se faz necessário uma avaliação extrínseca do conhecimento através do uso desta base em tarefas

de PLN como Extração de Informação, Desambiguação do Sentido de Palavras, Resolução de Correferência, Análise de Sentimento, etc. Um forma de realizar está avaliação é executar o respectivo processador com base InferenceNet antes e depois do processo de aquisição podendo ser definido um escopo de domínio para um melhor controle do experimento.

- O método de aquisição automática de relações semânticas apresentou problemas em textos complexos entre links de artigos ou no caso de relações com maior aridade. É necessário o estudo de estruturas gramaticais e conhecimento semântico para evoluir o algoritmo proposto.

REFERÊNCIAS BIBLIOGRÁFICAS

AHN, L. Von; KEDIA, M.; BLUM, M. Verbosity: a game for collecting common-sense facts. In: ACM. *Proceedings of the SIGCHI conference on Human Factors in computing systems*. [S.l.], 2006. p. 75–78.

ALLEN, J. *Natural language understanding*. [S.l.]: Benjamin/Cummings Menlo Park, CA, 1987.

ANACLETO, J. et al. A common sense-based on-line assistant for training employees. *Human Computer Interaction-INTERACT 2007*, Springer, p. 243–254, 2007.

ANACLETO, J. et al. Can Common Sense Uncover Cultural Differences in Computer Applications? *Artificial Intelligence in Theory and Practice*, Springer, p. 1–10, 2006.

AUER, S. et al. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, Springer, p. 722–735, 2007.

BAADER, F. et al. (Ed.). *The Description Logic Handbook: Theory, Implementation, and Applications*. New York, NY, USA: Cambridge University Press, 2003. ISBN 0-521-78176-0.

BAKER, C.; ELLSWORTH, M.; ERK, K. SemEval'07 task 19: frame semantic structure extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 4th International Workshop on Semantic Evaluations*. [S.l.], 2007. p. 99–104.

BAKER, C.; FILLMORE, C.; LOWE, J. The Berkeley Framenet Project. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. [S.l.], 1998. p. 86–90.

BECKETT, D.; MCBRIDE, B. RDF/XML syntax specification (revised). *W3C recommendation*, v. 10, 2004.

BICK, E. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. [S.l.]: Aarhus University Press Aarhus, Denmark, 2000.

BOBROW, D. G. Natural Language Input for a Computer Problem Solving System. 1964.

BOLLACKER, K. et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: ACM. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. [S.l.], 2008. p. 1247–1250.

BONTCHEVA, K. et al. Twitie: An open-source information extraction pipeline for microblog text. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics*. [S.l.: s.n.], 2013.

BRANDOM, R. *Articulating Reasons: An Introduction to Inferentialism*. [S.l.]: Harvard University Press, 2001.

CANKAYA, H.; MOLDOVAN, D. Method for extracting commonsense knowledge. In: ACM. *Proceedings of the Fifth International Conference on Knowledge Capture*. [S.l.], 2009. p. 57–64.

CHE, W. et al. Multilingual dependency-based syntactic and semantic parsing. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. [S.l.], 2009. p. 49–54.

COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, ACM, v. 39, n. 1, p. 80–91, 1996.

FADER, A.; SODERLAND, S.; ETZIONI, O. Identifying Relations for Open Information Extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2011. p. 1535–1545.

FOGAROLLI, A. Wikipedia as a Source of Ontological Knowledge: State of the Art and Application. *Intelligent Networking, Collaborative Systems and Applications*, Springer, p. 1–26, 2011.

FRALEY, C.; RAFTERY, A. E. How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis. *Technical Report*, Department of Statistics University of Washington, v. 41, n. 329, p. 578–588, 1998.

FRANCO, W. et al. Aquisição de relações semânticas a partir de textos da wikipédia. 2013.

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. Dependency-based Open Information Extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*. [S.l.], 2012. p. 10–18.

GILDEA, D.; JURAFSKY, D. Automatic Labeling of Semantic Roles. *Computational linguistics*, MIT Press, v. 28, n. 3, p. 245–288, 2002.

GONÇALVES, E. M. N. et al. Uma abordagem para especificação de conhecimento para sistemas multiagentes cognitivos. Florianópolis, SC, 2012.

