UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JONATHAN DARLAN CUNEGUNDES MOREIRA

ENRIQUECIMENTO SEMÂNTICO DE PERFIL DE USUÁRIO PARA APOIO A UM MODELO DE APRENDIZAGEM INFORMAL NO CONTEXTO DA SAÚDE

MOSSORÓ – RN
2015

JONATHAN DARLAN CUNEGUNDES MOREIRA

ENRIQUECIMENTO SEMÂNTICO DE PERFIL DE USUÁRIO PARA APOIO A UM MODELO DE APRENDIZAGEM INFORMAL NO CONTEXTO DA SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação – associação ampla entre a Universidade Federal Rural do Semi-Árido e a Universidade do Estado do Rio Grande do Norte, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Francisco Milton Mendes Neto – UFERSA.

Coorientador: Prof. Dr. Ricardo Alexandro de Medeiros Valentim – UFRN.

MOSSORÓ – RN
2015

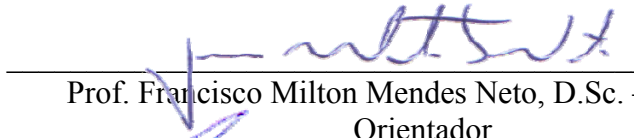
JONATHAN DARLAN CUNEGUNDES MOREIRA

ENRIQUECIMENTO SEMÂNTICO DE PERFIL DE USUÁRIO PARA APOIO A UM MODELO DE APRENDIZAGEM INFORMAL NO CONTEXTO DA SAÚDE

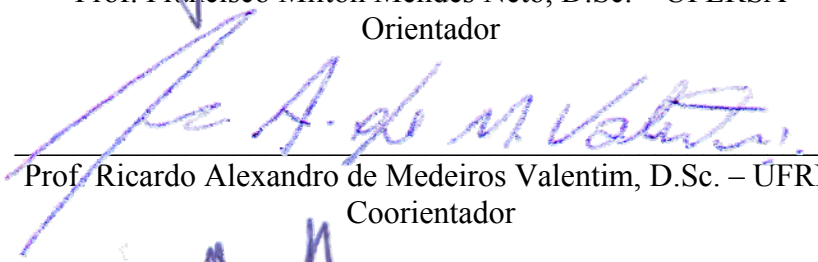
Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 18 / 03 / 2015 .

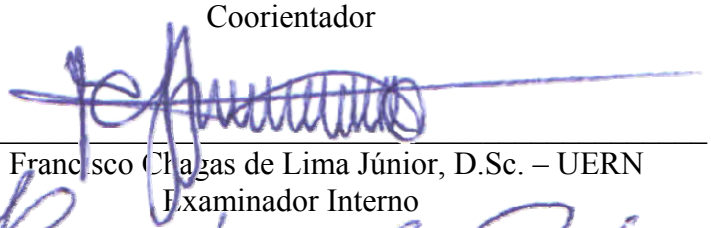
BANCA EXAMINADORA



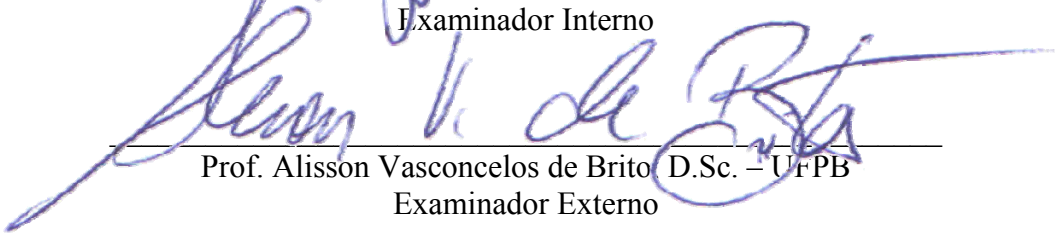
Prof. Francisco Milton Mendes Neto, D.Sc. – UFERSA
Orientador



Prof. Ricardo Alexandro de Medeiros Valentim, D.Sc. – UFRN
Coorientador



Prof. Francisco Chagas de Lima Júnior, D.Sc. – UERN
Examinador Interno



Prof. Alisson Vasconcelos de Brito, D.Sc. – UFPB
Examinador Externo

Dedico este trabalho especialmente a Deus, ao meu amado filho, Eduardo Moreira, a minha esposa, Renata Moreira, e aos meus pais, Francisco das Chagas e Maria das Graças Moreira.

AGRADECIMENTOS

A Deus por ter me dado o dom da vida e por me amparar em todos os momentos. Sem ele não conseguiria forças para seguir sempre em frente.

Aos meus pais pelo amor e apoio incondicional. Minha mãe Maria das Graças, minha heroína, que me ensinou o que é dedicação, determinação e doação. Meu pai Francisco das Chagas, um porto seguro, que sempre me ouviu e aconselhou de forma sensata e paciente, ensinando a contornar as adversidades da vida.

À minha família que me oferece suporte e motivação para atingir meus objetivos. À minha esposa Renata, pela dedicação e paciência, ajudando em todas as dificuldades. Ao nosso filho, Eduardo, que trouxe alegria e amor ao nosso lar. A eles devo tudo e por eles luto.

À minha irmã, Giza, ao cunhado, Clézio e ao meu sobrinho Miguel, pelo carinho, companheirismo e amizade.

Ao orientador Prof. Francisco Milton Mendes Neto, que soube guiar com destreza e discernimento. Orientando sempre e exigindo quando necessário. Que durante esse período buscou extrair o meu melhor e hoje posso chamá-lo de amigo.

Ao coorientador Prof. Ricardo Alexandro de Medeiros Valentim, pela disposição e suporte ao logo deste processo. Auxiliando sempre que possível para que pudesse atingir o objetivo final.

Aos amigos da SUTIC, companheiros de trabalho e que fizeram parte da minha formação, sempre dispondo seus conhecimentos para ajudar quando foi preciso. Aos colegas do LES, pela cooperação e compartilhamento de conhecimentos.

A esta universidade, seu corpo docente, direção e administração, pela competência, comprometimento e ética que tornaram possível essa conquista.

Por fim, a todos que direta ou indiretamente fizeram parte da minha formação. Para finalizar, aproveito para parafrasear Napoleão Bonaparte, que disse: "O entusiasmo é a maior força da alma. Conserva-o e nunca te faltará poder para conseguires o que desejas".

“Conhecer o homem - esta é a base de todo o sucesso.”

(Charles Chaplin)

RESUMO

Pessoas com doenças crônicas sofrem com limitações impostas por sua condição de saúde e aprender mais sobre a doença ajuda na melhoria de sua qualidade de vida. Esta aprendizagem deve acontecer dentro e fora do ambiente hospitalar, ocorrendo também por meio das experiências sociais cotidianas destes indivíduos. Os meios digitais favorecem a propagação do conhecimento, porém, ao mesmo tempo promovem o acesso a uma ampla quantidade e diversidade de conteúdos dificultando a localização de informações relevantes para as necessidades particulares de cada indivíduo. Sistemas de recomendação podem ajudar a fornecer o conhecimento certo (contextualizado), no momento certo, no entanto precisam conhecer bem o usuário e os conteúdos. Existe uma resistência dos usuários em prover informações sobre si mesmos e sobre conteúdos através do preenchimento de formulários. Contudo, não é uma tarefa trivial prover informações dinamicamente com base nas interações digitais do usuário sem a necessidade de sua interferência. Este trabalho propõe unir diferentes técnicas de enriquecimento semântico de modo que seja possível determinar os interesses do usuário e, assim, auxiliar na eficácia da recomendação de conteúdos relevantes às pessoas portadoras de doenças crônicas, favorecendo a aprendizagem informal em saúde.

Palavras-Chave: enriquecimento semântico, traços digitais, processamento de linguagem natural, ontologias, aprendizagem informal, saúde 2.0.

ABSTRACT

Persons with chronic diseases have limitations imposed by their health condition. Learning about their disease helps improving their life quality. This learning must occurs within and outside the hospital environment through everyday social experiences of these individuals. Digital media promotes the knowledge spread, but at the same time, promotes access to a wide range and diversity of content hindering find relevant information for individual particular needs. Recommender systems can help provide the right knowledge (contextualized), at the right time. However they need to know the users and contents. There is a resistance by user to provide information about themselves and contents through form filling. However, it is not a trivial task to provide information dynamically based on digital interactions user without their interferences. This work proposes uniting different semantic enrichment techniques so that it is possible to determine the user's interests. And thus help in the recommendation's effectiveness of relevant content to persons with chronic diseases, favoring the informal learning in health.

Keywords: semantic augmentation, digital traces, natural language processing, ontologies, informal learning, health 2.0.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS

ALSA	*Amyotrophic Lateral Sclerosis Association*
API	*Application Programming Interface*
EHR	*Electronic Health Record*
ELA	Esclerose Lateral Amiotrófica
GATE	*General Architecture for Text Engineering*
HL7	*Health Level Seven*
HTML	*Hypertext Markup Language*
HUOL	Hospital Universitário Onofre Lopes
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDF	*International Diabetes Federation*
ITU	*International Telecommunication Union*
LAIS	Laboratório de Inovação Tecnológica em Saúde
LES	Laboratório de Engenharia de Software
LN	Linguagem Natural
OMS	Organização Mundial de Saúde
OWL	*Ontology Web Language*
PHR	*Personal Health Record*
PLN	Processamento de Linguagem Natural
RDF	*Resource Description Framework*
SR	Sistema de Recomendação
TD	Traço Digital
TIC	Tecnologia da Informação e Comunicação

UFERSA	Universidade Federal Rural do Semi-Árido
UPO	*User Profile Ontology*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
W3C	*World Wide Web Consortium*
XML	*eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Uma pesquisa divulgada em 2009 pelo Instituto Brasileiro de Geografia e Estatística (IBGE) mostra que 29,9% da população são portadoras de pelo menos uma doença crônica, chegando a ser de 75% nos idosos. Apesar das limitações impostas por essas doenças, é possível promover uma melhor qualidade de vida a seus portadores, mas, para isso, deve ser seguido o tratamento recomendado por profissionais da saúde (IBGE, 2009).

Contudo, essa pesquisa também expõe que existe um alto número de pessoas que não aderem ao tratamento e que poderiam ter prevenido ou controlado a doença se cuidassem mais da sua saúde. Um dos fatores disso é a falta de conhecimento acerca da sua condição de saúde. É necessário o paciente conhecer a doença, aceitar e assumir a responsabilidade de proteção à sua saúde, tornando-se ativo no planejamento e execução de seu tratamento. Isso, aliado aos outros fatores, como tratamento, medicamentos, dietas, hábitos saudáveis, mudanças no estilo de vida e prática de exercícios, contribui para o controle e melhoria na qualidade de vida dos pacientes que possuem doenças crônicas (LUSTOSA; ALCAIRES; COSTA, 2011).

A aprendizagem informal é uma forma de fornecer conhecimento ao individuo fora do ambiente hospitalar, por meio de suas experiências sociais cotidianas. Esse tipo de aprendizagem aliada a tecnologias móveis e a internet remete ao conceito de Saúde 2.0, que visa estimular o individuo a ser responsável pelos cuidados relativos à sua saúde através da utilização de ferramentas da Web 2.0.

Estas tecnologias possibilitam interações sociais, o compartilhamento de vivências cotidianas e, consequentemente, a disseminação do conhecimento. Porém, ao mesmo tempo elas promovem o acesso a uma ampla quantidade e diversidade de conteúdo, o que pode acarretar em perda de desempenho dos indivíduos, por provocar a dispersão destes.

Uma forma de contornar esta situação é recomendando conteúdo de forma personalizada considerando as características particulares de cada usuário, com base nas informações resultantes de suas interações.

1.2. PROBLEMATICA

Elaborar um meio personalizado de aquisição de conhecimento requer a obtenção de informações relevantes do perfil do usuário, de forma a identificar seus interesses sem que haja a necessidade de sua intervenção. Contudo, as informações contidas no ambiente virtual não são fáceis de serem recuperadas, pois estão dispersas em diferentes formas e ferramentas e, na maioria das vezes, em linguagem natural. Normalmente também estão em ambientes com domínio não definido, sendo necessária uma contextualização dessas informações com domínios de conhecimento.

Nesse contexto, surge um novo problema: como determinar, de forma implícita, os interesses do usuário relacionados à sua saúde através de suas interações cotidianas? O contexto diário do usuário diz muito sobre seus interesses, e os Traços Digitais (TDs) possuem um grande potencial a ser explorado, trazendo experiências reais do usuário. Isto, aliado a técnicas de enriquecimento semântico, pode auxiliar na determinação dos interesses e necessidades reais do usuário, viabilizando a personalização do seu ambiente.

1.3. PROPOSTA DE SOLUÇÃO

Considerando a problemática apresentada na seção anterior, esta dissertação tem como objetivo o desenvolvimento de uma solução de enriquecimento semântico de perfil de usuário aplicado ao contexto da saúde, de modo que possa determinar os interesses do usuário em assuntos relacionados à sua condição de saúde. Este processo deve ocorrer de forma implícita, considerando o contexto diário do usuário e relacionando seus TDs a domínios de conhecimento, sem a necessidade de sua intervenção.

Para isso, o sistema analisa os TDs do usuário através de Processamento de Linguagem Natural (PLN) gerando relações entre o TD e ontologias de domínio de conhecimento na área da saúde. Estas relações permitem inferências sobre o perfil do usuário para determinar seus interesses nestes domínios.

O trabalho proposto irá compor, de forma integrada, a arquitetura do MobiLEHealth, que é um ambiente de aprendizagem informal no contexto de Saúde 2.0. Esse ambiente é destinado a pessoas portadoras de doença crônica e promove o conhecimento sobre a doença e, consequentemente, uma melhoria na sua qualidade de vida.

Como componente integrante deste ambiente, o sistema de enriquecimento semântico de perfil de usuário tem o objetivo de auxiliar no processo de seleção personalizada de conteúdos do MobiLEHealth. Para isto, utiliza as técnicas citadas anteriormente, e que serão detalhadas nas seções subsequentes, para processar os TDs previamente capturados pelo MobiLEHealth e armazenados em uma base de dados própria e centralizada. Como resultado deste processamento, busca-se extrair informações relevantes sobre os interesses do usuário em relação a sua saúde, fornecendo o apoio necessário ao MobiLEHealth para que possa promover a melhoria na qualidade de vida de pessoas portadoras de doenças crônicas.

1.4. ORGRANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada da seguinte forma: o Capítulo 2 descreve a base teórica para o desenvolvimento da proposta. No Capítulo 3, é abordada a proposta deste trabalho, apresentando uma descrição geral do MobiLEHealth, a arquitetura do ambiente e o escopo deste trabalho. No Capítulo 4 são discutidos os resultados obtidos na validação da proposta. Por fim, no Capítulo 5 são apresentadas as considerações finais e as perspectivas para trabalhos futuros.

2. REFERENCIAL TEÓRICO

2.1. APRENDIZAGEM INFORMAL NA SAÚDE

A educação em saúde constitui um importante recurso dos indivíduos para a determinação do seu bem-estar físico, psíquico e social. O *Health Promotion Glossary* (WHO, 2008), publicado pela Organização Mundial de Saúde (OMS), define a educação em saúde como sendo o conjunto de competências cognitivas e sociais que determinam a capacidade dos indivíduos para compreender e usar informações de modo que promovam e mantenham uma boa saúde.

Lustosa et al. (2011) afirmam que a adesão ao tratamento é o fator mais importante para o controle efetivo de muitas doenças, principalmente as crônicas. Porém um dos fatores que contribuem para uma elevada taxa de não adesão ao tratamento é a falta de informação por parte do paciente.

O conhecimento fornece recursos e meios para que cada indivíduo possa alcançar qualidade de vida no que se refere à sua saúde. Sem ele o indivíduo tem dificuldades de contornar os problemas, aprender a conviver com a doença, descrever sintomas, compreender instruções médicas, seguir o tratamento, compreender sua doença, prevenir-se, saber de suas limitações e possibilidades, comunicar-se com o profissional de saúde, conviver em sociedade, entre outros (BAKER *et al.*, 2008).

O desconhecido gera dúvidas, medos e anseios, fazendo com que o indivíduo se sinta coagido ao invés de agir e reagir perante a sua doença. Diante deste cenário, fica claro que a qualidade de vida de um portador de doença crônica está diretamente ligada ao seu conhecimento acerca de sua saúde.

Contudo o processo de aprendizagem não depende apenas de contextos formais de educação para saúde, mas também de contextos informais. Fazendo com que a aquisição do conhecimento por parte do individuo transcenda as barreiras do ambiente médico-hospitalar, tornando-se um processo integrado ao seu cotidiano.

A aprendizagem formal caracteriza-se como estruturada e apoiada institucionalmente. Ela ocorre sob a supervisão de um orientador que planeja, implementa e avalia as etapas do processo de aprendizagem (MERRIAM; CAFFARELLA; BAUMGARTNER, 2006).

Já a aprendizagem informal é o processo contínuo de aquisição do conhecimento por um indivíduo. Sendo este responsável pelo seu aprendizado e desenvolvimento, que ocorre através de suas experiências cotidianas (WANG; SHEN, 2012). Este tipo de aprendizagem está relacionado à busca por novos conhecimentos e habilidades de modo não planejado ou estruturado. De forma consciente ou inconsciente, ela emerge a partir de alguma demanda ou necessidade. Dentre as suas principais características estão (JIUGEN; RUONAN; XIAOQIANG, 2011):

- **Autonomia**: onde o indivíduo é responsável por sua aprendizagem e determina o objetivo a ser alcançado, o conteúdo a ser abordado e como este será processado;
- **Conhecimento**: obtido por meio das interações sociais e profissionais;
- **Diversidade**: obtida através de experiências diárias, recursos do ambiente, bibliotecas, redes sociais, dentre outros.

O processo de aprendizagem em ambientes informais relaciona-se essencialmente com aprendizagens sociais. Ele se entrelaça com a vida dos indivíduos conforme suas experiências. Este processo tem uma natureza abrangente, englobando domínios do desenvolvimento pessoal, social e cultural. Por isso, o contexto diário do individuo contribui fortemente para sua aprendizagem, que pode ocorrer através de atividades diárias, das interações sociais, por iniciativa própria, pela reflexão e pelos próprios erros (CASTLETON; GERBER; PILLAY, 2006).

Segundo Machles (2003), quando as pessoas interagem com o ambiente em que estão inseridas, desenvolvem codificações mentais de experiências que incluem maneiras particulares de percepções cognitivas e respostas a um conjunto de estímulos ou situações complexas. O processo de aprendizagem informal abrange esse conceito com base nas experiências únicas vivenciadas pelo indivíduo, mesmo que este tenha participado de um processo de aprendizagem formal em grupo.

Ambientes de aprendizagem informal em saúde favorecem a disseminação de informações relacionadas à condição de saúde de um individuo. Este conhecimento, em conjunto com o acompanhamento médico, contribui na qualidade de vida do portador de doença crônica. A Figura 1 projeta essa relação, exemplificando alguns dos elementos envolvidos.

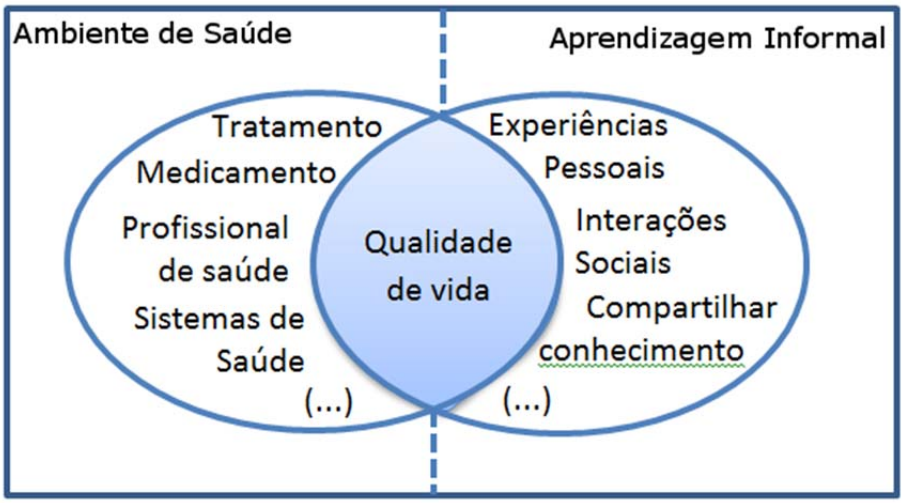


Figura 1 - Relação entre o ambiente de saúde e a aprendizagem informal
Fonte: Autoria própria

A transferência de parte da responsabilidade do profissional e das entidades de saúde para o individuo, em questões relacionadas à sua saúde, faz com que este seja responsável e ativo pelos cuidados relativos a ela. Isto o motiva a buscar o conhecimento e, como consequência, ele obtém uma melhor qualidade de vida. Um vez que, em uma perspectiva voltada para a atuação preventiva, o indivíduo consegue conviver melhor com sua doença, prevenir futuros problemas, superar suas limitações, manter as relações sociais ou, nos piores casos, amenizar o sofrimento.

2.2. SAÚDE 2.0

Segundo o Relatório Global de 2013 da *International Telecommunication Union* (ITU), há mais de 6 bilhões de celulares habilitados em todo o mundo e aproximadamente 3 bilhões com acesso à internet banda larga móvel (ITU, 2014). Esta realidade surge como uma forte aliada no contexto da saúde, pois, oferecem um ambiente favorável à aprendizagem informal em saúde, uma vez que fornecem ferramentas e apoio à criação e à disseminação coletiva do conhecimento (MARKKULA; SINKO, 2009).

O uso de tecnologia para beneficiar a área de saúde já vem sendo estudado há algum tempo, tanto que dele surgiram diversos termos, como:

- ***e-health***: abreviação de *electronic health*, que consiste no uso da internet ou outros meios eletrônicos de compartilhamento de dados e de serviços relacionados à saúde (DELLA MEA, 2001);
- **telemedicina:** que se utiliza das Tecnologias da Informação e Comunicação (TICs) para prestar cuidados clínicos a distância (DELLA MEA, 2001);
- ***m-health*:** abreviação de *mobile health*, usado para designar a prática da medicina e da saúde pública apoiada por dispositivos móveis (ISTEPANIAN *et al.*, 2005);
- ***u-health*:** abreviação de *ubiquitous health*, que se refere a serviços de saúde e informação incorporados ao dia-a-dia do indivíduo no conceito de *anywhere and anytime* (MOHAMMED; FIAIDHI, 2010);
- ***health 2.0:*** saúde 2.0 em português, que visa estimular o cidadão a ser ativo e responsável pela sua saúde e pelos cuidados relativos a ela, através de iniciativas mediadas pelas tecnologias da Web 2.0 para a construção coletiva do conhecimento (FERNANDEZ-LUQUE *et al.*, 2010).

Este último prega a saúde participativa, onde o foco é o indivíduo, apoiada por informação, software e comunidades on-line, com o objetivo de munir os indivíduos de conhecimento para que sejam parceiros eficazes nos cuidados da própria saúde (HUGHES; JOSHI; WAREHAM, 2008). O indivíduo não assume o lugar do profissional da saúde, mas se torna ativo neste processo, sendo mais autônomo e consciente. Além de adquirir o conhecimento necessário para participar nas tomadas de decisões.

Esta pesquisa está diretamente relacionada ao contexto da Saúde 2.0, uma vez que busca fornecer apoio a um ambiente de aprendizagem informal em saúde. Para isso considera a experiência cotidiana do usuário na web para determinar os seus interesses relacionados à sua saúde.

2.2.1. Web 2.0

A Web 2.0 é a base da Saúde 2.0 e alguns fatores contribuem para isto. Primeiramente por que a interatividade é uma das suas principais características. Isto permite o desenvolvimento de arquiteturas de socialização através de redes sociais, conexões entre os usuários, *wikis*, *podcasts*, *tags*, *blogs*, textos, imagens, músicas, vídeos, comentários, gostos,

relacionamentos e inúmeras outras. Ou seja, há um intenso compartilhamento de conhecimento, cenário que se encaixa na ideologia da Saúde 2.0, que busca a construção coletiva do conhecimento.

Outro fator importante é que as ações e interações dos usuários, comumente, são registradas em bases de dados, fazendo com que possam ser recuperadas e processadas. Segundo O'Reilly (2009), construir bases de dados que consigam gerir a informação que os utilizadores têm para acrescentar é a chave de produção da Web 2.0. Esta característica faz com que tenha um enorme potencial para recuperação e extração de informações em saúde quando aplicada técnicas apropriadas.

