

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

ELOISA TEIXEIRA MEDRADO

**Definição de grafos de conhecimento com
informações pessoais para completar discurso de
pessoas com distúrbios fonológicos**

Ribeirão Preto–SP

2018

ELOISA TEIXEIRA MEDRADO

Definição de grafos de conhecimento com informações pessoais para completar discurso de pessoas com distúrbios fonológicos

Versão Original

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Alessandra Alaniz Macedo

Ribeirão Preto–SP

2018

Resumo

A fala pode ser considerada o principal meio de comunicação do ser humano. Portanto, debilitações da fala podem ocasionar problemas de comunicação e, consequentemente, problemas nas relações interpessoais. O desenvolvimento da linguagem está relacionado com a aquisição de quatro componentes linguísticos: aquisição **lexical**, aquisição **morfossintática**, aquisição **fonológica** e aquisição **pragmática**. Somente os três primeiros componentes linguísticos serão tratados no presente trabalho. O objetivo do trabalho proposto é construir mecanismos para a definição de grafos de conhecimento pessoais centrados em diálogos conversacionais. Esses grafos serão utilizados por aplicativos personalizados que terão por objetivo auxiliar pessoas com problemas de fala a completar seus diálogos.

Os grafos construídos representarão as informações de pessoas com deficiência de fala, de modo a permitir a busca por informação correlata para a definição de relacionamentos entre o som emitido e o som pretendido na produção da fala. Os Grafos de Conhecimento, ou Knowledge Graph, fornecem uma representação poderosa do conhecimento, porém a construção automática desses grafos, a partir da linguagem falada possui inúmeros desafios.

Palavras-chave: construção de grafo de conhecimento, grafo de conhecimento pessoal, pesquisa de similaridade, predição de palavras, processamento de linguagem natural, reconhecimento de fala.

Abstract



Speech can be considered the main means of communication of the human being. Therefore, speech impairments can cause communication problems and, consequently, problems in interpersonal relationships. Language development is related to the acquisition of four language components, **lexical** acquisition, **morphosyntactic** acquisition, **phonological** acquisition, and pragmatic acquisition. Only the first three language components will be addressed in this paper. The objective of the proposed work is to build mechanisms for the definition of personal knowledge graphs centered on conversational dialogues. These graphs will be used by personalized digital assistants to help speech-impaired people complete their dialogues. That is, the KGs that will be built will represent the information of people with speech disabilities, in order to allow the search for related information to define relationships between the sound emitted and the intended sound in speech production. Knowledge Graphs provide a powerful representation of knowledge, but the automatic construction of these graphs from spoken language has numerous challenges.

Keywords: knowledge graph construction, personal knowledge graph, similarity research, word prediction, natural language processing, speech recognition.

Lista de figuras

Figura 1 – Exemplo de um grafo pessoal.	22
Figura 2 – Um grafo RDF com dois nós e uma tripla (Sujeito, Predicado e Objeto).	23
Figura 3 – Parte de um documento escrito em OWL.	24
Figura 4 – Parte de um discurso.	30
Figura 5 – Uma visão de várias camadas de uma arquitetura conceitual para sistemas de informação em saúde.	36

Lista de tabelas

Tabela 1 – Resultados Estudo de Caso 31

Tabela 2 – Cronograma de execução do plano de trabalho. 42

Lista de abreviaturas e siglas

AAC	Comunicação Alternativa e Aumentativa
GC	Grafo de Conhecimento
KG	Knowledge Graph
NLP	Processamento de Linguagem Natural
OWL	Web Ontology Language
PLN	Processamento de Linguagem Natural
PRIT	Processamento e Recuperação de Informação Textual para Computação Aplicada
RDF	Resource Description Framework

Sumário

1	INTRODUÇÃO	17
1.1	Objetivos	19
1.2	Contribuições	19
1.3	Organização do Documento	20
2	REFERENCIAL TEÓRICO	21
2.1	Grafos de Conhecimento	21
2.2	Web Semântica	22
2.2.1	RDF	23
2.2.2	OWL	23
2.3	Processamento de Linguagem Natural	25
2.4	Processamento Textual	25
2.5	Produtos Terminológicos	26
2.6	Trabalhos Relacionados	27
3	ESTUDO DE CASO	29
3.1	Dataset	29
3.2	Pré-processamento	30
3.3	Correção de palavras	30
3.4	Predição de palavras	31
3.5	Resultados e Considerações	31
4	PROPOSTA	33
4.1	Metodologia	33
4.2	Aquisição e Preparação dos Dados	35
4.3	Combinação e Integração de Conteúdo	37
4.4	Refinamento do Conhecimento	38
4.5	Atividades e Cronograma	39


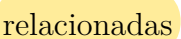
Referências	43
-----------------------	----

Introdução


A comunicação é uma das capacidades essenciais dos seres humanos para interagir e expressar sentimentos, pensamentos e necessidades. Indivíduos com habilidades eficazes de comunicação aumentam suas chances de sucesso tanto pessoalmente quanto profissionalmente, além disso possuem mais qualidade de vida (CAJAS-SALAZAR et al., 2003). Por outro lado, problemas de fala podem ocasionar em problemas de comunicação e, consequentemente, problemas nas relações interpessoais (Passadori, 2015).

As habilidades de comunicação podem ser afetadas devido a eventos de trauma, acidente vascular cerebral ou acidente cardiovascular, doenças neurodegenerativas como Parkinson, Alzheimer ou Esclerose Lateral Amiotrófica, ou mesmo devido a distúrbios genéticos, nos casos de Síndrome de Down e Autismo. Pessoas com distúrbios de linguagem como afasias sofrem preconceitos e apresentam enormes desafios para serem integradas na sociedade e para realizarem suas atividades de vida diária (PONTE; FEDOSSE, 2016). Esses problemas tornam objetivos básicos da vida muito mais complicados de serem alcançados.


A comunicação pode ocorrer com o uso de linguagem não verbal (gestos, expressões faciais e olhares) e/ou verbal, que acontece com a utilização das regras linguísticas relacionadas à forma, conteúdo e função (níveis semântico, morfossintático, fonético e fonológico, e pragmático). O desenvolvimento da linguagem está relacionado com a aquisição de quatro componentes linguísticos:

- Aquisição **lexical** é composta de palavras, ou seja, o vocabulário. Essas palavras são classificadas  em categorias semânticas (significado de substantivos, verbos e adjetivos) e gramaticais;
- Habilidades **morfossintáticas**  **relacionadas** com a elaboração de enunciados, as quais exigem o estabelecimento de relações entre palavras em uma frase, bem como o uso da inflexão nominal e verbal (NICOLIELO-CARRILHO et al., 2018);

- Aspecto **fonológico** relacionado com a capacidade de dominar as regras fonológicas de produção de fonemas, distribuição de fonemas, produção fonética e uso de fonemas em diferentes posições de palavras, bem como em diferentes contextos fonológicos e estruturas silábicas (VICK et al., 2012);
- Habilidades **pragmáticas** de linguagem funcional ~~para~~ análise de intenção e consequência da comunicação subjetiva. A ordem desses componentes linguísticos não deve ser aleatória, pois, por exemplo, a habilidade pragmática é empregada adequadamente quando os outros aspectos linguísticos (isto é, lexicais, morfossintáticos e fonológicos) foram dominados (Gianecchini; Maximino, 2018).

Durante as últimas três décadas foram criadas várias técnicas e estratégias de Comunicação Alternativa e Aumentativa, do inglês Augmented Alternative Communication (AAC)  porém, as tecnologias AAC apresentam desafios para sua adequação a pessoas com distúrbios de linguagem e fala. Normalmente, essas tecnologias são caras e genéricas, dificultando a obtenção e a personalização de tratamentos. Além disso, elas não evoluem à medida que ocorrem mudanças nas vidas e vocabulários das pessoas, não permitem comunicação em tempo real e, na maioria das vezes, elas são propostas apenas para fins terapêuticos.

