Um Analisador Semântico Inferencialista de Sentenças em Linguagem Natural

Vladia Pinheiro
Universidade Federal do Ceará
vladia@lia.ufc.br

Tarcisio Pequeno
Universidade Federal do Ceará
tarcisio@lia.ufc.br

Vasco Furtado
Universidade de Fortaleza e ETICE
vasco@unifor.br

Resumo

Este artigo descreve um raciocinador semântico para entendimento de linguagem natural que implementa um algoritmo que raciocina sobre o conteúdo inferencial de conceitos e padrões de sentenças – o Analisador Semântico Inferencialista (SIA). O SIA implementa um raciocínio material e holístico sobre a rede de potenciais inferências em que os conceitos de uma língua podem participar, considerando como os conceitos estão relacionados na sentença, de acordo com padrões de estruturas sintáticas. A medida de relacionamento inferencial e o processo de raciocínio do SIA são descritos. O SIA é usado como raciocinador semântico em um sistema de extração de informações sobre crimes – WikiCrimesIE. Os resultados obtidos e uma análise comparativa são apresentados e discutidos, servindo para a identificação de vantagens e oportunidades de melhoria para o SIA.

1. Introdução

Para o entendimento de linguagem natural por computadores, algumas questões de pesquisas são fundamentais e ainda estão em aberto: (i) *Qual o conhecimento semântico que deve ser expresso?* (ii) *Como se calcula ou infere o significado de uma expressão linguística?*.

Comumente, pesquisas e aplicações das áreas de Linguística Computacional (LC) e Processamento de Linguagem Natural (PLN) resolvem os problemas do nível semântico das linguagens naturais (responder perguntas sobre um texto, extrair informações, sumarizar textos, gerar textos etc) usando abordagens sintáticas. Dentre estas, podemos citar aquelas que consideram parâmetros morfossintáticos para identificar similaridade e relacionamento semânticos, por exemplo, a concordância de número para resolução de anáforas, freqüência de palavras em comum para fusão de textos, extração de informações a partir de padrões sintáticos de entidades nomeadas (endereços, cidades, empresas). Noutras abordagens, a intensão de um conceito (o "significado" de um conceito) é apreendida de sua extensão, expressa normalmente em um *corpus linguístico*. Em resumo, recorre-se a um processo de sintatização do nível semântico da linguagem que, claramente, é insuficiente para um completo entendimento de textos em linguagem natural.

Outros sistemas e aplicações usam conhecimento semântico onde a intensão dos conceitos é definida em bases de conhecimento (normalmente denominadas de ontologias) contendo classes, propriedades e atributos dos objetos referenciados pelos termos de uma língua natural. Outra característica destes sistemas é que eles normalmente adotam uma abordagem atomista para raciocínio semântico. Nesta abordagem, a interpretação semântica de um elemento é tratada de forma independente da atribuição semântica dos demais elementos de uma sentença. Estas características – a priorização de uma representação do mundo para definição do significado e um raciocínio semântico atomista - limitam a capacidade de entendimento de linguagem natural dos sistemas de PLN.

Frequentemente, as informações necessárias para o entendimento completo de textos por sistemas de PLN estão implícitas, e descobri-las requer a realização de inferências a partir do uso de conceitos em situações linguísticas. Por exemplo, quando lemos a notícia "*João assassinou sua esposa com dois tiros após uma discussão na Rua Solon Pinheiro.*", nós somos capazes de refutar uma afirmação que indicasse que o tipo de arma usada no crime foi "arma branca" (não foi usada arma de fogo), argumentar que o tipo de crime foi "homicídio" e que a causa do crime foi "crime passional". Estas conclusões são possíveis porque nós, usuários da língua natural, sabemos as condições nas quais os conceitos "tiro", "assassinar" e "esposa" podem ser usados e os compromissos que assumimos ao usá-los em uma sentença. Além disso, raciocinamos considerando o conteúdo individual dos conceitos de forma

conjunta com o conteúdo dos demais conceitos da sentença em que são usados.

O fato é que habilidades como argumentar sobre o texto, responder perguntas, extrair informacões explícitas e implícitas, refutar afirmações etc. são cada vez mais necessárias em tarefas de PLN que envolvem entendimento de linguagem natural.

Um caminho para melhoria da qualidade do processamento semântico de sistemas de PLN é buscar inspiração nas respostas que filósofos oferecem à questão *Em que consiste o significado de uma expressão linguística?*. Sellars (1980), Dummett (1973) e Brandom (1994)(2000) propuseram as teorias semânticas inferencialistas, que apresentam uma abordagem diferente para definir o conteúdo de conceitos e sentenças. Segundo estas teorias, a expressão do valor semântico de conceitos deve privilegiar o papel que estes desempenham em raciocínios, como premissas e conclusões, ao invés de seus referentes e suas características representacionais. Segundo Sellars (1980), compreender um conceito é ter o domínio prático sobre as inferências em que ele pode estar envolvido – saber o que segue da aplicabilidade do conceito e a partir de que situações ele pode ser aplicado.

Seguindo esta visão inferencialista, Pinheiro et al. (2008)(2009) propõem o *Semantic Inferentialism Model* (SIM) - um modelo computacional que define requisitos para expressão e raciocínio sobre conhecimento semântico linguístico. Suas bases de conhecimento semântico expressam conteúdo inferencial de conceitos e sentenças, ou seja, as condições e consequências de uso de conceitos e sentenças.

O componente principal do SIM é seu raciocinador semântico de textos em linguagem natural: Analisador Semântico Inferencialista – SIA. O SIA é responsável por gerar as premissas e conclusões das sentenças do texto de entrada. Estas premissas e conclusões habilitam os sistemas de PLN para dar razões sobre o texto, responder perguntas, extrair informacões explícitas e implícitas, refutar afirmações etc. As regras de inferência e a medida de relacionamento inferencial, implementadas pelo SIA, são responsáveis por um mecanismo de raciocínio semântico material e holístico. Raciocínio material no sentido de que as inferências são autorizadas e justificadas pelos conteúdos conceituais, e raciocínio holístico porque o SIA define a contribuição semântica dos conceitos considerando outros conceitos relacionados em uma sentença, de acordo com sua estrutura sintática.

Este artigo está estruturado da seguinte forma, A seção 2 discute os fundamentos teórico-filosóficos e apresenta a arquitetura e formalização do SIM. A seção 3 apresenta o algoritmo do SIA, seu processo de raciocínio, regras de inferência, e a medida de relacionamento inferencial. Na seção 4, tem-se a descrição de como o SIA é aplicado em um sistema para extração de informações sobre crimes – Extrator de Informações WikiCrimes (WikiCrimesIE), e a avaliação dos resultados obtidos. Na seção 5, os trabalhos relacionados e uma análise comparativa são discutidos e, finalmente, este artigo é concluído com a apresentação dos trabalhos em andamento e futuros.

2. Semantic Inferentialism Model (SIM)

2.1 Fundamentos do SIM

O *Semantic Inferentialism Model* (SIM) (Pinheiro et al, 2008) (Pinheiro et al, 2009) define os principais requisitos para expressar e manipular conhecimento semântico inferencialista de forma a capacitar os sistemas de linguagem natural para um entendimento mais completo de sentenças e textos.

SIM é fortemente inspirado nas teorias semânticas inferencialistas de Sellars (1980), Dummett (1973) e Brandom (1994)(2000). Para Dummett, saber o significado de uma sentença é saber a justificativa para o falante tê-la proferido: "Nós não explicamos o sentido de uma declaração estipulando seu valor-verdade em termos dos valores-verdade de seus constituintes, mas sim estipulando quando ela pode ser afirmada em termos das condições sobre as quais seus constituintes podem ser afirmados" (Dummett, 1978). Brandom (1994)(2000), por sua vez, sedimenta a visão inferencialista de Dummett e Sellars e reduz a visão pragmática da linguagem de Wittgenstein (1953) para um racionalismo pragmático, onde a tônica são os usos inferenciais de conceitos em jogos de pedir e dar razões (jogos racionais). Para Brandom, entendemos uma sentença quando sabemos defendê-la, argumentar a seu favor, dar explicações, e isto só é possível porque sabemos inferir as premissas que autorizaram seu proferimento e as conclusões de seu proferimento.

Seguindo esta visão inferencialista, SIM responde à questão (i) *Qual o conhecimento semântico que deve ser expresso?* definindo que expressar o conteúdo de um conceito requer expressar, tornando explícito, seus usos [do conceito] em inferências, como premissas ou conclusões de raciocínios. E, o que determina o uso de um conceito em inferências ou as potenciais inferências em que este conceito pode participar são:

- precondições ou premissas de uso do conceito – o que dá direito a alguém a usar o conceito e o que poderia excluir tal direito, servindo de premissas para proferimentos e raciocínios;
- pós-condições ou conclusões do uso do conceito – o que se segue ou as conseqüências do uso do conceito, as quais permitem saber com o que alguém se compromete ao usar um conceito, servindo de conclusões do proferimento em si e de premissas para futuros proferimentos e raciocínios.

Este conteúdo, denominado de *conteúdo inferencial*, define o importe ou competência inferencial de um conceito. Esta visão inferencialista de conteúdo conceitual se contrapõe à visão representacionalista, segundo a qual os sistemas de PLN deveriam expressar uma representação do mundo *a priori*. Eco (2001) assinala que qualquer classificação ou caracterização do mundo (qualquer representação do mundo) é conjectural e arbitrária, mesmo que consensual em uma comunidade ou área de conhecimento. Portanto, não se pode delimitar o poder de entendimento dos sistemas de PLN a este "muro ontológico" e a um método cartesiano de raciocínio semântico, no qual, a partir de hipóteses (uma representação do mundo), seguimos concluindo isso ou aquilo através de regras formais. Como conseqüência, a verdade de nossas conclusões herda as limitações da organização artificial do mundo, ou seja, tudo o que se pode entender de textos em linguagem natural já está condicionado *a priori* nas hipóteses assumidas.

Em contraposição, o que precisa ser expresso sobre um conceito deve ser expresso a partir de seus usos em práticas linguísticas. Isto é concernente com a idéia de que conceitos surgem dentro da prática linguística de uma comunidade, sociedade ou de uma área de conhecimento, e são apreendidos pelos usuários de uma língua a partir de seus usos e não porque existem *a priori* no mundo com tais e tais características. Para ilustrar a natureza do conteúdo inferencial de conceitos, imaginemos uma criança que, pela primeira vez, presencie o uso do conceito "egoísta" em uma discussão entre seus pais. Provavelmente, ela usará este conceito em uma situação de disputa com um colega de escola , ou seja, para ela, "alguém fazer algo que não gosto" é condição suficiente para que ela empregue o conceito. Na medida de seu amadurecimento na linguagem, perceberá que existem outras precondições e que, nem sempre, quando duas pessoas discutem, ela poderá usar o conceito "egoísta".

