PROCESSAMENTO DE LINGUAGEM NATURAL (PLN): FERRAMENTAS E DESAFIOS

Lucas Matheus Santos Andrade, Rafael Couto Barros, Marcelo Anderson Batista dos Santos

Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano-campus Salgueiro; lucasmatheusif@outlook.com.br, rafaelcoutobarros10@gmail.com, marcelo.santos@ifsertao-pe.edu.br

RESUMO: Processamento de Linguagem Natural (PLN) é uma das áreas da Inteligência Artificial que busca compreender a linguagem utilizada naturalmente pelos humanos. Essa busca para tornar o computador uma máquina que compreende e se comunica de forma semelhante a humanos vem ganhando cada vez ênfase na academia e indústria. O uso de PLN propicia a criação de interfaces de comunicação mais intuitiva para os usuários. Nesse cenário, diversas ferramentas e tecnologias surgiram nos últimos anos proporcionando uma maior facilidade para o desenvolvimento de aplicações com o uso de PLN. No entanto, o uso de linguagem natural ainda não é popular em softwares, pois existe uma barreira devido à complexidade para o desenvolvimento de aplicações que fazem uso de tal abordagem. Dessa forma, esse artigo tem como objetivo realizar uma análise das ferramentas existentes que auxiliam no desenvolvimento de aplicações com PLN, evidenciando seus pontos positivos, serviços oferecidos, deficiências, características e desafios.

Palavras–chave: PLN, Inteligência Artificial e Ferramentas de PLN

INTRODUÇÃO

Sempre se fez necessário na sociedade moderna a integração de mecanismos e plataformas capazes de desempenhar atividades semelhantes às realizadas pelo homem. Hoje em dia já temos plataformas capazes de interagir e até mesmo se comunicar com os seres humanos. Essa interação entre pessoas e plataformas está sendo cada vez mais comum na sociedade atual. O processamento de linguagem natural (PLN) é uma ferramenta que facilita essa interação, consiste em uma subárea da inteligência artificial, que desenvolve melhor a forma de interpretação da linguagem humana em diferentes dispositivos.

No entanto, há diversos desafios computacionais ainda em aberto para o funcionamento eficiente do processamento de linguagem natural, tendo em vista que a linguagem é uma estrutura complexa e possui variações de acordo com cada idioma. Entre as principais dificuldades se destacam, por exemplo: (1) A linguagem não padrão, (2) expressões idiomáticas e (3) conhecimento

sobre o mundo. A **linguagem não padrão** é quando utilizamos no nosso modo de escrita símbolos, erros gramaticais e abreviações que dificultam a interpretação semântica da frase. **Expressões idiomáticas**: São variações da língua criando expressões que são conhecidas geralmente pelo contexto cultural de uma região. Como exemplo, é comum falarmos que alguém pisou na bola quando fez algo que não deveria. Nesse contexto, temos um desafio de interpretação da semântica correta pelo computador. **Conhecimento sobre o mundo**: ocorre quando uma palavra tem duplo sentido e somente podemos saber o sentido analisando o contexto, temos como exemplo a palavra "ponto". Em um dado contexto a palavra "ponto" pode significar pontos em um jogo de basquete, um desenho de um ponto numa folha de papel ou ser apenas um sinal de pontuação. O problema reside na imprevisibilidade de combinações das palavras fazendo com que assumam diferentes contextos. No exemplo anterior a máquina não diferenciaria o sentido da palavra "ponto".No campo de PLN diversas ferramentas surgiram resolvendo problemas pontuais e oferecendo uma quantidade de serviços cada vez maior, como por exemplo, o reconhecimento por voz. Dessa forma, este artigo tem como objetivo fazer o levantamento de bibliotecas e ferramentas que auxiliem no processamento de linguagem natural, suas principais características com seus respectivos pontos fortes e deficiências. Assim, espera-se criar um guia para pesquisadores e programadores que desejam utilizar PLN para algum propósito.