Na área da saúde podemos destacar, dentro do contexto deste trabalho, as ferramentas de Registros Pessoais de Saúde (PHR), do Inglês *Personal Health Record*, que armazenam informações relacionadas à saúde que são documentadas e mantidas pelo próprio indivíduo (TOLEDO, 2013). Diferentemente do Histórico Eletrônico de Saúde (EHR), do inglês *Eletronic Health Record*, que é mantido por um profissional da saúde e mantém o histórico médico de uma pessoa (MALIK; SULAIMAN, 2013).

Os PHRs mantêm informações sobre alergias, histórico familiar, imunizações, visitas a profissionais de saúde, internações hospitalares, medicações, entre outras, todas mantidas pela pessoa que detém a informação, ou seja, o usuário.

Levando em consideração que o foco da Saúde 2.0 é o indivíduo, estas informações, em conjunto com outras ferramentas e tecnologias da Web 2.0 cujo foco também é o indivíduo, mostram um grande potencial a ser explorado para se determinar os interesses do usuário, com o objetivo a apoiar a aprendizagem informal em saúde.

2.3. TRAÇOS DIGITAIS

Os TDs são rastros das pessoas em ambientes virtuais que contêm suas interações sociais. Eles evoluem como consequência dos ambientes em que estão incorporados e estabelecem uma relação com o mundo real do usuário, representando seus pontos de vistas, interesses, experiências e emoções (DESPOTAKIS; LAU; DIMITROVA, 2011).

A análise dos TDs possibilita estabelecer relações geralmente difíceis de serem estudadas em ambientes convencionais, possibilitando a medição do comportamento humano coletivo com base em conjuntos de grandes sistemas sociais (KLEINBERG, 2008).

Para estabelecer essas relações, devem-se considerar algumas características dos TDs, como:

- A identificação, que permita a localização dentro de uma infraestrutura de informação;
- A sensibilidade, que é a capacidade de controlar e responder a alterações no contexto do usuário;
- A comunicabilidade, que representa a interação e o compartilhamento entre as entidades sociais;
- A memorização, possibilitando lembrar os resultados das interações;
- A rastreabilidade, que é a capacidade inter-relacionar cronologicamente eventos e entidades no tempo e no espaço;
- A associabilidade, que associa informações relacionadas a entidades, eventos e lugares, representando pontos ou eventos de interação através da plataforma na qual colaboram.

Segundo Kleingberg (2008), os dados sobre as relações sociais podem fornecer informações importantes que permitem desenvolver e avaliar modelos complexos de fenômenos sociais. Redecker e Punie (2010) afirmam que as tendências apontam que a mídia social irá ter um forte impacto sobre a aprendizagem informal, fornecendo conteúdo gerado pelo usuário para auxiliar o processo de aprendizagem relacionado a experiências do mundo real.

Estes traços coletados, combinado com aplicativos móveis e inteligentes, podem ser transformados em uma imagem contínua e em tempo real da nossa saúde pessoal (CHEN *et al.*, 2012).

2.4. PROCESSAMENTO DE LINGUAGEM NATURAL

Os seres humanos são capazes de estabelecer um processo de comunicação complexo, utilizando a forma oral, escrita e sinais, formando a Linguagem Natural (LN). Essa capacidade é possível devido ao conhecimento sobre a representação da linguagem utilizada e do entendimento de mundo em relação aos elementos envolvidos. Entretanto, não é uma tarefa trivial quando se trata de agentes computacionais, sendo necessário um processamento

da informação através de técnicas e ferramentas que proporcionem um entendimento sobre a LN (CAMBRIA; WHITE, 2014).

O PLN é um ramo da Inteligência Artificial (IA) que busca converter ocorrências das linguagens naturais em representações mais formais e manipuláveis por programas de computador, com o objetivo de compreender a informação. Ele analisa os radicais e origem das palavras para determinar e encontrar suas variações, como conjugações de verbos, flexões de adjetivos, graus nos substantivos, etc. Ele também isola frases e períodos para buscar as relações sintáticas e semânticas. Dessa forma, as informações podem ser utilizadas e o tratamento a ser dado depende da aplicação que está sendo usada (RUSSELL; NORVIG, 2009).

Segundo Balsa (2004), o PLN é considerado uma das áreas designadas por IA-Completa[1], já que para alcançar o seu objetivo final corresponde a conseguir resolver os problemas da IA, de representação do conhecimento, aprendizagem e raciocínio envolvendo conhecimento arbitrário sobre o mundo real.

O PLN possui um subconjunto de entrada e/ou saída codificado em uma linguagem natural e o processamento da entrada e/ou a geração da saída deve ser baseado no conhecimento, sobre aspectos sintáticos, semânticos e/ou pragmáticos, de uma linguagem natural.

Aplicações de PLN são prejudicadas pela incompletude dos recursos linguísticos utilizados (não existe, por exemplo, um dicionário que contenha todas as palavras utilizadas pela língua portuguesa), pela complexidade das tarefas específicas do processamento, pela ausência de um rigor absoluto na utilização habitual de uma língua natural e pelo fato das línguas naturais serem dinâmicas (LOPES; ROCIO; SILVA, 1999).

Como os TDs são normalmente escritos em linguagem natural, o PLN se torna necessário para entender a sua estrutura e significado, possibilitando o tratamento e compreensão computacional (POHOREC *et al.*, 2012).

[1] IA-Completa supõe visão computacional, processamento de linguagem e tratamento de circunstâncias não previstas para solucionar problemas do mundo real. Estes problemas não podem ser solucionados por um algoritmo simples e são equivalentes a solucionar o problema central da Inteligência Artificial, tornando os computadores tão inteligentes quanto às pessoas.

2.5. ONTOLOGIA

Ontologia é certamente uma das tecnologias mais importantes da web semântica. Principalmente quando se trata de enriquecimento semântico. Só o PLN não é suficiente quando existe a necessidade de entender o conteúdo. No processo de entendimento está incluso o conhecimento de mundo e as inter-relações entre conceitos e significados e a ontologia auxilia neste processo.

Segundo Allemang e Hendler (2011), ontologia é uma especificação explícita e formal de uma conceptualização compartilhada, que permite criar modelos abstratos, através de um conjunto de entidades, relações, restrições, axiomas e vocabulários. As ontologias são descritas por linguagens com sintaxe e semântica bem definida e expressas em lógica descritiva, o que possibilita a inferência por agentes computacionais.

O *World Wide Web Consortium* (W3C) especificou o *Resource Definition Framework* (RDF), uma linguagem escrita em formato XML (*eXtensible Markup Language*) para definição de recursos para ontologias através de suas propriedades e respectivos valores (W3C, 2014). Estes recursos são chamados de *statements* e são armazenados em forma de triplas [Sujeito, Predicado, Objeto], por exemplo, [diabetes, é uma, doença crônica]. Neste esquema, os sujeitos e objetos representam os conceitos do domínio a ser modelado. Os predicados, também chamados de propriedades, definem as relações, características e restrições (POWERS, 2003).

O RDF é limitado na definição de conceitos. Por este motivo, um grupo do W3C criou a *Ontology Web Language* (OWL) (GROUP, 2012) como uma extensão ao RDF, mas sem as limitações do antecessor. A OWL permite o encadeamento de propriedades, tipagem, cardinalidade, restrições e anotações, fornecendo muito mais recursos para a realização de inferências eficientes sobre seus recursos (ANTONIOU *et al.*, 2012). Estas tecnologias incorporam parte da arquitetura (Figura 2) da Web Semântica recomendada pelo W3C, que é constituída por (BECHHOFER *et al.*, 2014):

- **XML:** sintaxe universal de marcação de documentos;
- **XML *Schema*:** define a estrutura dos documentos XML;
- **RDF:** modelo de dados para descrição de recursos;
- **RDF *Schema*:** vocabulário básico para definições de documentos RDFs, classes, propriedades e relações hierárquica;

- **OWL:** estende o vocabulário do modelo RDF para definição de classes e propriedades, incluindo cardinalidades, igualdade, relacionamentos entre classes e características de propriedades.

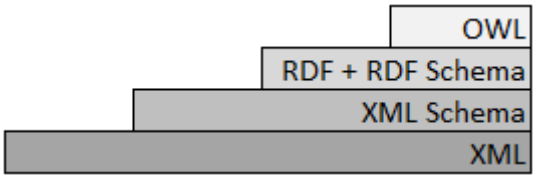


Figura 2 – Arquitetura da Web Semântica recomendanda pelo W3C
Fonte: (BECHHOFER *et al.*, 2014)

Todos os recursos das ontologias são identificados por um *Uniform Resource Identifier* (URI) que consiste de um *namespace* comum a toda a ontologia e um identificador de fragmento, separados por uma "#" (Figura 3). Através do URI, qualquer recurso da ontologia pode ser referenciado em uma fonte externa, já que o identifica unicamente, fornecendo uma estrutura flexível e poderosa para o enriquecimento semântico (GROUP, 2012).

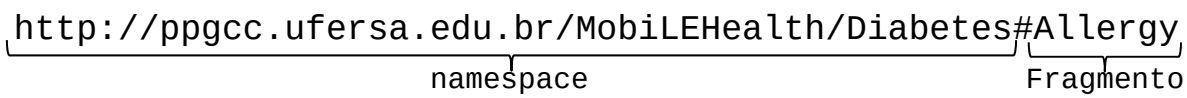


Figura 3 – Formato do URI
Fonte: Autoria Própria

Ontologias são criadas a partir de classes e propriedades que as relacionam. Ambas são derivadas do tipo *Resource* do RDF. As propriedades das ontologias, definidas pelo rdf:Property, definem as relações entre as classes e suas propriedades. A Figura 4 esquematiza a hierarquia dos elementos da OWL mostrando a relação com a camada do RDF.

A OWL possui basicamente três componentes distintos: classe, indivíduos e propriedades. As classes representam uma classificação dentro da ontologia permitindo uma hierarquização entre elas, adotando o conceito de herança. Também é possível gerar outras relações entre as classes, como a de igualdade ou disjunção. Os indivíduos representam uma concretização de um conceito, podendo este estar vinculado diretamente ou indiretamente a uma ou várias classes. As propriedades estabelecem restrição ou relação entre classes ou indivíduos e podem ser de dois tipos (i) de objeto (*Object Properties*), que definem relações

entre classes, que podem ser aplicáveis às classes ou às suas instâncias ou (ii) de dados (*Data Properties*), que definem relações de indivíduos com seus dados.

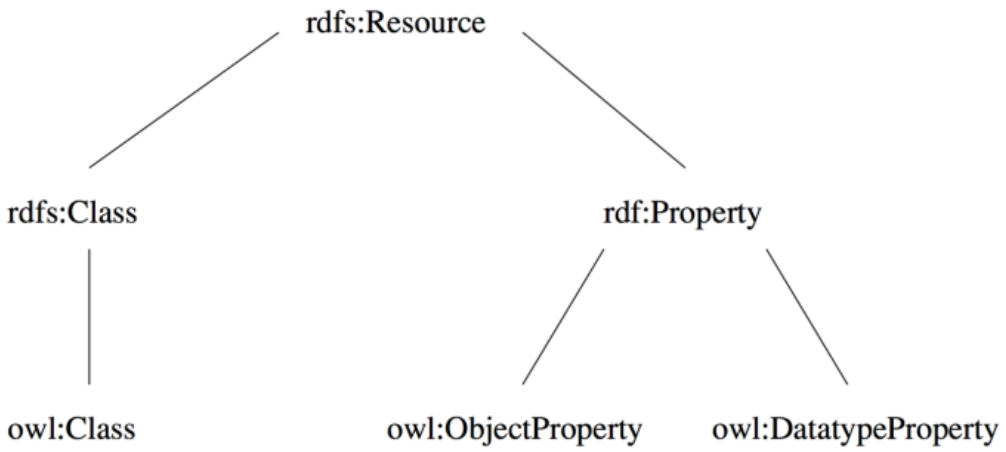


Figura 4 - Estrutura de subclasses da OWL/RDF
Fonte: (ANTONIOU *et al.*, 2012)

A OWL foi projetada para descrever uma ontologia de uma maneira que possa ser interpretada sem ambiguidade e usada por agentes de software. Dessa forma, a ontologia possibilita o uso de um vocabulário compartilhado criado sob a forma de taxonomia (JIANG; CONRATH, 1997) e um conjunto de regras de inferência, permitindo uma contextualização semântica entre agentes computacionais (HUHNS; STEPHENS, 1999).

Existem diversas pesquisas relacionadas à aplicação de ontologias na área de saúde, dentre elas podemos citar: (i) o trabalho apresentado por Nguyem el al. (2009) propõe o desenvolvimento de um modelo PHR em forma de ontologia, onde são descritos conceitos tais como informações demográficas, vacinas, alergias, diagnósticos, procedimentos e medicamentos; (ii) a pesquisa de Park et al. (2012), que propõe o uso de ontologia para integração entre bancos de dados heterogêneos da área de saúde, onde, baseado em um modelo de entidade-relacionamento sobre cada origem de ontologia, é gerado um esquema de banco de dados integrado e o mapeamento dinâmico entre essas fontes. Essas e outras pesquisas mostram o potencial do uso de ontologias na área de saúde, como forma de melhoria no atendimento e acompanhamento ao indivíduo.

O uso das ontologias para a modelagem do conhecimento permite que agentes possam fazer inferências sobre conteúdos, classificando-os e relacionando-os semanticamente, possibilitando sua compreensão e a extração de informações relevantes.

2.6. SISTEMAS DE RECOMENDAÇÃO

Algumas pessoas ainda não acham que a internet seja um instrumento totalmente confiável em razão da procedência e qualidade das informações que são fornecidas. Aliado a isso, a quantidade e a diversidade das informações disponíveis dificultam ainda mais a busca do conhecimento.

Os Sistemas de Recomendação (SRs) podem minimizar este problema por meio da recomendação personalizada de conteúdo, considerando as características particulares do usuário e as suas interações cotidianas (DA SILVA; MENDES NETO; JÁCOME JÚNIOR, 2011). Uma vez que possuam metadados suficientes sobre o conteúdo e o usuário, eles são capazes de analisar as relações existentes e selecionar conteúdos que se adequem às necessidades do usuário.

De acordo com Vieira e Nunes (2012), o aumento de meios de disponibilização de conteúdo, podendo ser serviços ou informação, através de sistemas web, provoca uma situação onde o usuário possui muitas opções de escolha antes mesmo de estar apto a selecionar uma opção que atenda suas necessidades. Os SRs buscam amenizar os impactos gerados por essa sobrecarga de informação.

Para tanto, os SRs utilizam repositórios de informação e dados de preferência dos usuários para direcionar conteúdos aos indivíduos com potenciais interesses. Um dos desafios destes sistemas é realizar a indicação de produtos, serviços e/ou informação que melhor atendam às expectativas dos usuários (CAZELLA *et al.*, 2012).

A personalização em SRs está fortemente relacionada em conhecer a quem se deseja recomendar e o Enriquecimento Semântico pode auxiliar estabelecendo uma similaridade entre usuários e domínios de conhecimento.

2.7. ENRIQUECIMENTO SEMÂNTICO

Uma das principais dificuldades dos sistemas de recomendação personalizada é a definição correta dos metadados a serem utilizados, que normalmente é feita de forma manual ou semiautomática. Esse procedimento se torna insuficiente quando se deseja recomendar de forma personalizada com base nas necessidades, preferências, interesses, características

sociais e psicológicas do usuário, direcionando conteúdo relevante a este (SHEN; TAN; ZHAI, 2005). É necessário que as mudanças sejam percebidas e ajustadas sem a intervenção humana, através do uso de agentes de software, que realizam análises semânticas no conteúdo coletado (LAKIOTAKI *et al.*, 2009).

Os TDs podem auxiliar neste sentido, já que estabelecem uma relação com o mundo real do usuário. Porém, normalmente estão dispersos em diversas ferramentas e sem um domínio de conhecimento bem-definido. Outra dificuldade é que, geralmente, eles estão escritos em linguagem natural, dificultando o seu processamento. Além disto, há também uma grande diversidade de usuários com interesses e necessidades diferentes, que nem sempre estão aptos ou dispostos a fornecerem informações sobre eles (HOEBEL; ZICARI, 2008).

Logo, para que a informação possa ser compreensível e tratável por um agente computacional, se faz necessário o uso de técnicas e ferramentas que proporcionem um entendimento mínimo sobre o conteúdo e que possibilitem detectar os interesses do usuário. O enriquecimento semântico possibilita estender a compreensão do domínio de um determinado TD.

O enriquecimento semântico, do inglês *Semantic Augmentation*, é o processo de anexar conceitos semânticos a partes específicas de um texto, provendo uma estrutura para interpretação automática de seu significado e a compreensão do domínio de um determinado TD (THAKKER *et al.*, 2012). Para isso, é realizada a identificação e o mapeamento de termos-chave a partir do conteúdo textual através do PLN (BALDAN; MENEZES, 2012). Estes termos são semanticamente associados a conceitos oriundos de uma ontologia de domínio, permitindo a compreensão do conteúdo (ZAPATER; MENDES NETO, 2014). Essas relações semânticas permitem uma exploração complexa e a descoberta de informações sobre um determinado recurso (SHETH; ARPINAR; KASHYAP, 2003).

Karanasios et al. (2013) propuseram uma arquitetura para lidar com grande parte da complexidade do PLN, do enriquecimento semântico e da ligação entre os TDs, ontologias e recursos a serem enriquecidos, conforme esquematizado na Figura 5.

Esta arquitetura serviu como base para a criação da arquitetura deste trabalho. O Extrator de Informação recebe o conteúdo textual, realiza o PLN e extrai termos conceituais deste, baseado em ontologias. O Indexador Semântico é responsável por realizar as marcações semânticas do conteúdo e retornar o conteúdo enriquecido semanticamente.

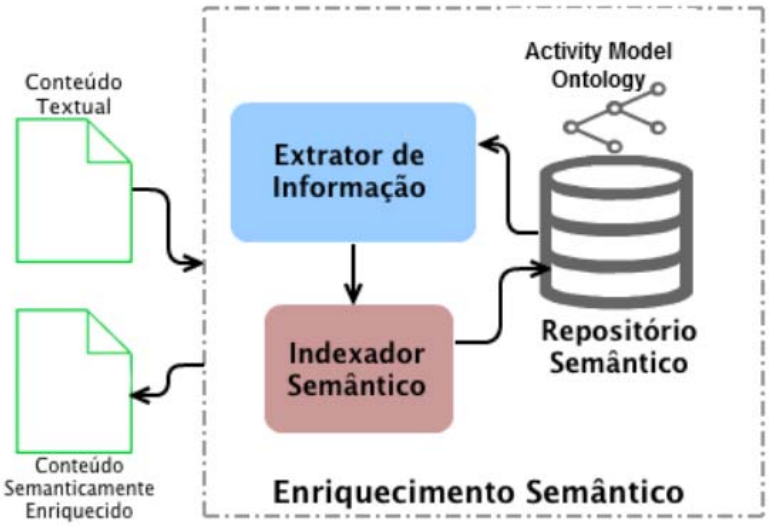


Figura 5 - Serviço de Enriquecimento Semântico
Fonte: (KARANASIOS *et al.*, 2013)

Uma vez que o conteúdo esteja enriquecido, ele deve ser relacionado às ações do usuário e as ontologias também contribuem neste sentido. Elas permitem criar perfis semânticos de usuários através das Ontologias de Perfil do Usuário, do inglês *User Profile Ontology* (UPO) (HECKMANN *et al.*, 2005). As ontologias de perfis de usuários mantêm relações semânticas entre usuário, conteúdo e conceitos de domínio, mantendo uma independência das ontologias de domínio. São estas relações que permitem mapear o perfil de um usuário.

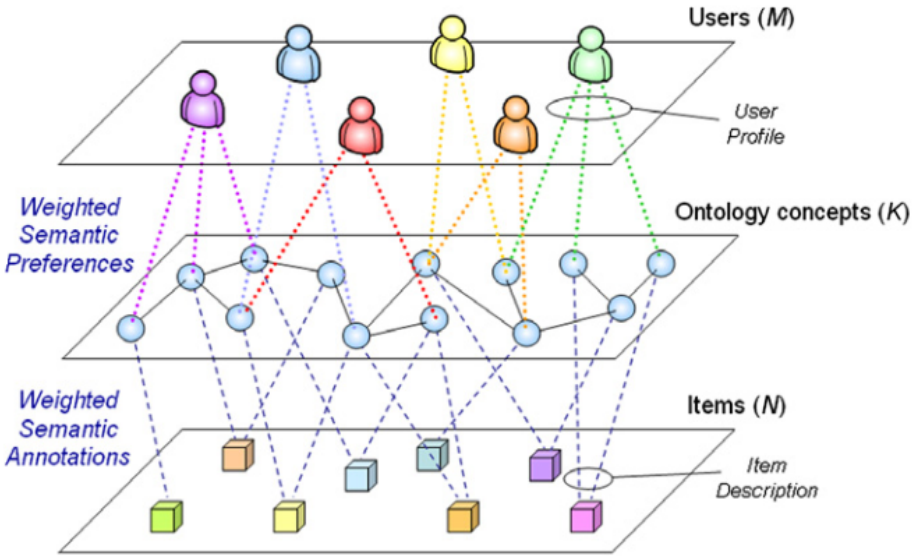


Figura 6 - Representação da relação semântica entre conteúdos e perfil do usuário.
Fonte: (CANTADOR; CASTELLS, 2011)

Um esquema destas relações é descrito na Figura 6. Os itens (N) estão relacionados aos TDs do usuário, que estabelecem uma relação com os conceitos (K) existentes nas ontologias de domínio, contextualizando o seu conteúdo. Se um TD é relacionado a um conhecimento e o mesmo possuem uma relação direta com os usuários, é possível realizar inferência semântica para estabelecer uma relação dos usuários (M) com os conceitos da ontologia, e criar um perfil do usuário relacionado ao conhecimento modelado (Ontologias de Domínio).

As UPOs, com os recursos enriquecidos, juntamente com as ontologias de domínio formam um repositório semântico (KARANASIOS *et al.*, 2013). Este repositório possibilita a agentes computacionais realizarem inferências eficientes capazes de (i) estabelecer uma relação de um recurso com um domínio de conhecimento e (ii) identificar itens relevantes, classificando-os conforme o grau de interesse do usuário (PAZZANI; MURAMATSU; BILLSUS, 1996).

2.7.1. Perfil de Usuário

O perfil do usuário representa seus interesses, suas crenças e experiências de vida. Ele é a base para sistemas de recomendação personalizada de conteúdos. Estes perfis podem ser construídos usando técnicas implícitas ou explícitas de coleta de informações. Na técnica explícita, o usuário é questionado diretamente sobre seus interesses ou solicita-se que o mesmo classifique o nível de interesse sobre algum assunto (JONG HWA KIM, 2012). No método implícito, os TDs (dados de navegação, localização, etc.) são analisados em busca de um padrão que denote os interesses do usuário (REFORMAT; GOLMOHAMMADI, 2009). O método explícito apresenta limitações, pois o usuário pode não se sentir confortável em fornecer informações sobre interesses pessoais em formulários web e estes podem se modificar com o tempo, bem como novos interesses podem surgir, tornando o método implícito mais adequado na maioria das situações.

Reformat e Golmohammadi (2009) propuseram um método baseado em técnicas implícitas para aprendizagem e atualização de perfis de usuários automaticamente. O método proposto analisa os dados dos *logs* de navegação em busca de conceitos relativos ao domínio em questão (música, no caso) e os compara com os conceitos registrados para o perfil do usuário. Para realizar tal comparação, é utilizado o conceito de similaridade semântica, onde

duas classes de ontologias diferentes são comparadas em duas ontologias diferentes, no caso, a de domínio e a do usuário.

Segundo Despotakis et al. (2011), o método de similaridade semântica nem sempre é eficiente, pois técnicas de personalização consideram a diversidade em prejuízo da similaridade. O processo de aprendizagem pode se tornar mais consistente e abrangente quando o indivíduo é exposto a informações que desafiem seu ponto de vista e conhecimento atual, o que não ocorreria com a similaridade. Porém no framework proposto por eles fica a cargo do usuário informar qual o tipo e teor do comentário que está sendo feito sobre um determinado recurso.

Já Karanasios et al. (2013) propuseram uma abordagem com uma ontologia construída nos pilares da Teoria da Atividade (KAPTELININ; NARDI, 2009). O software provê uma interface para busca e navegação pelos recursos (vídeos de entrevistas) com base na estrutura definida pela ontologia. Esta ontologia de domínio lida com aspectos da interação social e linguagem não verbal no contexto de entrevista de emprego. Porém, o usuário ainda precisa procurar itens de interesse através do sistema, condicionando o desempenho do sistema ao desempenho do usuário na utilização do mesmo.

O enriquecimento semântico proposto neste trabalho utiliza-se de informações do cotidiano do usuário, coletadas tanto de forma implícita como explícita. Estas informações são processadas a fim de gerar um perfil semântico do usuário que agrega as técnicas expostas anteriormente, porém sem que haja a necessidade do usuário classificar ou determinar uma informação.