Em outro exemplo, tem-se o conceito "saidinha bancária" que se originou dentro da prática linguística de se descrever assaltos em que os clientes são abordados após realizarem saques em agências bancárias. Não se originou, prioritariamente, pelas representações deste tipo de crime, mas pelas circunstâncias e as conseqüências que ditaram seu uso. Os usuários deste conceito aprenderam em que situações usá-lo e o que se segue do seu uso. Embora existam os conceitos "saidinha" e "bancária", a nova expressão linguística "saidinha bancária" denota um conteúdo com valor semântico distinto que foi moldado a partir de seus usos em sentenças.

Brandom (2000) apresenta exemplos onde um papagaio pode falar "Esta bola é vermelha!" na presença de uma bola da cor vermelha, e um termostato pode ligar o compressor de um ar-condicionado quando a temperatura está acima de 20ºC. Brandom discute a natureza da distinção entre estes relatos e quando os mesmos relatos são feitos por humanos. A resposta dada, à luz das teorias inferencialistas, é que tanto o papagaio quanto o termostato não sabem defender, dar razões, explicar seus relatos em situações de raciocínio – e isto é porque não conhecem as circunstâncias e conseqüências do uso dos termos "vermelho" e "quente" em situações linguísticas, não conhecem as potenciais inferências em que estes conceitos podem participar. Da mesma forma, uma criança que escuta um termo específico de uma área de conhecimento, por exemplo "inteligência artificial", provavelmente não saberá quando usá-lo e, se usá-lo, que conclusões podem ser inferidas. Isto implica dizer que a criança não sabe participar de situações linguísticas e raciocinar com este termo – não saberá defendê-lo, explicá-lo etc - ou seja, não entende este termo. Mesmo ontologias simples, que definem taxonomias, ou base semânticas mais complexas, que expressam conhecimento causal, funcional ou relativo a eventos, devem ser consideradas sob o ponto de vista inferencial e pragmático. Ou seja, seus conteúdos devem ser manipulados e qualificados em termos de precondições ou pós-condições de uso dos conceitos, em situações linguísticas, capturando, desta forma, o viés pragmático da linguagem natural. O argumento do SIM é que um modelo semântico, inspirado nas teorias inferencialistas, possibilita melhor habilidade aos sistemas de PLN que se proponham a entender linguagem natural.

O SIM, baseado no paradigma semântico-inferencialista, também apresenta resposta à questão

(ii) *Como se calcula ou infere o significado de uma dada sentença?*.

Os raciocinadores dos sistemas lógicos, usuais em PLN (Lógica Descritiva, PROLOG e Lógicas Intensionais), se resumem a realizar inferências formais, as quais geram novos fatos considerando apenas a forma das expressões lógicas e um raciocínio logicamente autorizado.

Acontece que muitas conclusões e respostas que humanos dão ao ler um texto são justificadas pelo conteúdo dos conceitos relacionados. Por exemplo, considere a inferência de "João é irmão de Pedro" para "Pedro é irmão de João". O conteúdo do conceito "irmão" é que torna esta inferência correta. Se substituirmos na primeira sentença o conceito "irmão" pelo conceito "pai", a inferência não pode ser realizada. Da mesma forma, a inferência "um relâmpago é visto agora" para "um trovão será ouvido em breve" é autorizada pelo conteúdo dos conceitos "trovão" e "relâmpago". Em outro exemplo mais complexo, a inferência de "João assassinou sua esposa com dois tiros" para "o tipo de arma usada foi arma de fogo" é autorizada pelo conteúdo dos conceitos "assassinar" e "tiro", analisados conjuntamente.

Para se realizar inferências desta natureza não se deve ter unicamente um mecanismo de raciocínio sobre a forma das sentenças, mas principalmente deve-se ter domínio dos conteúdos dos conceitos articulados nas sentenças e de como estes [os conteúdos dos conceitos] contribuem para o significado das mesmas. Daí a importância da natureza do conteúdo dos conceitos, expresso em bases semânticas.

Outro fato observável é a tradição atomista na semântica formal. Na abordagem atomista a atribuição de uma interpretação semântica a um elemento é tratada de forma independente da atribuição semântica dos demais elementos de uma sentença. Ao contrário, a semântica inferencialista é essencialmente holista: não se pode definir o valor semântico de um elemento sem considerar os outros elementos relacionados em uma sentença e como todos estão estruturados. Define-se "essencialmente holista" porque esta característica é uma conseqüência direta e simples da concepção inferencial do conteúdo de conceitos - "ninguém pode ter qualquer conceito a menos que tenha muitos conceitos" (Brandom, 2000, p.15-16). Ao expressar as potenciais inferências em que um conceito pode estar envolvido nada mais fazemos que expressar as relações inferenciais deste com outros conceitos e, na medida em que conhecemos um conceito, conhecemos vários.

Em contraposição à predominância, nos sistemas de PLN, de inferências formais e de raciocínio atomista, o SIM propõe o Analisador Semântico Inferencialista (SIA). O SIA implementa um raciocínio material e holístico sobre a rede de [potenciais] inferências em que conceitos podem participar (conteúdo inferencial), considerando como os conceitos estão relacionados na sentença, de acordo com sua [da sentença] estrutura sintática.

O raciocínio material possibilita a realização de inferências autorizadas pelo conteúdo (p.ex. de "um relâmpago é visto agora" para "um trovão será ouvido em breve", autorizada pelo conteúdo dos conceitos "trovão" e "relâmpago"), e argumento para refutar e validar inferências (p.ex. "A água é vermelha" é refutada pela precondição de uso do conceito "água" que define que este conceito só pode ser usado em sentenças onde não são associados ao mesmo uma cor).

O raciocínio holístico, por sua vez, considera o todo (sentença) e como suas partes (elementos subsentenciais) estão estruturalmente relacionadas a fim de definir a contribuição semântica de cada parte para com o todo (sentença). Nesta abordagem holística, as estruturas de sentenças assumem um papel importante porque, para determinar o valor semântico de um elemento subsentencial, devem-se considerar os outros elementos relacionados e é imprescindível levar em conta a estrutura que os organiza na sentença e que define suas formas e funções sintáticas. Tem-se, portanto, uma abordagem de raciocínio semântico de cima para baixo (ou *top-down*). Por exemplo, seja a sentença "Geison Santos de Oliveira foi executado com vários tiros", a qual segue uma estrutura de sentença que relaciona os conceitos "executar" (assassinar) e "tiro". Pelo conteúdo inferencial dos conceitos "executar" (assassinar) e "tiro" e como eles são articulados na sentença, é possível identificar similaridade inferencial entre ambos e definir que o conceito usado na sentença foi "executar", que alude a assassinar, e não "executar" com acepção a realizar algo. É importante salientar que o raciocínio holístico mantém a característica de composicionalidade da linguagem, no sentido de que o significado da sentença é obtido com base no conteúdo semântico dos elementos subsentenciais. No entanto, ao considerar como a estrutura da sentença articula seus elementos subsentenciais a fim de definir a contribuição semântica desses para com a sentença, tem-se uma abordagem holista de raciocínio.

As duas qualidades de raciocínio semântico em linguagem natural do SIA – material e holístico – completam o diferencial deste analisador semântico de textos em linguagem natural.

2.2 Formalização do SIM

A Figura 1 apresenta a arquitetura do SIM. O SIM contém três bases para expressão de conhecimento semântico:

- Base Conceitual — que contém conceitos da língua natural e seus conteúdos inferenciais;
- Base de Sentenças-Padrão — que contém sentenças-padrão e seus conteúdos inferenciais; e
- Base de Regras para Raciocínio Prático — que contém a expressão de conhecimento prático oriundo da cultura de uma comunidade ou de uma área de conhecimento.

Além de bases de conhecimento, o SIM inclui um componente responsável pelo raciocínio semântico de sentenças e textos em linguagem natural:

- Analisador Semântico Inferencialista (SIA) — que recebe o texto de entrada em linguagem natural e a árvore de dependência sintática do texto, gerada por um analisador (ou *parser*) morfossintático, e, a partir do conteúdo expresso nas três bases semânticas, gera a rede inferencial das sentenças do texto.

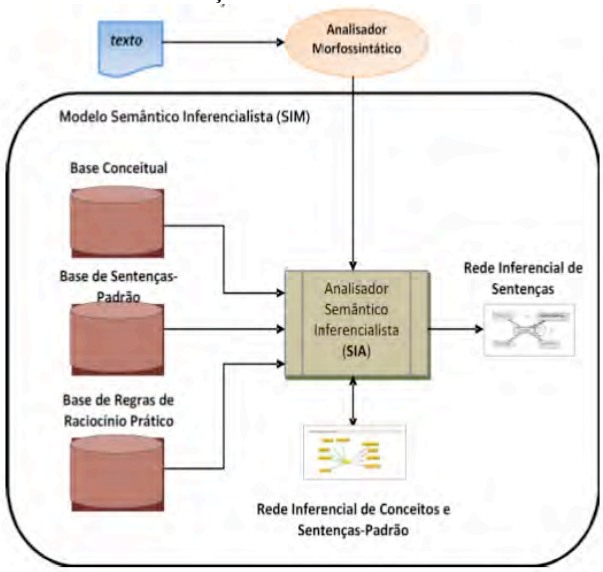

Figura 1: Arquitetura do *Semantic Inferentialism Model* - SIM.

A **Base Conceitual** contém o conteúdo inferencial de conceitos em língua natural, definidos e acordados em uma comunidade ou área de conhecimento. De acordo com a visão inferencialista, o conteúdo de um conceito *c* são as potenciais inferências em que *c* pode participar e o que determina esta participação são suas relações inferenciais com outros conceitos, na forma de suas precondições e pós-condições de uso. A base conceitual é um grafo direcionado $G_c(V,E)$, onde:

- V = conjunto não vazio de conceitos c_i (vértices do grafo). Um conceito em V pode ser representado na base conceitual por termos simples, que pertencem às classes abertas de palavras - nomes, verbos, adjetivos, advérbios (p.ex, 'crime', 'morte'); ou por expressões compostas de mais de um termo, ligados ou não por palavras das classes fechadas - preposições e conjunções (p.ex. 'prova de matemática', 'saidinha bancária').
- E = conjunto de arestas rotuladas por uma variável *tipo_rel* que expressa a relação binária entre conceitos de V. As relações *tipo_rel* são predefinidas como expressando as duas relações inferenciais: precondição ou pós-condição de uso de um conceito. Por exemplo, tem-se as relações "*CapazDe*" e "*EfeitoDesejávelDe*", onde a primeira define uma precondição de uso e a segunda uma pós-condição de uso de conceitos. Neste trabalho, usamos o formato $tipo\_rel(c_1,c_2)$ para expressar a relação inferencial *tipo_rel* entre c_1 e c_2, ambos conceitos em V, a qual pode ser interpretada como "c_1 possui *tipo_rel* em relação a c_2". Por exemplo, *CapazDe('crime','envolver violência')* é interpretada como "*crime é CapazDe envolver violência*".

Como se trata de um digrafo, $G_c(V,E)$ possui duas funções *s* e *t* onde:

- $s:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito de origem em V;
- $t:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito alvo em V.

A Base Conceitual permite expressar as relações inferenciais de um conceito com outros, obedecendo à visão holista de que conhecer um conceito é conhecer suas relações, na forma de premissas ou conclusões, com outros conceitos. A figura 2 apresenta o grafo inferencial G_{crime} do conceito '*crime*', o qual expressa as precondições e pós-condições de uso do conceito através de relações com outros conceitos.