Existem diferentes **abordagens** computacionais para tecnologias AAC, como por exemplo a utilização do Aprendizado de Máquina, das Redes Neurais e dos Dicionários Linguísticos. A fim de fazer uma prova de conceito com algumas abordagens, foi feito um Estudo de Caso na disciplina Processamento e Recuperação de Informação Textual para Computação Aplicada do programa PPG-CA. Nesse estudo de caso foram utilizadas as seguintes ferramentas: biblioteca SpellChecker¹ do Python para verificação e correção de ortografia. Dicionário WordNet² para confirmação da existência e ortografia de palavras. E por fim foram utilizados o modelo pré-treinado BERT e o algoritmo do Word2Vec para predição de alguns palavras.


Os resultados do Estudo de Caso são mostrados no Capítulo 2. Após o estudo dos resultados foi possível perceber diferentes pontos fracos das abordagens utilizadas, alguns deles já citados anteriormente, como por exemplo, são computacionalmente dispendiosas, dificuldade na personalização de tratamentos, e elas não evoluem à medida que ocorrem mudanças nas vidas e vocabulários das pessoas. Por isso, a proposta desse trabalho de mestrado é partir para outra abordagem, a fim de alcançar a personalização das tecnologias de AAC por meio da criação de técnicas e métodos para geração de grafos pessoais de conhecimento aliados a produtos terminológicos adequados. 

¹ <https://pypi.org/project/pyspellchecker/>

² <https://wordnet.princeton.edu/>

Para personalização das tecnologias de AAC, é imprescindível a manipulação de fontes de informação pessoais, que são comumente segmentadas em várias fontes distribuídas e heterogêneas de informação e armazenadas como pequenos e heterogêneos objetos de dados. É necessário que essas informações sejam manipuladas e integradas para fornecimentos de entradas linguísticas que podem ser úteis na comunicação de pessoas com distúrbios de linguagem e fala. As entradas linguísticas podem ser organizadas facilitando a busca por termos, uma vez que na apraxia verbal, informação similar pode complementar contextos na comunicação.

1.1 Objetivos

O **objetivo** do presente trabalho é desenvolver uma investigação teórico-prática para criação de técnicas e métodos para geração de grafos pessoais de conhecimento, a partir de informações dos usuários dispostas em diferentes fontes de informação, aliadas a produtos terminológicos adequados. O grafo pessoal será utilizado no sistema SICI a ser desenvolvido posteriormente no decorrer do projeto da orientadora. 

O SICI (Sistema de Informação de apoio a Comunicação Inteligível por fala) terá por objetivo apoiar a recuperação, o tratamento e a comunicação de pessoas diagnosticadas com afasia (um distúrbio de linguagem causado por danos neurológicos no cérebro que afetam os processos de compreensão e expressão da linguagem). Ele incluirá aspectos fonológicos, a partir da consulta do grafo em buscas de informações correlatas para complementação de conteúdo na produção de discursos.

Para o desenvolvimento da proposta, mecanismos de Extração da Informação irão localizar e extrair informações relevantes expressas em linguagem natural, de maneira automática, e converter as informações extraídas em estruturas morfossintáticas que facilitem a manipulação e a análise das mesmas. A Mineração de Texto, composta por conhecimentos de áreas como Extração de Informação, Processamento de Linguagem Natural, Aprendizado de Máquina e Linguística Computacional (FELDMAN et al., 2007), promoverá a análise e a busca por informações relevantes não-estruturadas ou estruturadas.



1.2 Contribuições

A associação de informações a partir de diferentes fontes poderá contribuir com a geração de conhecimentos que potencialmente deverão ser utilizados para a promoção da saúde e do bem estar de pacientes. Até o momento, poucos trabalhos relacionados foram encontrados.

Um limitante esperado é o foco nos componentes linguísticos de aquisição lexical,

de habilidades morfosintática e da tentativa de sintetização do discurso. O componente pragmático está muito relacionado a habilidades cognitivas e fogem do alcance desse trabalho.

1.3 Organização do Documento


O documento está organizado da seguinte forma: o Capítulo 2 detalha a fundamentação teórica que norteia o desenvolvimento do projeto. O Capítulo 3 resume o Estudo de Caso desenvolvido e suas considerações. Já o Capítulo 4 apresenta a proposta do trabalho, bem como a metodologia de desenvolvimento, as atividades que serão realizadas e o cronograma a ser seguido.

Referencial Teórico

Para a construção do grafo de conhecimento com informações pessoais será necessário uma fusão de informações pessoais e isso irá demandar algumas atividades como a manipulação de ~~seus~~ dados e informações, a partir de diferentes fontes de informações. Essas atividades serão apoiadas por tecnologias como Processamento de Linguagem Natural, Processamento Textual e da Web Semântica.

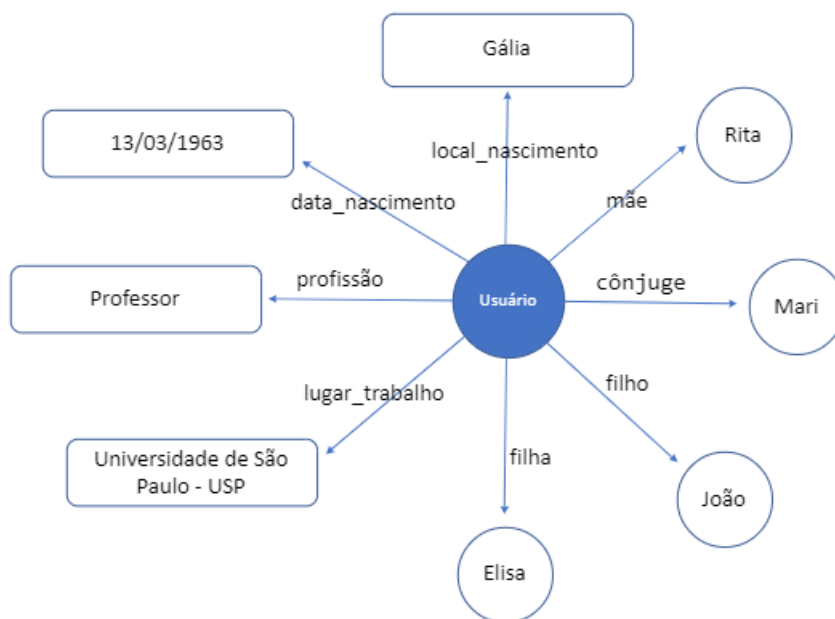
Os Grafos de Conhecimento ~~serão~~ representados utilizando RDF, uma tecnologia da Web Semântica, assim caso necessários eles poderão ser agregados a outros grafos. ~~Desse modo,~~ os fundamentos teóricos sobre Grafos de Conhecimento, Processamento Textual, Processamento de Linguagem Natural, RDF e OWL são imprescindíveis para o entendimento da proposta desse projeto e são apresentados ~~nesta seção~~.

2.1 Grafos de Conhecimento

Um Grafo de Conhecimento é basicamente uma Base de Conhecimento na estrutura de um grafo. O GC é uma tecnologia usada para armazenar informações estruturadas e não estruturadas entendíveis por um sistema de computador (FENSEL, 2001). Ele pode ser considerado um conjunto de dados estruturados, normalizado e conectado. Como exemplo de Grafos de Conhecimento temos o DBpedia (LEHMANN et al., 2015) e o Freebase (BOLLACKER et al., 2008). 

Os objetos que formam um grafo são chamados de nós ou vértices e o relacionamento entre os objetos é chamado de arestas. No caso das bases de dados e ontologias, é possível pensar nos nós como substantivos e nos seus relacionamentos como verbos, dando assim, semântica ao grafo (NEEDHAM, 2019). Assim, um grafo $G = (V, E)$ é uma estrutura de dados composta por um conjunto finito de nós, V e um conjunto de arestas E . O tamanho (ou ordem) de um grafo G é definido como o número de nós em G .

Figura 1 – Exemplo de um grafo pessoal.



Fonte: Eloisa Medrado, 2019

Na Figura 1, é apresentado um exemplo de Grafo de Conhecimento centrado no usuário, ele segue o esquema do GC semântico do Freebase ¹. ~~Esse exemplo será seguido para o desenvolvimento do trabalho proposto.~~

O conceito de Grafo de Conhecimento foi muito difundido a partir do surgimento da Web Semântica, que trouxe também diversas ferramentas para o mapeamento e representação dos CGs. Na próxima Seção o conceito de Web Semântica é introduzido, bem como as ferramentas que serão utilizadas para o mapeamento e representação dos grafos criados.