2.8. TRABALHOS RELACIONADOS

O presente trabalho está relacionado ao enriquecimento semântico de perfis de usuários para o contexto da saúde, de forma a favorecer a aprendizagem informal. Contudo, pode-se citar quatro subdivisões da pesquisa usadas para alcançar o objetivo proposto: (i) semântica em sistemas de informação na área da saúde, (ii) enriquecimento semântico, (iii) semântica para perfis de usuários e (iv) conhecimento informal por meio de traços digitais. Todas elas possuem uma ampla área de pesquisa com diversos trabalhos desenvolvidos, dentre os quais podemos destacar alguns.

Nguyen, Fuhrer e Pasquier-Rocha (2009) propõem o desenvolvimento de um modelo PHR em forma de ontologia, onde são descritos conceitos tais como informações demográficas, vacinas, alergias, diagnósticos, procedimentos e medicamentos. Esse PHR visa à integração com os sistemas públicos de saúde, podendo fornecer informações relativas ao paciente aos profissionais da saúde. Apesar de não possuir o foco no individuo, nem na definição de perfil de usuário, este trabalho contribuiu pela forma como são abordadas as técnicas de inferências sobre as ontologias.

Camous, Mccann e Roantree (2008) propõem a utilização de enriquecimento semântico no monitoramento da saúde no esporte, a partir de sensores, integração e gerenciamento de dados, fornecendo aos usuários uma interface de consulta. Duas abordagens desta pesquisa foram uteis para o presente trabalho (i) a forma que fazem inferência sobre dados valorados e (ii) a perspectiva do uso de ontologias para destinar a acompanhamento personalizado do usuário.

Já Dung e Kameyama (2007) propõem a extração de informações de saúde a partir de conteúdos publicados na internet através do enriquecimento semântico, realizando inferências sobre ontologias de domínio utilizando dois algoritmos desenvolvidos: (i) algoritmo de extração de elementos semânticos e (ii) algoritmo de aprendizagem de novos elementos semânticos. O foco desta pesquisa diverge do presente trabalho, porém as técnicas aplicadas para processamento de conteúdo para estabelecer relação com ontologias foram essenciais e utilizadas para este.

Stan et al. (2008) propõem um modelo de perfil de usuário de base ontológica que permite aos usuários terem uma rede social baseada na percepção da situação, controlando como eles são acessíveis para categorias específicas de pessoas em uma determinada situação. Esta pesquisa aborda diretamente o uso de ontologias para traçar perfis de usuários, o que foi benéfico ao presente trabalho para estabelecer essas relações a partir de conceitos e ideias concebidas por esta pesquisa.

Thakker et al. (2012) apresentam uma abordagem para enriquecimento semântico de conteúdos obtidos de traços digitais a serem utilizados em um sistema de busca, com o objetivo de facilitar a aprendizagem informal em domínios mal definidos. Apesar de não abordar perfil semântico de usuário, esta pesquisa contribuiu diretamente pela abordagem no processamento de conteúdos extraídos da internet. Estas abordagens foram utilizadas para fazer o processamento dos TDs do usuário, como por exemplo o uso de PLN e a relação conceitual com ontologias.

Abel (2011) propõe modelar o interesse dos usuários com base nos comentários do *twitter*, enriquecendo-os semanticamente e contextualizando-os a artigos de notícias relacionadas. O objetivo é personalizar conteúdos ofertados ao usuário com base no que é publicado no *twitter*, fazendo-o se interessar pela notícia. Esta pesquisa contribuiu no que se refere ao mapeamento de perfil de usuário para recomendar conteúdo. As relações estabelecidas, por esta pesquisa, para fazer essa determinação foram essenciais para determinar a forma em que o perfil semântico de usuário deste trabalho foi modelado.

Aliança Neto, Mendes Neto e Moreira (2014) propõem um mecanismo genérico de Enriquecimento Semântico de Perfil de Usuário para determinação de interesses baseado em Traços Digitais. O mecanismo necessita que seja criada a ontologia, que pode ser para qualquer domínio. Uma vez criada à ontologia, um texto pode ser enriquecido e vinculado ao perfil do usuário. Contudo o mesmo apresenta algumas limitações, dentre as quais podemos destacar:

- A oneração a ontologia de perfil de usuário realizando inúmeras anotações que inviabilizaram o uso em um ambiente real;
- Não possui uma interface amigável, uma vez que, ao solicitar o grau de interesse do usuário em um assunto, é necessário passar a URI especifica do assunto contido na ontologia de domínio que se deseja a relação;
- Processa apenas conteúdos em inglês e
- Não realiza processamento de dados não textuais, como, por exemplo, dados oriundos do PHR.

Desta pesquisa foram utilizadas como base para o presente trabalho a UPO e parte do mecanismo de PLN, onde foram realizadas adaptações para contornar as limitações encontradas e supracitadas.

O presente trabalho apresenta uma abordagem para determinar os interesses dos usuários relacionados à sua saúde em um ambiente de aprendizagem informal. Para isto utiliza um conjunto de conceitos e técnicas para realizar inferências semânticas e gerar relações de usuários e conteúdos com domínios de conhecimento.

Não foi encontrado na literatura um trabalho que propusesse a determinação dos interesses do usuário relacionados à sua saúde baseado no seu contexto diário, principalmente quando se refere a fazer de forma automática através do uso de enriquecimento semântico, agregando diversas técnicas para dirimir a limitação da necessidade da interação do usuário. Os interesses do usuário impactam diretamente na seleção de conteúdos adequados e

relevantes quando se trata de recomendação personalizada de conteúdos para a aprendizagem informal.

3. ENRIQUECIMENTO SEMÂNTICO DE PERFIL DO USUÁRIO

Este capítulo descreve detalhes do Sistema de Enriquecimento Semântico de Perfil do Usuário voltado para o contexto da saúde. O objetivo deste sistema é prover meios para determinar os interesses do usuário relacionados à sua saúde, considerando o seu contexto diário.

Como explicitado anteriormente, a solução foi projetada para apoiar um ambiente de aprendizagem informal, nomeado de MobiLEHealth, do inglês *Mobile Learning Environment for Health*. Apesar da concepção deste ambiente não fazer parte deste trabalho, é necessário contextualizá-lo para fornecer uma visão geral e da aplicabilidade deste e, assim, obter uma melhor compreensão do seu escopo e as funcionalidades do Sistema de Enriquecimento de Perfil de Usuário.

3.1. MOBILEHEALTH

Este ambiente foi desenvolvido por um grupo de pesquisa do Laboratório de Engenharia de Software (LES) da Universidade Federal Rural do Semi-Árido (UFERSA). Três grandes áreas de conhecimento foram utilizadas: (i) monitoramento ubíquo do usuário, (ii) enriquecimento semântico de perfil de usuário, e (iii) recomendação personalizada de conteúdos. Cada uma destas áreas originou pesquisas paralelas, independentes e ao mesmo tempo funcionando de forma integrada para conceber o MobiLEHealth.

O MobiLEHealth é um ambiente de aprendizagem informal no contexto da Saúde 2.0 destinado a portadores de doenças crônicas capaz de adequar-se às características particulares dos usuários, fornecendo conteúdo adequado às suas necessidades de saúde, sem interferir na sua rotina, interação social e profissional. Para isso leva em consideração o perfil do usuário, o seu contexto atual e o seu histórico de interações no meio virtual (MENDES NETO *et al.*, 2014a). O objetivo deste ambiente é fornecer um maior conhecimento sobre a doença e, consequentemente, melhoria da qualidade de vida destas pessoas.

Este ambiente monitora os usuários de forma dinâmica, autônoma e transparente, através do uso de seus dispositivos móveis, disponibilizando serviços web e registros pessoais de saúde. A captura das informações do usuário leva em consideração o seu contexto, como

localização, aplicativos utilizados, *status* do dispositivo, entre outros. Estas informações podem ser oriundas de diversas fontes, como conteúdos acessados ou publicados pelos usuários, interações nas redes sociais e informações pessoais relativas à saúde.

Os dados capturados do cotidiano do usuário, mediante a sua autoização, são armazenadas na base de dados do MobiLEHealth. Algumas destes passam pelo processamento de enriquecimento semântico, que relaciona-os a domínios de conhecimentos modelados em ontologias e gera um perfil semântico do usuário (MOREIRA et al., 2014). A partir deste perfil, determinam-se quais os domínios e conceitos de interesses do usuário.

Com base no contexto do usuário, o mecanismo de recomendação realiza a análise dos dados históricos do usuário e os seus interesses. Com estas informações aplica um conjunto de técnicas para selecionar conteúdos que mais se adequem ao usuário (COSTA et al., 2014).

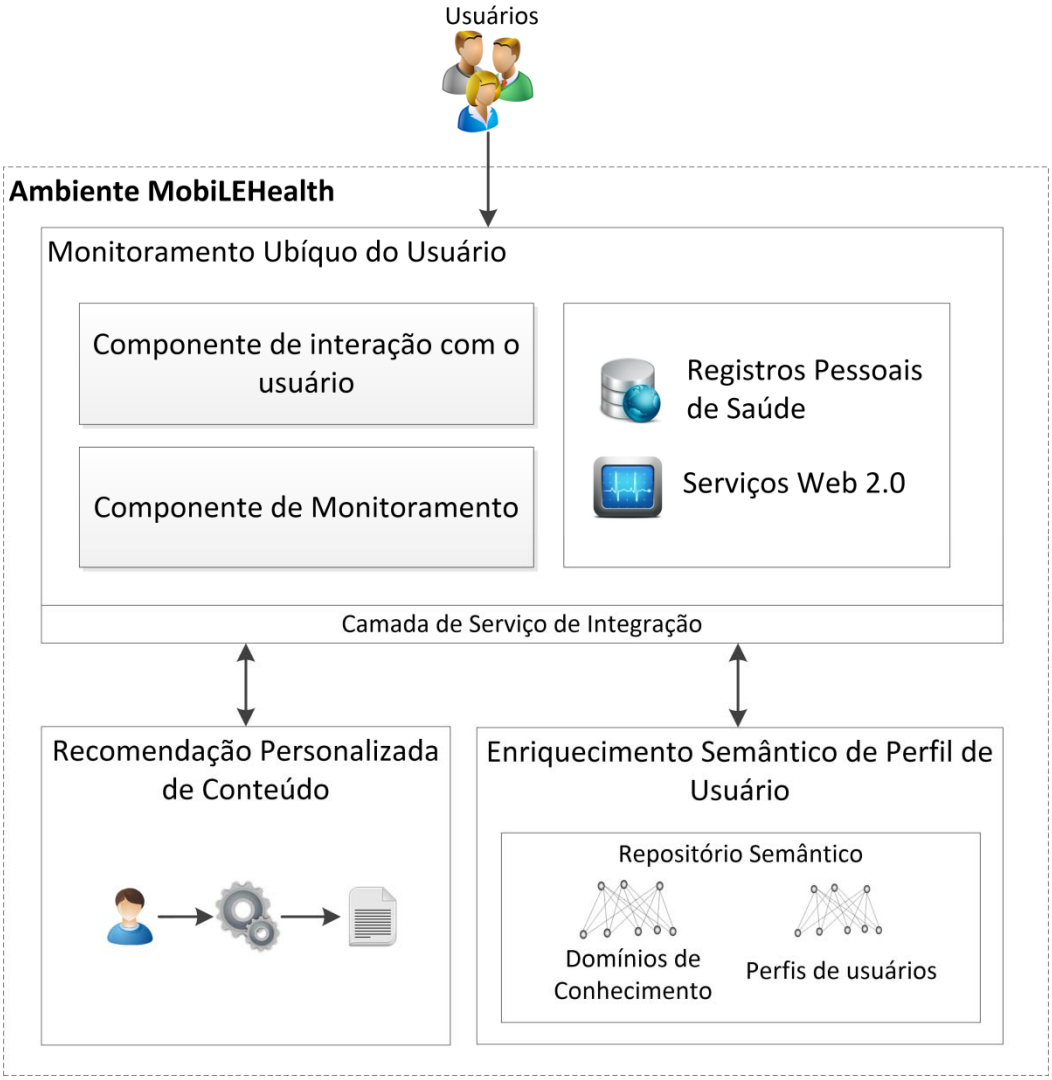


Figura 7 - Modelo arquitetural do MobiLEHealth.
Fonte: Adaptado de (MENDES NETO *et al.*, 2014a)

O ambiente direciona ao usuário os conteúdos para ele selecionados pelo sistema de recomendação. O usuário pode aceitar ou recusar, e ainda avaliar a recomendação que lhe foi destinada. Estas ações também influenciarão na próxima recomendação ao usuário. Cada um destes componentes estão demonstrado no modelo arquitetural do MobiLEHealth apresentado na Figura 7 (MENDES NETO *et al.*, 2014b).

A arquitetura do MobiLEHealth foi modelada para que os componentes estejam integrados, mas que funcionem de forma independente e autônoma. A Figura 8 mostra o diagrama de componentes do MobiLEHealth. Cada componente atua de forma isolada, desempenhando o seu papel para alcançar um objetivo comum.

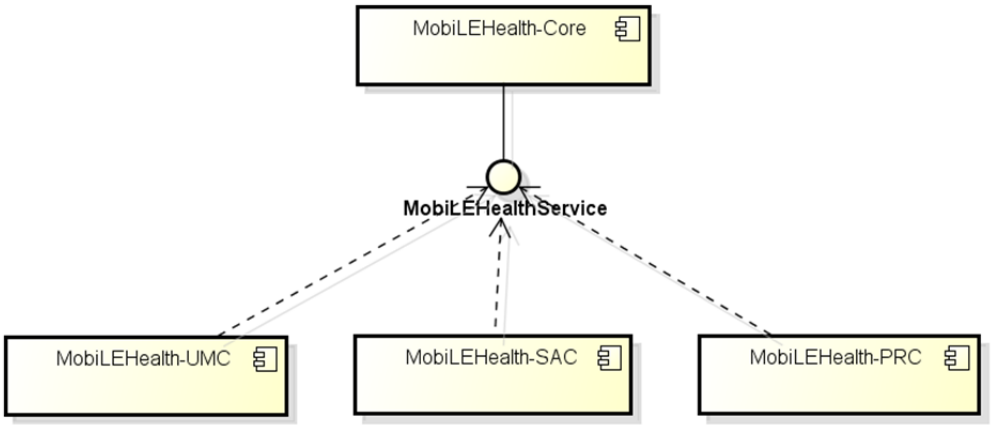


Figura 8 - Diagrama de Componentes do MobiLEHealth
Fonte: Autoria Própria

O *Core* é o núcleo central do MobiLEHealth e dispõe de uma interface de serviços para a integração dos componentes, que não se comunicam diretamente entre si. Este é responsável por garantir a padronização e a integridade das informações manipuladas pelos demais sistemas. O UMC (*User Monitoring Component*) contém as aplicações móveis responsáveis por interagir com o usuário e monitorar suas atividades diárias em seus dispositivos móveis, enviando estas informações ao *Core* para que sejam armazenadas na base de dados do MobiLEHealth. O PRC (*Personalized Recommendation Component*) analisa a base de dados e o perfil semântico para gerar as recomendações apropriadas aos usuários.

Por fim, o SAC (*Semantic Augmentation Component*) é o componente responsável por realizar o processamento semântico das informações contidas no MobiLEHealth e gerar o perfil semântico do usuário. Este componente é a parte do qual este trabalho faz parte. Sua função é estabelecer, de modo automático, a contextualização dos TDs com domínios de conhecimento e a relação de interesses do usuário relativos à sua saúde, de forma a auxiliar o

ambiente MobiLEHealth na recomendação personalizada de conteúdos adequados ao usuário. Nas próximas seções iremos detalhar o Enriquecimento Semântico de Pefil de Usuário.

3.2. ESCOPO

O principal objetivo do Sistema de Enriquecimento é determinar a relação de interesse de um usuário a domínios de conhecimentos no contexto da saúde, com base em seus TDs. Para isso é necessário contextualizar os TDs a estes domínios.

Este sistema funciona como um componente com interfaces externas bem definidas de forma a não sofrer interferências externas diretas, ou seja, toda a integração com os demais componentes ocorre por meio destas interfaces. Ele se integra ao MobiLEHealth de duas formas: (i) através de um serviço, consumido por este, que disponibiliza métodos que retornam a relação do usuário ou conteúdo com um domínio e (ii) através de agentes computacionais que vasculham a base de dados do MobiLEHealth, por meio de interfaces deste, buscando informações que necessitam ser enriquecidas e acrescentadas ao perfil semântico do usuário.

Não faz parte do escopo do Sistema de Enriquecimento Semântico a captura dos dados do usuário, que é feita pelo próprio MobiLEHealth. O acesso a base de dados é controlado pela *engine* do MobiLEHealth, que garante a integridade e a consistência dos dados.

3.3. ARQUITETURA

O Sistema de Enriquecimento Semântico apresenta uma arquitetura (Figura 9) modular. Ela fornece uma interface que se integra ao MobiLEHealth de forma coesa, isolando o componente e garantindo a consistência do seu funcionamento. Este sistema tem a função de enriquecer semanticamente o perfil de usuário com os dados coletados pelos componentes de interação com usuário e o de monitoramento ubíquo, e com base nos domínios de conhecimentos registrados no repositório semântico. Esta arquitetura foi estruturada com base no modelo proposto por Karanasios et al. (2013), sendo que este trata apenas o enriquecimento do conteúdo.

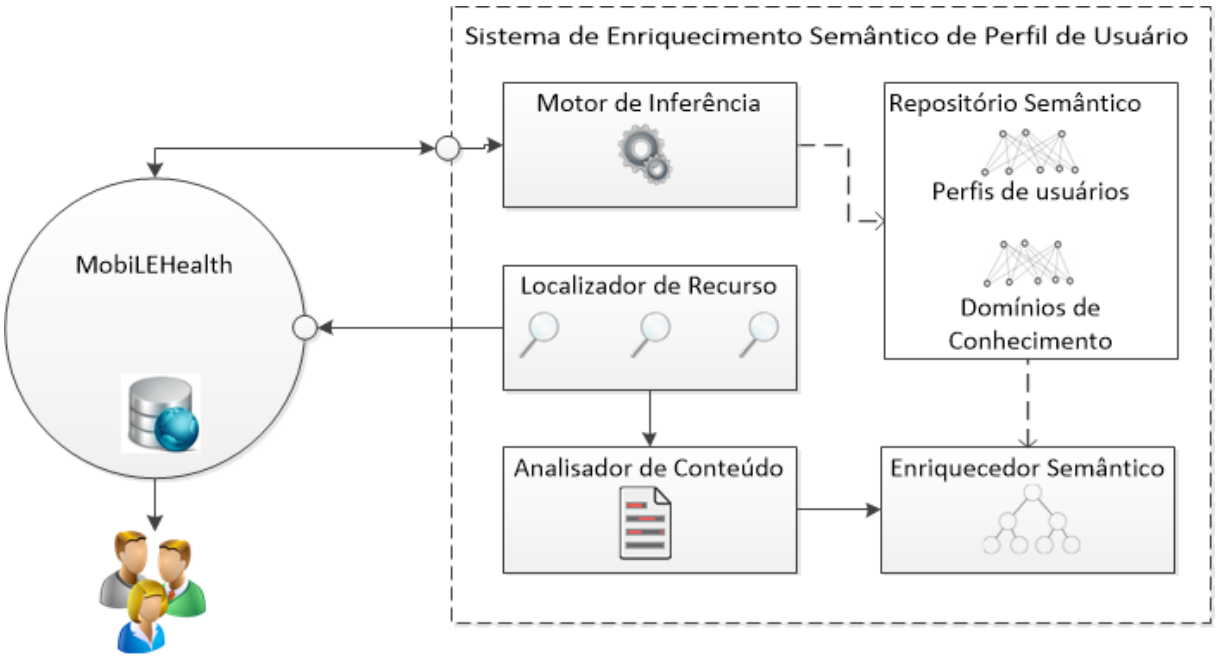


Figura 9 - Modelo da Arquitetura do Sistema de Enriquecimento Semântico.
Fonte: Autoria própria.

Através de uma interface disponibilizada pelo MobiLEHealth, Localizadores de Recursos especializados na base de dados do MobiLEHealth identificam TDs que necessitam do processamento semântico. Os TDs são carregados, onde é realizado um tratamento prévio do recurso, verificando a sua integridade e a conformidade com os pré-requisitos do sistema.

O Analisador de Conteúdo realiza um processamento do conteúdo, por vezes aplicando técnicas de PLN, e busca por termos semelhantes nas ontologias de domínio. O resultado desse processamento é enviado para o Enriquecedor Semântico, que gera as anotações semânticas para as correspondências encontradas e as armazena na ontologia de perfil do usuário.

O Repositório Semântico fornece ontologias de domínio, que representam a modelagem dos conceitos relacionados às doenças crônicas que compõem o sistema, e a ontologia de perfil de usuário. O motor de inferência é responsável por responder as solicitações oriundas do MobiLEHealth. Estas solicitações podem ser (i) para um usuário, onde o sistema deverá determinar qual a relação entre o usuário e um domínio de conhecimento, ou (ii) para um conteúdo, definindo a relação entre este e um domínio de conhecimento.

Nas próximas seções serão detalhadas as funcionalidades de cada uma dos componentes supracitados.

3.4. REPOSITÓRIO SEMÂNTICO

O Repositório Semântico é a base para o funcionamento do Sistema de Enriquecimento Semântico de Perfil de Usuário. Conhecer a sua estrutura e funcionamento é essencial para entender os processos envolvidos neste sistema. Por isto ele será detalhado antes dos demais componentes.

Alguns cuidados estruturais e relacionais foram tomados para garantir a flexibilidade promovida pelas ontologias, sem perder a capacidade de resposta do sistema no processamento destas.

O Repositório Semântico é composto por uma única ontologia de perfil e pode conter várias ontologias de domínio de conhecimento. Para cada domínio tratável no sistema, é necessário criar uma ontologia que represente esse conhecimento. A separação entre estas ontologias fornece flexibilidade ao sistema.

A ontologia de perfil está diretamente ligada à estrutura do sistema e mudanças em sua estrutura, provavelmente, irá requerer mudanças no mesmo. Já a ontologia de domínio não esta diretamente ligada ao sistema e mudanças na sua estrutura ou adição de novas ontologias não impactam diretamente neste. Contudo, é necessária a adoção de alguns padrões para sua correta utilização.

3.4.1. Ontologia de Domínio

A ontologia de domínio é uma modelagem de um conhecimento específico. Ela deve conter conceitos e relações entre estes que formalizem o conhecimento. Esta formalização permite que agentes computacionais possam realizar inferências e ter um entendimento sobre aquele conhecimento. Isto permite ao Sistema de Enriquecimento Semântico contextualizar conteúdos a domínios modelados.

O mecanismo utilizado permite certa flexibilidade e extensibilidade às ontologias de domínio. A forma como são estruturadas não interferem diretamente no sistema, exceto pelo fato de que algumas boas práticas devem ser seguidas para um bom funcionamento. A capacidade do sistema está ligada à qualidade da ontologia de domínio, por isto, uma ontologia mal definida impacta em uma contextualização não tão eficiente do conteúdo.

O sistema tem a capacidade de lidar com múltiplas ontologias que podem ser adicionadas, retiradas ou modificadas a qualquer momento. O sistema possui a capacidade de percorrer toda a estrutura da ontologia de domínio em busca de conceitos a serem processados. Estas ontologias podem ser criadas com qualquer hierarquia e categorização desde que sigam algumas regras pré-definidas:

- Deve ser utilizado o formato RDF/OWL;
- Deve existir uma ontologia de domínio modelada para cada conhecimento a ser utilizado pelo sistema;
- A ontologia deve ter sempre apenas uma classe principal (raiz). Esta classe é o ponto de partida para o algoritmo do sistema e a partir dela as inferências são realizadas;
- Os conceitos da ontologia devem estar sempre inseridos como indivíduos, pois o algoritmo do sistema busca apenas os indivíduos ao realizar as inferências. As classes devem ser apenas usadas para categorização dos conceitos, pois esta é a sua principal funcionalidade;
- Os indivíduos devem ter uma ou mais anotações do tipo *label* com termos que descrevam o seu conceito. Deve ser evitado usar *label* para escrever textos longos, pois estes devem estar nas anotações do tipo *comment*, que não são utilizadas na identificação dos conceitos;
- Devido à portabilidade a vários idiomas, na definição do *label* deve ser definido o idioma a que se refere, usando o conceito de internacionalização. Caso não seja definido, o *label* será utilizado em qualquer idioma. A Figura 10 mostra um exemplo de um individuo, em OWL, com dois *labels* definidos: *glucose* para o inglês e glicose para o português.

```
<owl:NamedIndividual rdf:about="http://(...)/Diabetes#glucose">
    <rdf:type rdf:resource="http://(...)/Diabetes#Biomolecule"/>
    <rdfs:label xml:lang="en">glucose</rdfs:label>
    <rdfs:label xml:lang="pt">glicose</rdfs:label>
</owl:NamedIndividual>
```

Figura 10 - Exemplo em OWL para definição de labels em vários idiomas
Fonte: Autoria Própria

Estas boas práticas foram adotadas pelo Sistema de Enriquecimento Semântico para que se criem ontologias de domínios reutilizáveis.