HAVASI, C.; SPEER, R.; ALONSO, J. ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: *Recent Advances in Natural Language Processing*. [S.l.: s.n.], 2007. p. 27–29.

KAY, M. Introduction to Computational Linguistics. *Mitkov, R.(ed). The Oxford Handbook of Computational Linguistics*, v. 30, n. 1, p. 17–22, 2003.

KIPPER, K.; DANG, H. T.; PALMER, M. Class-based construction of a verb lexicon. In: *AAAI*. [S.l.: s.n.], 2000. p. 691–696.

KITTUR, A.; KRAUT, R. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In: ACM. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. [S.l.], 2008. p. 37–46.

KOHL, K. et al. Representing Verb Alternations in Wordnet. *WordNet. An Electronic Lexical Database*, p. 153–178, 1998.

LEE, G. G. et al. Siteq: Engineering high performance qa system using lexico-semantic pattern matching and shallow nlp. In: *TREC*. [S.l.: s.n.], 2001.

LEHMANN, J. et al. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2013.

LEMLE, M. *Análise sintática: teoria geral e descrição do português*. [S.l.]: Editora Atica, 1984.

LENAT, D. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, ACM, v. 38, n. 11, p. 33–38, 1995.

LEUF, B.; CUNNINGHAM, W. The wiki way: quick collaboration on the web. Addison-Wesley Professional, 2001.

LI, H. et al. Using Graph Based Method to Improve Bootstrapping Relation Extraction. In: *Computational Linguistics and Intelligent Text Processing*. [S.l.]: Springer, 2011. p. 127–138.

LIU, H.; SINGH, P. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, v. 22, n. 4, p. 211–226, 2004. Disponível em: <http://publication.wilsonwong.me>.

MARRAFA, P. et al. WordNet.Pt - Uma Rede Léxico-conceptual do português on-line. In: *XXI Encontro da Associação Portuguesa de Linguística, Porto, Portugal*. [S.l.: s.n.], 2005.

MILLER, G. WordNet: a Lexical Database for English. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995.

MILNE David; WITTEN Ian H. Learning to Link with Wikipedia. In: *CIKM*. [S.l.: s.n.], 2008. p. 509–518.

MINGHELLI, T. D.; BERTOLDI, A.; CHISHMAN, R. O subframe sentença no complexo frame processo de conhecimento no direito processual civil. 2013.

MINSKY, M. A Framework for Representing Knowledge. 1974.

MINSKY, M. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. [S.l.]: Simon and Schuster, 2007.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, John Benjamins publishing company, v. 30, n. 1, p. 3–26, 2007.

NAKAYAMA, K.; HARA, T.; NISHIO, S. A Thesaurus Construction Method from Large Scaleweb Dictionaries. In: IEEE. *Advanced Information Networking and Applications, 2007. AINA'07. 21st International Conference on*. [S.l.], 2007. p. 932–939.

NASTASE, V. et al. Wikinet: A Very Large Scale Multi-lingual Concept Network. In: *Proceedings of International Conference on Language Resources and Evaluation (LREC)*. [S.l.: s.n.], 2010. v. 10.

OVCHINNIKOVA, E. *Integration of World Knowledge for Natural Language Understanding*. [S.l.]: Springer, 2012. 15 p.

PADRÓ, L.; STANILOVSKY, E. Freeling 3.0: Towards Wider Multilinguality. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. *Proceedings of Language Resources and Evaluation (LREC)*. [S.l.], 2012.

PAIVA, V. d.; RADEMAKER, A.; MELO, G. d. Openwordnet-pt: An open brazilian wordnet for reasoning. COLING 2012, 2012.

PARDO, T.; CASELI, H.; NUNES, M. Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais. In: *The Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology-STIL*. [S.l.: s.n.], 2009. p. 1–21.

PEI, M. et al. Constructing a Global Ontology by Concept Mapping Using Wikipedia Thesaurus. In: IEEE. *Advanced Information Networking and Applications-Workshops, 2008. AINAW 2008. 22nd International Conference on*. [S.l.], 2008. p. 1205–1210.