2.2 Web Semântica

~~Atualmente, a World Wide Web (Web) ainda em sua maior parte é formada por documentos e informações não estruturadas e elas são disponibilizadas principalmente para as pessoas.~~

~~O que se pretende com a Web Semântica é proporcionar o entendimento da informação também pelas máquinas, por meio do uso de linguagens de marcação similares às utilizadas atualmente.~~ Com o desenvolvimento da Web Semântica, ~~seria~~ possível, então, que as máquinas se comunicassem de forma universal, sendo a ação do usuário cada vez menor e mais natural. Nas próximas subseções serão descritas algumas ferramentas da Web Semântica que se fizeram necessárias para a desenvolvimento e fortalecimento da mesma.

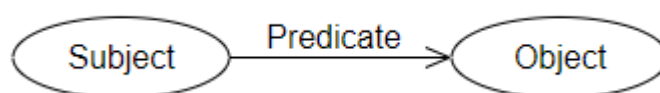
¹ <https://developers.google.com/freebase/>

Por exemplo, o RDF se tornou padrão para o mapeamento de dados e ontologias. As ontologias, por sua vez, passaram a ser representadas pelos conceitos da Web Ontology Language (OWL), que em sua essência funciona da mesma forma que o RDF, porém com um poder representativo um pouco maior. Outra tecnologia essencial para a Web Semântica ~~que vem ganhando cada vez mais espaço~~ é a linguagem de consulta SPARQL, que permite a realização de buscas em arquivos RDF.

2.2.1 RDF

O *Resource Description Framework* (RDF) é uma estrutura utilizada para representar dados na Web Semântica. Como uma ferramenta de representação de conhecimento, é necessário que o RDF seja capaz de conectar conceitos. Para isso, o **framework** foi estruturado a partir da ideia de triplas (Consortium et al., 2014). ~~Como na formação mais básica de uma oração em linguagem natural, as triplas são compostas de um sujeito, aquele que toma ação, um predicado, que representa a ação, e um objeto, o alvo da ação (Consortium et al., 2014)~~

Figura 2 – Um grafo RDF com dois nós e uma tripla (Sujeito, Predicado e Objeto).



Fonte: Consortium et al. (2014)

Um conjunto de triplas é chamado de grafo RDF, o qual pode ser visualizado como um diagrama de nós direcionados, onde cada trio é representado como um link nó-arco-nó. Na Figura 2, é possível observar que os nós podem ser ora sujeito, ora o objeto. E o arco (ou predicado), partem dos sujeitos e apontam para os objetos (Consortium et al., 2014).

Todos os documentos RDF já são também OWL (Linguagem de Ontologia da Web) Full, conceito explicado na próxima subseção. Portanto os GC pessoais gerados serão representados de ambas as formas, ou seja, serão mapeados utilizando RDF e representados pela linguagem OWL.

2.2.2 OWL

A Linguagem de Ontologia da Web (do inglês, *Ontology Web Language* OWL) é uma linguagem da Web Semântica projetada para definir e instanciar conhecimento rico e complexo de ontologias e KGs. Ela é uma linguagem baseada em lógica computacional, ou seja, conhecimento representado em OWL pode ser explorado por programas de

computador. Normalmente as ontologias são expressas nessa linguagem (Consortium et al., 2009).

Bases de conhecimento escritas com a linguagem OWL podem, também, ser comparadas aos grafos RDF. Ao passo que todos os documentos RDF são também documentos OWL Full. Como a sintaxe OWL é criada sobre as tecnologias da Web Semântica, alguns de seus atributos básicos são herdados do RDF Schema. Dentre eles estão:

- *Class*: atributo responsável por agrupar indivíduos com características semelhantes;
- *rdfs:subClassOf*: responsável por definir a hierarquia das relações das classes;
- *rdf:Property*: utilizada para a declaração de relacionamentos entre os indivíduos;
- *rdfs:subPropertyOf*: adiciona uma propriedade a um indivíduo de acordo com a existência de outra propriedade (*rdf:Property*);
- *rdfs:domain*: delimita que o indivíduo da propriedade relacionada a este atributo seja, necessariamente, do tipo por ele definido;
- *rdfs:range*: atributo responsável por definir a classe um indivíduo relacionado a uma propriedade;
- *Individuals*: é um membro, ou seja, representação de um indivíduo do grupo definido por aquela classe.


Na Figura 3 é possível observar um exemplo de parte de um documento escrito com a linguagem OWL. 

Figura 3 – Parte de um documento escrito em OWL.

```
<owl:Class rdf:ID="WineDescriptor" />

<owl:Class rdf:ID="WineColor">
  <rdfs:subClassOf rdf:resource="#WineDescriptor" />
  ...
</owl:Class>

<owl:ObjectProperty rdf:ID="hasWineDescriptor">
  <rdfs:domain rdf:resource="#Wine" />
  <rdfs:range rdf:resource="#WineDescriptor" />
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="hasColor">
  <rdfs:subPropertyOf rdf:resource="#hasWineDescriptor" />
  <rdfs:range rdf:resource="#WineColor" />
  ...
</owl:ObjectProperty>
```

Fonte: Consortium et al. (2009)

2.3 Processamento de Linguagem Natural


Processamento de Linguagem Natural (do inglês, *Natural Language Processing* - NLP) possui um sentido amplo que abrange qualquer tipo de manipulação computacional da linguagem natural. Por linguagem natural, entende-se uma linguagem humana que é utilizada para comunicação cotidiana entre pessoas (S.Bird; K.Ewan; L.Edward, 2009).


Em outras palavras a NLP é formada pelo desenvolvimento de modelos computacionais para a realização de atividades que dependem de informações apresentadas em alguma língua natural. Conforme Covington, Nute e Vellino (1996), a pesquisa em NLP está voltada, essencialmente, a três aspectos da comunicação em língua natural: (1) som: fonologia; (2) estrutura: morfologia e sintaxe e; (3) significado: semântica e pragmática.

A NLP é uma subárea da Ciência da Computação, da Inteligência Artificial e da Linguística. Ela pode assumir extremos de complexidade: ela pode ser tão simples quanto contar frequência de palavras em determinado texto para contrastar tipos de escrita; e tão complexa a ponto de envolver "compreensão" de expressões humanas e de ser capaz de dar respostas úteis a elas (S.Bird; K.Ewan; L.Edward, 2009).

Para a criação dos GCs pessoais, mecanismos de extração de informação irão localizar e extrair informações relevantes expressas em linguagem natural e converter essas informações em estruturas morfossintáticas que facilitem a manipulação e a análise das mesmas ~~através~~ do processamento textual.

2.4 Processamento Textual

O processamento de texto é uma tarefa bastante comum nas ferramentas atuais. Transformar texto em algo que um algoritmo pode processar é uma tarefa complicada (NABI, 2018). 

A primeira etapa do processamento textual consiste na seleção e **recuperação de documentos** ou textos. A segunda etapa para uma boa aplicação de processamento textual é o **pré-processamento** dos dados, ~~onde~~ são feitas atividades como remoção de caracteres especiais e pontuação, quebra do texto em palavras ou tokenização, remoção de termos comuns do documento ou *stop words*, redução de palavras ao seu radical conhecida como *stemming* e lematização (NABI, 2018). Essa etapa consiste na limpeza, tratamento e padronização dos dados, além da redução dos atributos a serem processados nas próximas etapas (BAEZA-YATES; RIBEIRO-NETO, 1999). 

Já a terceira etapa resume-se na **extração de conhecimento**. A análise para extração de conhecimento em texto não estruturado pode ser feita em diferentes níveis,

como o sintático, o semântico e o morfológico (BAEZA-YATES; RIBEIRO-NETO, 1999). A análise sintática é a avaliação da função e relação de um termo aos demais termos da sentença. A análise semântica avalia o significado de elementos como as palavras, as frases e os seus relacionamentos. Já a análise morfológica avalia a gramática de cada palavras da frase isoladamente com abordagens de análises morfo/sintáticas (BAEZA-YATES; RIBEIRO-NETO, 1999). A quarta e última etapa do processamento textual é a **avaliação e interpretação dos resultados**.