3.4.2. Ontologia de Perfil de Usuário

A UPO do sistema possui uma estrutura fixa (Figura 11) e diretamente relacionada ao sistema de enriquecimento semântico. O modelo proposto por (ALIANÇA NETO; MENDES NETO; MOREIRA, 2014) é a base desta ontologia. Ela é composta por uma classe raiz chamada *Profile*, cujas classes filhas são:

- *User*: Contém informações do usuário;
- *Resource*: Armazena informações sobre os recursos enriquecidos. Esta classe possui uma classe filha chamada *Link*, que estabelece a relação entre o conteúdo do recurso e as ontologias de domínio; e
- *Access*: Armazena informações temporais das ações dos usuários sobre os *Resources*.

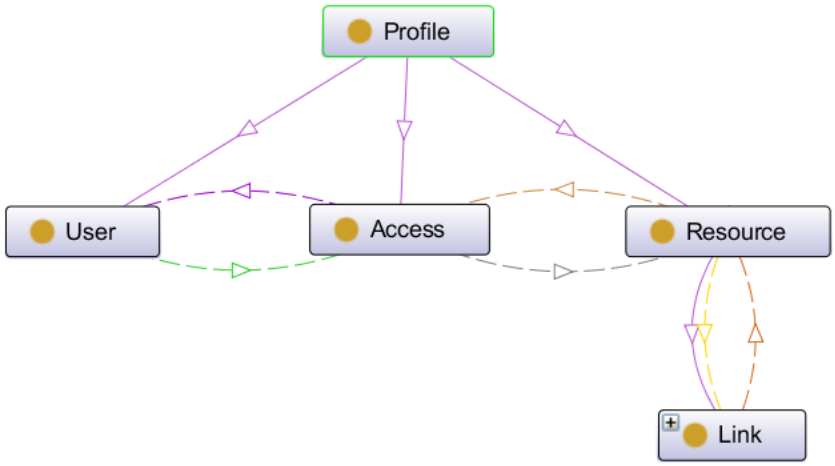


Figura 11 - Estrutura da Ontologia de Perfil do Usuário
Fonte: Autoria própria.

Estas classes se relacionam com base no princípio de que os usuários (*User*) estão ligados por seus acessos (*Access*) a um determinado recurso (*Resource*), que por sua vez está ligado às ontologias de domínios por meio de *links* (*Link*). Estas relações são estabelecidas na ontologia pelas *Object Properties*, através de *statements* formados pela tripla [sujeito, relação com, objeto]. Cada *Object Property* possui três definições importantes para os motores de inferência:

- *Domain*, que restringe que o sujeito de uma relação deve ser um indivíduo pertencente à classe especificada;

- *Range*, que faz a mesma restrição para os objetos; e
- *Inverse*, que define uma *Object Property* inversa, ou seja, o *Domain* de um será o *Range* da outra e o *Range* será o *Domain* da outra.

Essas definições são restrições automaticamente interpretáveis pelos motores de inferência. A Tabela 1 exibe as relações para as *Objects Properties* usadas pelo sistema.

Tabela 1 - *Object Properties* da Ontologia de Perfil do Usuário.

Object Property	***Domain***	***Range***	***Inverse of***
hasAccess	*User*	*Access*	*accessedBy*
accessedFor	*Resource*	*Access*	*hasResource*
hasLink	*Resource*	*Link*	*isLink*

Estas relações são definidas no nível de classe, mas são estabelecidas no nível de indivíduo. A Figura 12 ilustra como são estabelecidas essas relações entre indivíduos. Indivíduos da classe *Access* mantêm uma relação de um para muitos com cada uma das classes *Resource* e *User*, ou seja, só podem estar relacionados a um indivíduo de *Resource* e um de *User*. Já um indivíduo de *User* ou *Resource* pode estar relacionado com vários indivíduos de *Access*.

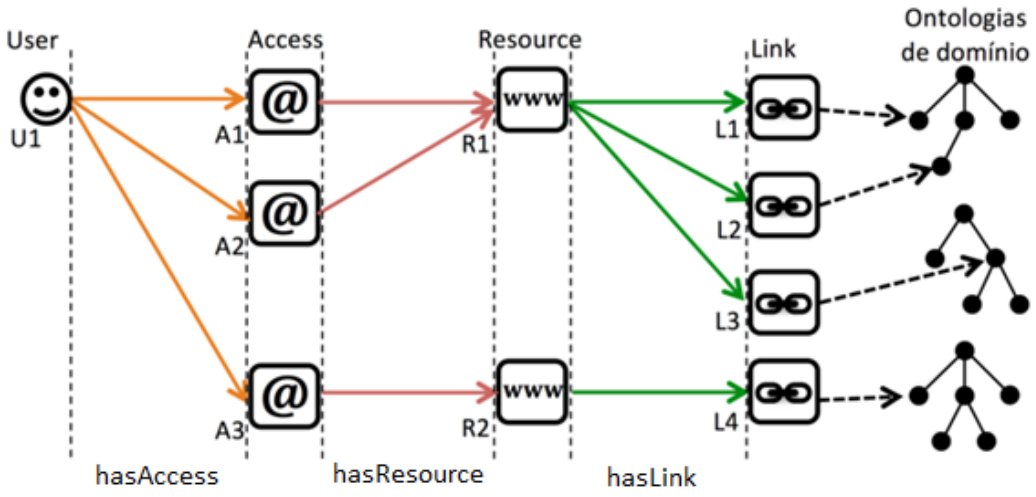


Figura 12 - Relacionamento entre indivíduos na UPO.
Fonte: Adaptado de (ALIANÇA NETO; MENDES NETO; MOREIRA, 2014)

Além das relações, estes indivíduos contêm informações que os identificam. As informações são atribuídas através de *Data Properties*. As definições de *Domain*, *Range* e *Inverse* também são aplicáveis aos *Data Properties*. Uma vez estabelecidas, são interpretadas pelos motores de inferência. Por exemplo, se um *Data Property* tiver como domínio a classe

Resource, esta não pode ser utilizada por um indivíduo da classe *User*. A Tabela 2 mostra as principais *Data Properties* da UPO.

Tabela 2 - Principais *Data Properties* da UPO.

Data Property	***Domain***	***Range***	**Descrição**
Schema	*Resource*	*String*	Identifica o *schema* do recurso na base de dados do MobiLEHealth.
tableName	*Resource*	*String*	Identifica a tabela do recurso na base de dados do MobiLEHealth.
fieldName	*Resource*	*String*	Identifica o campo chave do recurso na base de dados do MobiLEHealth.
fieldValue	*Resource*	*Int*	Valor do campo chave do recurso na base de dados do MobiLEHealth.
quantityTokens	*Resource*	*Int*	Quantidade total de elementos textuais de um recurso. Os elementos são contabilizados mesmo que não gerem *Links*. Esta quantidade serve para determinação da relação de um recurso com o domínio.
Root	*Link*	*String*	Identifica o elemento textual enriquecido.
domainURI	*Link*	*String*	URI da ontologia de domínio para o qual o elemento textual está relacionado.
quantity	*Link*	*String*	Quantidade de vezes que a relação [root, domainURI] aparece no recurso. Atribui um peso ao conceito no cálculo da relação com o domínio.
Id	---	*Int*	Identificador. Pode ser utilizado por qualquer indivíduo.
Date	---	*Date Time*	Identificador temporal. Pode ser utilizado por qualquer indivíduo.

A forma mais comum de estabelecer uma relação entre indivíduos é através das *Object Properties*. Contudo, não é possível o emprego destas para estabelecer a relação dos indivíduos da classe *Link* com os indivíduos das ontologias de domínio, pois os indivíduos estão em ontologias diferentes e uma não conhece a estrutura da outra. Por isto, adotou-se uma *Data Property* do tipo s*tring* para armazenar a URI do conceito da ontologia de domínio. Desta forma é possível estabelecer a relação entre ontologias diferentes, além de garantir a compatibilidade com qualquer ontologia.

3.5. LOCALIZADOR DE RECURSOS

Como citado anteriormente, os TDs do usuário são coletados pelo MobiLEHealth e armazenados em um banco de dados relacional. A integração para acesso a esta base de dados é realizada através de uma interface disponibilizada no *core* da aplicação (MobiLEHealth-Core). Esta interface permite que entidades persistentes possam ser acessadas e analisadas pelo Sistema de Enriquecimento Semântico.

Para realizar esta integração, foram projetados Localizadores de Recursos com especialidades específicas em relação à base de dados do MobiLEHealth. Estes têm a função de realizar constantes verificações em busca de recursos que precisam ser enriquecidos. Para isto procuram registros novos, modificados ou, considerando uma variação temporal, que precisam ser reprocessados.

Além de ter a função de buscar, os localizadores mapeiam os TDs que estão em entidades do MobiLEHealth e os transformam em um nível de abstração utilizado pelo sistema. Nesse mapeamento mantem-se uma relação que identifica os registros na base de dados. Esta abstração dá independência e flexibilidade ao sistema em relação ao MobiLEHealth, criando uma camada intermediária entre esses sistemas. Esta camada permite que alterações na base de dados não impactem diretamente no Sistema de Enriquecimento Semântico, mas que seja possível contemplá-las modificando apenas essa camada.

Basicamente o sistema busca as interações digitais e os dados de saúde do usuário. Para realizar esses mapeamentos foram criados três localizadores: Localizador de Conteúdo, Localizador de Acesso e Localizador de Dados Pessoais de Saúde.

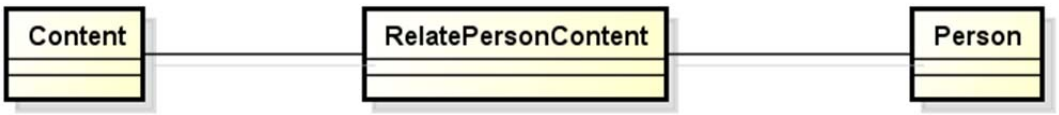


Figura 13 - Diagrama de Classe - Entidades que contêm as interações digitais do usuário
Fonte: Autoria Própria

Os dois primeiros são especializados em localizar as interações digitais do usuário. A Figura 13 mostra o Diagrama de Classe das entidades que incorporam a especialidade destes localizadores. Estas entidades estão descritas abaixo:

- *Person*: Contém informações da entidade pessoa, como identificador, *e-mail*, nome, gênero, data de aniversário, etc. Cada utilizador do sistema tem um registro na classe *Person*.
- *Content*: Representa os conteúdos, como páginas web, conteúdos publicados, recursos compartilhados, postagem em redes sociais, entre outros. E contém informações do tipo título, autor, descrição, URL (*Uniform Resource Locator*), taxa de aceitação, avaliações, etc.
- *RelateContentPerson*: São as interações do usuário com os conteúdos (*Content*). Mantêm informações sobre a ação do usuário indicando o *status*, a data e a hora, a avaliação do usuário e a relação entre usuário e conteúdo.

O Localizador de Conteúdo é especializado na entidade *Content*. Ele busca qualquer TD que não seja oriundo do PHR do MobiLEHealth. Ele percorre a base de dados e filtra conteúdos selecionando aqueles que podem ser processados, desde que atendam os pré-requisitos do sistema.

Todo e qualquer conteúdo é transformado em um recurso. Quando possível, o conteúdo é convertido em dado textual. Por exemplo, se for uma URL, é realizado o *download* da página, já se for uma publicação do usuário, são concatenados os metadados (titulo, autor, ...) e a descrição do conteúdo. Tudo isso gerando uma relação entre a entidade *Content* e o recurso.

Já o Localizador de Acesso é especializado nas entidades *RelatePersonContent*. Ele mapeia esta entidade para um acesso que contém a relação entre um conteúdo e um usuário. Nesta relação é mantida a temporalidade da ação, tornando possíveis inferências contextuais em relação ao usuário.

A Figura 14 exibe um diagrama de sequência da ação do Localizador de Conteúdo. Inicialmente ele obtém todos os conteúdos através do serviço do MobiLEHealth. Para cada conteúdo é verificado se já foi processado ou se a validade do processamento expirou. Esta validade garante, através de uma métrica de tempo, que o enriquecimento do conteúdo seja atualizado, pois este pode ter sofrido modificações. Caso o conteúdo necessite de processamento, este é mapeado para um objeto do tipo *Resource* (abstração criada para o sistema). Caso o conteúdo seja uma URL, o conteúdo do *site* é baixado. Todos os dados textuais do conteúdo são concatenados ao texto do recurso em formato de caractere. Em seguida o recurso é enviado ao Enriquecedor Semântico.

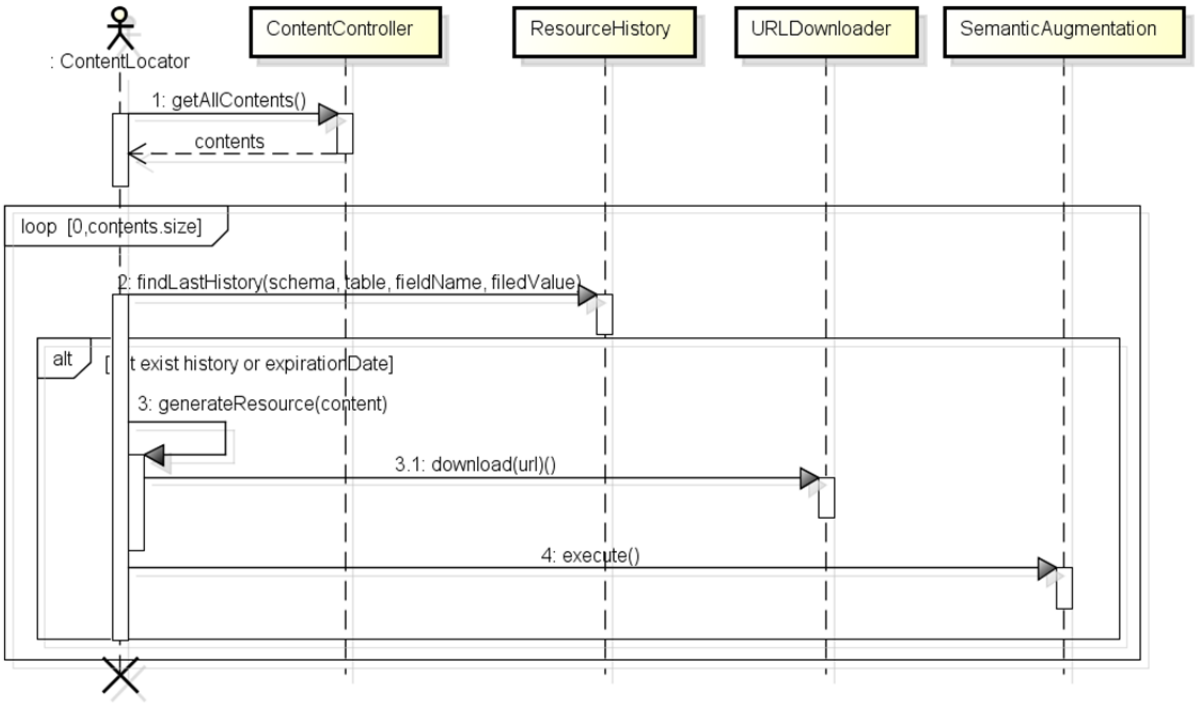


Figura 14 - Diagrama de Sequencia do Localizador de Conteúdo
Fonte: Autoria própria.

Este processamento é semelhante para o Localizador de Acesso, com a diferença de que, após a etapa 3, é gerado um mapeamento de *Person* para *User* e outro de *RelatePersonContent* para *Access*.

O MobiLEHealth mantém um PHR do usuário. Neste estão armazenadas informações sobre a saúde do usuário, mantidas pelo mesmo. O Localizador PHR é especializado em localizar estes dados do usuário e selecionar aqueles que podem ou precisam ser enriquecidos. Além disto, realiza o mapeamento dos dados do PHR para um Recurso.

A Figura 15 mostra um diagrama de classe com as entidades que formam o PHR. Estas tabelas foram classificadas em duas categorias: histórico e medidas. As entidades classificadas como histórico são comumente compostas de dados textuais, salvo algumas exceções, como no caso de *Lab Test Result* que contém valores dos resultados de exames. A Tabela 3 detalha esta classificação.

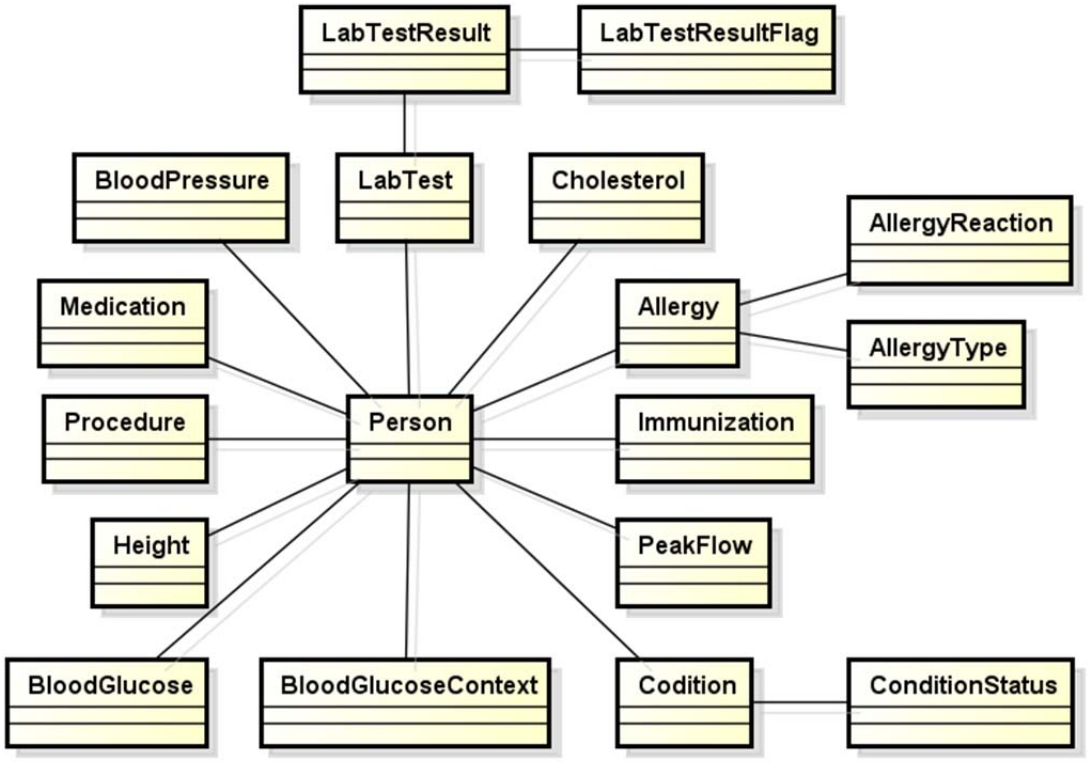


Figura 15 - Diagrama de Classe das Entidades do PHR do MobiLEHealth
Fonte: autoria Própria

A classificação foi feita porque ela impacta diretamente na forma em que é feito o processo de enriquecimento semântico do conteúdo. Os dados classificados como histórico de saúde são mapeados para um recurso e passam pelo processamento textual do Sistema de Enriquecimento Semântico.

Tabela 3 - Classificação das Entidades do PHR.

Classificação	**Entidade**
Histórico	*Allergy*
	Allergy Reaction
	Allergy Type
	Condition
	Condition Status
	Immunization
	Lab Test
	Lab Test Result
	Lab Test Result Flag
	Medication

	Procedure
Medidas	*Blood Glucose*
	Blood Glucose Context
	Blood Pressure
	Cholesterol
	Height
	Peak Flow

Já os dados classificados com medidas não possuem texto para ser processado, por se tratar de valores, como, por exemplo, o nível de glicose do usuário. Eles precisam de uma inferência semântica aprimorada para realizar o seu enriquecimento. O recurso é criado passando uma *string* contendo o identificador do recurso, como, por exemplo, *"Blood Glucose"* e um conjunto de parâmetros mapeados no formato {chave, valor}. A chave contém a informação do que se trata a medida e o valor é a medição desta chave. Considerando a medição da pressão sanguínea, são informados no mapeamento os parâmetros: sistólica, diastólica, pulso e arritmia cardíaca. Para exemplificar, esta estrutura ficaria da seguinte forma: *["Blood Pressure"*, [{ *"systolic"*, 120 }, { *"diastolic"*, 80}, { *"pulse"*, 120}, { *"irregularHeartbeat"*, *false*}]. Esta estrutura é vinculada ao recurso que é enviado para o processamento semântico.

3.6. ANALISADOR DE CONTEÚDO

O Analisador de Conteúdo é responsável por processar o conteúdo do recurso gerando, como resultado final, uma estrutura de dados que relaciona o recurso a conceitos das ontologias de domínio. Algumas tarefas são necessárias conforme o tipo e o conteúdo do recurso, conforme pode ser visto na Figura 16.

A primeira etapa do Analisador de Conteúdo é um pré-processamento do recurso, onde é validado o recurso e realizadas algumas operações e identificações necessárias em seu conteúdo. Se esse recurso não passar nas restrições do validador, ele é descartado.

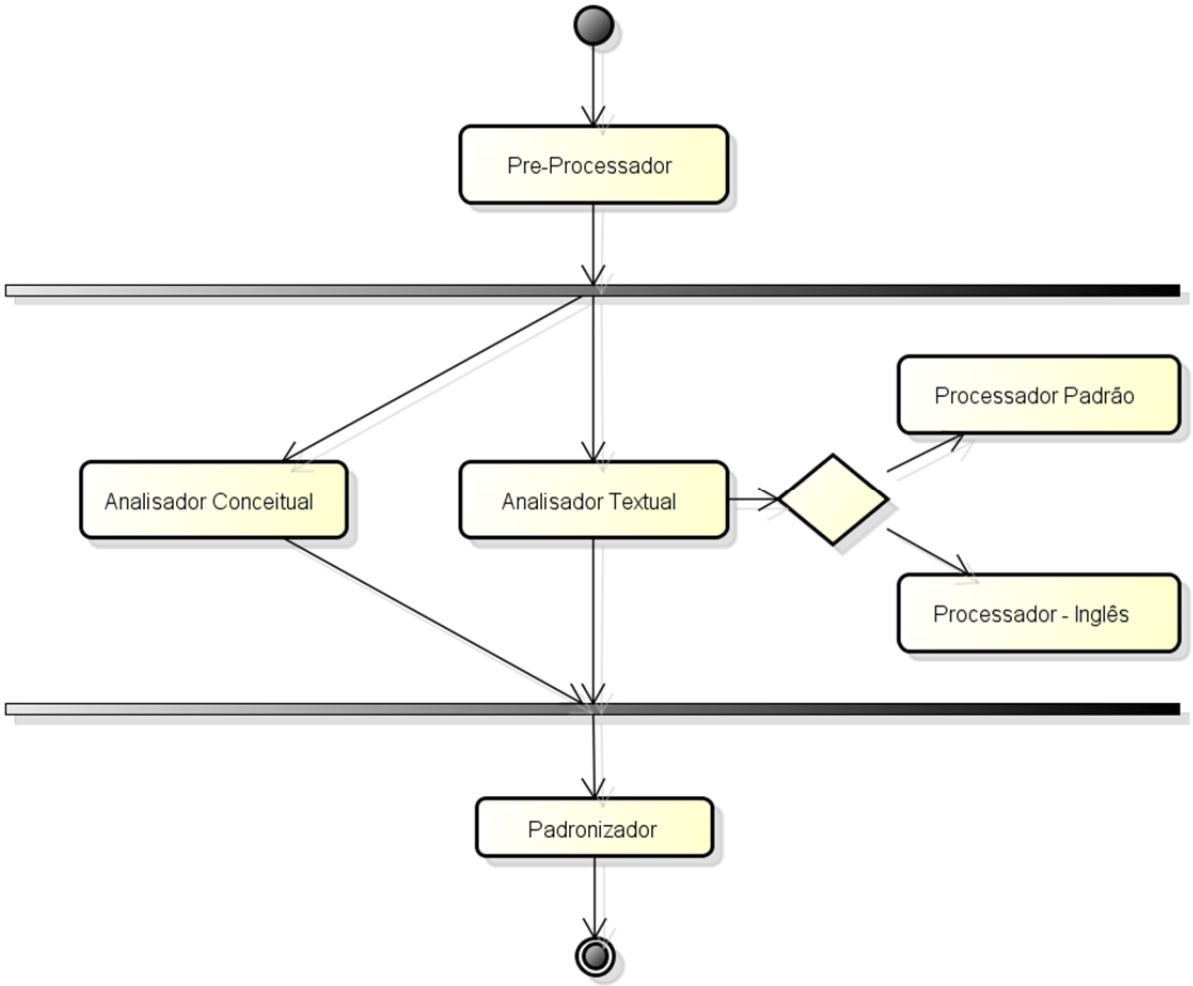


Figura 16 - Diagrama de Atividades das etapas do processador de conteúdo.
Fonte: Autoria Própria

Uma vez validado, o recurso pode ser processado pelo Analisador Conceitual ou pelo Analisador Textual. O Primeiro é quando se tratar de inferências semânticas sobre um determinado valor do recurso, por exemplo, os dados de medidas do PHR. Já o segundo quando forem dados escritos em linguagem natural. Neste ainda existem duas especializações: um processador para o inglês e um processador padrão que pode ser aplicado a qualquer linguagem. Detalhes deste processamento estão descritos na Subseção 3.6.2.