A **Base de Sentenças-Padrão** contém sentenças genéricas que seguem uma dada estrutura sintática e que funcionam como *templates*, cujos *slots* podem ser preenchidos com termos de uma língua natural. Uma sentença-padrão segue certa estrutura de sentença, com algumas partes variáveis a serem

preenchidas (*slots*) por conceitos da base conceitual ou por outros elementos subsentenciais (nomes próprios, artigos, preposições, conjunções etc). Por exemplo, '*X ser assassinar por Y*' segue a estrutura <sentença> ::= <sn> <sv> <sp>[1]. O sintagma nominal (sn), representado por *X*, pode ser preenchido por 'uma mulher' e o sintagma complementar (sp), representado por 'por *Y*', pode ser complementado por 'seu amante'. Assim teremos a sentença '*uma mulher ser assassinar por seu amante*', gerada a partir do padrão '*X ser assassinar por Y*'.

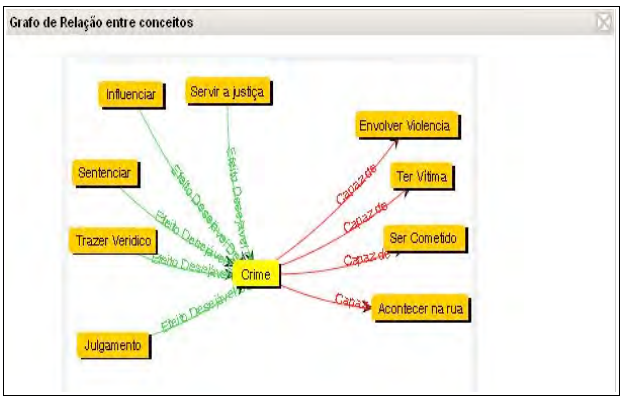

Figura 2: Grafo Inferencial do conceito 'crime'.

A importância de uma base de sentenças-padrão para análise semântica é a seguinte: algo do que se pode inferir ao se ler uma sentença não advém direta e unicamente, pelo menos de forma eficiente, do conteúdo dos conceitos da sentença, mas destes juntos e articulados sob determinada estrutura de sentença. Por exemplo, uma pessoa ao ler a sentença "*uma mulher foi assassinada por seu amante*" rapidamente é capaz de responder quem foi a vítima ('uma mulher') e quem foi o assassino ('seu amante'). Um mecanismo de raciocínio para gerar estas conclusões poderia até raciocinar sobre o conteúdo do conceito 'assassinar' - precondições de uso 'existir um assassino' e 'existir uma vítima' - porém, identificá-las de forma direta e eficiente na sentença exigiria conhecimento sobre como articular este conteúdo e sobre o elemento estruturador 'por'. Este conhecimento é justamente o que a base de sentenças-padrão provê: expressar conteúdo inferencial (premissas e conclusões) das sentenças-padrão. Este conteúdo inferencial consiste de conhecimento que não pode, pelo menos de forma eficiente, ser inferido do conteúdo dos conceitos. A base de sentenças-padrão consiste, portanto, de um grafo direcionado $G_s(V,E)$, onde:

- V = conjunto não vazio de sentenças-padrão s_j e conceitos c_i (vértices do grafo);
- E = conjunto de arestas rotuladas pela variável *tipo_rel*, que expressa uma relação binária entre uma parte (nominal, verbal ou complementar) da sentença-padrão p_j e um conceito c_i, ambos elementos de V, sempre na direção da sentença p_j para o conceito c_i. As relações *tipo_rel* são predefinidas como expressando as duas relações inferenciais: precondição ou pós-condição da sentença-padrão p_j. Neste trabalho, usamos o formato *tipo_rel(parte(p_j),c_i)* para expressar a relação inferencial *tipo_rel* entre uma parte de p_j e c_i. Esta relação é interpretada como "a parte de p_j possui *tipo_rel* em relação a c_i". Por exemplo, a precondição *ehUm(sn('X ser assassinar por Y'), 'pessoa')* é interpretada como "*X ehUm pessoa*", e a pós-condição *ehUm(sp('X ser assassinar por Y'), 'assassino')* é interpretada como "*Y ehUm assassino*".

Da mesma forma que no grafo da base conceitual (G_c), $G_s(V,E)$ possui duas funções *s* e *t* onde:

- $s:E \rightarrow V$ é uma função que associa uma aresta de E a uma sentença-padrão em V;
- $t:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito alvo em V.

A **Base de Regras para Raciocínio Prático** possibilita a expressão de conhecimento prático oriundo da cultura de uma comunidade, através de regras. Cada regra ρ_i visa combinar as premissas e conclusões já geradas para uma sentença original s_i, gerando novas premissas e conclusões para s_i. As regras ρ_i são cláusulas Horn da forma ($A_1 \wedge A_2 \wedge \ldots \wedge A_n \rightarrow$ *tipo_rel*(s_i, s_j)). Por exemplo, há o consenso oriundo da cultura das grandes cidades, que o local provável do crime é o local onde o cadáver foi encontrado. Para expressar este conhecimento pode-se usar a seguinte regra:

$\forall x,y,z$ (((encontrar(x,y) $\wedge$ encontrarEm(y,z) $\wedge$ ehUm(y,'cadáver')) $\rightarrow$ ('Pos', s_i,ehUm(z,'local do crime')))

A construção destas bases de conhecimento semântico inferencialista é um desafio, principalmente pelo seu caráter inovador. Uma proposta de construção das bases está descrita em (Pinheiro et al., 2010) e deu origem ao primeiro recurso linguístico com conteúdo semântico inferencialista para língua portuguesa – InferenceNet.BR – contendo em torno de 190.000 conceitos, 700.000 relações inferenciais entre conceitos, 6000 sentenças-padrão e 1500 relações inferenciais de sentenças-padrão. O sítio www.inferencenet.org possui funcionalidades para consulta, evolução e disseminação deste recurso linguístico.

1 <SN>: sintagma nominal; <SV>: sintagma verbal; <SP>: sintagma complementar

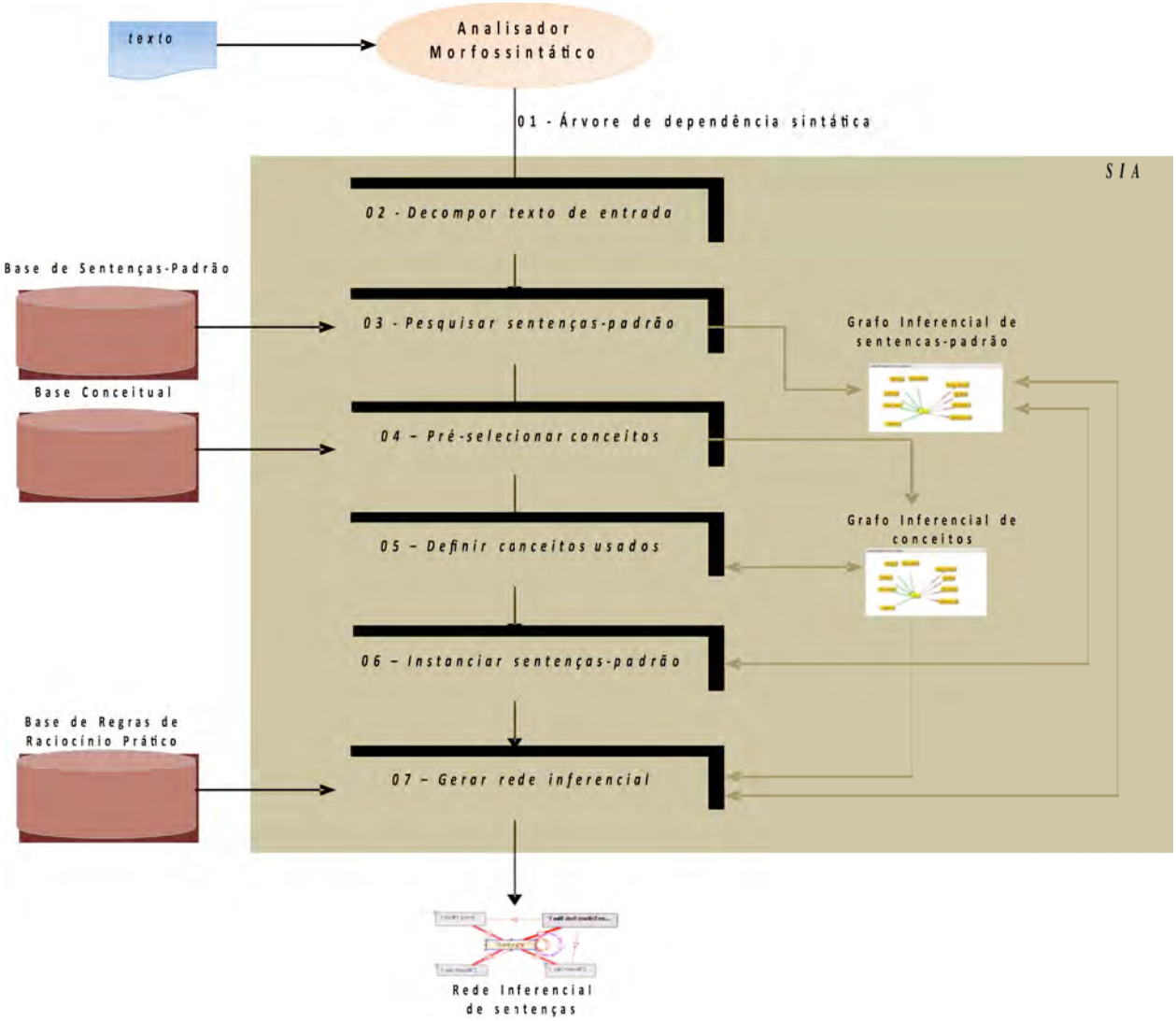

Figura 3: Visão gráfica do processo de raciocínio semântico do SIA.

3. Analisador Semântico Inferencialista (SIA)

O componente principal do SIM é seu raciocinador semântico de textos em linguagem natural - o Analisador Semântico Inferencialista – SIA. Em linhas gerais, um analisador semântico tem como objetivo descobrir o significado das expressões em linguagem natural e realizar o entendimento de sentenças em linguagem natural (Vieira e de Lima, 2001). De acordo com a teoria semântica do SIM (Pinheiro et al., 2008), o significado de uma sentença em linguagem natural é o conjunto de suas premissas (precondições) e conclusões (pós-condições), geradas a partir do conteúdo inferencial de seus conceitos articulados em uma dada estrutura de sentença (sentença-padrão).