2.5 Produtos Terminológicos

Segundo Sager (1990), a terminologia pode ser entendida como um conjunto de termos técnicos ou científicos. Algumas características das terminologias são:

1. Ciência essencialmente multidisciplinar, porém com objetivo de ordem linguística;
2. O que importa ao terminólogo é o significado em uso atual dos termos e o sistema de conceitos;
3. A forma escrita dos termos é o que interessa a essa ciência, uma vez que eles são internacionalmente unificados;
4. É formada a partir da estrutura da língua comum, nos níveis morfossintático e semântico.

Os Produtos Terminológicos tem o objetivo de organizar termos, conceitos e siglas, favorecendo a recuperação, acesso, divulgação e disseminação das informações. Eles podem apresentar-se na forma de:

- Dicionários terminológicos (monolíngues, bilíngues, multilíngues);
- Glossários, que são listas de termos técnicos de alguma especialidade, ordenadas alfabeticamente, providas de definições;
- Tesouros, que são listas de termos técnicos, de alguma especialidade, estruturadas como sistemas abertos de relações conceituais e designações.

As ontologias, por exemplo, podem ser vistas como produtos terminológicos à medida que são desenvolvidas para fornecer uma semântica processável por máquinas e que pode ser comunicada entre diferentes agentes (RUSSELL; NORVIG, 2009). Ontologias possibilitam um entendimento (ou definição) comum e compartilhado de determinados domínios e que podem ser comunicado sem ambiguidade entre pessoas e sistemas (FENSEL, 2001).



2.6 Trabalhos Relacionados

Em busca de trabalhos relacionados à proposta, esta proponente em 17 de junho realizou buscas por artigos científicos relacionados nas seguintes fontes de informação: Scopus ², ACM/DL ³ e IEEE Xplorer ⁴. Inicialmente, a **string** de busca foi formulada com as seguintes palavras-chave “personal knowledge graph” OR “user-centric graph” no intervalo de publicação dos últimos 5 anos. Na base Scopus, os seguintes trabalhos foram retornados: Sun e Zhang (2019) e Li et al. (2014).

Em (SUN; ZHANG, 2019), os autores propõem uma nova estrutura de grafos de conhecimento especificamente para combater problemas de ausência de conhecimento, ambiguidade, conflitos e conhecimento errôneo quando aplicados à manipulação de informações pessoais. Isso se deve ao fato de que essas informações têm propriedades exclusivas, como sua natureza volátil específica do usuário e disponibilidade de dados limitada. A contribuição é um método de construção de grafos conhecimento pessoal baseados em entrada do usuário.

Em (Li et al., 2014), os autores manipulam a construção de grafos de conhecimento a partir da linguagem **Freebase** usando uma linguagem estatística para sua formação. Segundo os autores **Freebase** essas informações têm o potencial de entender melhor as solicitações dos usuários, cumprindo-as e habilitando outras tecnologias, como o desenvolvimento de melhores inferências ou interações proativas. O conhecimento codificado em grafos semânticos como o Freebase demonstrou beneficiar a análise semântica e a interpretação de enunciados em linguagem natural. Assim, **como primeira etapa**, ~~exploramos~~ a relação factual pessoal tripla do Freebase para fragmentos de linguagem natural com um mecanismo de pesquisa e os **snippets** resultantes contendo pares de entidades relacionadas para criar os dados de treinamento.

Por outro lado, para IEEE Xplorer não foram retornados artigos relacionados à proposta. Para a ACM/DL, foi retornado (DENG et al., 2019). Esse trabalho trata da construção de grafos de conhecimento para a definição de um plano de educação baseada em segurança cibernética. Em uma análise do trabalho, esse artigo pode ser desconsiderado do conjunto de trabalhos retornados da pesquisa realizada para o projeto. Essa análise motivou alterações na **string** de busca.

Quando a *string* foi alterada para "user-centric knowledge graph"AND speech, apenas o segundo artigo foi retornado para Scopus. Para ACM/DL nada foi retornado

² scopus.com

³ dl.acm.org

⁴ <http://ieeexplore.ieee.org>

para a segunda string. No caso da IEEE Xplorer, os artigos retornados são: Li et al. (2014), Agarwal et al. (2017) e Shen et al. (2019).

Em Li et al. (2014), os autores apresentam uma abordagem da organização do conhecimento falado para palestras de curso para uma aprendizagem personalizada eficiente. A interconexão entre a estrutura semântica e a estrutura temporal, juntamente com a detecção de termos falada, oferece aos alunos formas eficientes de navegar pelo conhecimento do curso com caminhos de aprendizagem personalizados, considerando seus interesses pessoais, tempo disponível e conhecimento prévio. Um sistema protótipo preliminar também foi desenvolvido.

Em Agarwal et al. (2017), os autores propõem uma abordagem de domínio agnóstico que permite ao sistema endereçar consultas referentes ao passado usando uma abordagem de recuperação de informações para classificar várias entidades para uma determinada consulta. Esses autores, assim como a proposta deste trabalho, exploram enriquecimento semântico aumentando as entidades com informações de um grafo de conhecimento para ampliar as possibilidades de recuperação. Finalmente em (Shen et al., 2019), os autores desenvolveram o SliQA-I que é um gerador de perguntas iterativo humano a partir da linguagem falada para um sistema de Q&A do tipo assistente pessoal.



Estudo de caso


Como dito anteriormente, existem diversas abordagens computacionais para tecnologias AAC, como exemplo temos a utilização do Aprendizado de Máquina, das Redes Neurais e dos Dicionários Linguísticos. O estudo de caso descrito ~~nessa Seção~~ teve como objetivo ~~testar~~ e analisar algumas dessas abordagens. O estudo foi feito utilizando um conjunto de discursos retirados do trabalho de Howell, Davis e Bartrip (2009).

Nesse estudo de caso foram utilizadas as seguintes ferramentas: expressão regular, biblioteca SpellChecker do Python para verificação e correção de ortografia. Em seguida cada uma das palavras corrigidas foi buscada no dicionário WordNet¹ para confirmar sua existência e sua ortografia. Por fim foram utilizados o modelo pré-treinado BERT e o algoritmo do Word2Vec para predição das palavras faltantes palavras.

As ~~Seções~~ desse ~~Capítulo~~ descrevem as etapas realizadas no Estudo de Caso, na ordem em que foram feitas. A última ~~Seção~~ mostra os resultados alcançados e as considerações sobre as abordagens utilizadas e objetivo do trabalho de mestrado proposto.

3.1 Dataset

O dataset utilizado para ~~esse trabalho~~ foi retirado do trabalho de UCLASS2009, desenvolvido na University College London - UCL por vários anos. Em homenagem a universidade, o dataset recebeu o nome de *UCL Archive of Stuttered Speech* (UCLASS). O conjunto de dados possui duas versões, a versão utilizada nesse trabalho é a primeira.


O conjunto de dados da primeira versão possui ~~tanto~~ a versão falada (áudio) ~~quanto~~ escrita (transcrição) de discursos monólogos de pessoas entre cinco e quarenta e sete anos. As gravações são, principalmente, de crianças que foram encaminhadas para clínicas em Londres para avaliação da gagueira. As crianças eram incluídas no estudo caso o diagnóstico de gagueira fosse confirmado por um clínico especializado nesse distúrbio. 

¹ <https://wordnet.princeton.edu/>

Figura 4 – Parte de um discurso.


{T1} and yjeu went off. I seen um IN I IN it um they had animation and, um computer work. Um it was quite FU FUNNY. {T2} Tell me about the story. {T1} Um, it was about this um girl WHO WHO was a PRI PRINCESS but then got SSSSSSPLIT up with her FA FAMILY, and. And she was trying to um get BA CK to them. Um. {T2} How did she get split up from her family? {T1} Um someone had um u um she didn't quite make it to um the train. And she um FE FELL OFFFFFFFFF the train. Um and she went to ANNNNNNN ORPHANA. Um. Where she um LIVVVVVVED. Um. {T2} Then what happened? {T1} Then she was thrown out of the um ORFFFFFFFANA GE. Um. And she um had to MAKE UM MAKE do with her OWN OWN life.