Após a conclusão do processamento do conteúdo, é acionado um padronizador para organizar a estrutura de dados a ser utilizada pelo Sistema de Enriquecimento Semântico. Esse padronizador garante a extensão no que se refere ao uso outras técnicas ou ferramentas para o processamento do conteúdo.

3.6.1. Pré-Processador

O pré-processador prepara o recurso para o processamento realizando validações e tratamentos no conteúdo. É verificado se o recurso pode ser processado pelo Sistema de Enriquecimento Semântico, caso não possa, o recurso é descartado. Nesta etapa são verificados:

- **Tipo do conteúdo:** Verifica se existe conteúdo textual, pois apenas imagem ou vídeo, sem seus metadados, não podem ser tratados pelo sistema;

Languages detected							
Symbol	Name	Symbol	Name	Symbol	Name	Symbol	Name
af	Afrikaans	et	Estonian	it	Italian	sk	Slovak
am	Amharic	eu	Basque	ka	Georgian	sl	Slovenian
ar	Arabic	fa	Persian	ko	Korean	sq	Albanian
be	Belarusian	fi	Finnish	la	Latin	sr	Serbian
bg	Bulgarian	fr	French	lt	Lithuanian	sv	Swedish
br	Breton	fy	Western Frisian	lv	Latvian	sw	Swahili
bs	Bosnian	ga	Irish	mr	Marathi	ta	Tamil
ca	Catalan/Valencian	gd	Scottish Gaelic	ne	Nepali	th	Thai
cs	Czech	gv	Armenian	nl	Serbian	tl	Tagalog
cy	Welsh	he	Hebrew (Modern)	no	Norwegian	tr	Turkish
da	Danish	hi	Hindi	pl	Polish	uk	Ukrainian
de	German	hr	Croatian	pt	Portuguese	vi	Vietnamese
el	Greek (Modern)	hu	Hungarian	qu	Quechua	yi	Yiddish
en	English	hy	Armenian	ro	Romanian	??	Unknown
eo	Esperanto	id	Indonesian	ru	Russian		
es	Spanish	is	Icelandic	sa	Sanskrit		

Figura 17 - Idiomas identificados pelo *Language Identification* API
Fonte: (TEXTALYTICS, 2014)

- **Idioma:** A identificação é realizada utilizando a *Language Identification* API do Textalytics (TEXTALYTICS, 2014). Trata-se de um serviço que faz a identificação automática do idioma de um texto. Ele utiliza técnicas de estatísticas baseadas na avaliação *N-grams*[2] (SIDOROV *et al.*, 2013) e possui suporte a mais de 60 idiomas

[2] Campo da linguística computacional e probabilidade. Um n-grama é uma sequência contígua de n itens de uma determinada sequência de texto ou de voz. Os itens podem ser fonemas, sílabas, letras, palavras ou pares de bases de acordo com a aplicação. Os n-gramas normalmente são coletados a partir de um texto ou discurso (http://nlpwp.org/book/chap-ngrams.xhtml).

(Figura 17). Esta identificação é importante na hora de relacionar os conceitos da ontologia e na hora do processamento textual;

- **Suporte ao Idioma:** É verificado se existe definições nas ontologias de domínio para o idioma do recurso. Isto é feito para evitar processamento desnecessário no sistema, evitando uma sobrecarga deste. Sempre que uma ontologia é carregada ou alterada, o sistema realiza uma varredura verificando os idiomas suportados pela ontologia e armazena na base de dados os idiomas que contêm conceitos;
- **Linguagem de Marcações:** As marcações de linguagens web são retiradas do conteúdo textual, como HTML (*Hypertext Markup Language*), XML, Javascript e outras.

3.6.2. Analisador Textual

Após as validações, os recursos que possuem conteúdo textual são encaminhados ao Analisador Textual. Este tem a função de processar o conteúdo do recurso e localizar conceitos na ontologia de domínio. O recurso é submetido ao processamento com todas as ontologias de domínio que estão registradas no repositório semântico.

Devido a algumas restrições encontradas nas ferramentas de PLN pesquisadas, foi necessário criar duas especializações do analisador textual. Um para o idioma inglês e outro padrão para os demais. Estas limitações encontradas estão detalhadas no Apêndice A.

3.6.2.1. Analisador Inglês

O conteúdo é analisado morfologicamente e sintaticamente, isolando os tipos de palavras, suas flexões, radicais e funções semânticas. Com base nestes dados, uma busca por termos semelhantes nas ontologias de domínio é realizada. As correspondências encontradas são armazenadas em estruturas de dados juntamente com as informações das análises sintática e morfológica. Um analisador morfológico é aplicado, identificando e eliminando variações de uma palavra base, como variações de gênero, número, uso de prefixos e sufixos. As classificações identificadas pela ferramenta de PLN são exibidas no Apêndice B.

O Analisador Inglês utiliza o *General Architecture for Text Engineering* (GATE) (CUNNINGHAM *et al.*, 2014) como ferramenta de PLN e anotação. Esta ferramenta requer uma sequencia de etapas, como pode observado na Figura 18.

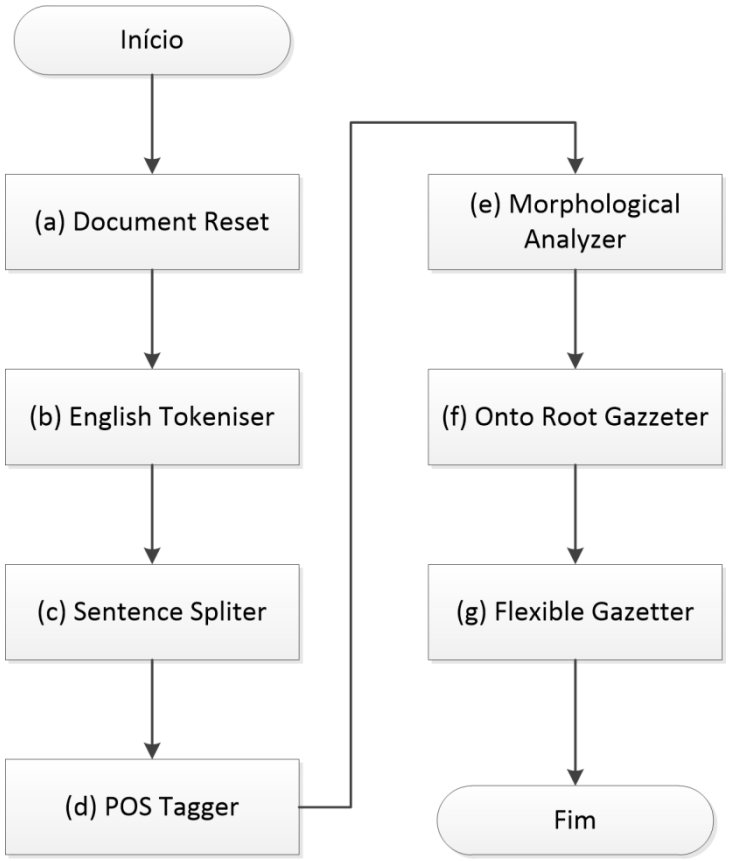


Figura 18 - Fluxograma do PLN do GATE
Fonte: Autoria Própria

As etapas apresentadas na Figura 18 são (CUNNINGHAM *et al.*, 2014):

a) *Document Reset* garante que o documento a ser anotado esteja no estado inicial sem que nenhuma definição ou anotações estejam definidas;
b) *English Tokeniser* é aplicada para produzir o grupo de anotações *Tokens*, que serão utilizadas nas ferramentas seguintes. Esta ferramenta identifica se o texto está estruturado no padrão da língua inglesa e cria *tokens* a partir das palavras e sinais de pontuação presentes no texto. Nesse caso, palavras de origem não saxônica são marcadas como *tokens* do tipo *"unknow"*. Os *tokens* produzidos contêm informações sobre o tipo, ou seja, se é um número, palavra, pontuação, etc. No caso de palavras, armazena o comprimento e como está capitalizada, ou seja, identifica se alguma letra é maiúscula, se todas o são, etc.;

c) *Sentence Splitter* gera o grupo de anotações *Sentence*, que contém informações sobre as frases do texto. Neste caso é armazenado o *offset* (o deslocamento em termos de caracteres em relação ao início do texto) de início e fim de cada frase e um identificador único;

d) *Part Of Speech Tagger*, ou simplesmente *POS Tagger*, categoriza as palavras como substantivos, adjetivos, verbos, etc. Esta função é particularmente importante na fase de análise e mineração de informações. Ao encontrar um adjetivo, pode concluir que o mesmo se refere a um substantivo presente no texto e, assim, conferir alguma qualidade ao mesmo. Qualidade essa que pode ser indexada e tratada posteriormente;

e) *Morphological Analyzer* tem o papel de encontrar variações de gênero, número e grau nas palavras e isolá-las, registrando a palavra de origem no *token*. Essa etapa é necessária para que a busca de termos na ontologia se torne mais robusta. Caso contrário, se a ontologia conter uma palavra qualquer no singular e o texto conter a mesma palavra no plural ou no diminutivo, a mesma pode não ser devidamente registrada;

f) A busca de termos no texto que tenha relacionamentos com termos na ontologia não é realizada diretamente. Na prática, um processador de texto cria arquivos específicos que são utilizados para a busca. O processador é chamado de *Onto Root Gazetteer*. Os arquivos, chamados de *gazetteers*, contêm termos presentes nas ontologias, como nomes de classes, instâncias e *labels* e seus respectivos *links* e URIs. Opcionalmente, é possível inserir pesos nos termos estabelecidos nos arquivos de *gazetteers*;

g) A busca e o casamento entre os *gazetteers* e os termos contidos no texto são realizados pelo *Flexible Gazetteer*, que, ao encontrar uma correspondência, insere uma anotação com os dados referentes ao casamento no grupo de anotações *Lookup*, como a URI, o tipo de recurso e suas propriedades.

Ao final são gerados grupos de anotações. Cada um desses grupos contém informações diferentes e complementares. Para o Sistema de Enriquecimento Semântico, os grupos necessários são: *Token*, *Sentence* e *Lookup*.

O grupo *Token* contém informações sobre a morfologia e sintaxe. Através destas anotações é possível identificar a categoria (verbo, pronome, adjetivo, ...), o tipo do *token* (palavra, pontuação, numeral, ...), tamanho, maiúscula/minúscula, palavra que se refere, a palavra raiz, posição no texto, e plural/singular. Estas anotações são particularmente

importantes, pois são elas que proporcionam o poder para um eficiente enriquecimento semântico, possibilitando a detecção de palavras na ontologia independente da variação em que está escrita. Por exemplo, na ontologia de domínio existe a referência para *"increase"*, mas no texto está escrito como *"increased"*. Estas anotações permitem estabelecer essa relação.

O grupo *sentence* contém cada uma das sentenças identificadas no conteúdo. Elas dividem o texto para efetuar o processamento. E, por fim, o grupo *Lookup* contém as relações dos conceitos existentes no texto que foram localizados nas ontologias de domínio. Este grupo identifica o tipo (classe, individuo e anotações), a classe, e o identificador do elemento da ontologia, onde foi localizado o termo. A Tabela 4 apresenta exemplos de anotações nos três grupos.

Tabela 4 - Exemplos das anotações geradas pelo GATE

Grupo	**Início**	**Fim**	**Anotação**
Sentence	162	258	{}
Sentence	0	158	{}
Token	0	8	{category=NNP, kind=word, length=8, orth=upperInitial, root=diabetes, string=Diabetes}
Token	9	17	{category=JJ, kind=word, length=8, orth=lowercase, root=mellitus, string=mellitus}
Token	18	19	{category=(, kind=punctuation, length=1, position=startpunct, root=(, string=(}
Token	19	21	{category=NNP, kind=word, length=2, orth=allCaps, root=dm, string=DM}
Token	21	22	{category=), kind=punctuation, length=1, position=endpunct, root=), string=)}
Lookup	241	257	{URI=(...)/MobiLEHealth/Diabetes#ExtremeHunger, classURI=(...)/MobiLEHealth/Diabetes#Sympton, classURIList=[(...)/MobiLEHealth/Diabetes#Sympton], heuristic_level=0, majorType=, propertyURI=http://www.w3.org/2000/01/rdf-schema#label, propertyValue=Increased hunger, type=instance}
Lookup	199	217	{URI=(...)/MobiLEHealth/Diabetes#FrequentUrination, classURI=(...)/MobiLEHealth/Diabetes#Sympton, classURIList=[(...)/MobiLEHealth/Diabetes#Sympton], heuristic_level=0, majorType=, propertyURI=http://www.w3.org/2000/01/rdf-schema#label, propertyValue=Frequent Urination, type=instance}
Lookup	115	120	{URI=(...)/MobiLEHealth/Diabetes#blood,

			classURI=(...)/MobiLEHealth/Diabetes#Parts, classURIList=[(...)/MobiLEHealth/Diabetes#Parts], heuristic_level=0, majorType=, propertyURI=http://www.w3.org/2000/01/rdf-schema#label, propertyValue=Blood, type=instance}
Lookup	115	120	{URI=(...)/MobiLEHealth/Diabetes#blood, classURI=(...)/MobiLEHealth/Diabetes#Parts, classURIList=[(...)/MobiLEHealth/Diabetes#Parts], heuristic_level=0, majorType=, type=instance}

Apesar da separação das anotações, elas estão indiretamente interligadas. Para estabelecer estas relações, é utilizado o *offset* do *token*, que se trata de uma identificação por posicionamento estabelecendo o início e fim de uma anotação no texto. Esta correspondência pode retornar que uma anotação *Lookup* é referenciada por dois *Tokens*.

Um algoritmo de seleção de dados é executado sobre os grupos de anotações para filtragem dos termos relevantes. Este algoritmo utiliza todas as anotações de *Lookup*, advindas do casamento de termos do conteúdo com os termos da ontologia e as enriquece com o conteúdo relevante das anotações dos outros grupos.

Esta operação é repetida para cada ontologia de domínio registrada no sistema, procurando correspondências do recurso sendo enriquecido com todos os domínios de conhecimento descritos no sistema. As anotações geradas pela análise semântica são armazenadas em estruturas de dados do tipo dicionário, com estruturas do tipo *chave* $\rightarrow$ *valor*.

Por exemplo, a chave *sentence* armazena a frase em que a correspondência foi encontrada e *offset* armazena o deslocamento da primeira letra da palavra em relação ao início do texto. Os dicionários com as anotações são armazenados como valores de outro dicionário, que contém como chave os identificadores das ontologias.

A análise sintática e semântica do texto provê robustez e eficácia ao enriquecimento semântico. Ele possibilita incrementar o mecanismo de contextualização do conteúdo com domínio de conhecimento. Isto é possível devido à análise das variações das palavras e suas categorias e funções.

3.6.2.2. Analisador Padrão

O processador padrão foi criado para suprir as limitações de ferramentas de PLN para os demais idiomas. Dessa forma é possível realizar um enriquecimento semântico de qualquer outra linguagem. Embora de forma não tão eficiente.

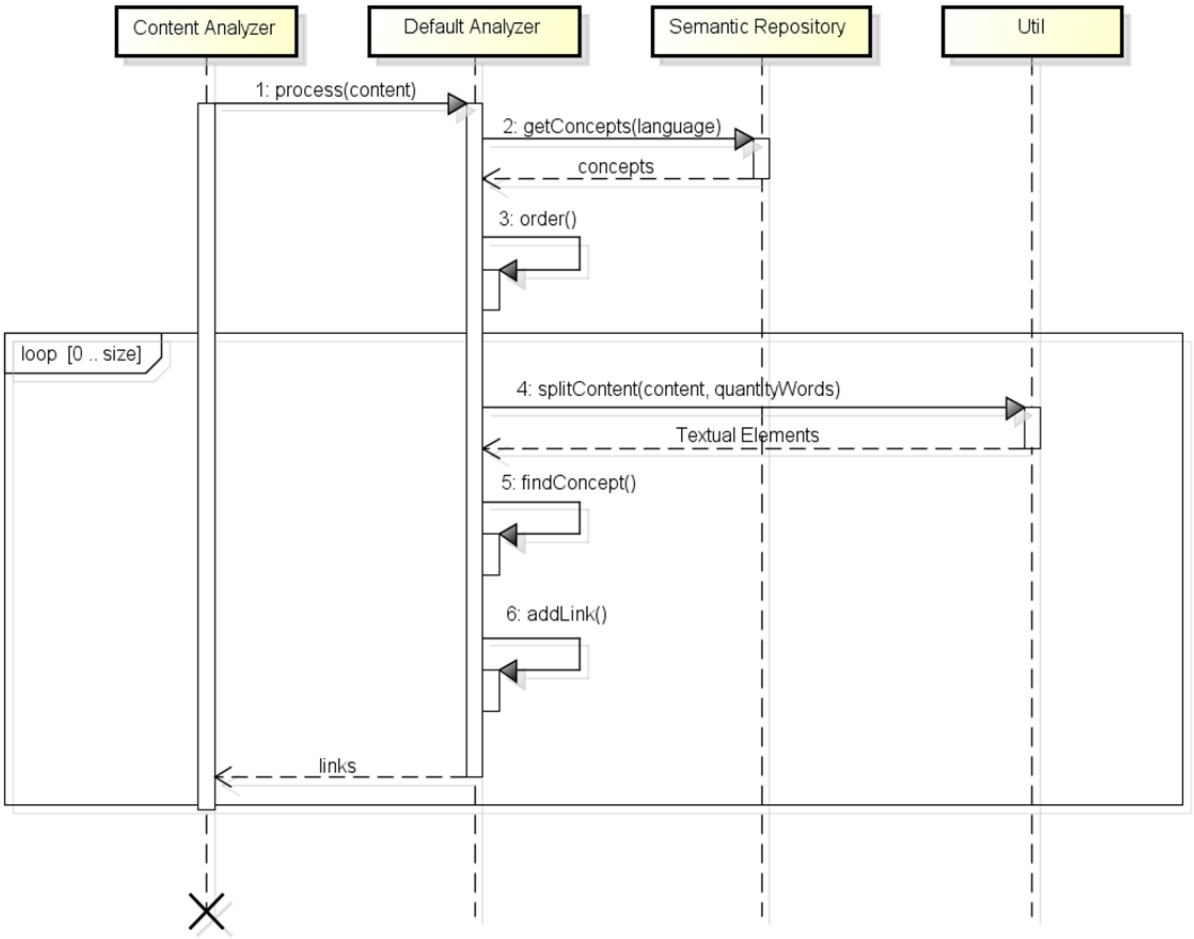


Figura 19 - Diagrama de Sequência do Analisador Padrão.
Fonte: Autoria Própria

A Figura 19 mostra um diagrama de sequência deste processamento. O processador utiliza uma forma inversa para realizar a busca de conceitos. Ao invés de analisar inicialmente o texto, ele busca, para o idioma desejado, os conceitos existentes em todas as ontologias de domínio. Estes são ordenados e agrupados por quantidade de palavras. Para cada agrupamento, o conteúdo é dividido criando um conjunto de elementos textuais. Com base nessa estrutura, os conceitos são buscados nos elementos textuais do recurso. Este processo se

repete para todos os agrupamentos. No fim do processamento serão encontradas as relações do recurso com a ontologia de domínio.

Por exemplo, considerando que a ontologia de domínio contenha os conceitos “diabetes” e “doença metabólica”, e que esteja processando o texto “Diabetes é uma doença metabólica”. Inicialmente é realizada a busca pelo conceito “diabetes”, separando o texto no conjunto de elementos textuais: {diabetes, é, uma, doença, metabólica}. Será encontrada então uma combinação. Em seguida é realizada a busca pelo conceito “doença metabólica”, separando o texto no conjunto de elementos textuais: {diabetes é, é uma, uma doença, doença metabólica}. Que também encontrará uma combinação.

Este analisador tem um fator limitador quando comparado ao analisador do inglês. Por exemplo, se na ontologia existir o conceito “medicamento” e for processado o texto “a medicação utilizada é a insulina”, neste analisador esta combinação não será encontrada. Porém no analisador do inglês é possível determinar que medicação e medicamento são o mesmo conceito, devido ao uso do PLN que permite a localização de combinações entre ontologias e conteúdo mesmo que contenham palavras com variações divergentes, conforme explicado na seção 3.6.2.1.

3.6.3. Analisador Conceitual

O Analisador Conceitual é o componente responsável por processar conteúdos não textuais do PHR do usuário, como, por exemplo, nível de glicose, pressão sanguínea, etc. A sua função é estabelecer uma relação conceitual destes dados com domínios de conhecimento sem gerar dependência da ontologia de domínio. Ou seja, sem que o processamento do dado esteja amarrado a uma classe ou instância específica da ontologia, e sim que dependa apenas da sua especificação.

Exemplificando, uma das medidas do nível de glicose do sangue é a mg/dL (miligrama por decilitro). Valores acima de 100 mg/dL são considerados como um estado hiperglicêmico. Já os valores abaixo de 70 mg/dL são considerados como um estado hipoglicêmico. Esta relação deve estar definida na ontologia de domínio. Considerando isto, através do histórico de nível de glicose do usuário, pode-se estabelecer uma relação deste registro com o conceito hiperglicemia ou hipoglicemia. Esta inferência é realizada de forma

automática e sem relação com tipo de dado, implementação do sistema e ontologia. O Sistema não busca diretamente o conceito e sim realiza uma inferência para descobrir esta relação.

Primeiramente, para poder estabelecer esta relação, é necessário que esteja definido na ontologia. A Figura 20 mostra um exemplo desta definição.

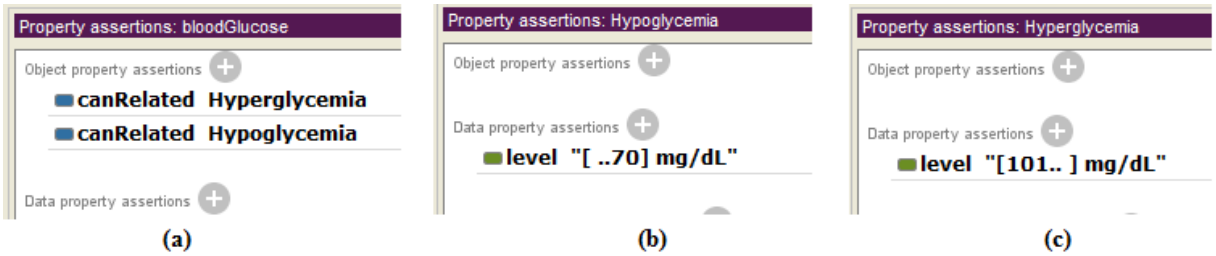


Figura 20 - Exemplificação de relações entre conceitos na ontologia de domínio.
Fonte: Autoria Própria.

De acordo com a Figura 20: no indivíduo (a) (*Blood Glucose*) são definidos os axiomas [*canRelated*, *Hyperglycemia*] e [*canRelated*, *Hypoglycemia*], que estabelecem uma relação com estes dois conceitos; no indivíduo (b) (*Hypoglycemia*) são definidos na *Data Property* os limites do nível de glicose no sangue considerados como hipoglicemia; no indivíduo (c) (*Hyperglycemia*) também existe essa definição. Através destas definições, o algoritmo consegue identificar as relações e estabelecer uma relação destes conceitos com o registro de saúde do usuário.

As definições da Figura 20 são ilustrativas. Devido à independência da ontologia, esta definição fica a cargo da modelagem do conhecimento, não precisando ser feita necessariamente desta forma. Contudo, o algoritmo não consegue identificar onde buscar a informação. Ele parte do principio que as definições estão corretas e conforme padrões normalmente adotados para os PHR, como as sugeridas pela *Health Level Seven* (HL7)[3], realizando a busca na ontologia conforme a nomenclatura do PHR.