PROCESSAMENTO DE LINGAGUEM NATURAL E SEUS DESAFIOS

PLN vem sendo estudada desde 1946 com objetivando que o computador possa se comunicar naturalmente através da linguagem humana. Ao longo do tempo ocorreu várias evoluções sobre o tema, houve a criação de várias ferramentas para desenvolvimento de aplicações de PLN, aplicações como análise fonética, analise morfológica, análise sintática e análise semântica.

Análise fonética é o reconhecimento de sons que está presente nas palavras, utilizada frequentemente no processamento da voz humana para texto. **Análise morfológica** realiza a análise da estrutura, formação e classificações das palavras para que seja possível inferir a que se refere determinada palavra, determinando, por exemplo, se uma palavra é um substantivo ou adjetivo. **Análise Sintática** tem por objetivo identificar como se relacionam as palavras entre si, e o seu papel na oração. **Análise semântica** analisa o significado das palavras e seu relacionamento na frase, para construir as sentenças, buscando identificar o contexto em que uma palavra está inserida. (GONZALEZ, 2003).

Nesse contexto, foram criados vários aplicativos para conversação humana com máquinas, como exemplo, os assistentes pessoais virtuais que estão presente em quase todos sistemas operativos (OS) atualmente, tais como Google Now[1], Cortana[2] e Siri[3]. Muitos deles são capazes de reconhecer sua fala e escrita, respondendo através da linguagem natural. Um outro de ferramenta é o Powerset[4], um buscador que pretende revolucionar a forma de pesquisar na web fazendo uso da habilidade de interpretação de linguagem natural em mecanismo de busca. O objetivo da Powerset é trazer uma resposta clara ao contrário da busca da busca do Google traz uma lista de páginas, ou seja, o usuário poderá fazer uma pergunta clara e o mecanismo Powerset processa essa pergunta trazendo uma resposta objetiva. Em 2008, o Powerset foi comprado pela Microsoft por 100 milhões de dólares, mostrando o valor que PLN possui.

Quando se imagina possíveis aplicações para PLN, encontramos diversas aplicações utilizadas nas mais diferentes áreas. Pode-se, por exemplo, ajudar pessoas com deficiência física que possuem limitações motoras para digitar em um teclado através do reconhecimento de voz, auxiliando também pessoas com alguma deficiência visual. O desafio está em conseguir compreender diversas linguagens diferentes, diversos tons de vozes, diversas regras gramáticas que geram um número de combinações extraordinárias. Consequentemente, a precisão de PLN em alguns casos precisa ser melhorada restringindo o espaço de possibilidades (ROSA).

METODOLOGIA

O trabalho realiza uma pesquisa qualitativa das ferramentas que auxiliam no desenvolvimento de softwares para processamento de linguagem natural também conhecido como PLN. Dessa forma, realizou-se uma revisão sistemática das ferramentas existentes e suas características. Primeiramente iniciou-se a fase de planejamento de revisão para identificação dos tópicos necessários a serem estudados. A segunda fase foi composta pela condução da revisão, seleção dos trabalhos a serem analisados e síntese do estudo realizado. Por fim, o resultado final foi elaboração deste trabalho.

[1] https://www.google.com/intl/pt-BR/landing/now/
[2] https://developer.microsoft.com/pt-br/cortana
[3] http://www.apple.com/br/ios/siri/
[4] https://en.wikipedia.org/wiki/Powerset_(company)

FERRAMENTAS PARA PLN: UMA VISÃO DE SEU FUNCIONAMENTO

Nesta seção é feita uma análise de várias ferramentas e bibliotecas desenvolvidas objetivando o processamento de linguagem natural em diferentes plataformas. Existem ainda diversos objetivos quando se deseja utilizar PLN. Dessa forma, está seção tem como objetivo exibir as principais características, pontos fortes e deficiências de ferramentas de PLN.