Fonte: Howell, Davis e Bartrip (2009)


Esse conjunto de discursos possui temas gerais falados por diferentes pessoas com gagueira. Na Figura 4 é possível ver um exemplo de um discurso retirado do dataset, as fases da gagueira podem ser vistas onde o discurso está em caixa alta. As transcrições estão em formas ortográficas e fonéticas legíveis por máquina e utilizam uma ortografia regular, ou seja, os momentos de gagueira são transcritos à medida que são ouvidos.  CLASS2009.



3.2 Pré-processamento

A primeira etapa para realização do estudo foi o pré-processamento, ou seja, a limpeza dos textos dos discursos. No pré-processamento foi feita a retirada das falas dos interlocutores, das *tags* que foram utilizadas pelo grupo de pesquisa de Howell, Davis e Bartrip (2009) e também de diversos caracteres especiais e pontuações. Após a limpeza textual, as frases onde ocorre a gagueira foram encontradas e separadas dos discursos. 

3.3 Correção de palavras

A princípio as frases em que a pessoa apresenta a gagueira foram corrigidas por meio da aplicação de expressão regular para retirada das letras repetidas, depois foi utilizada a biblioteca SpellChecker² do Python para verificação e correção de ortografia.  n seguida cada uma das palavras corrigidas foi buscada no dicionário WordNet³ para confirmar sua existência e sua ortografia, as palavras que não foram encontradas no dicionário foram pesquisadas em uma lista de *stop words*, pois o WordNet não inclui *stop words*.



² <https://pypi.org/project/pyspellchecker/>

³ <https://wordnet.princeton.edu/>

3.4 Predição de palavras

A predição de palavras foi a etapa final do estudo de caso. Essa etapa foi realizada apenas para as frases cujas palavras não foram corrigidas ou encontradas nas etapas anteriores e para as palavras que já estavam incompletas no discurso original. Foram utilizados para a predição o modelo pré-treinado BERT e modelo Word2Vec.

O modelo do BERT está disponível na *web* já treinado, basta fazer a instalação da biblioteca PyTorch do Python⁴ e utilizá-lo. Já o modelo Word2Vec foi treinado com o corpus do ROCStories⁵ utilizando a arquitetura Skip-Gram-. O corpus do ROCStories é uma coleção de alta qualidade de histórias da vida cotidiana que formam relações de senso comum causal e temporal entre os eventos diários.

3.5 Resultados e Considerações

Os resultados de cada etapa do estudo podem ser vistos na Tabela 1 entre todas as 350 palavras em caixa alta, 275 delas foram corrigidas corretamente com a aplicação da expressão regular e do corretor ortográfico SpellChecker, ou seja, 78,51% delas. Das 350 palavras já corrigidas (corretamente ou não), 191 foram encontradas no dicionário WordNet, porém dessas 191 palavras somente em 134 casos, ou seja, 70,16%, foi encontrada a palavra correta.

Isso quer dizer que as palavras que estão no maior nível de complexidade são as que não foram corrigidas corretamente com a aplicação da expressão regular e do corretor ortográfico, e as que não foram encontradas corretamente no WordNet e nem na lista de *stop word*, que soma um total de 118 palavras, ou seja, 33,71%.

Tabela 1 – Resultados Estudo de Caso

	Corretas	Erradas	Total	Acertos
Correção ortográfica	275	75	350	78,51%
WordNet	134	57	191	70,15%
Predição BERT	11	107	118	9,32%
Predição Word2vec	19	99	118	16,10%

Fonte: Eloisa Medrado, 2019

Para cada uma das 118 palavras, foi extraída dos discursos a frase completa onde a palavra se encontrava. Dentre todas as 118 frases testadas no modelo BERT, somente 11

⁴ <https://pytorch.org/>

⁵ <https://www.cs.rochester.edu/nlp/rocstories/>

tiveram a predição correta, ou seja, 9,31%. Já no modelo treinado do Word2vec, somente 19 tiveram a predição correta, ou seja, 16,10%.

Esse estudo de caso teve como objetivo comprovar a necessidade de uma nova abordagem para as tecnologias AAC, visto que as tecnologias estudadas não alcançam de maneira descomplicada o principal objetivo desse trabalho de mestrado, que é a personalização das tecnologias de AAC.

Após o estudo dos resultados foi possível perceber diferentes pontos fracos das abordagens utilizadas, alguns deles já citados anteriormente, como por exemplo, são computacionalmente dispendiosas, dificuldade na personalização de tratamentos, e elas não evoluem à medida que ocorrem mudanças nas vidas e vocabulários das pessoas. Por isso, a proposta desse trabalho de mestrado é partir para outra abordagem, a fim de alcançar a personalização das tecnologias de AAC por meio da criação de técnicas e métodos para geração de grafos pessoais de conhecimento aliados a produtos terminológicos adequados.

Proposta

Neste projeto, propõe-se uma investigação teórica-prática do tipo exploratória para definição de grafos pessoais de conhecimento a partir de informações dispostas em diferentes fontes de informação dos usuários, apoiadas por produtos terminológicos. A seguir, a proposta é detalhada em perspectivas de pesquisa e desenvolvimento para o alcançar o objetivo citado.

Os Grafos de Conhecimento (GC) fornecem uma representação poderosa de entidades e as relações entre **eles**, mas a construção de grafos, a partir de enunciados de linguagem escrita e falada sugere novidades e inúmeros desafios. As etapas pretendidas para a construção de grafos pessoais de conhecimento são descritas nas subseções a seguir. Essas etapas serão suportadas por componentes da arquitetura conceitual para criação de sistemas de informação em saúde da Figura 5. Os componentes são relacionados e descritos nas Seções de 4.2 a 4.4. Nesse capítulo também são apresentadas a metodologia de desenvolvimento, as atividades propostas e o cronograma para realização das atividades.

4.1 Metodologia

Para a construção dos grafos de conhecimento com informações pessoais será necessária a fusão de informações pessoais e isso irá demandar algumas atividades como a manipulação de dados e informações, a partir de fontes de informações.

O grafo de conhecimento será gerado a partir de etapas de desenvolvimento. São elas: Aquisição, Modelagem, Extração de conteúdo, Anotação semântica, Extração de padrões e Armazenamento. Cada uma dessas etapas será descrita a seguir, seguindo a arquitetura de (BULCÃO; SILVA; MACEDO, 2019).

Aquisição manipula detalhes de baixo nível que são profundamente influenciados pelos requisitos de heterogeneidade e aquisição e análise como a decisão por métodos de aquisição baseados em responsabilidade (pull ou push), frequência (coleta instantânea

ou baseada em intervalo), tipo de fonte de dados (por exemplo, sensor físico, arquivo de dados, banco de dados ou mídia social) e tipo de dados adquiridos (por exemplo, texto ou série temporal). O requisito segurança e privacidade também deve ser tratado durante a coleta de dados e informações de várias fontes de informações.

Modelagem representa o esquema subjacente dos dados adquiridos. Esse esquema pode ser construído automaticamente ou por meio de uma intervenção do designer humano. Neste último caso, o esquema pode ser continuamente refinado à medida que novos dados são adquiridos. A semântica importada do MTCS pode enriquecer a representação do esquema da informação.

Extração de conteúdo primeiro, este componente extrai o conteúdo de dados de integridade previamente adquiridos e armazenados. Depois disso, ele insere esses conteúdos (como entidades e relacionamentos) no modelo construído pelo componente de modelagem. Este módulo deve considerar o contexto de heterogeneidade e capacitar o modelo com informações incorporadas de da semântica latente.

Pré-processamento um sistema de informação na área de saúde geralmente requer uma etapa de pré-processamento para que o conteúdo do GC não contenha dados desnecessários, bem como dados de menor qualidade. Por exemplo, no caso de dados de saúde textuais, palavras não-discriminantes (por exemplo, preposições, artigos e conjunções) podem ser eliminadas. Por sua vez, um componente de pré-processamento para dados de saúde baseados em sensores deve fornecer mecanismos para detectar e superar os problemas inerentes de qualidade desse tipo de dados, para que esses dados possam ser processados posteriormente para tomada de decisões precisas, por exemplo. Este módulo pode aplicar técnicas de pré-processamento, principalmente de processamento de linguagem natural, que devem ser escolhidas de acordo com o tipo de dado manipulado, levando em consideração a heterogeneidade informação. Por exemplo, dados podem ser anonimizados para manter privacidade, técnicas de limpeza podem ser aplicadas para melhorar a qualidade da informação, e entidades podem ser nomeadas para facilitar a anotação semântica.