Por exemplo, caso sejam processadas as taxas de colesterol do histórico do usuário, o sistema só irá localizar estas definições se existir na ontologia o conceito colesterol. E a partir deste buscar as definições dos conceitos relacionados. A ausência deste conceito não implica em erro no sistema, mas sim no não processamento deste dado.

O diagrama de sequencia da Figura 21 mostra o funcionamento deste algoritmo. O analisador recebe, do localizador PHR, um mapeamento no formato {conceito, {atributo, valor} }, conforme explicitado na Seção 3.5. O conceito é buscado nas ontologias de domínio.

[3] http://www.hl7.org/

Caso nada seja encontrado, o processamento é abortado. Se o conceito for encontrado, os indivíduos relacionados a este conceito são retornados pelo Repositório Semântico. Para um melhor entendimento chamaremos estes indivíduos de ind_1.

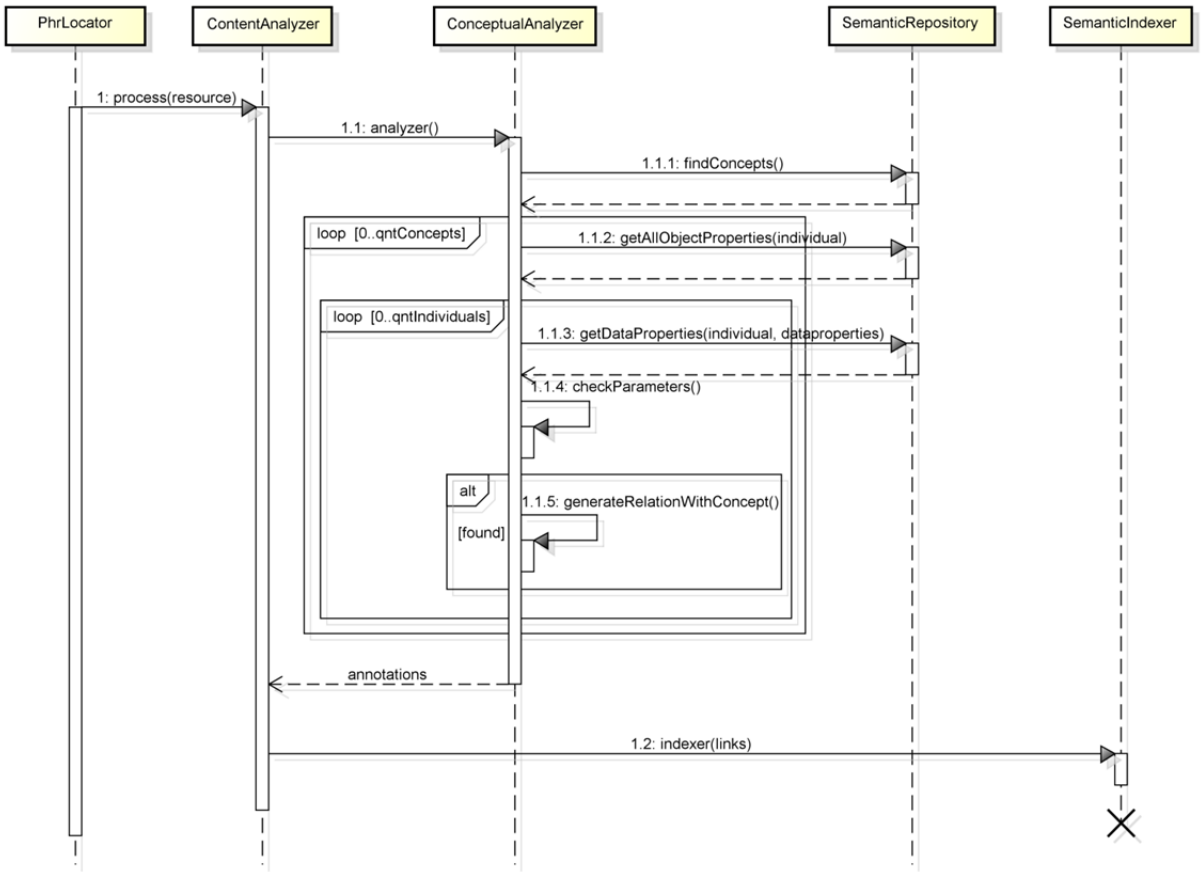


Figura 21 - Diagrama de Sequência do Analisador Conceitual.
Fonte: Autoria Própria.

Para cada ind_1 são obtidas todas as relações deste com outros indivíduos através de suas *Object Properties*, gerando uma segunda lista de indivíduos, que nomearemos de ind_2. Se nenhuma *Object Property* for encontrada, o individuo de ind_1 é descartado, passando para o subsequente.

Caso existam indivíduos em ind_2, para cada individuo são obtidas as suas definições, que estão descritas através dos suas *Data Properties*. Para cada atributo definido no mapeamento {atributo, valor}, é buscada uma relação com as *Data Property* do indivíduo. Por exemplo, no caso de pressão sanguínea é verificado se no individuo existe alguma *Data Property* chamada *diastolic* e *systolic*. Se nenhuma *Data Property* for encontrada, o indivíduo é descartado, passando para o subsequente. Se o atributo for encontrado, o seu valor é

comparado com a definição da *Data Property*. Em caso de uma combinação positiva, é gerada uma relação com este individuo da ontologia de domínio.

Ao fim do processamento, tem-se uma estrutura de dados que relaciona os itens processados e todas as relações encontradas com a ontologia de domínio. Esta estrutura contém apenas os itens para os quais foram encontradas relações, sendo os demais descartados. Esta estrutura é encaminhada ao Padronizador.

3.6.4. Padronizador

O padronizador realiza o resumo das combinações geradas pelos analisadores. Este resumo é feito para reduzir o tamanho da ontologia e melhorar a eficiência dos motores de inferência. Basicamente ele computa as combinações agrupando por elemento textual e conceito. Por exemplo, { {medicação, URI<Medicamento>, 3}, {diabetes, URI<diabetes>, 6} }. Este conjunto indica que para a palavra raiz "medicação" foram geradas três relações no texto com o conceito medicamento. Já no caso da palavra "diabetes" foram geradas seis relações. Estas anotações são convertidas em *Links* e atribuídas ao recurso.

3.7. ENRIQUECEDOR SEMÂNTICO

O Enriquecedor Semântico tem a função de processar os Recursos e seus *Links*, já com relações vinculadas, e gerar os elementos semânticos que serão gravados na ontologia de perfil do usuário. As anotações semânticas, geradas pelas etapas anteriores, são indexadas semanticamente e gravadas na ontologia de perfil de usuário, contextualizando semanticamente o recurso com domínios de conhecimento.

O seu processamento ocorre em duas etapas, como pode ser visto no Digrama de Atividade exposto na Figura 22. Na primeira, o Indexador Semântico analisa os *Links* do recurso e gera os axiomas, estabelecendo as relações de classe, instâncias e propriedades. Posteriormente, o Anotador Semântico mapeia os axiomas para a ontologia de perfil do usuário, sendo este responsável por fazer as devidas verificações para uma correta gravação.

Ao fim deste processamento, tem-se o perfil do usuário enriquecido de forma a permitir a determinação do interesse do usuário nos domínios de conhecimento registrados no sistema.

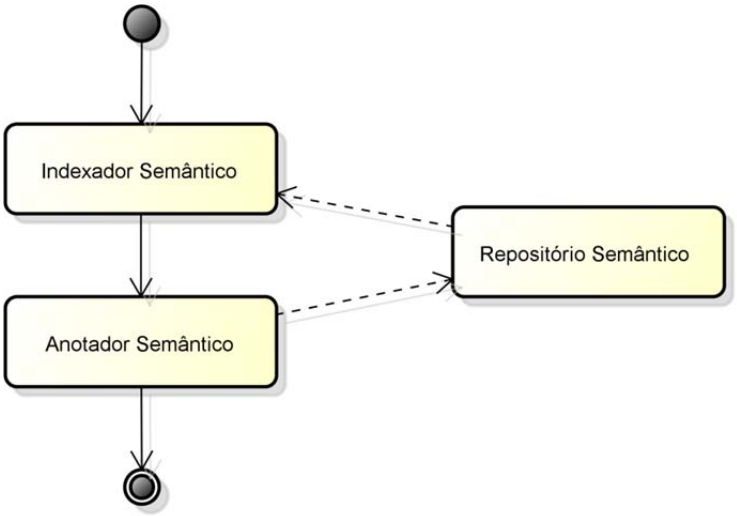


Figura 22 - Diagrama de Atividade do Enriquecedor Semântico.
Fonte: Autoria Própria.

Tanto o Indexador quanto o Anotador utilizam a OWL API[4] que é uma API Java para criação, manipulação e serialização de ontologias. Esta API foi adotada por se adequar ao formato OWL adotado pelo Sistema de Enriquecimento Semântico. Outras ferramentas foram testadas, porém não gravavam no formato desejado e recomendado pelo W3C. Além disto, esta API possui um *reasoner* nativo e também permite o uso de outros, como FaCT++, JFact, HermiT, Pellet, RacerPro, ELK. Um *reasoner* é um *software* capaz de inferir consequências lógicas a partir de um conjunto de fatos ou axiomas, normalmente usando lógica de primeira ordem para realizar o raciocínio.

3.7.1. Indexador Semântico

A principal função do Indexador Semântico é gerar os elementos semânticos das relações estabelecidas pelas camadas anteriores. Estes elementos representam os axiomas, a serem gravados na UPO, que mapeiam os dados oriundos do Analisador Semântico. Este

[4] http://owlapi.sourceforge.net/

mapeamento consiste na definição dos objetos, propriedades e suas relações, obedecendo às regras e à formatação da UPO.

O Indexador possui dois processamentos principais, um para *Resource* e outro para *Access*. No processamento do *Resource*, o primeiro passo é obter os objetos do tipo *Class, Object Property e Data Property* que serão utilizados para estabelecer as relações entre os elementos semânticos. Em seguida é gerado o individuo que representa o recurso e criado um axioma para estabelecer a relação entre o individuo e a classe *Resource*. As *Data Properties* são definidas gerando um axioma para cada uma.

Uma vez criado o individuo que representa o recurso e suas relações, é analisado cada um dos *link*s atribuídos ao recurso seguindo os seguintes passos:

- Gera-se o individuo que representa o *Link*;
- Cria-se o axioma que relaciona este individuo à classe *Link*;
- Estabelece-se as definições das *Data Properties* deste individuo;
- Gera-se o axioma que estabelece a relação entre o *Link* e o *Resource*, como também a sua relação inversa.

Por fim é executado o *reasoner* para fazer a verificação dos axiomas, garantindo que nenhuma definição errada foi estabelecida. Ao fim da execução é gerada uma lista de axiomas com todas as relações dos elementos semânticos, definidas e validadas, a serem armazenados na ontologia.

O processamento do *Access* é semelhante ao do *Resource*, contendo apenas algumas etapas a mais e a utilização de *Object e Data Properties* diferentes. Assim como no *Resource*, o primeiro passo é a obtenção dos objetos do tipo *Class, Object Property e Data Property* que serão utilizados para estabelecer as relações entre os elementos semânticos.

Em seguida é gerado o individuo *User*, que representa o usuário contido no *Access*, e gerado o axioma que contém a relação deste indivíduo com a sua classe. Uma vez que o individuo esteja criado, as suas *Data Properties* são definidas.

Como um *Access* também contém um recurso, é executado o processamento do *Resource* explicado nos parágrafos anteriores. Uma vez existindo os axiomas que comporão o *Access*, o individuo deste é criado e gerado o axioma que o relaciona à sua classe. Também são definidos as suas *Data Properties*.

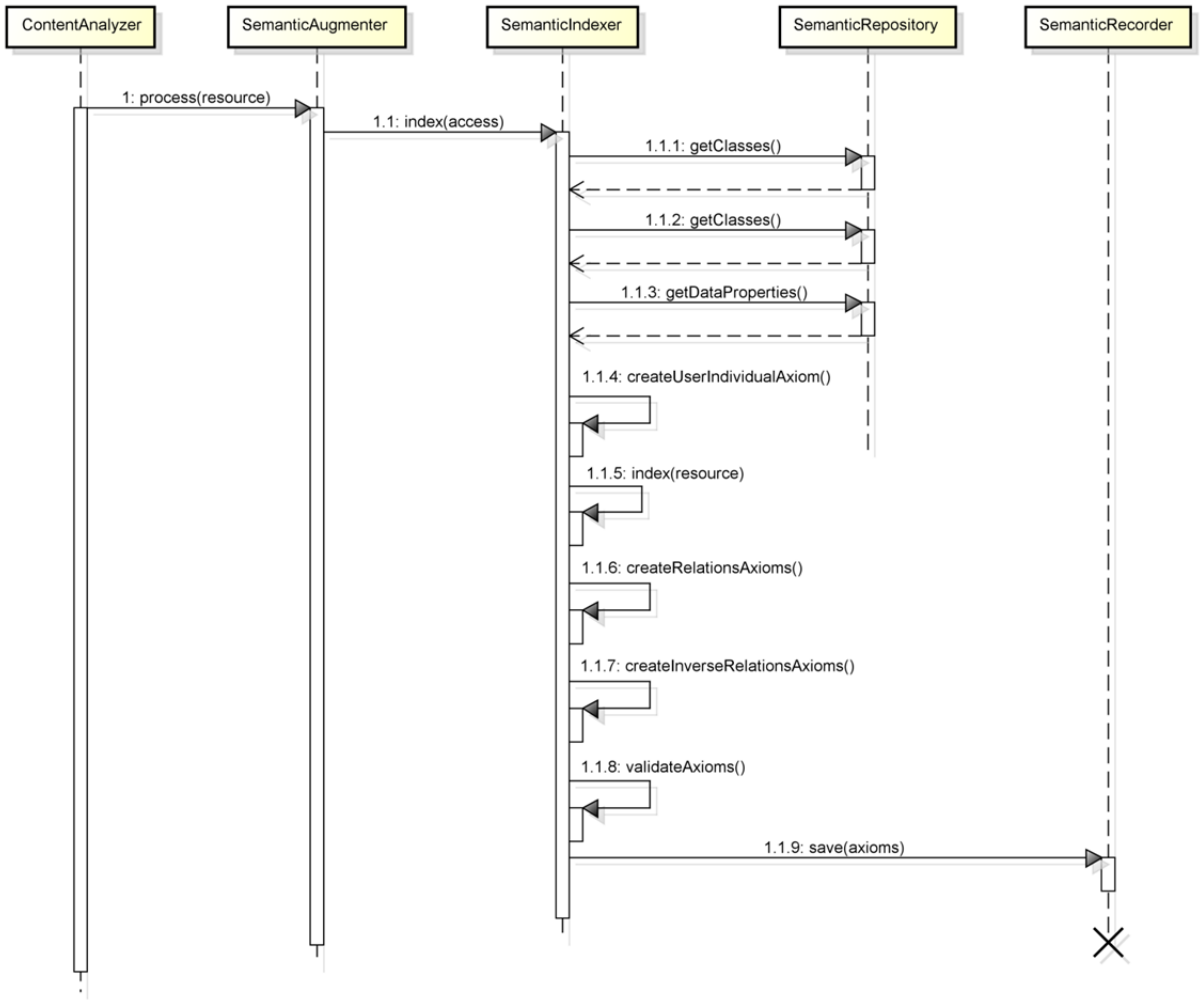


Figura 23 - Digrama de Sequência do Indexador Semântico.
Fonte: Autoria Própria.

Na etapa seguinte são gerados os axiomas que estabelecem a relação entre *User* → *Access* e *Resource* → *Access*, como também suas relações inversas *User* ←*Access* e *Resource* ←*Access*. Por fim, é executado o *reasoner* para fazer a verificação dos axiomas, garantindo que nenhuma definição errada foi estabelecida.

A lista de axiomas gerada pelo indexador contém todas as anotações semânticas necessárias para armazenar os dados enriquecidos na ontologia e estabelecer uma relação temporal com o usuário. A Figura 23 mostra o digrama de sequência do indexador semântico conforme explicado nesta subseção.

3.7.2. Anotador Semântico

O Anotador Semântico atua como uma camada de persistência do Sistema de Enriquecimento Semântico para a UPO. Ele realiza validações dos elementos semânticos a serem anotados na ontologia e realiza a sua gravação/atualização no formato utilizado pelo sistema. Este componente recebe como entrada a lista de axiomas gerada pelo Indexador e verifica a validade do axioma. Para isto tem-se como pré-requisito:

- A definição dos três elementos da relação: sujeito, predicado e objeto;
- A existência destes três elementos na ontologia;
- Se não existir, estes elementos devem ser criados e, em caso da não possibilidade desta criação, o processamento é abortado;
- Por meio de um *reasoner*, deve-se validar a relação.

Se o axioma não existir, ele é gravado na ontologia. Caso já exista, é verificado se precisa ser atualizado.

O processamento do *Access* passa por todas as etapas do Anotador, uma vez que incorpora todos os elementos da UPO: *User*, *Access*, *Resource* e *Link*. Ao receber uma solicitação para anotação de um *Access*, o sistema realiza as validações descritas acima e posteriormente verifica se já existem na ontologia os indivíduos do *Resource* e o *User* relacionado. Caso não existam, são gravados todos os axiomas relacionados a cada um destes, incluindo os *Links* de *Resource*. Caso existam, as *Object e Data Properties* são atualizadas, o que inclui: inserir os novos, excluir os que não existem mais e atualizar os que ainda existem.

Quando se tratar de um *Resource*, todos os *Links* são apagados e recriados com base nas novas anotações. Isto é feito para garantir a correta anotação deste. Apesar de poder estar relacionado a vários usuários, um recurso só é gravado uma única vez na ontologia para evitar duplicidade da informação. Contudo é atualizado sempre que necessário, tendo em vista que um recurso pode estar em constante mudança. Isto garante a correta anotação do recurso.

Após isto, verifica-se a existência do individuo *Access* na ontologia. Para isto, considera-se a relação usuário, recurso e data/hora. Ou seja, para considerar se um *Access* já existe, tem que existir uma relação de um determina usuário a um determinado recurso naquele exato momento. Se esta relação não existir, trata-se de um novo *Access*. Se existir, os dados do individuo *Access* são atualizados.

Outra função do Anotador é garantir que recursos que não possuem nenhum acesso relacionado a ele não fiquem armazenados na ontologia, onerando-a. Para isto, são verificados se existem recursos que não estão relacionados a nenhum *Access* e, quando encontrado, o recurso e seus relacionamentos são apagados da ontologia.

3.8. MOTOR DE INFERÊNCIA

O Motor de inferência do Sistema de Enriquecimento Semântico tem a função de processar as informações contidas na ontologia e extrair informações relevantes ao domínio da aplicação. Este componente fornece uma interface externa, disponibilizando um serviço que é consumido pelo ambiente do MobiLEHealth, sempre que este demandar. Este serviço tem a função de responder para este ambiente: (i) a relação de um conteúdo com um domínio de conhecimento e (ii) a relação de um usuário com um domínio de conhecimento.

Estas relações estabelecem um percentual de relação do usuário/recurso com o domínio de conhecimento. Por exemplo, pode-se determinar que um determinado recurso possui 70% de possibilidade de ter relação com o domínio diabetes ou que um usuário pode ter 90% de relação com este mesmo domínio, porém ter uma relação de apenas 0,15% com outro domínio que não é do seu interesse.

A interface externa do Sistema de Enriquecimento Semântico disponibiliza quatro serviços ao MobiLEHealth, que fornecem estas respostas (Figura 24).

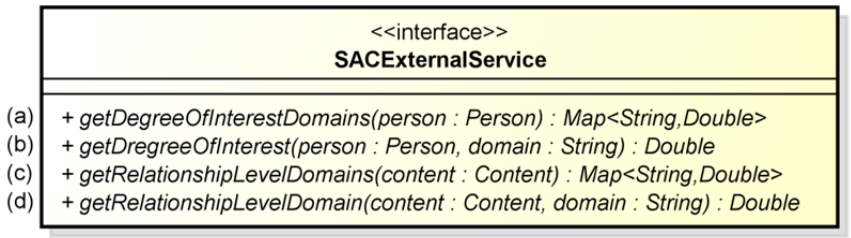


Figura 24 - Interface Externa do Componente de Enriquecimento Semântico.
Fonte: Autoria Própria.

Como pode ser visto na Figura 25, os serviços (a) *getDegreeOfInterestsDomains* e (b) *getDegreeofInterest* respondem a relação do conteúdo com um domínio, sendo retornado para o primeiro o domínio especificado e para o segundo uma lista com a relação do conteúdo para todos os domínios registrados. Já o (c) *getRelationshipLevelDomains* e o (d)

getRelationshipLevelDomain fazem o mesmo, sendo que para o usuário. Para (a) e (c), caso seja informado um domínio não existente, é retornado o valor zero como resposta.

Nas subseções abaixo são detalhados os cálculos das relações do conteúdo e do usuário com o domínio de conhecimento. As fórmulas são de autoria deste trabalho, onde, as métricas que as compõe foram elencadas com base no impacto para a análise de conteúdo e perfil de usuário, conforme os conceitos extraídos das pesquisas realizadas.

3.8.1. Relação do Conteúdo com Domínio

A Relação do Conteúdo com Domínio (Rc) trata-se de um índice que estabelece o grau de relação que possivelmente um conteúdo possui com um determinado domínio de conhecimento. Este índice varia de zero a um, representando o percentual desta relação. O calculo do índice está representado pela Equação 1.

$$Rc = \frac{Mr + Mc}{2} \quad (1)$$

O Rc trata-se de uma média aritmética baseada em outras duas médias: A média de *links* do recurso com o domínio (Mr) e a média de conceitos do recurso com os conceitos do domínio (Mc).

O calculo do Mr, mostrado na Equação 2, calcula a representatividade das anotações semânticas para o domínio que está sendo calculado em relação a todas anotações do recurso. Para isto, é realizada uma inferência sobre a UPO para obter os elementos semânticos do recurso que está sendo processado. Estes elementos são analisados contabilizando as relações do recurso com domínios de conhecimento, estabelecidas por suas *Object Properties* do tipo *hasLink*. Para contabilizar cada *link*, é multiplicado pelo *Data Property quantityTokens* que contém a quantidade de vezes que aquele *link* apareceu no recurso. Por fim, é calculada uma média aritmética entre a quantidade de relações com domínio especificado e a quantidade total de relações do recurso.

$$Mr = \frac{\sum link_{domínio}}{\sum link_{recurso}} \quad (2)$$

Neste cálculo são considerados apenas os elementos do recurso que geraram relações com algum domínio. Por exemplo, se um recurso possuir quatro relações com os domínios e se três delas forem para o domínio diabetes, considera-se uma representatividade de 75%, independente do tamanho do recurso.

A Equação 3 mostra o cálculo do Mc, que é a representatividade dos conceitos do recurso relacionados ao recurso em relação a todo o domínio. Neste calculo são consideradas apenas as anotações relacionadas ao domínio em questão, desconsiderando as outras anotações. Isto porque esta média representa o quanto um recurso incorpora conceitualmente um determinado domínio.

$$Mc = \frac{R_1 + R_2 + R_3}{3} \tag{3}$$

A representatividade conceitual do recurso no domínio (R_1) é explicitada na Equação 4, que calcula o percentual de conceitos que o recurso abrange do total de conceitos da ontologia. Para isto é contabilizada apenas a quantidade de conceitos únicos nas anotações. Por exemplo, se um recurso contém dez anotações com a ontologia de domínio, mas estas apontam para apenas três conceitos, é contabilizado apenas três. Esta quantidade de conceitos únicos é divida pelo total de conceitos descritos na ontologia de domínio.

$$R_1 = \frac{\sum conceito_{recurso}}{\sum conceitos_{dominio}} \tag{4}$$

Contudo apenas a quantidade de conceitos não é suficiente para determinar esta relação. A capacidade que o recurso possui de estabelecer relação com o domínio de conhecimento também deve ser considerada. Por isto é considerada a representatividade das anotações semânticas do recurso para o domínio em questão em relação ao tamanho do recurso (R_2), destacada na Equação 5, como também a capacidade máxima de um conteúdo gerar relação com a ontologia (R_3), destacada na Equação 6.