PINHEIRO, V. *SIM: Um Modelo Semântico Inferencialista para Expressão e Raciocínio em Sistemas de Linguagem Natural. Phd Thesis, Universidade Federal do Ceará.*

PINHEIRO, V. et al. Aquisição de conhecimento de senso comum e inferencialista. 2011.

PINHEIRO, V. et al. Towards a common sense base in portuguese for the linked open data cloud. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer, 2012. p. 128–138.

PINHEIRO, V. et al. A Semi-Automated Method for Acquisition of Common-sense and Inferentialist Knowledge. *Journal of the Brazilian Computer Society*, Springer, p. 1–13, 2013.

PINHEIRO, V. et al. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. In: *PROPOR*. [S.l.]: Springer, 2010. (Lecture Notes in Computer Science, v. 6001), p. 90–99. ISBN 978-3-642-12319-1.

PUSTEJOVSKY, J. The Generative Lexicon. *Computational linguistics*, MIT press, v. 17, n. 4, p. 409–441, 1991.

RUPPENHOFER, J. et al. *FrameNet II: Extended Theory and Practice*. 2006.

SALOMÃO, M. FrameNet Brasil: Um Trabalho em Progresso. *Calidoscópio*, p. 171–182, 2009.

SCHANK, R. C. *Conceptual information processing*. [S.l.]: Elsevier Science Inc., 1975.

SCOTT, A.; CLAYTON, J.; GIBSON, E. *A practical guide to knowledge acquisition*. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1991.

SILVA, B. C. D. da et al. Introdução ao processamento das línguas naturais e algumas aplicações. 2007.

SILVA, B. D. da; FELIPPO, A. D.; HASEGAWA, R. Methods and Tools for Encoding the WordNet.Br sentences, Concept Glosses, and Conceptual-Semantic Relations. In: *PROPOR*. [S.l.: s.n.], 2006. v. 3960, p. 120–130.

SILVA, C.; RIBEIRO, B. The Importance of Stop Word Removal on Recall Values in Text Categorization. In: IEEE. *Neural Networks, 2003. Proceedings of the International Joint Conference on*. [S.l.], 2003. v. 3, p. 1661–1666.

SILVA, J. W. Franco da. FACIL: Uma Ferramenta de AquisiÃ§Ã£o de Conhecimento Inferencialista. *Trabalho de Conclusão de Curso. UNIVERSIDADE ESTADUAL DO CEARÃ. Fortaleza, CE*, 2010.

SILVA, M.; KOCH, I. *Lingüística aplicada ao português: sintaxe*. [S.l.]: Cortez, 1989.

SINGH, P. et al. Open mind common sense: Knowledge acquisition from the general public. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, Springer, p. 1223 – 1237, 2002.

SOON, W. M.; NG, H. T.; LIM, D. C. Y. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, MIT Press, v. 27, n. 4, p. 521–544, 2001.

SPEER, R. Open Mind Commons: An inquisitive approach to learning common sense. In: *Workshop on Common Sense and Intelligent User Interfaces*. [S.l.: s.n.], 2007.

SPEER, R.; HAVASI, C. Representing general relational knowledge in Conceptnet 5. In: *International Conference on Language Resources and Evaluation (LREC)*. [S.l.: s.n.], 2012. p. 79–86.

SPEER, R. et al. An Interface for Targeted Collection of Common Sense Knowledge Using a Mixture Model. In: ACM. *Proceedings of the 14th international conference on Intelligent user interfaces*. [S.l.], 2009. p. 137–146.

STEVENSON, M.; WILKS, Y. Word Sense Disambiguation. *The Oxford Handbook of Comp. Linguistics*, p. 249–265, 2003.

STOUTENBURG, S.; KALITA, J.; HAWTHORNE, S. Extracting Semantic Relationships between Wikipedia Articles. In: *Proc. 35th International Conference on Current Trends in Theory and Practice of Computer Science*. [S.l.: s.n.], 2009.

SUCHANEK, F.; KASNECI, G.; WEIKUM, G. Yago: A Large Ontology from Wikipedia and Wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 6, n. 3, p. 203–217, 2008.

SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. Yago: a core of semantic knowledge. In: ACM. *Proceedings of the 16th international conference on World Wide Web*. [S.l.], 2007. p. 697–706.