Anotação semântica enriquece a estrutura de dados pré-processados e os conteúdos de saúde com uma semântica adequada e consensual no contexto dos modelos de informação em saúde. Recursos linguísticos, incluindo ontologias relacionadas à saúde e sistema de codificação padronizada, devem anexar o significado apropriado de cada conceito de maneira consensual. Além disso, a implantação de informações padronizadas representar semântica é crucial para fornecer dados de assistência médica que sejam compreensíveis pela máquina e fáceis de trocar com um sistema externo ou com um componente interno dessa arquitetura.

Extração de padrões é responsável por extrair padrões considerando os dados heterogêneos. Classificação, regressão, categorização, estatística e outros métodos podem ser aplicados aos dados de entrada para a descoberta de conhecimentos úteis. Esse conheci-

mento pode estar presente na semântica da informação lexicamente expressa e codificada por uma terminologia, uma ontologia ou um produto terminológico. O conhecimento gerado deve ser explorado por ferramentas ou serviços. Integração (Integration) - integra múltiplas descrições semânticas interligadas de entidades no GC como uma estrutura de representação uniforme e formal. Em tal estrutura, cada descrição de entidade (nó ou aresta) é processável por máquina e representa parte das especificações de outras entidades relacionadas a ela. A relação semântica da informação revelada por uma terminologia codificada suporta a interoperabilidade semântica.

Armazenamento gerencia a persistência de dados e estruturas necessária para criar e atualizar o GC, que inclui os dados de entrada adquiridos das fontes, o esquema subjacente a esses dados, os recursos linguísticos (terminologias) que transmitem a semântica dos dados e o próprio GC. Vários mecanismos de armazenamento que dependem da escolha do designer, incluindo sistema de arquivos (por exemplo, para dados adquiridos), memória interna, banco de dados relacional e banco de dados não-relacional (por exemplo, para o GC) são permitidos. Consequentemente, este módulo manipula dados e estrutura heterogêneos e também deve considerar as preocupações de segurança e privacidade dos dados.

Os grafos gerados serão representados utilizando tecnologia da Web Semântica RDF, de modo que eles sejam universalmente acessíveis, online e automaticamente navegáveis. Além disso, eles poderão ser agregados a outros grafos.

4.2 Aquisição e Preparação dos Dados

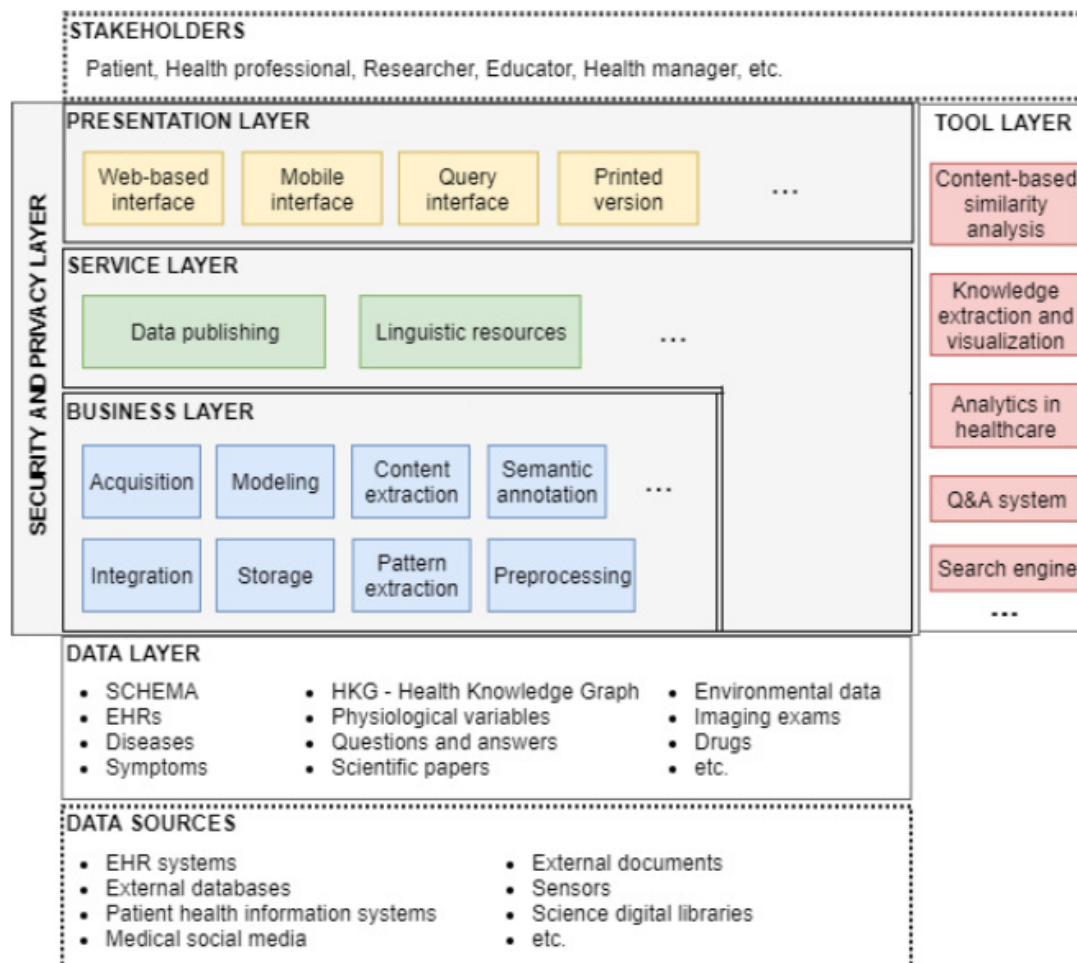
O estabelecimento de vocabulários de contexto e pessoais, por meio da aquisição e do processamento de dados de fontes heterogêneas de informação é fundamental nesta proposta. Deverá ser criado um repositório de vocabulários, representando as informações de contexto do usuário sobre as atividades do contexto a ser modelado no GC. As fontes de informação serão armazenadas na Data Layer da Figura 5.

Para aquisição de dados, deve ocorrer a seleção de fontes de informação. Os detalhes de aquisição serão tratados pelo componente Acquisition. Já a preparação de dados envolve a extração de informação, por meio métodos de Processamento de Linguagem Natural (PLN) e relacionamento de entidades.

A preparação dos dados envolve o armazenamento dos mesmos de maneira uniforme, análise com identificação e isolamento, transformação e padronização de dados, identificação dos campos a serem manipulados (schema matching e mapeamento para ontologia) realizados pelos componentes Modeling, Content Extraction e Pre-processing da Figura 5. No mapeamento do esquema, poderão ser utilizadas linguagem declarativa, ou

linguagem estatística (DO; RAHM, 2002), ou mesmo o apoio de Produtos Terminológicos de Informações Médica como nos componentes Modeling e Semantic annotation.

Figura 5 – Uma visão de várias camadas de uma arquitetura conceitual para sistemas de informação em saúde.



Fonte: Bulcão, Silva e Macedo (2019)

A extração de informação ou mesmo a aquisição de conhecimento para construção de GC deverá considerar cobertura, precisão, conhecimento verificável, intervenção humana eficiente, e alta manutenção. Atualmente, alguns procedimentos para extrair dados estruturados designados pelo usuário são apoiados por regras (expressões regulares, linguagens de programação de wrapper etc.), por árvores de segmentação DOM, e por aprendizado de máquina. Porém, os extratores de conteúdo da Web degradam com o tempo (WENINGER et al., 2016), como, por exemplo, o uso de JavaScript e de CSS para HTML (estático) tornou a extração pouco confiável. Em um futuro, a extração deveria ter renderização visual.

A extração de informação em fontes desestruturadas continua sendo o maior desafio para pesquisadores. As fontes desestruturadas de conhecimento incluem newswire, fóruns

de discussão, e-mails, calendários, mídias sociais e elas podem ser multilíngues. Instituições e eventos como o NIST e o TAC KBP organizam frequentemente desafios para extração de informação em fontes desestruturadas. Esse também será um desafio da proposta.