SIA implementa um mecanismo de inferência sobre os grafos G_c e G_s (base conceitual e base de sentenças-padrão) e da base de regras ρ_i, com o objetivo de gerar a rede inferencial (premisssas e conclusões) das sentenças do texto, a qual consiste em um grafo direcionado $G_N(V,E)$, onde:

- V = conjunto de sentenças s_i (vértices do grafo);
- E = conjunto de arestas rotuladas por uma variável *tipo_rel* que indica o tipo de relação inferencial (precondição (*pre*) ou pós-condição (*post*)) entre uma sentença original do texto s_i e outra sentença s_j, que expressa uma premissa ou conclusão de s_i. Estas sentenças são inferidas a partir do processo de raciocínio do SIA (ver seção 3.2). Neste trabalho, usamos o formato *tipo_rel*(s_i, s_j) que é interpretado como "s_j é *tipo_rel* de s_i". Por exemplo, a pós-condição *post("João comeu","João ganhar energia")* é interpretada como " *'João*

ganhar energia' é pós-condição (ou conclusão) de s_j = 'João comeu' ".

Similarmente aos grafos G_c e G_s, o grafo $G_N(V,E)$ possui as funções *s* e *t* que associa, a uma aresta em *E*, seus elementos de origem e destino (sentenças) em *V*.

A figura 3 apresenta a visão gráfica do algoritmo implementado pelo SIA, que define os seguintes passos:

(1) Inicialmente, o algoritmo recebe a árvore de dependência sintática do texto de entrada, a qual foi gerada por um analisador morfossintático.

(2) É realizado um pré-processamento para decomposição dos períodos do texto em sentenças simples. Sentenças simples são sentenças que seguem a estrutura <sentença> ::= <SN> <SV> <SP>. Este pré-processamento é realizado pelo SIA com base na árvore de dependência sintática gerada pelo analisador morfossintático e gera uma sentença simples para cada ocorrência distinta de sujeito, verbo (ou locução verbal) e complemento verbal. Este passo é necessário porque os períodos compostos dificultam a combinação com sentenças-padrão. Por exemplo, o texto "*Na noite do último sábado, um jovem identificado como Geison Santos de Oliveira foi executado com vários tiros na Rua Titan, 33*" contém um período composto por três sentenças simples (s_1="*Um jovem identificado... foi executado na noite do último sábado*; s_2="*Um jovem identificado... foi executado com vários tiros*"; s_3="*Um jovem identificado... foi executado na Rua Titan, 33*").

(3) São pesquisadas e combinadas as estruturas das sentenças simples do texto com sentenças-padrão da Base de Sentenças-Padrão do SIM, gerando grafos G'_s (subgrafos de G_s) para cada sentença-padrão identificada.

(4) São pré-selecionados os conceitos da Base Conceitual do SIM que combinam literalmente com os termos usados nas sentenças do texto. Este passo é necessário porque existe um ou mais conceitos na base conceitual que são homônimos. Por exemplo, "executar" no sentido de realizar ou fazer, e "executar" com acepção a assassinar ou fuzilar.

(5) Neste passo, são definidos, dentre os conceitos pré-selecionados, quais conceitos foram usados, utilizando a ordem definida pela Medida de Relacionamento Inferencial, descrita na seção 3.1. Neste ponto, o algoritmo elimina os conceitos homônimos pré-selecionados que possuem menor proximidade inferencial com os demais conceitos da sentença s_i em que são usados. Para cada conceito *c* definido, gera grafos G'_c (subgrafo de G_c, a base conceitual do SIM, tal que $c \in V'$, conjunto de vértices de G'_c). Em seguida, é definida a contribuição semântica dos conceitos para a sentença s_i. A contribuição semântica de um conceito *c* usado em uma sentença s_i é o subgrafo de G'_c, gerado pela eliminação de G'_c das precondições e pós-condições que não influenciaram na proximidade inferencial de *c* com demais conceitos de s_i.

(6) Cada sentença-padrão em G'_s é instanciada com os elementos subsentenciais da sentença original (conceitos, preposições e outros elementos de ligação).

(7) Finalmente, neste passo é gerada a rede inferencial $G_N^{si}(V,E)$ de premissas e conclusões de cada sentença s_i do texto original. É o método principal do SIA, pois implementa formas de raciocínio que endossam inferências a partir de(a): (i) contribuição semântica dos conceitos usados em s_i, de acordo com a estrutura da sentença-padrão que os articula; (ii) contribuição semântica da sentença-padrão correspondente a s_i. Ambas as contribuições foram definidas a partir dos conteúdos inferenciais (pré e pós-condições) dos conceitos e sentenças-padrão e estão expressas nos subgrafos de G'_c e G'_s; (iii) regras pragmáticas da Base de Regras de Raciocínio Prático. As formas de raciocínio do SIA são detalhados na seção 3.2. Opcionalmente, objetivos da aplicação cliente são considerados para filtrar as premissas e conclusões geradas. A definição destes objetivos e como eles são usados pelo SIA são detalhadas na seção 3.2.

3.1 Medida de Relacionamento Inferencial

Cada vez mais, aplicações em Linguística Computacional requerem uma medida de relacionamento ou parentesco semântico entre dois conceitos e muitas abordagens têm sido sugeridas (Budanitsky e Hirst, 2001). A despeito de qualquer discussão filosófica e psicológica sobre a existência de uma medida numérica para a noção intuitiva de relacionamento semântico, a importância de uma

medida é que ela define uma relação de ordem (c_1 é mais similar a c_2 do que a c_3).

De acordo com a visão inferencialista e holística do SIM, o relacionamento entre conceitos não deve ser dissociado da sentença em que são usados e deve tomar como base o conteúdo inferencial compartilhado entre os conceitos articulados. Nesse sentido, dois conceitos usados em uma sentença estarão mais "inferencialmente relacionados" quanto mais o conjunto das precondições (ou das pós-condições) de um conceito é igual ao conjunto das precondições (ou das pós-condições) do outro conceito. A hipótese é que quanto mais as circunstâncias e consequências de uso de dois conceitos são semelhantes mais eles [os conceitos] podem ser usados em fluxos de raciocínio semelhantes.

São definidas, então, três formas de proximidade inferencial entre dois conceitos c_1 e c_2. Para cada uma das formas, tem-se um conjunto de relações inferenciais de c_1 e c_2 que satisfazem às condições de proximidade inferencial. As formas de proximidade inferencial são:

- **Proximidade por Relação Direta —** quando uma precondição (ou pós-condição) de c_1 expressa uma relação direta com c_2, ou vice-versa. Por exemplo, no caso dos conceitos c_1 = "crime" e c_2="roubo", e o conceito "roubo" possui a relação inferencial *éUm(roubo,crime)*.
- **Proximidade por Relação em Comum—** quando c_1 e c_2 expressam o mesmo tipo de relação semântica com um mesmo conceito. Por exemplo, no caso dos conceitos c_1 ="crime" e c_2="roubo" e ambos possuírem as relações inferenciais *capazDe(crime, ter vítima)* e *capazDe(roubo, ter vítima)*.
- **Proximidade por Relação de mesma Natureza —** quando c_1 e c_2 expressam relações inferenciais de mesma natureza (relações funcionais, causais, de eventos etc) com um mesmo conceito. Por exemplo, no caso dos conceitos c_1="tiro" e c_2="dedo" e ambos possuírem as relações inferenciais *usadoPara(tiro,ferir)* e *capazDeReceberAcao(dedo,ferir)*, onde as relações semânticas "usadoPara" e "capazDeReceberAcao" são de mesma natureza.

A medida de relacionamento inferencial $\theta_{c1,c2}$, entre dois conceitos c_1 e c_2 é calculada pela fórmula a seguir.

$$\theta_{(c_1,c_2)} = (F_1 w_1 + F_2 w_2 + F_3 w_3)\,\mu_{(c_1,c_2)}$$

Onde,

- F_1, F_2, F_3 são os somatórios das forças das relações inferenciais de c_1 e c_2 que satisfazem às três formas de proximidade inferencial, definidas acima;
- w_1, w_2, w_3 são os pesos, atribuídos por parâmetro, das três formas de proximidade inferencial, definidas acima; e
- $\mu_{(c1,c2)}$ é o fator de normalização entre os conceitos c_1 e c_2, calculado pela fórmula a seguir.

$$\mu_{(c_1,c_2)} = \frac{(n+m+p)}{|R_{(c_2)}|}$$

onde:

- (n+m+p) é o total de relações inferências de c_1 e c_2 que são semelhantes nas três formas de proximidade inferencial acima; e
- $|R_{c2}|$ é a cardinalidade do conjunto de relações inferenciais de c_2.

O fator de normalização serve para evitar que um conceito c_1 seja considerado mais inferencialmente relacionado a c_2 do que a c_3, somente porque c_2 possui maior número de relações inferenciais e, por isso, provavelmente maiores serão os valores de F_1, F_2 e F_3 , calculados entre c_1 e c_2.

A medida de relacionamento inferencial é utilizada no SIA para:

- desambiguação de termos homônimos;
- definição da contribuição semântica de um conceito *c* para a sentença *s*, pelo descarte de pré e pós condições de *c* que são irrelevantes para definição da proximidade inferencial de *c* com demais conceitos da sentença *s*;
- seleção de premissas e conclusões a serem geradas na rede inferencial da sentença *s*, a partir dos conceitos relacionados aos objetivos da aplicação cliente.

3.2 Raciocínio Inferencial do SIA

O SIA implementa três formas de raciocínio semântico para geração da rede inferencial $G_N^{si}(V,E)$, contendo premissas e conclusões das sentenças s_i do texto de entrada.

A primeira forma de raciocínio gera premissas e conclusões das sentenças do texto de entrada com base no conteúdo inferencial de conceitos usados nas sentenças. Definimos regras genéricas de introdução e eliminação de conceitos que podem ser instanciadas para cada conceito. A inspiração para estas regras vem do padrão de definição de conectivos lógicos de Gentzen (1935). Para Gentzen, um conectivo lógico é definido através de regras de introdução, que especificam sob quais circunstâncias o conectivo pode ser introduzido em um teorema; e através de regras de eliminação, que especificam sob quais condições o conectivo pode ser eliminado de um teorema. Dummett (1978) transpôs este modelo de definição para os conceitos de uma língua: um conceito é definido especificando-se regras de introdução do conceito (precondições de uso do conceito ou condições suficientes para uso do conceito) e regras de eliminação do conceito (pós-condições de uso do conceito ou conseqüências necessárias do uso do conceito).

A seguir, são apresentadas a interpretação e a sintaxe[2] das regras de introdução e de eliminação de conceitos em sentenças, e como estas regras são usadas pelo SIA para geração de premissas e conclusões das sentenças do texto de entrada.

(1) **A regra (I-c)** define que, se uma precondição de um conceito for satisfeita, a qual atende a uma precondição de uma sentença-padrão, então o conceito pode ser usado na parte da sentença que segue a estrutura da sentença-padrão (a parte é definida na precondição da sentença-padrão). Formalmente,

$$\frac{tipo\_rel(c_1, c_2), tipo\_rel(parte(p_1), c_2)}{s(parte(s)|c_2),\ p_1 \in P_s}(I-c)$$

Onde:

- p_1 é uma sentença-padrão;
- *parte(s)* é uma das partes nominal (sn), verbal (sv) ou complementar (sp) de *s*; e
- P_s é o conjunto das sentenças-padrão que determinam a estrutura sintática de *s*.