OpenNLP (OPENNLP, 2016): OpenNLP é uma biblioteca que consiste de várias ferramentas para aprendizagem da máquina para o processamento de linguagem natural. Escrita em na linguagem Java (Java,2016). Possui suporte as principais tarefas da PLN, tais como uso de tarefas extração de entidade nomeada, etiquetagem morfossintática, extração de sintagmas dentre outras. Desenvolvida pela Apache Project (Apache,2016), com última atualização feita no ano de 2015.

Vantagens:

- Solução código abeto
- Material de apoio ao uso da biblioteca desenvolvida pela própria desenvolvedora
- Realização das principais funções PLN

Desvantagens:

- Não contém suporte ao reconhecimento a língua portuguesa
- Não possui fórum de debate sobre a plataforma

UIMA (UIMA, 2016): A biblioteca apache UIMA é um framework onde pode ser projetado um sistema de software para análise de informações não estruturadas, tais como texto simples e identificador de entidades. Tem suporte de tarefas como detecção de entidade, detecção de relações, extração de tokens, lematização e etiquetagem morfossintática. Inicialmente foi desenvolvida pela IBM (IBM,2016), e agora pertence ao Apache Project, licenciando com código aberto, teve usa última atualização no ano de 2016.

Vantagens:

- Suporte para reconhecimento a língua portuguesa.
- Capaz de analisar grandes volumes de informações.
- Solução código aberto

- Possui material de apoio para uso da plataforma.
- Realização das principais funções PLN

Desvantagens:

- Não possui fórum de discursão sobrea plataforma.

Natural Language Toolkit (NLTK, 2016): NTLK é uma plataforma para criação de programas de linguagem natural escrito na linguagem de programação python (Python,2016). Possui suporte para português e fórum para discursão entre usuários e desenvolvedores. Tem suporte para tarefas tokenization, decorrente, marcação entre outros. Desenvolvida pela NTLK, com última atualização em 09 de abril de 2016.

Vantagens:

- Suporte para reconhecimento da escrita na língua portuguesa
- Solução código aberto
- Fórum de debate sobre a ferramenta
- Manual de uso da plataforma

Desvantagens:

- Não realizar todas as principais tarefas da PLN

GATE (GATE, 2016)**:** GATE é um software para realização de tarefas comuns de processamento de linguagem natural, tais como tarefas como anotação de texto com base em ontologia, etiquetagem morfossintática, reconhecimento de entidades e tratamento de anáforas. Desenvolvida pela GATE, tendo sua última atualização em 27 maio de 2016.

Vantagens:

- Solução código aberto
- Fórum de debate sobre a ferramenta
- Realização das principais funções PLN

Desvantagens:

- Não possui suporte de reconhecimento da língua portuguesa

- Não possui manual de utilização da plataforma

Apache Lucene (LUCENE, 2016): Apache Lucene é software que tem como principal tarefa se utilizado como motor de buscar ou simplesmente pesquisa de texto, escrito em Java tem o código aberto. Desenvolvido também pela Apache Project. Teve última atualização no ano de 2016.

Vantagens:

- Solução código aberto
- Manual de utilização da plataforma
- Realização das principais funções PLN

Desvantagens:

- Não possui fórum de debate sobre a plataforma
- Não contem reconhecimento da língua portuguesa

Stanford CoreNLP (CORENLP, 2016)**:** Stanford CoreNLP é um software que contém várias ferramentas para análise de linguagem natural, escrito em Java. Desenvolvida pelo um grupo de pesquisa na universidade de Stanford, com última atualização no 09 de dezembro de 2015.

Vantagens:

- Solução código aberto

Desvantagens:

- Não possui material de apoio de utilização da plataforma
- Não possui fórum de debate sobre a plataforma
- Não a suporte para reconhecimento da língua portuguesa
- Não Realizar das principais funções PLN

CoGrOO (CoGrOO, 2016): CoGrOO é um software que possui função principal de correção gramatical. Esta ferramenta está implementada no LibreOffice. Desenvolvida pela CoGrOO, com última atualização no ano de 2013.