O R_2 calcula o percentual que os *links*, para o domínio que está sendo tratado, representam em relação ao tamanho total do conteúdo. Por exemplo, um conteúdo pode conter apenas duas anotações semânticas, porém estas podem representar 60% do seu tamanho total. Como também um conteúdo pode conter vinte anotações semânticas, mas representar apenas

10% do seu tamanho total. Por isto o tamanho do recurso impacta diretamente na determinação da sua relação com um domínio.

$$R_2 = \frac{\sum Link_{dominio}}{size_{resource}} \quad (5)$$

Outro fator considerado é a capacidade máxima de quantidade de *links* que um conteúdo pode gerar para um domínio, levando em conta o limite mínimo da relação do item de um recurso com o domínio (R_3). Ou seja, considerando que um item do recurso pode estabelecer uma relação com o domínio. Por exemplo, se um recurso tem tamanho seis, é considerado que ele poderia estabelecer seis relações com um domínio, através da proporção de um para um nesta relação.

$$R_3 = \frac{size_{resource}}{\sum conceitos_{dominio}} \quad (6)$$

Para entender melhor o cálculo do Mc, vamos considerar o conteúdo "a Diabetes é uma doença crônica". E que ele gerou uma relação (*Link*) para o conceito "Diabetes Mellitus" da ontologia de domínio de Diabetes. Supondo também que esta ontologia possui cinquenta definições conceituais, estabelece-se um valor para Mr de 100% (um *link* divido por um *link* - total).

Para o R_1 tem-se uma relação conceitual de 2% (um *link* divido por cinquenta conceitos). Já para o R_2 tem-se uma representatividade de 16,67% (um *link* com o domínio dividido por seis palavras – tamanho do recurso) das anotações em relação ao recurso. O R_3 representa uma capacidade de 12% (seis palavras dividido por cinquenta conceitos) que o recurso tem de estabelecer algum vínculo com a ontologia. Com isso chega-se a um percentual de Mc de 10,22% de representatividade. Por fim obtém-se um Rc de 55,11% de relação deste conteúdo com o domínio diabetes.

3.8.2. Relação do Usuário com Domínio

A Relação do usuário com Domínio (Ru) trata-se de um índice que estabelece o grau de interesse que possivelmente um usuário possui com um determinado domínio de conhecimento. Este índice varia de zero a um, representando o percentual desta relação. O cálculo do índice está representado pela Equação 7.

$$Ru = Ma\ x\ \frac{\sum_1^i Rc_i}{i} \tag{7}$$

O cálculo do Ma, apresentado na Equação 8, determina o percentual de acesso do usuário a recursos do domínio em questão. Para isto, contabiliza a quantidade de *Access* relacionados ao usuário e calcula o Rc de todos eles para o domínio em questão. Em seguida são contabilizados a quantidade de *Access* cujo recurso possui um Rc maior ou igual a 30%[5] e é calculada a média aritmética dos seus Rc. Um recurso que possua um índice abaixo disto não é considerado como relacionado ao domínio. O valor de Ma é multiplicado pela média dos Rc para obter-se o índice de relação do usuário com o domínio.

$$Ma = \frac{\sum access_{recurso}}{\sum access} \tag{8}$$

Exemplificando, considerando que um usuário possua cinco acessos, cujos Rc para o domínio Diabetes são 50%, 9%, 75%, 90%, 40%. Determina-se que 80% (quatro acessos acima de 30% divido por cinco acessos totais) dos acessos deste usuário estão relacionados ao domínio diabetes, através do cálculo do Ma. A média dos Rc relacionados à diabetes é de 63,75% ((50 + 75 + 90 + 40) / 4). Com isso é estabelecida uma relação de 51% deste usuário com o domínio diabetes, através do cálculo do Ru (80 * 63,75 / 100).

[5] Testes no Sistema de Enriquecimento Semântico mostraram que percentuais iguais ou maiores que 30% representam que o recurso possui alguma relação com domínio.

4. VALIDAÇÕES E RESULTADOS

Para os experimentos de validação do sistema foram selecionados como domínios da aplicação as doenças crônicas Diabetes e a Esclerose Lateral Amiotrófica (ELA). Estas doenças foram escolhidas (i) pela abrangência em que a doença atinge a população, no caso da Diabetes, (ii) pelo impacto que causam no cotidiano das pessoas portadoras, no caso da ELA, e (iii) pelo fato de que a falta de informação dificulta ainda mais a convivência com essas doenças. Outro ponto levado em consideração foi o fato do Laboratório de Inovação Tecnológica em Saúde (LAIS) do Hospital Universitário Onofre Lopes (HUOL), cujo coordenador faz parte da equipe deste projeto, trabalhar com estas duas doenças.

O processo de validação passou por quatro etapas:

i. Criação das ontologias de domínio para realizar o processo de enriquecimento semântico;
ii. Seleção dos conteúdos e o processamento semântico;
iii. Definição e geração dos perfis de usuários;
iv. Análise dos resultados.

Na primeira etapa foram criadas duas ontologias de domínio, uma para diabetes e outra para ELA, que descrevem os conceitos de cada uma das doenças crônicas. Para a criação foram selecionados alguns conteúdos de portais conceituados sobre os domínios em questão como, por exemplo, os portais da *International Diabetes Federation* (IDF) (www.idf.org) e da *Amyotrophic Lateral Sclerosis Association* (ALSA) (www.alsa.org). Desses conteúdos foram extraídos termos comuns em mais de 90% dos conteúdos selecionados e categorizados para a criação da ontologia. Desse processo foram selecionados 25 termos para o domínio diabetes e 51 termos para o domínio ELA, classificados em dezesseis categorias, que para os domínios selecionados foram definidos de forma idêntica. Estas categorias podem ser vistas na Figura 25.

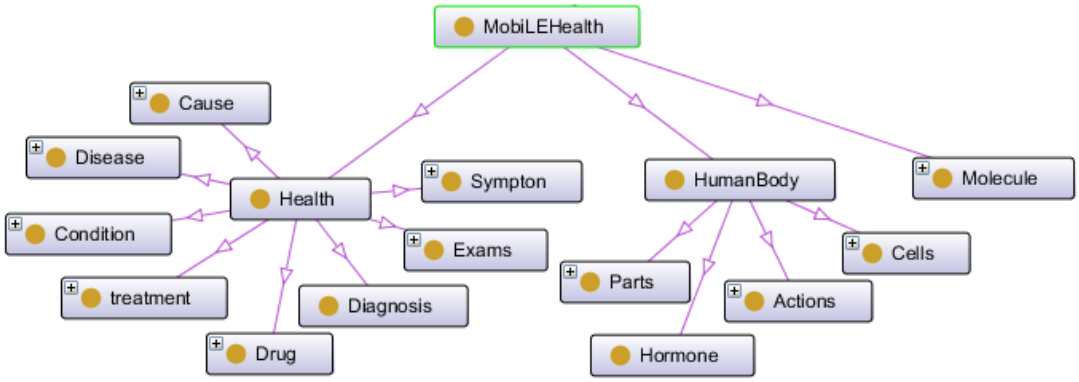


Figura 25 - Estrutura das ontologias de domínio criadas para simulação.
Fonte: Autoria Própria

Já na segunda etapa, para realizar o enriquecimento semântico, foram selecionados 240 conteúdos para cada um dos domínios. A seleção ocorreu pela internet e de forma manual, com o intuito de garantir:

- Que o conteúdo era fortemente relacionado ao domínio;
- A qualidade e confiabilidade do conteúdo.

Além disso, foram selecionados aleatoriamente 240 conteúdos de assuntos diversos, como economia, esporte, jogos, etc., pois acessos a conteúdos diversos fazem parte do cotidiano dos usuários.

Após a análise semântica de todos os conteúdos, foi verificado o percentual de acerto das anotações semânticas. Uma anotação é considerada "certa" se fizer referência à ontologia de domínio correlata ao assunto do conteúdo. Qualquer anotação semântica gerada para um conteúdo de assuntos diversos é considerada "errada". Como resultado obteve-se uma precisão média de 75,91% com um desvio padrão de 6,035%. Esses dados podem ser observados na Tabela 5.

Tabela 5 - Percentual de acertos por domínio.

Domínio	**Qtd Certos**	**Qtd Errados**	**% Acerto**
Diabetes	17.281	3.818	81,94%
ELA	8.662	3.735	69,87%
Média	25.943	7.553	75,91%

Na terceira etapa, os perfis de usuários foram definidos baseados nas possibilidades de interesses dos usuários considerando os três assuntos dos conteúdos (diabetes, ELA e

diversos), onde foram definidos quatro perfis: (i) diabetes, (ii) ELA, (iii) diabetes e ELA, (iv) nenhum.

Para cada perfil foram simulados cinquenta usuários e para cada usuário foram realizados trinta acessos a conteúdos, distribuídos conforme demonstrado na Figura 26. Os percentuais definidos nesta tabela foram definidos para garantir a definição correta dos perfis na simulação. Ao todo foram simulados duzentos perfis de usuários e seis mil acessos a conteúdos.

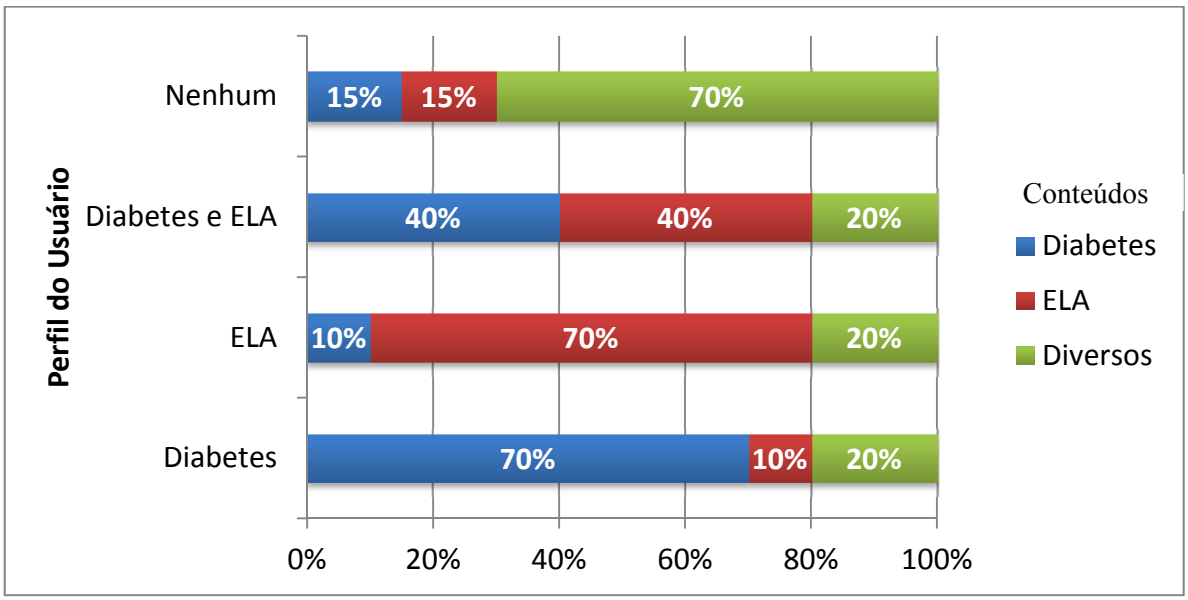


Figura 26 - Distribuição em percentual dos acessos para cada perfil dos dados simulados.
Fonte: Autoria Própria.

Por fim, foram analisados os resultados obtidos com o processamento dos dados simulados. Para avaliar a eficácia do sistema de enriquecimento semântico, foram realizadas (i) a relação do conteúdo com o domínio e (ii) a relação do usuário com o domínio.

Em ambos os casos citados, o cálculo da eficácia considera que os índices abaixo de 30% representam a não existência de relação do conteúdo ou usuário com o domínio. A variação de 30% a 100% representa nível de relação com o domínio e será utilizada pelo sistema de recomendação do MobiLEHealth como peso para selecionar os conteúdos mais apropriados ao usuário.

Para analisar os resultados, foram consideradas as métricas Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN), onde VP são os índices que submetidos a um domínio tiveram os valores na margem esperada, enquanto que VN são os índices que não retornaram valores

dentro da margem esperada. O índice de eficácia (Ei) é avaliado pela representatividade do VP, como mostrado na Equação 9.

$$\mathrm{Ei} = \frac{\sum \mathrm{VP}}{\sum(\mathrm{VN} + \mathrm{VP})} \tag{9}$$

Foi avaliada a relação recurso-domínio para cada um dos 240 conteúdos dos três assuntos (diabetes, ELA e outros) para cada um dos dois domínios (diabetes e ELA). Considerando a metodologia de avaliação descrita anteriormente, foi obtida uma eficácia de 91,42% do sistema de enriquecimento semântico ao determinar a qual domínio um conteúdo pertence. Os detalhes podem ser observados na Figura 27.

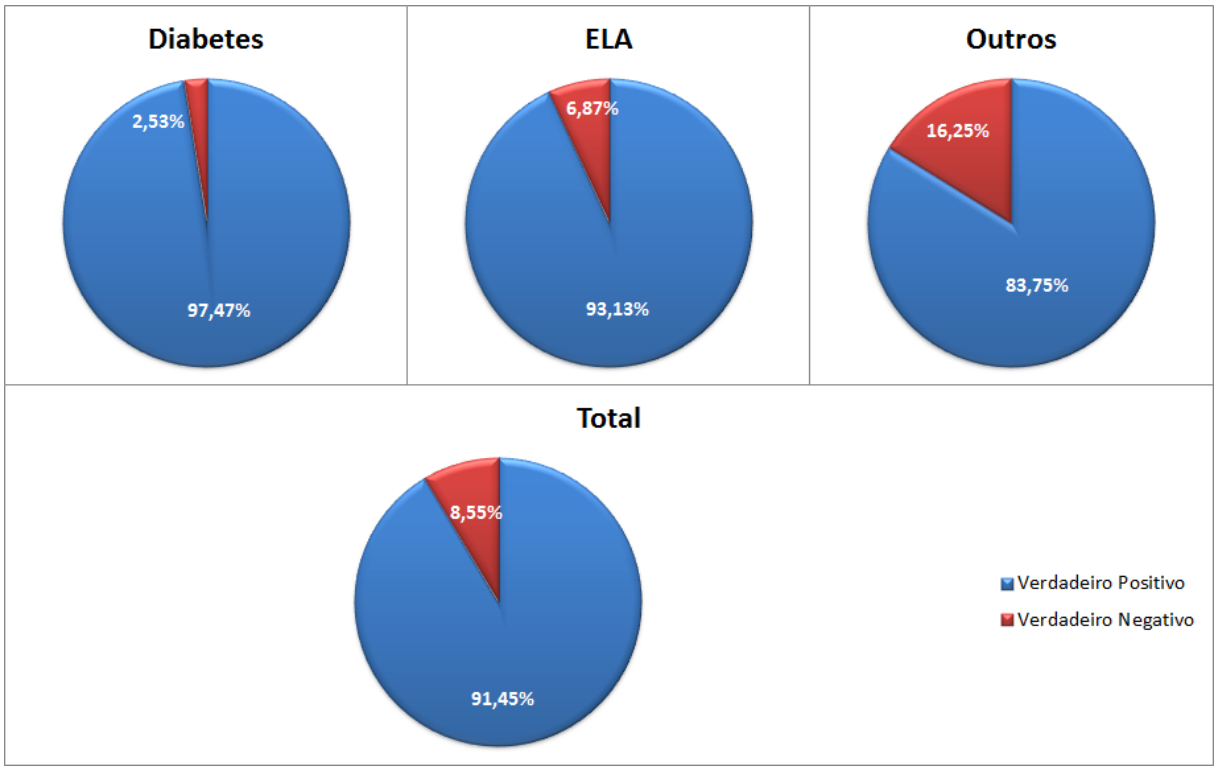


Figura 27 - Resultado da análise dos índices da relação do conteúdo com o domínio.
Fonte: Autoria Própria.

Na validação do índice de interesse do usuário, foi analisado cada usuário para cada um dos domínios e foi obtida uma eficácia de 86,54% ao determinar qual o domínio de interesse do usuário. O detalhamento desses resultados por perfil de usuário pode ser observado na Figura 28.

Como pôde ser comprovado pelos índices de eficácia, o sistema de recomendação apresentou resultados satisfatórios ao determinar a relação de um conteúdo com um domínio e

o interesse do usuário em um domínio. Porém percebeu-se que a precisão semântica depende das ontologias de domínio, que precisam ser bem definidas.

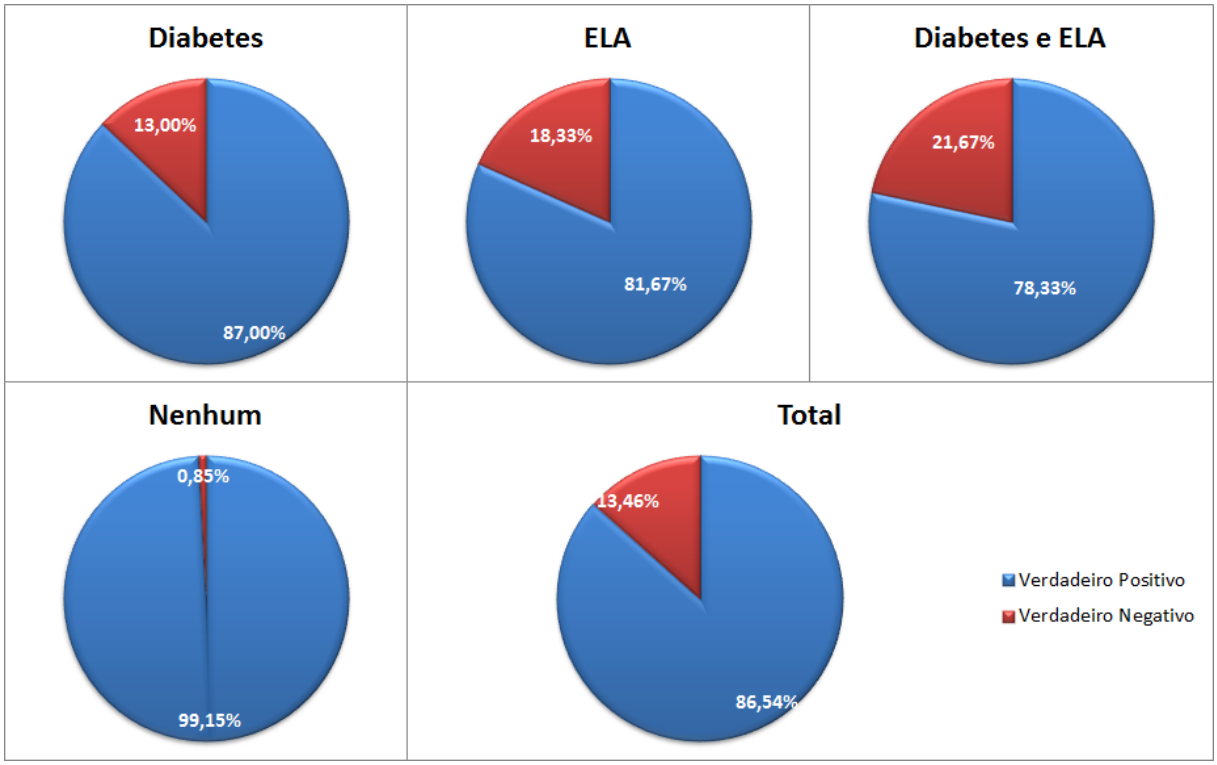


Figura 28 - Resultado da análise dos índices de domínio de interesse do usuário.
Fonte: Autoria Própria.

A separação do perfil do usuário das ontologias de domínio foi outro ponto relevante, uma vez que se mostrou eficaz e flexível, permitindo a extensão da aplicação a outros domínios sem a necessidade de alterações no código do sistema.

5. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Este trabalho demonstrou que, através da contextualização de traços digitais (TDs) dos usuários, é possível estabelecer uma relação de interesse relacionada à saúde deste. Sendo isto, estabelecido de forma implícita e automatizada através de relações semânticas com domínios de conhecimento. A pesquisa demonstrou também ser possível traçar o perfil do usuário com base em dados coletados de interações em seu cotidiano.

O sistema de enriquecimento semântico desenvolvido foi integrado ao MobiLEHealth, auxiliando-o na recomendação personalizada de conteúdos relacionados à saúde de pessoas portadoras de doenças crônicas, visando uma melhoria na qualidade de vida destas pessoas.

O Sistema de Enriquecimento Semântico foi validado através de dados simulados, testando-o de forma isolada e em um ambiente controlado. Os resultados destes testes responderam às expectativas esperadas, fornecendo relações precisas sobre o usuário. Os resultados alcançados resultaram nas publicações/submissões listadas abaixo.

Como trabalho futuro pretende-se avaliar o Sistema de Enriquecimento Semântico com dados reais, obtidos através de testes com o MobiLEHealth em um ambiente com usuários reais. Esta validação tem como objetivo comprovar a eficácia do Sistema de Enriquecimento Semântico integrado ao MobiLEHealth.

Além disto, o perfilhamento de usuários por meio de TDs tem muito a ser explorado. Como, por exemplo, (i) o processamento de linguagem natural eficiente para outras linguagens além do inglês e (ii) a inclusão da localização à ontologia de perfil do usuário para determinação do seu interesse considerando a sua localização atual.

REFERÊNCIAS BIBLIOGRÁFICAS

ABEL, F. *et al.* Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. ESWC'11, 2011, Berlin, Heidelberg. *Anais*... Berlin, Heidelberg: Springer-Verlag, 2011. p. 375–389.

ALIANÇA NETO, A. S. DE; MENDES NETO, F. M.; MOREIRA, J. D. C. Uma Nova Abordagem para o Perfilamento de Usuários em Ambientes de Aprendizagem Ubíqua. *Anais do Simpósio Brasileiro de Informática na Educação*, v. 25, n. 1, p. 1243–1252, 2014.

ALLEMANG, D.; HENDLER, J. *Semantic Web for the Working Ontologist, Second Edition: Effective Modeling in RDFS and OWL*. 2. ed. Waltham, MA: Morgan Kaufmann, 2011.

ANTONIOU, G. *et al. A Semantic Web Primer*. Cambridge, MA: MIT Press, 2012. (3a Edtiton).

BAKER, D. W. *et al.* Health Literacy, Cognitive Abilities, and Mortality Among Elderly Persons. *Journal of General Internal Medicine*, PMID: 18330654PMCID: PMC2517873, v. 23, n. 6, p. 723–726, jun. 2008.

BALDAN, M. A.; MENEZES, C. S. Um Ambiente para Construção de Perfis a Partir de Textos Pessoais. *Anais do Simpósio Brasileiro de Informática na Educação*, v. 23, n. 1, 2012.

BALSA, J. *Uma Arquitectura Multiagente para um Sistema de Processamento de Línguas Naturais Robusto e Evolutivo*. 2004. Department of Informatics, University of Lisbon, 2004. Disponível em: <http://www.di.fc.ul.pt/tech-reports/04-20.pdf>.

BECHHOFER, S. *et al. OWL Web Ontology Language Reference*. . [S.l.]: W3C, 2014. Disponível em: <http://www.w3.org/TR/2004/REC-owl-ref-20040210/>. Acesso em: 13 abr. 2013.

CAMBRIA, E.; WHITE, B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, v. 9, n. 2, p. 48–57, maio 2014.

CAMOUS, F.; MCCANN, D.; ROANTREE, M. Capturing Personal Health Data from Wearable Sensors. In: INTERNATIONAL SYMPOSIUM ON APPLICATIONS AND THE INTERNET (SAINT), 2008, Turku, Finland. *Anais*... Turku, Finland: IEEE Computer Society, 2008. p. 153–156.

CANTADOR, I.; CASTELLS, P. Extracting multilayered Communities of Interest from semantic user profiles: Application to group modeling and hybrid recommendations. *Computers in Human Behavior*, Social and Humanistic Computing for the Knowledge Society. v. 27, n. 4, p. 1321–1336, jul. 2011.

CASTLETON, G.; GERBER, R.; PILLAY, H. (Org.). *Improving Workplace Learning: Emerging International Perspectives*. New York: Nova Science Publishers Inc, 2006.

CAZELLA, S. C. *et al.* Recomendando Objetos de Aprendizagem Baseado em Competências em EAD. *RENOTE*, v. 9, n. 2, 2 jan. 2012. Disponível em: <http://seer.ufrgs.br/index.php/renote/article/view/25123>.

CHEN, C. *et al.* Making Sense of Mobile Health Data: An Open Architecture to Improve Individual- and Population-Level Health. *Journal of Medical Internet Research*, v. 14, n. 4, p. e112, 9 ago. 2012.

COSTA, A. A. L. *et al.* Recomendação personalizada de conteúdo para suporte à aprendizagem informal no contexto da saúde. *Revista Novas Tecnologias na Educação*, n. 12, 2014.

CUNNINGHAM, H. *et al. Developing Language Processing Components with GATE Version 8 (a User Guide*. Disponível em: <http://gate.ac.uk/sale/tao/tao.pdf>. Acesso em: 25 mar. 2014.