SYED, Z.; FININ, T. Unsupervised Techniques for Discovering Ontology Elements from Wikipedia Article Links. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. [S.l.], 2010. p. 78–86.

TABOADA, M.; ANTHONY, C.; VOLL, K. Methods for creating semantic orientation dictionaries. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy*. [S.l.: s.n.], 2006.

TRASK, R. L. *Dicionário de Linguagem e Linguística*. [S.l.: s.n.], 2004.

TRYON, R. C.; BAILEY, D. E. *Cluster Analysis*. [S.l.]: New York:McGraw-Hill, 1970.

VEYSSIERES, M.; PLANT, R. E. Identification of Vegetation State and Transition Domains in CaliforniaÃ¢ÂÂs Hardwood Rangelands. *University of California*, 1998.

VIEIRA, R.; LIMA, V. L. Lingüística computacional: princípios e aplicações. In: *Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial*. [S.l.: s.n.], 2001. v. 3, p. 47–86.

VÖLKEL, M. et al. Semantic wikipedia. In: ACM. *Proceedings of the 15th international conference on World Wide Web*. [S.l.], 2006. p. 585–594.

WEIZENBAUM, J. Eliza: A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, ACM, v. 9, n. 1, p. 36–45, 1966.

WITBROCK, M. et al. An Interactive Dialogue System for Knowledge Acquisition in CYC. In: *Proceedings of the Workshop on Mixed-Initiative Intelligent Systems*. [S.l.: s.n.], 2003. p. 138–145.

WU, F.; WELD, D. S. Open Information Extraction using Wikipedia. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. [S.l.], 2010. p. 118–127.

XAVIER, C.; LIMA, V. de. A Semi-automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia Categories. *Advances in Artificial Intelligence–SBIA 2010*, Springer, p. 11–20, 2011.

XAVIER, C.; LIMA, V. L. de; SOUZA, M. Open Information Extraction Based on Lexical-Syntactic Patterns. In: SBC. *Proceedings of Brazilian Conference on Intelligent Systems*. [S.l.], 2013. [To appear].

XAVIER, C. C.; LIMA, V. L. S. de. A Method for Automatically Extracting Domain Semantic Networks from Wikipedia. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer, 2012. p. 93–98.

ZANG, L.-J. et al. A Survey of Commonsense Knowledge Acquisition. *Journal of Computer Science and Technology*, Springer, v. 28, n. 4, p. 689–719, 2013.

6 ANEXO A

Tabela 6.1: Baseline para "crime passional"

nomePre	papel	nomePos
crime passional	CapableOf	pagar
crime passional	CapableOf	não pagar
crime passional	CapableOf	ser cometido
crime passional	CapableOf	ter vítima
crime passional	CapableOf	envolver força ou fraude
crime passional	CapableOf	envolver violência
crime passional	CapableOf	acontecer na rua
crime passional	CapableOf	provocar disputa
crime passional	CapableOf	carregar penalidade dura
crime passional	CapableOf	pretender significar desejo
crime passional	CapableOfReceivingAction	aumento
crime passional	CapableOfReceivingAction	quebrar regra da sociedade
crime passional	CapableOfReceivingAction	cometer por pessoa
crime passional	CapableOfReceivingAction	cometer com arma
crime passional	CapableOfReceivingAction	medir como taxa
crime passional	CapableOfReceivingAction	evitar por polícia
crime passional	CapableOfReceivingAction	punir com penalidade
crime passional	CapableOfReceivingAction	informado por pessoa
crime passional	DefinedAs	entretenimento de bobo
crime passional	DefinedAs	forma do plural de crime
crime passional	DesirousEffectOf	conduzir o julgamento
crime passional	DesirousEffectOf	trabalho
crime passional	DesirousEffectOf	julgar a pessoa
crime passional	DesirousEffectOf	servir à justiça
crime passional	DesirousEffectOf	destruir inimigo da pessoa
crime passional	DesirousEffectOf	combater inimigo
crime passional	DesirousEffectOf	trazer veredicto
crime passional	DesirousEffectOf	proferir sentença
crime passional	DesirousEffectOf	iniciar guerra
crime passional	DesirousEffectOf	avancçar na batalha
crime passional	DesirousEffectOf	fazer parte do juri
crime passional	DesirousEffectOf	esmurrar a pessoa
crime passional	EffectOf	culpa
crime passional	EffectOf	sofrimento
crime passional	EffectOf	consciência culpada
crime passional	EffectOf	horrível
crime passional	EffectOf	retribuir
crime passional	FirstSubeventOf	escolher a vítima
crime passional	IsA	cometido por pessoa anormal
crime passional	IsA	problema social
crime passional	IsA	transgressão contra a sociedade
crime passional	IsA	violação da lei
crime passional	LastSubeventOf	ir preso
crime passional	LocationOf	metro
crime passional	LocationOf	cadeia
crime passional	LocationOf	demonstração
crime passional	MotivationOf	dinheiro
crime passional	MotivationOf	vingança