Em termos de e-mails e calendários, a extração de informação pode trazer conhecimentos sobre informações pessoais e profissionais da pessoa vinculada a e-mails (GAO; DREDZE; OARD, 2017), informações de menções de uma organização (GAO; DREDZE; OARD, 2016), vinculações de menções de reuniões de e-mails a calendários (GAO; DREDZE; OARD, 2018), “tópicos” através de clustering e expertise (TANG et al., 2014), traços de solução de problemas em e-mails profissionais (FRANCOIS; NADA; HASSAN, 2015).

Alguns trabalhos em extração de informação desestruturada incluem: extração de eventos e atributos (WANG; FINK; AGICHTEIN, 2015), extração de perfis de usuário (LI; RITTER; HOVY, 2014), extração de eventos de segurança do computador (RITTER et al., 2015), extração de entidades emergentes usando sementes (BRAMBILLA et al., 2017), extração de informações quantitativas a partir de dados sociais (ALONSO; SELAM, 2018), geração de grafos de conhecimento de produto (catálogos). Esse último tipo de pesquisa é muito interessante, uma vez que não existem fontes importantes para apoiar a organização de conhecimento de alguns produtos como existe Wikipedia para conhecimento geral e UMLS para a área da saúde.

A extração de informação para mídias de áudio (fala), imagens e vídeo complementa algumas áreas diferentes de Visão Computacional, que estão contempladas normalmente pelos mesmos eventos. Em 2017, o ImageCLEF (IONESCU et al., 2017) tratou da recuperação e sumarização de dados do Lifelogging; em imagens médicas para textual, e da descoberta de informações desconhecidas de imagens de observação da Terra. O TACKBP 2018 avaliou sistemas para extrair e agregar conhecimento de fontes heterogêneas tais como fontes multimídia multilíngues, incluindo texto, fala, imagens, vídeos e arquivos PDF.

Como nos trabalhos correlacionados e citados, a extração é uma das atividades mais desafiadoras desta proposta, uma vez que ela manipulará fontes desestruturadas de informação em mídia de áudio e vídeo, por exemplo. A leitura e acompanhamento de tecnologias e trabalhos relacionados será fundamental.

4.3 Combinação e Integração de Conteúdo

Para integração de conteúdos, deve ocorrer a combinação e a integração de informação, por meio, por exemplo, da detecção de entidades. A detecção de entidades correspondentes é um problema bem conhecido, pois envolve a identificação e a descoberta de instâncias

referentes à mesma entidade do mundo real. No GC da proposta, as entidades deverão ser termos encontrados nas fontes de informação e na pretensão de fala.

Pretende-se enriquecer os dados, melhorar sua qualidade, identificar e remover duplicados, e suportar exatidão dos fatos ao mesclar dados de várias fontes. Outras denominações para a tarefa de Identificação de Entidades são: Linking de Entidades, Resolução de Entidades, Reconciliação de Referência, e Deduplicação. A detecção de entidade auxilia em tarefas de qualidade de dados e desambiguação. Desafios para esse tipo de atividade são a quantidade de dados, os diferentes domínios e os diferentes tipos de representação e de conexão de dados.

A conexão entre instâncias de uma mesma entidade pode acontecer pelo conteúdo usando funções genéricas de comparação de conteúdo como funções baseadas em caracteres, em tokens, em fonética, funções específicas (ex. comparação de código postal, datas etc) ou mesmo funções que manipulam recursos linguísticos como proposto em (BULCÃO; SILVA; MACEDO, 2019). Nessa etapa, os componentes Pattern extraction e Integration da Figura 5 deverão ser explorados e estendidos.

4.4 Refinamento do Conhecimento

O refinamento do conhecimento deverá ocorrer após a detecção de entidades e integração. Essa etapa pretende resolver conflitos e erros, por meio, por exemplo, de inferência factual. Em termos de fusão de conhecimento, as principais abordagens são: mesclar os nós da entidade no grafo para acabar com conflitos de fatos e de conexões, resolver fatos buscando a verdade, realizar votação por maioria, identificar fontes autorizadas, e reunir, avaliar e prever evidências de diferentes fontes.

Em termos de detecção de erros, costuma-se usar regras de qualidade de dados, análise de dependência funcional e sua variação condicional por exemplo, analisar inconsistência, detectar outliers, e usar recursos de PLN. Para inferência, normalmente realizam-se métodos de enriquecimento e de complementação de dados, os quais podem ser internos ou externos com produtos terminológicos, por exemplo. Contudo o maior desafio para refinar o conhecimento de GC é a falta de dados de treinamento em domínios específicos para validar GC. Esse será outro grande desafio da proposta.

A maioria das abordagens para criação de grafos de conhecimento ainda depende de dados textuais. A chave para o desenvolvimento de GC é obter fluxo constante de dados de treinamento de alta qualidade com o mínimo de intervenção humana. Os componentes da Business Layer da Figura 5 deverão ser estendidos para tarefas de refinamento e qualidade dos conteúdos.

4.5 Atividades e Cronograma

O plano de trabalho obedecerá a seguinte sequência de atividades:

1. Cumprimento dos 40 créditos obrigatórios exigidos pelo PPG-CA. As disciplinas Engenharia de Software e Introdução ao Aprendizado de Máquina, com 10 créditos cada foram cursadas, ambas com conceito **A**. No segundo semestre de 2019, foram cursadas as disciplinas Estatística para Computação Aplicada e Processamento e Recuperação de Informação Textual para Computação Aplicada, com 12 e 8 créditos respectivamente, ambas também com conceito **A**. Desse modo, o cumprimento total de créditos ocorreu no final de 2019;
2. Estudo do referencial teórico;
3. Ampliação do mapeamento sistemático sobre o tema;
4. Estudo de caso
5. Escrita e Defesa da Qualificação;
6. Aprofundamento teórico dos conceitos e ferramentas envolvidas com a proposta, como: Web Semântica, arquivos RDF e OWL, Processamento de Linguagem Natural e Textual.
7. Pesquisa de repositórios disponíveis de termos, conceitos e sentenças da vida cotidiana em diferentes domínios. Produtos terminológicos deverão ser também consultados.
8. Aquisição das fontes de informação e produtos terminológicos selecionados. Nesse caso, deverão ser utilizados métodos de coleta baseados em responsabilidade (pull ou push) e frequência da coleta (instantânea ou baseada em intervalo).
9. Modelagem para geração de um esquema subjacente aos dados adquiridos. Esse esquema poderá ser construído automaticamente ou por meio de uma intervenção humana. A semântica importada do produtos terminológicos poderá enriquecer a representação do esquema da informação. A criação de um corpus de domínios do paciente deverá utilizar fontes de dados fornecidos por paciente e os repositórios da A1, usando métodos do grupo. As abordagens de marcação serão usadas para atribuir marcadores de classe a palavras no corpus. Abordagens para marcações de termos/conceitos são: taggers baseados em regras e/ou estocásticos (JURAFSKY; MARTIN, 2010).
10. Extração do conteúdo dos repositórios e produtos terminológicos previamente adquiridos e armazenados. Os conteúdos (como entidades e relacionamentos) deverão

ser inseridos no modelo construído pelo componente Modeling. Este módulo deve considerar o contexto de heterogeneidade e capacitar o modelo com informações incorporadas da semântica latente.

11. Armazenamento do corpus em um modelo.
12. Pre-processamento, a partir da aplicação de técnicas de PLN.
13. Realização de anotação semântica, a partir de consultas de produtos terminológicos.
14. Armazenamento do corpus anotado.
15. Seleção de métodos de identificação de entidades. Métodos para modelagem de linguagem baseada em contexto definidos na área de PLN, uso de produtos terminológicos e inferências serão vistos.
16. Análise de abordagens de seleção e integração das entidades identificadas. O método mais viável será selecionado para definir grafos pessoais de conhecimento para pacientes afásicos.
17. Construção e armazenamento de grafos de conhecimento para o sistema SICI a ser desenvolvido pelo bolsista TT-III.
18. Estabelecimento do estudo de caso e protocolos experimentais. Nesse caso, pode ser necessário a aprovação do Comitê de ética dos parceiros da área de Fonoaudiologia e Terapia ocupacional do grupo SofiaFala e da FMRP-USP.
19. Condução do estudo de caso e experimentos.
20. Apresentação de relatórios de avaliação.
21. Escrita, submissão e divulgação de resultados publicando e apresentando trabalhos científicos em congressos e periódicos nacionais e internacionais de grande impacto, principalmente na área de Sistemas de Informação.
22. Estudos de conceitos relacionados à proposta como processamento de linguagem natural, grafos de conhecimento, extração de informação, enriquecimento semântico de informações e demais necessários para o projeto. Essa atividade deverá ser realizada com a consulta de artigos de conferências e periódicos disponibilizados gratuitamente na USP e na biblioteca do laboratório da pesquisadora, a qual contém livros atuais e relacionados a áreas relacionadas.
23. Escrita e submissão de artigo;
24. Escrita da dissertação;
25. Defesa da dissertação.