Exemplo 01:
Sejam
- c_1 = "jovem"
- precondição de c_1: éUm('jovem','pessoa')
- p_1 = "<X> <ser assassinar>"
- precondição de p1: éUm(sn(p_1), 'pessoa')
Logo, por (I-c), pode ser gerada sentença
s: <Um jovem> <ser assassinar>, a qual segue a estrutura sintática de p_1 e o conceito c_1 foi usado na parte nominal de s (sn(s)).

(2) **A regra (E_1-c)** define que, se um conceito é usado em uma sentença, então as precondições do conceito podem ser usadas para gerar precondições da sentença. A sentença $s(c_1|c_2)$ é a precondição na qual o conceito c_1 foi substituído por c_2. Formalmente,

$$\frac{tipo\_rel(c_1, c_2), s(c_1)}{('Pre', s(c_1), s(c_1|c_2))}(E_1-c)$$

A regra (E_1-c) autoriza o SIA a gerar sentenças s_j que expressam premissas das sentenças s_i do texto de entrada. A geração da premissa s_j ($s(c_1|c_2)$) depende da função sintática do conceito c_1 em s_i.

Se conceito c_1 = nucleo(sn(s_i)) (núcleo do sintagma nominal de s_i), então a premissa s_j é gerada da forma "*<reescrita(nome_relacao)> <c_2> <sv(s_i)> <sp(s_i)>*".

Exemplo 02:
Sejam
- s_1="O crime ocorreu na Rua Titan, 33"
- c_1="crime" = nucleo(sn(s_1))
- precondição de c_1: éUm('crime', 'violação da lei')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2 = "<Um(a)> <violação da lei> <ocorreu> <na Rua Titan, 33>"

Se conceito c_1 = nucleo(sv(s_i)) (núcleo do sintagma verbal de s_i), então a premissa s_j é gerada da forma "*<sn(s_i)> <"realizou ação que"|"sofreu ação que"> < reescrita(nome_relacao)> <c_2>*".

Exemplo 03:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="executar" = nucleo(sv(s_1))
- precondição de c_1: usadoPara('executar','vingança')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Um jovem> <sofreu ação que> <é usada para> <vingança>"

Se conceito c_1 = nucleo(sp(s_i)) (núcleo do sintagma complementar de s_i), então a premissa s_j é gerada da forma "*<sn(s_i)> <sv(s_i)> <preposicao> < reescrita(nome_relacao)> <c_2>*".

2 A formalização das regras de inferência do SIA segue o padrão de formalização das regras de inferência do sistema lógico de Dedução Natural de Prawitz (1965).

Exemplo 04:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="tiro" = nucleo(sp(s_1))
- precondição de c_1: usadoPara('tiro','ferir')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Um jovem> <foi executado> <com> <algo usado para> <ferir>"

(3) **A regra (E_2-c)** define que, se um conceito é usado em uma sentença, as pós-condições do conceito podem ser usadas para gerar pós-condições da sentença. A sentença $s(c_1|c_2)$ é a pós-condição na qual o conceito c_1 foi substituído por c_2. Formalmente,

$$\frac{tipo\_rel(c_1, c_2), s(c_1)}{('Pos', s(c_1), s(c_1|c_2))}(E_2 - c)$$

A regra (E_2-c) autoriza o SIA a gerar sentenças s_j que expressam conclusões das sentenças s_i do texto de entrada. A geração da conclusão s_j ($s(c_1|c_2)$) depende da função sintática do conceito c_1 em s_i.

Se conceito c_1 = nucleo(sn(s_i)) (núcleo do sintagma nominal de s_i), então a conclusão s_j é gerada da forma "*<reescrita(nome_relacao)> <c_2> <sv(s_i)> <sp(s_i)>*".

Exemplo 05:
Sejam
- s_1="O crime ocorreu na Rua Titan, 33"
- c_1="crime" = nucleo(sn(s_1))
- pós-condição de c_1: efeitoDe('crime', 'sofrimento')
Logo, por (E_2-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2 = "<Algo que tem efeito de> <sofrimento> <ocorreu> <na Rua Titan, 33>".

Se conceito c_1 = nucleo(sv(s_i)) (núcleo do sintagma verbal de s_i), então a conclusão s_j é gerada da forma "*<sn(s_i)> <"realizou ação que"|"sofreu ação que"> < reescrita(nome_relacao)> <c_2>*".

Exemplo 06:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="executar" = nucleo(sv(s_1))
- pós-condição de c_1: efeitoDe('executar','morte')
Logo, por (E_2-c), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Um jovem> <sofreu ação que> <tem efeito de> <morte>"

Se conceito c_1 = nucleo(sp(s_i)) (núcleo do sintagma complementar de s_i), então a conclusão s_j é gerada da forma "*<sn(s_i)> <sv(s_i)> <preposicao> < reescrita(nome_relacao)> <c_2>*".

Exemplo 07:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="tiro" = nucleo(sp(s_1))
- pós-condição de c_1: efeitoDe('tiro','ferir')
Logo, por (E_2-c), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Um jovem> <foi executado> <com> <algo que tem efeito de> <ferir>"

A segunda forma de raciocínio gera premissas e conclusões das sentenças do texto de entrada com base no conteúdo inferencial das sentenças-padrão correspondentes. Definimos regras genéricas para premissas e conclusões de uma sentença-padrão, as quais podem ser instanciadas para cada sentença-padrão usada nas sentenças do texto de entrada. A seguir, são apresentadas a interpretação e a sintaxe das regras de premissa e de conclusão de sentenças-padrão, e como estas são usadas pelo SIA para geração de premissas e conclusões das sentenças do texto de entrada.

(4) **A regra (P-p)** define que, se uma sentença é usada conforme a estrutura de uma sentença-padrão, então as precondições da sentença-padrão podem ser usadas para gerar precondições da sentença. Formalmente,

$$\frac{tipo\_rel(parte(p_1), c_1), p_1 \in P_{si}}{('Pre', s_i, tipo\_rel(parte(s_i), c_1))}(P - p)$$

A regra (P-p) autoriza o SIA a gerar sentenças s_j que expressam premissas das sentenças s_i do texto de entrada. A geração da premissa s_j depende da parte de p_1 que é o domínio da precondição de p_1.

Se parte(p_1) = sn(p_1) (a parte nominal da sentença-padrão p_1 é o domínio da precondição), então a premissa s_j é gerada da forma "*<sn(s_i)> <reescrita(nome_relacao)> <c_1>*".

Exemplo 08:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- precondição de p_1: éUm (sn(p_1), 'pessoa')
Logo, por (P-p), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Maria da Rocha> <é um(a)> <pessoa>"

Se parte(p_1) = sp(p_1) (a parte complementar da sentença-padrão p_1 é o domínio da precondição), então a premissa s_j é gerada da forma "*<sp(s_i)> < reescrita(nome_relacao)> <c_1>*".

Exemplo 09:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- precondição de p_1: éUm(sp(p_1), 'pessoa')
Logo, por (P-p), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Seu amante> <é um(a)> <pessoa>"

(5) **A regra (C-p)** define que, se uma sentença é usada conforme a estrutura de uma sentença-padrão, então as pós-condições da sentença-padrão podem ser usadas para gerar pós-condições da sentença. Formalmente,

$$\frac{tipo\_rel(parte(p_1), c_1),\ p_1 \in P_{si}}{('Pos', s_i, tipo\_rel(parte(s_i), c_1))}(C-p)$$

A regra (C-p) autoriza o SIA a gerar sentenças s_j que expressam conclusões das sentenças s_i do texto de entrada. A geração da conclusão s_j depende da parte de p_1 que é o domínio da pós-condição de p_1.

Se parte(p_1) = sn(p_1) (a parte nominal da sentença-padrão p_1 é o domínio da pós-condição), então a conclusão s_j é gerada da forma "*<sn(s_i)> < reescrita(nome_relacao)> <c_1>*".

Exemplo 10:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- pós-condição de p_1: éUm(sn(p_1), 'vítima')
Logo, por (C-p), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2:
"<Maria da Rocha> <é um(a)> <vítima>"

Se parte(p_1) = sp(p_1) (a parte complementar da sentença-padrão p_1 é o domínio da pós-condição), então a conclusão s_j é gerada da forma "*<sp(s_i)> <reescrita(nome_relacao)> <c_1>*".

Exemplo 11:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- pós-condição de p_1: éUm(sp(p_1), 'assassino')
Logo, por (C-p), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Seu amante> <é um(a)> <assassino>"

A terceira forma de raciocínio do SIA consiste na geração de premissas e conclusões das sentenças do texto de entrada com base na aplicação das regras expressas na Base de Regras de Raciocínio Prático do SIM. A seguir, são apresentadas a interpretação e a sintaxe de uma regra de raciocínio prático ρ_i, e como esta é usada pelo SIA para geração de premissas e conclusões das sentenças do texto.

(6) **A regra (I-r)** define que se os antecedentes de uma regra de raciocínio prático forem satisfeitos, então a conclusão da regra pode ser gerada para a sentenca do texto de entrada, que está sob análise. Os antecedentes da regra são comparados às premissas e conclusões da sentença do texto de entrada e às relações inferenciais de conceitos e sentenças-padrão, usados na sentença do texto de entrada. Formalmente, seja ρ_i uma cláusula de Horn da forma ($A_1 \wedge A_2 \wedge ... \wedge A_n \rightarrow tipo(s_i, s_j)$), então,

$$\frac{A_1, A_2, ..., A_n}{(tipo, s_i, s_j)}(I-r)$$

Exemplo 12:
Sejam
$\rho_1 = \forall x,y,z$ (((encontrar(x,y) $\wedge$ encontrarEm(y,z) $\wedge$ éUm(y,'cadáver'))$\rightarrow$ ("Pos",s,éUm(z,'local do crime'))).
s_1="<Os policiais Leandro e Vitor> <encontrar> <o corpo>"
s_2="<o corpo> <é um(a)> <cadáver>"
s_3="<o corpo> <foi encontrar> <Rua Titan,33>

Logo, por (I-r), os antecedentes de ρ_1 foram satisfeitos e pode ser gerada a seguinte conclusão de s_1: éUm('Rua Titan,33','local do crime').

Como visto, as três formas de raciocínio do SIA são responsáveis por gerar inferências endossadas pelo conteúdo que autoriza o uso dos conceitos e das conseqüências deste uso, bem como de premissas e conclusões de sentenças-padrão, as quais expressam conteúdo que não pode ser direta e eficientemente extraído dos conceitos, tomados individualmente. Todo este conteúdo prioritariamente inferencialista, tornado explícito, serve de base para responder perguntas, argumentar, refutar afirmações, extrair informações etc. Além

disso, como as bases semânticas do SIM são flexíveis para expressão de conhecimento de senso-comum e pragmático da língua natural, inferências mais interessantes sobre estes conteúdos são realizadas. Todas estas características completam o diferencial do uso do SIM em sistemas de PLN.

Outro componente particularmente importante no processo de raciocínio do SIA são os objetivos da aplicação cliente. Uma aplicação cliente é uma aplicação ou sistema de PLN que utiliza o SIA como raciocinador semântico de textos em linguagem natural. O uso de objetivos possibilita que o SIA direcione as premissas e conclusões geradas conforme necessidades de informações específicas, por exemplo, o local do crime. Portanto, somente as inferências relacionadas aos conceitos que expressam tais objetivos são potencialmente relevantes.