Tabela 6.2: Baseline para "violencia policial"

nomePre	papel	nomePos
violencia policial	CapableOf	ler
violencia policial	CapableOf	votar
violencia policial	CapableOf	ajudar a pessoa
violencia policial	PropertyOf	perigoso
violencia policial	CapableOf	pegar joe
violencia policial	DesireOf	não morrer
violencia policial	CapableOf	matar com rifle
violencia policial	CapableOfReceivingAction	treinar
violencia policial	CapableOfReceivingAction	cancelar
violencia policial	CapableOfReceivingAction	braço
violencia policial	CapableOf	relatar o crime
violencia policial	CapableOf	vestir o uniforme
violencia policial	CapableOf	correr através de ponte
violencia policial	CapableOf	ter revolver
violencia policial	UsedFor	assustar a pessoa
violencia policial	DesirousEffectOf	mover carro
violencia policial	CapableOf	proteger a pessoa
violencia policial	CapableOf	chorar
violencia policial	CapableOf	confirmar a lei
violencia policial	CapableOf	prisioneiro livre
violencia policial	CapableOf	usar o revolver
violencia policial	CapableOf	pegar criminoso
violencia policial	CapableOf	gostar de pessoa
violencia policial	CapableOf	reforçar a lei
violencia policial	CapableOf	ser criminoso
violencia policial	CapableOf	servir a pessoa
violencia policial	CapableOfReceivingAction	procurar por ladrão de pintura
violencia policial	CapableOf	prenda metade de quadrilha
violencia policial	CapableOf	prenda joe
violencia policial	CapableOf	prender john
violencia policial	CapableOfReceivingAction	autorize usar arma de fogo debaixo de protocolo específico
violencia policial	CapableOf	trazer miséria
violencia policial	CapableOfReceivingAction	chamar policial
violencia policial	CapableOf	carregar arma de fogo enquanto no trabalho
violencia policial	CapableOf	carregar arma carregada
violencia policial	CapableOf	pegar ladrão
violencia policial	CapableOf	cobrar roxanne
violencia policial	CapableOf	existir para minimizar roubo
violencia policial	CapableOf	existir para prevenir crime
violencia policial	CapableOf	sentir ameaçado pelo sujeito
violencia policial	CapableOf	dar ingresso de velocidade
violencia policial	CapableOf	molestar hector
violencia policial	CapableOf	ajudar a vítima
violencia policial	CapableOf	segurar suspeito
violencia policial	CapableOf	lei vigente do homem
violencia policial	CapableOfReceivingAction	instruir para defender a própria vida

Tabela 6.3: Baseline para "má iluminação publica"

nomePre	**papel**	**nomePos**
má iluminacao publica	LocationOf	teatro
má iluminacao publica	LocationOf	escritório
má iluminacao publica	PropertyOf	perigoso
má iluminacao publica	PropertyOf	melhor
má iluminacao publica	PropertyOf	elétrico
má iluminacao publica	PropertyOf	quente
má iluminacao publica	CapableOf	aparecer para estar por paerto da direita
má iluminacao publica	CapableOf	vir com trovão
má iluminacao publica	CapableOf	acontecer durante a tempestade
má iluminacao publica	PropertyOf	iluminando
má iluminacao publica	CapableOf	significar começar a queimar
má iluminacao publica	CapableOf	atingir objeto mais alto
má iluminacao publica	CapableOf	não produzir leite