A Tabela 2 apresenta o cronograma de execução para o plano de trabalho de acordo com o tempo padrão para conclusão do mestrado em dois anos (quatro semestres), com início no primeiro semestre de 2019 e fim no segundo semestre de 2020.

Tabela 2 – Cronograma de execução do plano de trabalho.

Semestre	Atividade																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1/2019																									
2/2019																									
1/2020																									
2/2020																									
1/2021																									
2/2021																									

Fonte: Eloisa Medrado, 2019

Referências

Agarwal, V. et al. Remembering what you said: Semantic personalized memory for personal digital assistants. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2017. p. 5835–5839. ISSN 2379-190X.

ALONSO, O.; SELAM, T. Quantitative information extraction from social data. In: ACM. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. [S.l.], 2018. p. 1005–1008.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 020139829X.

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

BRAMBILLA, M. et al. Extracting emerging knowledge from social media. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 26th International Conference on World Wide Web*. [S.l.], 2017. p. 795–804.

BULCÃO, R.; SILVA, R.; MACEDO, A. A conceptual architecture for modern healthcare information systems based on medical terminologies and coding systems support. *Submitted to Journal of Intelligent Information Systems*, Springer US, p. 25, 2019.

CAJAS-SALAZAR, N. et al. Combined effect of mpo, gstm1 and gstm1 polymorphisms on chromosome aberrations and lung cancer risk. *International Journal of Hygiene and Environmental Health*, v. 206, n. 6, p. 473 – 483, 2003. ISSN 1438-4639. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1438463904702448>>.

Consortium, W. W. W. et al. *Rdf 1.1 concepts and abstract syntax*. [S.l.]: World Wide Web Consortium, 2014.

Consortium, W. W. W. et al. *OWL Web Ontology Language Guide*. World Wide Web Consortium, 2009. Disponível em: <<https://www.w3.org/TR/owl-guide/>>.

COVINGTON, M. A.; NUTE, D.; VELLINO, A. *Prolog Programming in Depth*. USA: Prentice-Hall, Inc., 1996. ISBN 013138645X.

DENG, Y. et al. Knowledge graph based learning guidance for cybersecurity hands-on labs. In: *Proceedings of the ACM Conference on Global Computing Education*. New York, NY, USA: ACM, 2019. (CompEd '19), p. 194–200. ISBN 978-1-4503-6259-7. Disponível em: <<http://doi.acm.org/10.1145/3300115.3309531>>.

DO, H.-H.; RAHM, E. Coma: A system for flexible combination of schema matching approaches. In: *Proceedings of the 28th International Conference on Very Large Data Bases*. VLDB Endowment, 2002. (VLDB '02), p. 610–621. Disponível em: <<http://dl.acm.org/citation.cfm?id=1287369.1287422>>.

FELDMAN, R. et al. *The text mining handbook: Advanced approaches in analyzing unstructured data*. [S.l.: s.n.], 2007.

FENSEL, D. Ontologies. In: _____. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 11–18. ISBN 978-3-662-04396-7. Disponível em: <https://doi.org/10.1007/978-3-662-04396-7_2>.

FRANCOIS, R.; NADA, M.; HASSAN, A. How to extract knowledge from professional e-mails. In: *2015 11th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. [S.l.: s.n.], 2015. p. 687–692.

GAO, N.; DREDZE, M.; OARD, D. Knowledge base population for organization mentions in email. In: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*. [S.l.: s.n.], 2016. p. 24–28.

GAO, N.; DREDZE, M.; OARD, D. Enhancing scientific collaboration through knowledge base population and linking for meetings. In: . [S.l.: s.n.], 2018.

GAO, N.; DREDZE, M.; OARD, D. W. Person entity linking in email with nil detection. *Journal of the Association for Information Science and Technology*, v. 68, n. 10, p. 2412–2424, 2017. Disponível em: <<https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23888>>.

Giannecchini, T.; Maximino, L. *Programa de Intervenção Prático-produtivo para Indivíduos com Transtorno Fonológico*. [S.l.: s.n.], 2018. ISBN 9788565027939.

HOWELL, P.; DAVIS, S.; BARTRIP, J. The uclass archive of stuttered speech. *Journal of speech, language, and hearing research : JSLHR*, 03 2009.

IONESCU, B. et al. Overview of imageclef 2017: Information extraction from images. In: . [S.l.: s.n.], 2017. v. 10456, p. 11–14.

LEHMANN, J. et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, IOS Press, v. 6, n. 2, p. 167–195, 2015.

LI, J.; RITTER, A.; HOVY, E. Weakly supervised user profile extraction from twitter. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2014. p. 165–174.

Li, X. et al. Personal knowledge graph population from user utterances in conversational understanding. In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. [S.l.: s.n.], 2014. p. 224–229.

NABI, J. Machine learning — text processing. *Towards Data Science*, 09 2018. Disponível em: <<https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958>>.

NEEDHAM, A. E. H. M. *Graph Algorithms: Practical Examples in Apache Spark Neo4j*. First edition. [S.l.]: O'Reilly Media, Inc., 2019.

NICOLIELO-CARRILHO, A. P. et al. Relationship between phonological working memory, metacognitive skills and reading comprehension in children with learning disabilities. *Journal of Applied Oral Science*, scielo, v. 26, 00 2018. ISSN 1678-7757. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-77572018000100486&nrm=iso>.

Passadori, R. Problemas mais comuns de comunicação. In: . [s.n.], 2015. Disponível em: <<https://www.rhportal.com.br/artigos-rh/problemas-mais-comuns-de-comunicao/>>.

PONTE, A. S.; FEDOSSE, E. Characterization of individuals with acquired brain injury in working age. *Revista CEFAC*, SciELO Brasil, v. 18, n. 5, p. 1097–1108, 2016.

RITTER, A. et al. Weakly supervised extraction of computer security events from twitter. In: *Proceedings of the 24th International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015. (WWW '15), p. 896–905. ISBN 978-1-4503-3469-3. Disponível em: <<https://doi.org/10.1145/2736277.2741083>>.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136042597, 9780136042594.

Sager, J. *A Practical Course In Terminology Processing*. First edition. [S.l.]: John Benjamins Publishing CO, 1990. ISBN 9789027220776.

S.Bird; K.Ewan; L.Edward. *Natural Language Processing with Python*. Second edition. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 978-0-596-51649-9.

Shen, Y. et al. Sliqa-i: Towards cold-start development of end-to-end spoken language interface for question answering. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2019. p. 7195–7199. ISSN 2379-190X.

SUN, X.; ZHANG, S. User input-based construction of personal knowledge graphs. *Advances in Intelligent Systems and Computing*, v. 787, p. 339–345, 2019. Cited By 0. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049668459&doi=10.1007%2f978-3-319-94229-2_33&partnerID=40&md5=44be99f9a725160e34f155107bc20078>.

VICK, J. C. et al. Distinct developmental profiles in typical speech acquisition. *Journal of Neurophysiology*, American Physiological Society Bethesda, MD, v. 107, n. 10, p. 2885–2900, 2012.

WANG, Y.; FINK, D.; AGICHTEIN, E. Seef: Planned social event discovery and attribute extraction by fusing twitter and web content. In: *Ninth International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2015.

WENINGER, T. et al. Web content extraction: A metaanalysis of its past and thoughts on its future. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 17, n. 2, p. 17–23, fev. 2016. ISSN 1931-0145. Disponível em: <<http://doi.acm.org/10.1145/2897350.2897353>>.