Os objetivos da aplicação cliente funcionam como *templates* que contém campos a serem preenchidos, com base nas inferências geradas pelo SIA. Cada *template* representa um objetivo. Por exemplo, o *template* "O local do crime é _____" representa o objetivo "Encontrar o local do crime". Cada objetivo é definido por: (i) um conceito que expressa o assunto do objetivo (por exemplo, o conceito *'crime');* (ii) uma lista de conceitos relacionados, que definem a informação requerida sobre o assunto do objetivo (por exemplo, o '*local*', a '*vítima*', o '*horário*', o '*tipo*' etc); (iii) um lista de conceitos que definem cada opção de resposta possível. Este último parâmetro de objetivos é opcional, pois algumas informações requeridas pela aplicação cliente possuem uma faixa de valores possíveis como resposta (por exemplo, o tipo de crime pode ser, alternativamente, um '*assassinato*', um '*roubo*', um '*furto*' etc).

Os conceitos envolvidos na definição de um objetivo são expressos na base conceitual do SIM. Todos eles são considerados para selecionar as premissas e conclusões, geradas pelo SIA, que são relevantes para responder ao objetivo. O critério de seleção é o melhor resultado da medida de relacionamento inferencial entre os conceitos das premissas/conclusões geradas e os conceitos relacionados ao objetivo.

4. Extrator de Informações para WikiCrimes

O SIA e as bases semânticas InferenceNet.Br (Pinheiro et al., 2010) são usadas em uma aplicação real: o Extrator de Informações para o sistema colaborativo WikiCrimes (Furtado et al., 2009). WikiCrimes[3] provê um ambiente colaborativo e interativo na Web para que as pessoas possam reportar e monitorar crimes ocorridos. Uma necessidade urgente do projeto WikiCrimes era fornecer a seus usuários uma ferramenta que os assistisse no registro de crimes a partir de notícias reportadas na Web e, desta forma, promovesse um estímulo à colaboração.

Esta necessidade existe para os sistemas colaborativos em geral. De um lado, tem-se a Web como uma fonte rica de informações sobre qualquer domínio, seu conteúdo é vasto e, em sua maioria, está na forma não estruturada e em linguagem natural. De outro lado, sistemas colaborativos dependem da iniciativa dos usuários para geração do conteúdo e de uma inteligência coletiva. No entanto, não é motivador deixar para os usuários a tarefa de ler os textos da Web, extrair as informações e ainda registrar manualmente no sistema. Portanto, existe a necessidade crescente de ferramentas que auxiliem a captura rápida, de forma simples, semi-automática e interativa de informações para registro em sistemas colaborativos e, além disso, que saibam manipular conteúdo em linguagem natural.

Para atender a esta necessidade, foi desenvolvido um sistema Extrator de Informações para o WikiCrimes - WikiCrimesIE – para extrair informações de crimes descritos em língua portuguesa, em jornais da Web, e gerar os registros do crime na base de dados de WikiCrimes.

O diferencial do uso do SIA como raciocinador semântico de WikiCrimesIE é sua melhor capacidade para entendimento completo de textos em linguagem natural. Algumas das informações requeridas pelo sistema WikiCrimes não estão comumente explícitas no texto, por exemplo, tipo e causa do crime, tipo de vítima e tipo de arma utilizada.

A figura 4 apresenta a interface de WikiCrimesIE, dividida em quadros, conforme segue:

A) Texto selecionado de um sítio da web, a partir do qual as informações sobre o crime relatado serão extraídas.
B) Mapa geoprocessado onde é localizado o endereço do crime.
C) Dados analíticos sobre o endereço do crime.
D) Dados do crime: data e horário da ocorrência, quantidade de criminosos e vítimas, relação do usuário com o crime e informação para polícia.
E) Dados especiais sobre o crime: tipo do crime, tipo de vítima, arma utilizada, motivos ou causas do crime.

3 www.wikicrimes.org, acessado em 18/12/2009

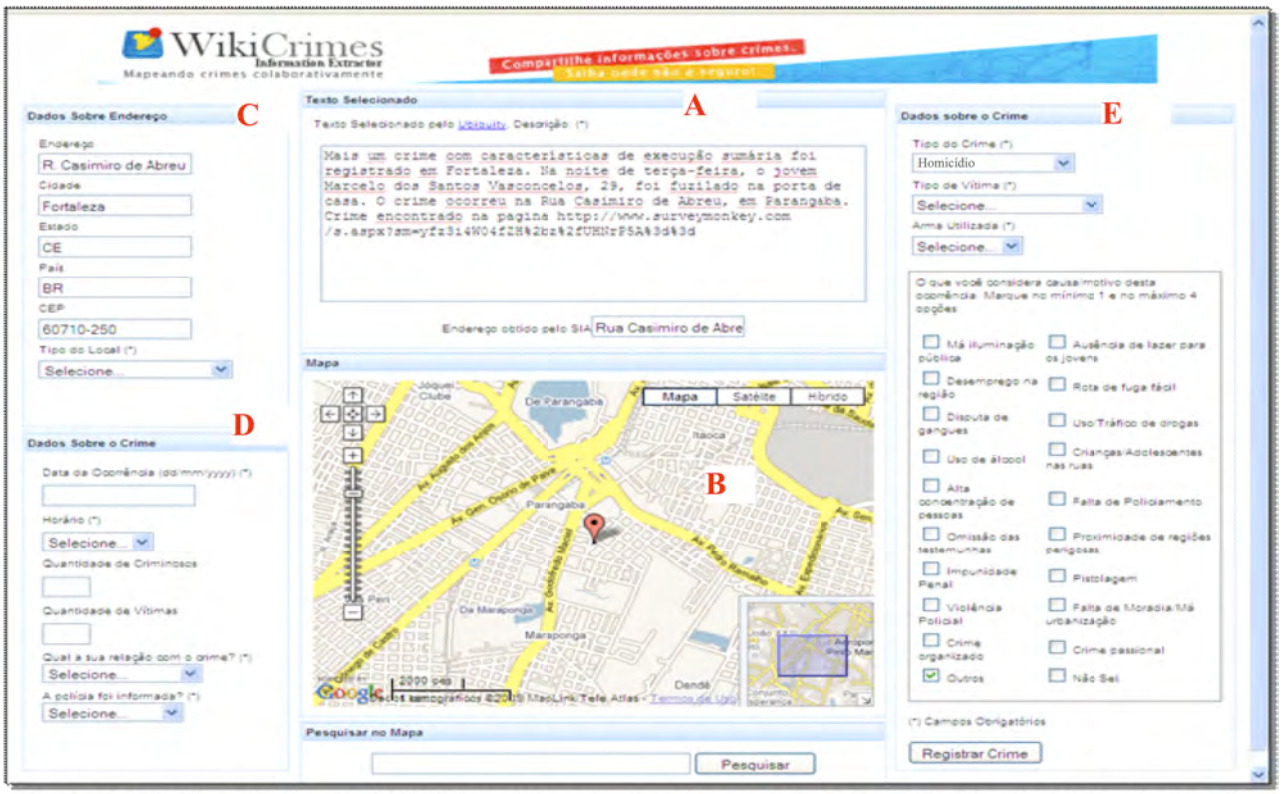

Figura 4: Interface do sistema WikiCrimesIE apresentando as informações extraídas do texto selecionado: local do crime ("Rua Casimiro de Abreu, Parangaba") e tipo do crime ("homicídio"). O endereço correspondente ao local do crime foi localizado no mapa geoprocessado.

4.1 Funcionamento de WikiCrimesIE

O processo de WikiCrimesIE para extração de informações sobre um dado crime, a partir de um texto descritivo em língua portuguesa, segue os seguintes passos:

(1) o usuário seleciona um texto de um sítio da web e executa o comando *mapcrimes*, desenvolvido na ferramenta *Ubiquity (*Nogueira et al., 2009). O Ubiquity é um plug-in do Mozilla Firefox e consiste em uma ferramenta de programação orientada ao usuário. Através de sua linguagem é possível implementar comandos que realizam a integração e *mashups* de aplicações Web. Nogueira et al. (2009) desenvolveram o comando *mapcrimes* que seleciona um texto de um sítio qualquer da Web e o envia ao sistema WikiCrimesIE;

(2) WikicrimesIE envia o texto selecionado ao analisador morfossintático PALAVRAS (Bick, 2000);

(3) WikicrimesIE instancia os objetivos da aplicação. Para cada objetivo devem ser especificados o conceito do assunto principal (por exemplo, 'crime'), conceitos relacionados à informação requerida (por exemplo, 'local', 'endereço' etc), e, no caso de existirem respostas alternativas, conceitos relacionados a cada opção de resposta (por exemplo, 'assassinato', 'roubo', 'furto');

(4) WikicrimesIE envia o texto analisado pelo parser PALAVRAS e os objetivos de extração de informação para o SIA;

(5) O SIA realiza a análise semântica das sentenças do texto e gera a rede de inferências para cada sentença, filtrando as premissas e conclusões pelos objetivos de extração. Para o caso de objetivos abertos, ou seja, objetivos sem opções de respostas predefinidas (por exemplo, local do crime), o SIA retorna ao WikicrimesIE a parte da sentença s_i (sintagma nominal, verbal ou complementar de s_i) que contém a resposta e a sentença s_j (premissa ou conclusão gerada pelo mecanismo de raciocínio do SIA) que justifica a resposta. Para o caso de objetivos com respostas predefinidas (por exemplo, tipo do crime), o SIA retorna a WikicrimesIE uma ou mais respostas selecionadas e as respectivas sentenças s_j (premissas ou conclusões geradas pelo

mecanismo de raciocínio do SIA), que justificam as respostas.

(6) WikicrimesIE interpreta as respostas dos objetivos, retornadas pelo SIA, e apresenta-as na interface (Figura 4);

(7) O usuário tem a opção de aceitar as respostas dadas pelo SIA ou alterá-las antes de registrar o crime na base de dados de Wikicrimes. Para fins de avaliação dos níveis de precisão e cobertura do SIA, é armazenado o log dos resultados retornados pelo SIA e as alterações realizadas pelos usuários.

4.1.1. Extração do Local e Tipo do Crime

A seguir será exemplificado o processo de raciocínio do SIA para extração do local do crime e do tipo de crime descrito no texto (ver quadro A da Figura 4): "*Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba.*"

O sistema WikicrimesIE instancia dois objetivos.

OBJETIVO1. "*Encontrar o local do crime*":

i. conceito que expressa o assunto principal do objetivo: *'crime'*;
ii. lista de conceitos relacionados, que definem a informação requerida sobre o assunto principal: '*local*' , '*endereço*' , '*cidade*', '*bairro*';
iii. lista de respostas predefinidas: não se aplica para este objetivo.