DA SILVA, L. C. N.; MENDES NETO, F. M.; JÁCOME JÚNIOR, L. MobiLE: Um ambiente Multiagente de Aprendizagem Móvel para Apoiar a Recomendação Sensível ao Contexto de Objetos de Aprendizagem. *Anais do Simpósio Brasileiro de Informática na Educação*, v. 1, n. 1, 2011.

DELLA MEA, V. What is e-Health (2): The death of telemedicine? *Journal of Medical Internet Research*, v. 3, n. 2, p. e22, 22 jun. 2001.

DESPOTAKIS, Dimoklis; LAU, Lydia; DIMITROVA, Vania. Capturing the semantics of individual viewpoints on social signals in interpersonal communication. *Journal of Web Semantics, Special Issue on Personal and Social Semantic Web*, 2011.

DUNG, T. Q.; KAMEYAMA, W. A Proposal of Ontology-based Health Care Information Extraction System: VnHIES. In: IEEE INTERNATIONAL CONFERENCE ON RESEARCH, INNOVATION AND VISION FOR THE FUTURE, 2007, Hanoi, Vietnam. *Anais*... Hanoi, Vietnam: IEEE, 2007. p. 1–7.

FERNANDEZ-LUQUE, L. *et al.* Personalized health applications in the Web 2.0: The emergence of a new approach. In: 2010 ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC), ago. 2010, Buenos Aires, Argentina. *Anais*... Buenos Aires, Argentina: IEEE, ago. 2010. p. 1053–1056.

GROUP, O. W. *OWL 2 Web Ontology Language Document Overview*. Disponível em: <http://www.w3.org/TR/owl2-overview/>. Acesso em: 10 maio 2014.

HECKMANN, D. *et al.* Gumo – The General User Model Ontology. In: ARDISSONO, L.; BRNA, P.; MITROVIC, A. (Org.). . *User Modeling 2005*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2005. p. 428–432.

HOEBEL, N.; ZICARI, R. V. Creating User Profiles of Web Visitors Using Zones, Weights and Actions. In: 2008 10TH IEEE CONFERENCE ON E-COMMERCE TECHNOLOGY AND THE FIFTH IEEE CONFERENCE ON ENTERPRISE COMPUTING, E-COMMERCE AND E-SERVICES, jul. 2008, Washington, USA. *Anais*... Washington, USA: IEEE Computer Society, jul. 2008. p. 190–197.

HUGHES, B.; JOSHI, I.; WAREHAM, J. Health 2.0 and Medicine 2.0: Tensions and Controversies in the Field. *Journal of Medical Internet Research*, v. 10, n. 3, p. e23, 6 ago. 2008.

HUHNS, M. N.; STEPHENS, L. M. Multiagent Systems. In: WEISS, G. (Org.). . 1. ed. Cambridge, MA, USA: MIT Press, 1999. p. 79–120.

IBGE. *Indicadores Sociodemográficos e de Saúde no Brasil*. Pesquisa, nº 25. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2009. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/indic_sociosaude/2009/indicsaude.pdf>.

ISTEPANIAN, R. *et al. M-Health: Emerging Mobile Health Systems*. 2006 edition ed. New York, N.Y: Springer, 2005.

ITU. *World Telecommunication/ICT Indicators database*. Disponível em: <http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx>. Acesso em: 13 nov. 2014.

JIANG, J. J.; CONRATH, D. W. Semantic similarity based on corpus statistics and lexical taxonomy. 1997, [S.l: s.n.], 1997.

JIUGEN, Y.; RUONAN, X.; XIAOQIANG, H. Constructing informal learning mode based on social software. In: 2011 6TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE EDUCATION (ICCSE), ago. 2011, [S.l: s.n.], ago. 2011. p. 1227–1230.

JONG HWA KIM, H. J. L. Extraction of user profile based on workflow and information flow. *Expert Syst. Appl.*, v. 39, p. 5478–5487, 2012.

KAPTELININ, V.; NARDI, B. A. *Acting with Technology: Activity Theory and Interaction Design*. Cambridge (Mass.); London: The MIT Press, 2009.

KARANASIOS, S. *et al.* Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, v. 64, n. 12, p. 2452–2467, 1 dez. 2013.

KLEINBERG, J. The Convergence of Social and Technological Networks. *Commun. ACM*, v. 51, n. 11, p. 66–72, nov. 2008.

LAKIOTAKI, K. *et al.* User profiling based on multi-criteria analysis: the role of utility functions. *Operational Research*, v. 9, n. 1, p. 3–16, 1 maio 2009.

LOPES, G. P.; ROCIO, V.; SILVA, J. B. DA. Overcoming lexical information incompleteness (Superando a incompletude da informação lexical). P. Marrafa e M. A. Mota ed. [S.l.]: Edições Colibri, 1999. p. 121–149.

LUSTOSA, M. A.; ALCAIRES, J.; COSTA, J. C. DA. Adesão do paciente ao tratamento no Hospital Geral. *Revista da SBPH*, v. 14, n. 2, p. 27–49, dez. 2011. Acesso em: 13 nov. 2014.

MACHLES, D. L. Situated Learning. *Professional Safety*, v. 98, n. 9, p. 22–28, 2003. Acesso em: 13 nov. 2014.

MALIK, M. S. A.; SULAIMAN, S. Doctor's perspective for use of EHR visualization systems in public hospitals. In: SCIENCE AND INFORMATION CONFERENCE (SAI), 2013, out. 2013, London, UK. *Anais*... London, UK: IEEE, out. 2013. p. 86–92.

MARKKULA, M.; SINKO, M. Economias baseadas no conhecimento e sociedades inovadoras evoluem em torno da aprendizagem. *Inovação e criatividade*, v. 13, 2009.

MENDES NETO, F. M. *et al.* An Approach for Recommending Personalized Contents for Homecare Users in the Context of Health 2.0. EATIS '14, 2014, New York, NY, USA. *Anais*... New York, NY, USA: ACM, 2014. p. 33:1–33:2.

MENDES NETO, F. M. *et al.* Content's Personalized Recommendation for Implementing Ubiquitous Learning in Health 2.0. *Revista IEEE América Latina*, v. 12, n. 8, p. 1507–1514, 2014.

MERRIAM, S. B.; CAFFARELLA, R. S.; BAUMGARTNER, L. M. *Learning in Adulthood: A Comprehensive Guide*. 3 edition ed. San Francisco: Jossey-Bass, 2006.

MOHAMMED, S.; FIAIDHI, J. *Ubiquitous Health and Medical Informatics: The Ubiquity 2.0 Trend and Beyond*. First edition ed. Hershey, PA: Medical Information Science Reference, 2010.

MOREIRA, J. D. C. *et al.* Um sistema de enriquecimento semântico de perfil de usuário baseado em traços digitais para apoio à aprendizagem informal no contexto da saúde. *Revista Novas Tecnologias na Educação*, n. 12, 2014.

NGUYEN, M. T.; FUHRER, P.; PASQUIER-ROCHA, J. Enhancing E-health information systems with agent technology. *Int. J. Telemedicine Appl.*, v. 2009, p. 1:1–1:13, jan. 2009.

O'REILLY, T. *What is Web 2.0*. 1 edition ed. [S.l.]: O'Reilly Media, 2009.

PARK, J. *et al.* Health-connect: An ontology-based model-driven information integration framework and its application to integrating clinical databases. In: 2012 IEEE 13TH INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION (IRI), 2012, Las Vegas, USA. *Anais*... Las Vegas, USA: IEEE, 2012. p. 393–400.

PAZZANI, M.; MURAMATSU, J.; BILLSUS, D. Syskill & Webert: Identifying Interesting Web Sites. AAAI'96, 1996, Portland, Oregon. *Anais*... Portland, Oregon: AAAI Press, 1996. p. 54–61.

POHOREC, S. *et al.* Natural language processing resources: Using semantic web technologies. In: PROCEEDINGS OF THE ITI 2012 34TH INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY INTERFACES (ITI), jun. 2012, Cavtat / Dubrovnik, Croatia. *Anais*... Cavtat / Dubrovnik, Croatia: IEEE, jun. 2012. p. 397–402.

POWERS, S. *Practical RDF*. 1st edition ed. Beijing ; Sebastopol: O'Reilly Media, 2003.

REDECKER, C.; PUNIE, Y. Learning 2.0 Promoting Innovation in Formal Education and Training in Europe. In: PROCEEDINGS OF THE 5TH EUROPEAN CONFERENCE ON TECHNOLOGY ENHANCED LEARNING CONFERENCE ON SUSTAINING FROM

INNOVATION TO LEARNING AND PRACTICE, 2010, Berlin, Heidelberg. *Anais*... Berlin, Heidelberg: Springer-Verlag, 2010. p. 308–323.

REFORMAT, M.; GOLMOHAMMADI, S. K. Updating user profile using ontology-based semantic similarity. In: IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, 2009. FUZZ-IEEE 2009, 2009, Jeju Island, Korea. *Anais*... Jeju Island, Korea: IEEE, 2009. p. 1062–1067.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3 edition ed. Upper Saddle River: Prentice Hall, 2009.

SHEN, X.; TAN, B.; ZHAI, C. *UCAIR: Capturing and Exploiting Context for Personalized Search*. Lisbon, Portugal: ACM New York, 2005.

SHETH, A.; ARPINAR, I. B.; KASHYAP, V. Relationships at the heart of semantic web: Modeling, discovering, and exploiting complex semantic relationships. 2003, [S.l.]: Springer-Verlag, 2003. p. 63–94.

SIDOROV, G. *et al.* Syntactic Dependency-Based N-grams as Classification Features. In: BATYRSHIN, I.; MENDOZA, M. G. (Org.). . *Advances in Computational Intelligence*. Lecture Notes in Computer Science. [S.l.]: Springer Berlin Heidelberg, 2013. p. 1–11.

STAN, J. *et al.* A User Profile Ontology For Situation-Aware Social Networking. *3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence*, jul. 2008. Disponível em: <http://liris.cnrs.fr/publis/?id=3502>. Acesso em: 18 jun. 2014.

TEXTALYTICS. *Language Identification API*. Disponível em: <http://textalytics.com/core/lang-info>. Acesso em: 15 jul. 2014.

THAKKER, D. *et al.* Taming Digital Traces for Informal Learning: A Semantic-driven Approach. 2012, Berlin, Heidelberg. *Anais*... Berlin, Heidelberg: Springer-Verlag, 2012. p. 348–362.

TOLEDO, S. S. DE. Proposta de Personal Health Record (PHR) para o NUTES: um sistema de informações sobre saúde voltado ao projeto. 27 ago. 2013. Disponível em: <http://dspace.bc.uepb.edu.br:8080/xmlui/handle/123456789/1733>. Acesso em: 19 nov. 2014.

VIEIRA, F. J. R.; NUNES, M. A. S. N. DICA: Sistema de Recomendação de Objetos de Aprendizagem Baseado em Conteúdo. *Scientia Plena*, v. 8, n. 5, 6 ago. 2012.

W3C. *World Wide Web Consortium*. Disponível em: <http://www.w3.org/Consortium/>. Acesso em: 16 abr. 2014.

WANG, M.; SHEN, R. Message design for mobile learning: Learning theories, human cognition and design principles. *British Journal of Educational Technology*, v. 43, n. 4, p. 561–575, 1 jul. 2012.

WHO. *Health Promotion Glossary*. . Geneva, Switzerland: World Health Organization, 2008. Disponível em: <http://www.who.int/healthpromotion/about/HPR%20Glossary%201998.pdf?ua=1>.

ZAPATER, J. J. S.; MENDES NETO, F. M. *Uso de tecnologías semánticas en diferentes dominios de aplicación: Entorno educativo y sistemas de información de tráfico vial*. Saarbrücken: Editorial Académica Española, 2014.

APÊNDICE A – FERRAMENTAS DE PLN PARA A LÍNGUA PORTUGUESA

Dentre as ferramentas de PLN analisadas detectou-se que nenhuma apresentava uma eficiência desejada quando utilizada para o processamento da língua portuguesa (Brasil). Os resultados gerados pelo processamento das mesmas, em grande parte das execuções, traziam resultados não condizentes com os dados esperados pela aplicação. As que mais se aproximaram de obter um resultado satisfatório, apesar de serem ferramentas livres, funcionavam em forma de serviço e possuíam cotas que podiam ser utilizadas para testes, porém para demandas maiores era necessário o pagamento pelo serviço. Abaixo são detalhadas as principais ferramentas testadas.

1. AlchemyApi

É uma ferramenta que se baseia no aprendizado de máquina para efetuar o PLN. Em sua especificação foi visto que a ferramenta deveria ser capaz de extrair dados semânticos do contexto, bem como informações sobre pessoas, lugares, companhias, tópicos, relações, etc.

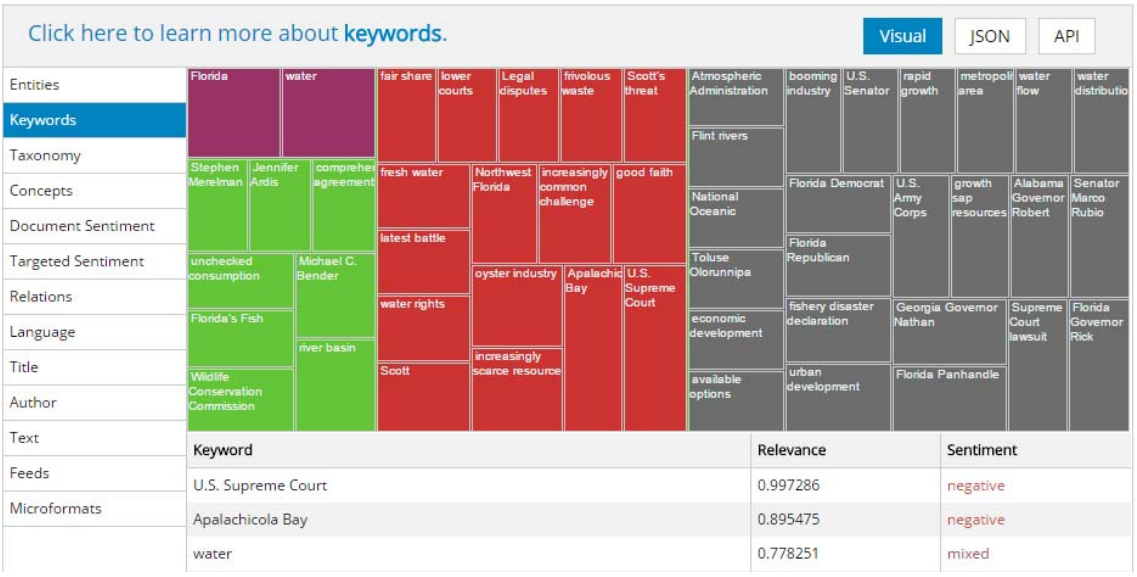


Figura 29 - Resultado do processamento de texto em inglês do AlchemyAPI.

Na prática os resultados foram diferentes do esperado. Para o processamento em inglês a ferramenta se mostrou bem mais eficaz do que para o processamento em português. Para cada seção (*Entities, Keywords, Taxonomy*, etc), a ferramenta conseguiu capturar do texto uma grande quantidade de informações resultantes do processamento da linguagem. A Figura 29 exibe o resultado do processamento do texto *"Florida to Sue Georgia in U.S. Supreme*

Court Over Water". Em cada seção da ferramenta foram obtidas inúmeras informações e classificações do texto.

Para o português a ferramenta trouxe resultados bem mais "simples" e indesejados. A Figura 30 exibe o resultado para o texto "Comunidade Ubuntu Brasil". Como pode ser percebido, as informações são mais restritas e por vezes ineficiente. Outro fator é que a ferramenta não é gratuita e trabalha com a ideia de créditos, onde cada funcionalidade consome certa quantidade de créditos.

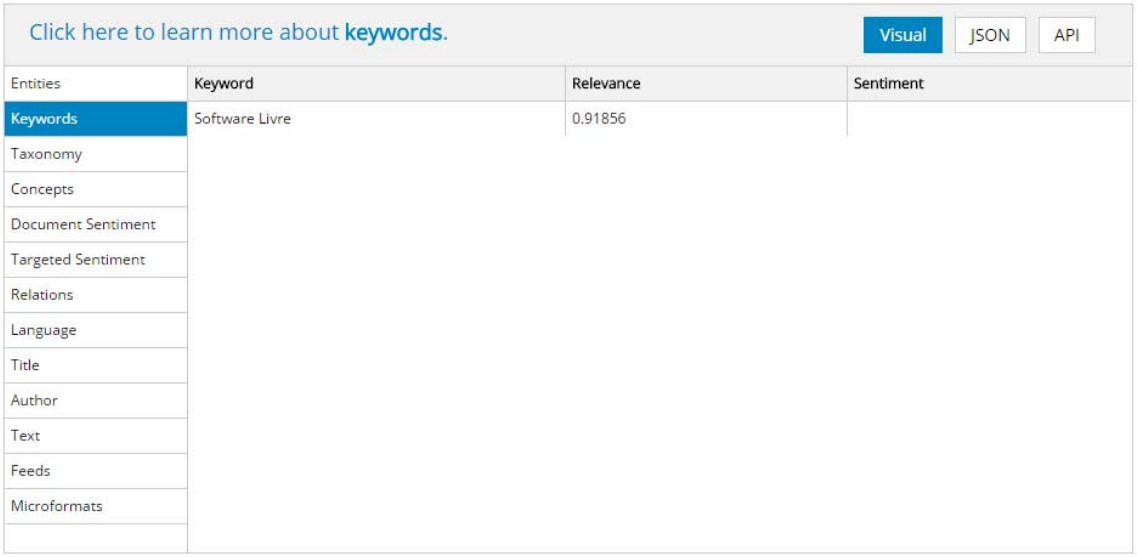


Figura 30 - Resultado do processamento de texto em português do AlchemyAPI.

2. TextalyticsAPI

Essa ferramenta tem muitas semelhanças com a AlchemyApi no quesito funcionalidades. A mesma se propõe a trazer classificação de texto, extração de tópicos, identificação da linguagem, lematização (*pos and parsing*)[6], etc.

Os resultados dos testes não foram tão bem quanto esperados. Para alguns casos, palavras como a palavra "como" foram classificadas como entidade, e várias palavras que eram somente simples substantivos foram marcadas como conceitos. A Figura 31 exibe o resultado da execução de um texto em português.

[6] Técnica usada por buscadores de palavras para abranger a quantidade de opções de palavras relacionadas. Para isto, reduz a palavra a sua forma raiz desconsiderando o tempo verbal, gênero, plural e etc.

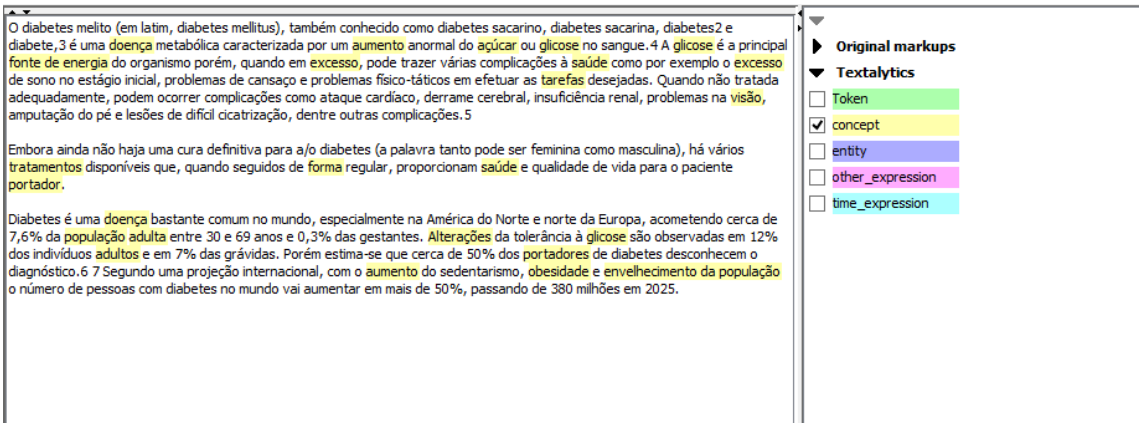


Figura 31 - Resultado do processamento de texto em português da TextalyticsAPI.

Na análise das palavras, se faz necessário que para uma dada palavra, por exemplo “originalmente”, o processamento pudesse identificar o radical dessa palavra que seria “original”, porém a ferramenta também falha. Para várias palavras variantes dos verbos isso funciona, porém não para todas as palavras que são necessárias para a aplicação. Vale salientar que esta ferramenta também não é gratuita, trabalhando com a ideia de créditos gastos por funcionalidade usada, portanto isso geraria uma dependência da aplicação junto a essa ferramenta.

3. Tree-Tagger

Essa ferramenta é uma ferramenta de PLN que trabalha com o conceito tido como *“part-of-speech”* desenvolvido por Helmut Schmid no projeto TC do instituto para computação linguística da universidade de Stuttgart. É necessário criar o modelo de treinamento da língua. Por exemplo, para o português deve-se criar o modelo para tratar a linguagem, podendo-se usar mais de um modelo. Porém não é viável despender tempo na criação e treinamento dos modelos para o processamento do português, motivo pelo qual a ferramenta não foi utilizada.

APÊNDICE B – CATEGORIAS IDENTIFICADAS PELO GATE

Abaixo estão listadas as categorias gramaticais, para o idioma inglês, identificadas pelo GATE durante o PLN.

Ident.	Descrição
CC	coordinating conjunction: ‘and’, ‘but’, ‘nor’, ‘or’, ‘yet’, plus, minus, less, times (multiplication), over (division). Also ‘for’ (because) and ‘so’ (i.e., ‘so that’).
CD	cardinal number
DT	determiner: Articles including ‘a’, ‘an’, ‘every’, ‘no’, ‘the’, ‘another’, ‘any’, ‘some’, ‘those’.
EX	existential ‘there’: Unstressed ‘there’ that triggers inversion of the inflected verb and the logical subject; ‘There was a party in progress’.
FW	foreign word
IN	preposition or subordinating conjunction
JJ	adjective: Hyphenated compounds that are used as modifiers; happy-go lucky
JJR	adjective - comparative: Adjectives with the comparative ending ‘-er’ and a comparative meaning. Sometimes ‘more’ and ‘less’.
JJS	adjective - superlative: Adjectives with the superlative ending ‘-est’ (and ‘worst’). Sometimes ‘most’ and ‘least’.
JJSS	nknown , but probably a variant of JJS
LRB	unknown
LS	list item marker: Numbers and letters used as identifiers of items in a list.
MD	modal: All verbs that don’t take an ‘-s’ ending in the third person singular present: ‘can’, ‘could’, ‘dare’, ‘may’, ‘might’, ‘must’, ‘ought’, ‘shall’, ‘should’, ‘will’, ‘would’.
NN	noun - singular or mass
NNP	proper noun - singular: All words in names usually are capitalized but titles might not be.
NNPS	proper noun - plural: All words in names usually are capitalized but titles might not be.
NNS	noun plural

NP	proper noun singular
NPS	proper noun plural
PDT	predeterminer: Determiner like elements preceding an article or possessive pronoun; ‘all/PDT his marbles’, ‘quite/PDT a mess’.
POS	possessive ending: Nouns ending in ‘’s’ or ‘’’.
PP	personal pronoun
PRP	unknown, but probably possessive pronoun,such as ‘my’, ‘your’, ‘his’, ‘his’, ‘its’, ‘one’s’, ‘our’, and ‘their’.
RB	adverb: most words ending in ‘-ly’. Also ‘quite’, ‘too’, ‘very’, ‘enough’, ‘indeed’, ‘not’, ‘-n’t’, and ‘never’.
RBR	adverb comparative: adverbs ending with ‘-er’ with a comparative meaning.
RBS	adverb superlative
RP	particle: Mostly monosyllabic words that also double as directional adverbs.
STAART	start state marker (used internally)
SYM	symbol: technical symbols or expressions that aren’t English words.
TO	literal “to”
UH	interjection: Such as ‘my’, ‘oh’, ‘please’, ‘uh’, ‘well’, ‘yes’.
VBD	verb - past tense: includes conditional form of the verb ‘to be’; ‘If I were/VBD rich...’.
VBG	verb - gerund or present participle
VBN	verb - past participle
VBP	verb - non
VB	verb - base form: subsumes imperatives, infinitives and subjunctives.
VBZ	verb - 3rd person singular present
WDT	‘wh’ determiner
WP$	possessive ‘wh’ pronoun: includes ‘whose’
WP	‘wh’ pronoun: includes ‘what’, ‘who’, and ‘whom’.
WRB	‘wh’
::	literal colon
,	literal comma
$	literal dollar sign
“	literal double quotes

´	literal grave
(	literal left parenthesis
.	literal period
#	literal pound sign
)	literal right parenthesis
’	literal single quote or apostrophe