OBJETIVO2. *"Encontrar o tipo do crime*":

i. conceito que expressa o assunto principal do objetivo: *'crime'*;
ii. lista de conceitos relacionados, que definem a informação requerida sobre o assunto principal: '*tipo*', '*espécie*';
iii. lista de respostas predefinidas[4]:
    1. **roubo** ('furto', 'violência')
    2. **tentativa de roubo** ('tentativa', 'furto', 'violência')
    3. **furto** ('furto')
    4. **tentativa de furto** ('tentativa, 'furto')
    5. **violência doméstica** ('violência', 'família', 'esposa', 'marido')
    6. **rixas ou brigas** ('luta')
    7. **homicídio** ('assassinato', 'morte')
    8. **tentativa de homicídio** ('tentativa', 'assassinato', 'morte')
    9. **latrocínio** ('roubo', 'morte', 'violência', 'furto')

O SIA executa os seguintes passos:

(1) Recebe a árvore sintática gerada pelo PALAVRAS.

(2) Decompõe as sentenças do texto em:
- s_1="*Mais um* ***crime*** *com características de* ***execução sumária foi registrado*** *em Fortaleza."*
- s_2="*O* ***jovem****... foi fuzilado na* ***noite*** *de* ***terça-feira****."*
- s_3="*O* ***jovem****... foi fuzilado na* ***porta*** *de* ***casa****."*
- s_4="*O* ***crime ocorreu*** *na Rua Casimiro de Abreu, Parangaba."*

(3) Combina sentenças-padrão com as sentenças decompostas. As respectivas sentenças-padrão são:
- p_1="*X ser registrar em Y.*"
- p_2="*X ser fuzilar em Y.*"
- p_2="*X ser fuzilar em Y.*"
- p_3="*X ocorrer em Y.*"

(4) Seleciona os conceitos possíveis da Base Conceitual que foram usados nas sentenças originais, correspondentes aos termos em negrito, destacados acima:
- *conceitos*(s_1)=(*crime, execução sumária, ser, registrar)*
- *conceitos*(s_2)=(*jovem, ser, fuzilar, noite, terça-feira)*
- *conceitos*(s_3)=(*jovem, ser, fuzilar, porta, casa)*
- *conceitos*(s_4)=(*crime, ocorrer)*

(5) Define todos os conceitos previamente selecionados, pois, neste exemplo, não há conceitos a desambiguar. Instancia, para cada sentença s_i, um grafo com os conteúdos inferenciais dos conceitos definidos (subgrafo G'_c , para cada conceito *c*) e outro grafo com os conteúdos inferenciais das sentenças-padrão definidas (subgrafo G'_s, para cada sentença-padrão *p*). No exemplo, não é possível eliminar pré e pós-condições porque não há conceitos a desambiguar. Por exemplo, para a sentença s_4, é gerado subgrafo $G'_{fuzilar}$ com aresta expressando a pós-condição *efeitoDe('fuzilar', 'morte')* e subgrafo G_{p3} com aresta expressando a precondição *ehUm(sp*(p_3)*, 'local')*.

(6) Instancia as sentenças-padrão p_1 a p_3 com os elementos subsentenciais das respectivas

4 A lista de respostas predefinidas consiste de uma lista de opções da forma ***tipo_crime*** *('conceito*$_1$*', conceito*$_2$*, ..., conceito*$_n$*)*, onde *conceito*$_i$ são os conceitos que definem o ***tipo_crime***.

sentenças originais s_1 *a* s_4. Por exemplo, a sentença-padrão p_3*="X ocorrer em Y"* é instanciada com os elementos subsentenciais e respectivos conceitos da sentença s_4: *X="o crime"* e *Y="a Rua_Casimiro_de_Abreu"*. Com isso, tem-se a estrutura das sentenças originais s_1 a s_4 e seus respectivos elementos subsentenciais com conceitos associados.

(7) Gera a rede inferencial de G_N^{si} para cada sentença s_1 a s_4, aplicando as formas de raciocínio semântico sobre os subgrafos G'_c e G'_s. Para cada premissa/conclusão gerada é calculada a medida de relacionamento inferencial entre os conceitos do objetivo e o conceito relacionado na premissa e conclusão. Vejamos o detalhe das inferências geradas e o cálculo da medida de relacionamento inferencial em relação a cada objetivo:

OBJETIVO1.
A premissa de s_4 *ehUm('a Rua_Casimiro_de_Abreu', 'local')* foi gerada em G_N^{s4} pela regra de inferência (P-p) sobre o grafo G_{p3}. A medida ***θ('crime','crime')*** (c_1='*crime*', assunto principal do objetivo; e c_2='*crime*', núcleo do sintagma nominal de s'_4) e ***θ('local','local')*** (c_1='*local*', conceito que define a informação requerida sobre *crime*; e c_2='*local*', conceito relacionado na premissa de s'$_4$) apresentaram valor máximo, indicando que esta premissa responde melhor ao objetivo. Com isso, o SIA retorna a WikiCrimesIE a premissa "*a Rua_Casimiro_de_Abreu é um(a) local"* como resposta ao objetivo "*Encontrar local do crime*".

OBJETIVO2.
A conclusão de s_2 "*O jovem sofreu ação que tem efeito de morte"* foi gerada em G_N^{s2} pela regra de inferência (E_1-c) sobre o grafo $G_{fuzilar}$. A medida ***θ('crime','crime')*** (c_1='*crime*', assunto principal do objetivo; e c_2='*fuzilar*', núcleo do sintagma verbal de s'_2) e ***θ('morte','morte')*** (c_1='*morte*', conceito que define o tipo de crime = ***homicídio***('assassinato', 'morte')); e c_2='*morte*', conceito relacionado na premissa de s'$_4$) apresentaram valores maiores comparados às outras premissas/conclusões. Estes resultados indicam que esta conclusão responde melhor ao objetivo e o tipo do crime = homicídio. Com isso, o SIA retorna a WikicrimesIE a resposta selecionada *tipo_crime* = **homicidio** e a conclusão "*O jovem sofreu ação que tem efeito de morte"* como sentença inferida que justifica a resposta.

Nos quadros A e B da Figura 4, tem-se, respectivamente, a resposta do OBJETIVO1 identificada (**Rua Casimiro de Abreu**) e sua localização no mapa geoprocessado. No quadro E da Figura 4, tem-se a caixa de seleção do tipo do crime = **homicídio**, conforme OBJETIVO2.

4.2 Avaliação dos Resultados do SIM

Ao avaliarmos os resultados do sistema WikiCrimesIE na tarefa de extração de informações a partir de textos em linguagem natural, o que estamos avaliando, de fato, é o desempenho do SIM como modelo para expressão e raciocínio semântico em sistemas de entendimento de linguagem natural.

A metodologia de avaliação seguiu os passos delineados na sequência.

(1) Foi elaborada uma Coleção Dourada (CD) com 100 textos jornalísticos, publicados nas páginas policiais de jornais brasileiros, na Internet. Estes textos foram coletados por pessoas que não participavam do projeto e de forma aleatória.

(2) Os textos da CD foram lidos por duas pessoas adultas, proficientes em língua portuguesa, e foi solicitado a elas que respondessem às duas perguntas abaixo e registrassem a resposta, para cada texto. Antes, as pessoas receberam orientações sobre os tipos de crimes que deveriam ser considerados e acerca das respostas a serem dadas. Por exemplo, que a resposta para a pergunta sobre o local do crime deveria ser descritiva, contendo o maior número de informações sobre a localização exata do crime (endereço, bairro, ponto de referência, cidade, localidade etc).
   - Qual o local do crime?
   - Qual o tipo de crime?

(3) As respostas das duas pessoas participantes foram comparadas e, em caso de divergência, uma terceira pessoa foi consultada sobre qual das respostas era a correta. Ao final, apenas uma resposta de cada pergunta foi anotada para cada texto da CD.

(4) Os textos da CD foram submetidos ao sistema WikiCrimesIE e as informações extraídas pelo sistema sobre o local e tipo de crime foram registradas, para cada texto.

(5) As respostas dos especialistas humanos e do WikiCrimesIE foram comparadas e analisadas manualmente. Para uma avaliação quantitativa do SIM, foi atribuído, para cada informação extraída pelo WikiCrimesIE, de cada texto, um valor numérico que correspondia ao resultado da comparação, conforme Tabela 1.

Valor Atribuído	Resultado da comparação
1	***Informação CORRETA***
2	***Informação PARCIALMENTE CORRETA.*** ***Obs.: Este valor é atribuído quando o sistema não identificou o endereço completo do local do crime.***
3	***Informação INCORRETA*** ***Obs.: Este valor é atribuído quando o sistema identificou a sentença do texto que justificava a informação correta, porém não inferiu a informação correta.***
4	***Informação INCORRETA***
5	***Informação NÃO EXTRAÍDA***
6	***Informação NÃO EXTRAÍDA por erro de processamento***

Tabela 1: Valores atribuídos na comparação das informações extraídas pelo WikiCrimesIE em relação às respostas dadas pelos especialistas humanos.

As medidas usadas na avaliação do SIM, para cada informação extraída (local e tipo de crime), foram:

- **precisão**, que mede o quanto da informação extraída (casos em que A=1,2,3 ou 4) foi corretamente extraída (A=1 ou 2). Esta medida indica o quanto o sistema WikiCrimesIE é confiável em extrair a informação;
- **cobertura**, que mede o quanto da informação que deveria ter sido extraída (casos em que A=1,2,3,4 ou 5) foi corretamente extraída (A=1 ou 2). Esta medida indica o quanto o sistema WikiCrimesIE é abrangente em extrair a informação;
- **medida-F**, que é a média harmônica das medidas de precisão e cobertura;
- **percentual de erros de processamento**, que mede o percentual de textos não analisados por erro de processamento (A=6), ocasionado por problemas relacionados à estrutura sintática do texto: sentenças mal formadas (sem sujeito), períodos complexos etc; e
- **percentual de erros do analisador morfossintático**, que mede o percentual de erros de análise morfossintática do PALAVRAS. Esta medida indica o quanto a dependência da análise sintática prejudica os resultados do sistema.

A Tabela 2 apresenta os resultados das medidas de avaliação do WikiCrimesIE na tarefa de extração do local do crime, do tipo de crime e de ambos.

	Local do crime	Tipo do crime	Geral
Precisão	*87.00%*	*72.00%*	*79.00%*
Cobertura	*71.00%*	*68.00%*	*69.00%*
Medida F	*78.00%*	*70.00%*	*74.00%*
%Erros processamento	*3.00%*	*8.00%*	*8.00%*
%Erros Análise Morfossintática	*2.00%*	*7.00%*	*7.00%*

Tabela 2: Resultados do WikiCrimesIE na extração do "Local do Crime" e "Tipo do Crime".

5. Trabalhos Relacionados e Análise Comparativa

Nesta seção, serão citados alguns trabalhos relacionados a tarefas de PLN que envolvem entendimento de linguagem natural. Uma análise comparativa com a tarefa realizada pelo SIM no sistema WikiCrimesIE não é trivial, devido a diferença entre a natureza das informações requeridas por esse sistema e o foco dos sistemas atuais de Extração de Informação (EI). Segundo Grishman (2003), as pesquisas em EI evoluem em duas linhas: extração de nomes (*Named Entity Recognition* – NER) e extração de relações entre entidades participantes de eventos.

Na tarefa de NER, os sistemas da Priberam (Amaral et al., 2008) e REMBRANDT (Cardoso, 2008) apresentam um algoritmo para reconhecimento de entidades mencionadas (REM) para língua portuguesa. Tais sistemas foram os que apresentaram os melhores resultados em termos de cobertura e Medida-F para a tarefa de REM do Segundo HAREM (Mota e Santos, 2008). Respectivamente, tem-se os resultados de 51,46%

(cobertura) e 57,11% (Medida-F) do sistema da Priberam, e 50,36% (cobertura) e 56,74% (Medida-F) do sistema REMBRANDT. Para a língua inglesa, o melhor resultado foi registrado no evento MUC-7 (1997) com 87% de Medida-F. É importante salientar que a tarefa executada por estes sistemas restringe-se à identificação de entidades mencionadas nos textos e à classificação destas em categorias/tipos e subtipos semânticos predefinidos.

Na tarefa de extração de eventos, tem-se o melhor sistema avaliado na tarefa 4 do evento SemEval-2007 com 72,40% de Medida-F, para língua inglesa (Girju, 2007). Para língua portuguesa, tem-se o melhor sistema avaliado na tarefa de Reconhecimento de RELações entre Entidades Mencionadas (ReRelEM) do Segundo HAREM (Freitas et al, 2008) – o sistema REMBRANDT, com 45,02% de Medida-F.

Em uma análise quantitativa, WikicrimesIE, com medida-F = 74% (resultado geral da Tabela 2), apresentou melhor resultado dentre todos os sistemas mais bem avaliados para língua portuguesa. Para língua inglesa, ficou apenas abaixo do melhor sistema na tarefa de NER da MUC-7. É importante salientar que esta comparação é ainda injusta nos seguintes sentidos:

- nestes eventos de avaliação, requerem-se informações explícitas no texto. Argumentamos que os sistemas participantes destes eventos não foram avaliados na extração ou anotação de informações implícitas no texto. O tipo de crime, por exemplo, na maioria das vezes, não é mencionado. Na CD desta avaliação, havia 72% de textos nos quais o tipo de crime não era mencionado, requerendo o mínimo de raciocínio semântico para inferir o tipo de crime. WikiCrimesIE conseguiu extrair corretamente 69% dos tipos de crime, nestes casos;
- o raciocínio do SIA não se baseia em técnicas de aprendizagem de máquina ou regras gramaticais, como é o caso dos sistemas de EI aqui comparados.

Em uma análise qualitativa, foram estudados todos os casos de imprecisão do SIA na extração do local ou tipo de crime e identificados os principais problemas:

- 71% dos casos de imprecisão na extração do local do crime decorreram do mesmo estar de forma indireta em outras sentenças que relatam ações do criminoso ou da vítima;
- 12% dos casos de imprecisão na extração do local do crime decorreram de o SIM não realizar resolução de correferências;
- 50% dos casos de imprecisão na extração do tipo do crime tiveram origem na falta de conceitos na Base Conceitual do SIM; e
- 39% dos casos de imprecisão na extração do tipo do crime decorreram de problemas nas heurísticas de comparação de conceitos relacionados, implementadas pelo WikiCrimesIE

Esta análise evidenciou pontos de melhorias para o SIM: número de níveis da rede inferencial de conceitos a ser considerado pelo SIA; relações n-árias entre conceitos; relações com teor negativo; integração de uma solução de resolução de referência (anáfora pronominal, anáfora conceitual etc); pesquisa de expressões linguísticas com múltiplas palavras; definição de padrões gramaticais para conceitos da Base Conceitual.

Difícil encontrar sistemas que se propõem a realizar inferências de natureza complexa com base em textos. Textual Inference Logic (TIL) (de Paiva et al., 2007) (Bobrow et al., 2005) fornece mecanismos de raciocínio, baseados em um léxico unificado WordNet/VerbNet (Crouch e King, 2005) e em lógica descritiva. Por estes mecanismos, TIL consegue responder se uma sentença pode ser deduzida de outra ou é uma contradição de outra. Por exemplo, se a sentença "*A person arrived in the city*" é concluída de "*Ed arrived in the city*".

As inferências realizadas pelo SIA são materiais (autorizadas pelo conteúdo inferencial de conceitos e sentenças-padrão). Esta característica associada ao conteúdo semântico expresso nas bases do SIM possibilita realizar inferências para identificar/classificar o local específico do crime relatado no texto e não apenas identificar/classificar quaisquer endereços e locais mencionados no texto. Este é o caso dos sistemas de REM ou NER em geral, e de abordagens como a de Borges et al. (2007) para descoberta de localizações geográficas, a qual define seis padrões sintáticos de endereçamento. Esta abordagem apresentou uma precisão de 99,60% em reconhecer endereços explícitos no texto. O predomínio de abordagens sintáticas para a tarefa de EI evidencia a pressuposição de uma equivalência entre sintaxe e semântica das linguagens naturais.

Para a tarefa de extração do tipo de crime, nenhuma abordagem dos trabalhos relacionados se aplica a extração de informações desta natureza, principalmente por se tratar de informação comumente implícita em textos. O SIM ao explicitar o conteúdo inferencial dos conceitos, o

qual expressa as situações de uso dos mesmos, permite ao sistema uma base para explicações, argumentações, refutações, as quais permitem inferir conhecimento implícito.

6. Conclusão

Neste artigo foi descrito um analisador semântico de sentenças – o SIA, baseado nas teorias semânticas inferencialistas. O diferencial do SIA está em seu processo de raciocínio material e holístico sobre conceitos e sentenças e o uso de conhecimento inferencialista. O SIA implementa um processo sistemático para gerar as premissas e conclusões de sentenças, provendo a base para boas argumentações, respostas e explicações. A aplicação do SIA em um sistema de extração de informações sobre crimes – WikiCrimesIE - está sendo o seu cenário de avaliação. Os resultados obtidos até aqui foram motivadores, principalmente quando analisados os resultados da extração do tipo do crime. A extração de informações desta natureza exigem uma melhor capacidade para entendimento de textos em linguagem natural por parte de sistemas de PLN, principalmente, por não estarem, em sua maioria, explícitas no texto. Por exemplo, em 72% dos textos avaliados, as informações sobre o tipo do crime, causa do crime, tipo de vítima e tipo de arma utilizada não estavam explícitas, sendo necessário uma compreensão das notícias para que elas pudessem ser extraídas.

Como trabalhos em andamento, estamos otimizando a implementação do SIA, incluindo uma solução para resolução de anáforas e estendendo os objetivos de extração do WikiCrimesIE (causa do crime, tipo de vítima e tipo de arma utilizada). Um trabalho futuro será a divulgação do portal InferenceNet.org para que as bases semânticas do SIM possam ser usadas pela comunidade de PLN e, com isso, possibilitar a evolução do conhecimento inferencialista expresso. Outros trabalhos futuros incluem a investigação de: novas regras de inferência que combinem de forma diferente o conteúdo inferencial de conceitos e sentenças; técnicas de aprendizado automático e/ou semiautomático de conteúdo inferencialista; mecanismos de inferência com base no conteúdo inferencial de duas ou mais sentenças do texto, os quais gerem uma rede inferencial do texto; complexidade do algoritmo SIA.

Referências

Amaral, C. et al. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. Em Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca.

Bick, E. The Parsing System ”Palavras”. 2000. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.

Bobrow, D.G. et al. 2005. A basic logic for textual inference. In: Proceedings of the AAAI Workshop on Inference for Textual Question Answering, Pittsburg, PA.

Borges, K. Laender, A., Medeiros, C. e Clodoveu, D.Jr. 2007. Discovering geographic locations in web pages using urban addresses. Proceedings of the 4th ACM workshop on Geographical Information Retrieval (GIR’07), p.31-36, Lisboa, Portugal.

Brandom, R.1994. Making it Explicit. Cambridge, MA, Harvard University Press.

Brandom, R.B. 2000. Articulating Reasons. In: An Introduction to Inferentialism. Harvard University Press, Cambridge.

Budanitsky, A e Hirst, G. 2001. Semantic distance in Wordnet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources, 2nd meeting of the NAACL, Pittsburgh, PA.

Cardoso, N. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca.

Crouch, R; King, T.H. 2005. Unifying lexical resources. In: Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbruecken, Germany.

De Paiva, V. et al. 2007. Textual Inference Logic: Take Two. In: Proceedings of the Workshop on Contexts and Ontologies, Representation and Reasoning, CONTEXT 2007, 27-36.

Dummett, M. 1973. Frege’s Philosophy of Language. Harvard University Press.

Dummett, M. 1978. Truth and Other Enigmas. Duckworth, London.

Eco, U. 2001. A Busca da língua perfeita na cultura européia. EDUSC, São Paulo.

Freitas, C. et al. 2008. ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista. In: Cristina Mota & Diana Santos (eds.). Desafios na avaliação conjunta do reconhecimento de

entidades mencionadas: O Segundo HAREM, 309-317, Linguateca.

Furtado, V et al. 2009. Collective intelligence in law enforcement – The WikiCrimes system. Information Sciences, In Press, Corrected Proof, Available online, August 2009. doi:10.1016/j.ins.2009.08.004.

Gentzen, G. 1935. Untersuchungen über das logische Schliessen. Mathematische Zeitschrift, 39, pp.176-210, pp. 405-431, 1935. Translated as 'Investigations into Logical Deduction',and printed in M. Szabo The Collected Papers of Gerhard Gentzen,Amsterdam: North-Holland,1969, 68–131.

Girju, R. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).

Grishman, R. Information Extraction. 2003. In: Mitkov. R. (ed). Oxford Handbook of Computational Linguistics, Oxford University Press, 545-559.

Mota, C. e Santos, D. (eds.) 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca. (ISBN: 978-989-20-1656-6)

Nogueira, D., Pinheiro, V., Furtado, V., Pequeno, T. 2009. Desenvolvimento de Sistemas de Extração de Informações para Ambientes Colaborativos na Web. In proceedings of the II International Workshop on Web and Text Intelligence (WTI – 2009), co-located with STIL 2009.

Pinheiro,V., Pequeno, T., Furtado, V., Assunção, T. e Freitas, E. 2008. SIM: Um Modelo Semântico-Inferencialista para Sistemas de Linguagem Natural. VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2008), WebMedia, Brasil.

Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D. 2009. Information Extraction from Text Based on Semantic Inferentialism. T. Andreasen et al. (Eds.): FQAS 2009, Springer Berlin / Heidelberg, LNAI 5822, pp. 333–344.

Pinheiro, V., Pequeno, T., Furtado, V., Franco, W. 2010. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. International Conference on Computational Processing of Portuguese Language (PROPOR) (to appear).

Prawitz, D. 1965. Natural Deduction: A Proof Theoretical Study. Stockholm:Almqvist & Wiksell.

Sellars, W. 1980. Inference and meaning (1950). Reprinted in Pure Pragmatics and Possible Worlds. Ed. J.Sicha. Reseda, California. Ridgeview Publishing Co.

Vieira, R. e De Lima, V.L.S. 2001. Linguística Computacional: Princípios e Aplicações. Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial. v.3. p. 47-86.

Wittgenstein, L. 1953. Philosophical Investigations. Tradução G.E.M.Anscombe, Oxford:Basil Blackwell.