Vantagens:

- Suporte para reconhecimento da escrita na língua portuguesa
- Solução código aberto

Desvantagens:

- Não possui material de apoio de utilização da plataforma
- Não possui fórum de debate sobre a plataforma
- Não realizar das principais funções PLN

LX-Center (LX-CENTER, 2016): É uma conjunto de ferramentas online e offline, que possui seguintes funções Silabificador, Lematizador Verbal, Conjugador Verbal, Flexionador Nominal, anotador categorial, reconhecedor nomes próprios, parser constituência, Parser Dependência, etiquetador de Papéis Semânticos, navegador Corpus, buscador Treebank, navegador MultiWordnet, Analisador Temporal. Desenvolvido pela universidade de Lisboa.

- Realização das principais funções PLN
- Não possui fórum de debate sobre a plataforma

Vantagens:

- Suporte para reconhecimento na língua portuguesa
- Possui material de apoio na utilização da plataforma

Desvantagens:

- Não possui solução código aberto

Cognitive Computation Group Demos(COGCOMP) (COGCOMP, 2016): É uma biblioteca desenvolvida em C++ para realizar as principais tarefas de processamento de linguagem natural, tais como etiquetagem morfossintática, analise sintática superficial e reconhecimento de entidade nomeada. Desenvolvida pela Cognitiva Computação Grupo(Cognitiva Computação Grupo, 2016), ainda na fase de testes com demos disponível para download.

Vantagens:

-Matérias de apoio para utilização da plataforma
- Solução código aberto
- Realização das principais funções PLN

Desvantagens:

- Não possui fórum de debate sobre a plataforma

Freeling (Freeling, 2016)**:** é uma biblioteca escrita em C++ [C++,2016], desenvolvida para dar suporte a análise morfológica, identificação da linguagem, tokenizing, detecção NE e outras funções. Desenvolvida por um grupo de pesquisar da universidade politécnica da Catalunha, com última atualização em 2016.

Vantagens:

- Solução de código aberto
- Possui material de apoio para utilização da plataforma

Desvantagens:

- Não possui reconhecimento para língua portuguesa
- Não possui fórum de debate sobre a plataforma

WordNET (WORDNET, 2016): é um grande banco de dados em inglês, nesse banco de dados possui classes gramaticais para formação das palavras, que são eles advérbios, substantivos, adjetivos e verbo. Desenvolvida pela universidade de Princeton, teve sua última atualização em 2006.

Vantagens:

- Solução código aberto
- Possui material de apoio

Desvantagens:

- Não possui suporte para língua portuguesa
- Não possui fórum de debate sobre a plataforma

LingPipe (LINGPIPE, 2016): é uma biblioteca em java, que é utilizado para realizar tarefas como reconhecimento de entidades, corretor ortográfico e classificação automática. Possui suporte para principais línguas do mundo.

Vantagens:

- Possui material de apoio da ferramenta

- Reconhecimento da língua portuguesa

Desvantagens:

- Não possui fórum de debate sobre a plataforma
- Não contem solução de código aberto

https://www.google.com/intl/pt-BR/landing/now/
https://developer.microsoft.com/pt-br/cortana
http://www.apple.com/br/ios/siri/
https://en.wikipedia.org/wiki/Powerset_(company)

ARK Syntactic & SemanticParsing (ARK SYNTACTIC, 2016)**:** São duas ferramentas unidas para análise da estrutura linguística , etiquetagem morfossintática, análise semântica da estrutura predicado-argumento e análise sintática de dependência. Desenvolvida pela universidade de Universidade Carnegie Mellon.

Vantagens:

- Solução de código aberto
- Realizar as principais funções da PLN

Desvantagens:

- Não possui suporte reconhecimento da língua portuguesa
- Não possui fórum de debate sobre a plataforma
- Não possui material de apoio da plataforma

FrameNET (FRAMENET, 2016): é um recurso semelhante ao WordNET com diferença que as palavras são usadas em sentenças, e possui cerca de 170 mil sentenças anotadas. Foi desenvolvida por um grupo de pesquisa da universidade Berkeley.

Vantagens:

- Solução de código aberto
- Possui material de apoio

Desvantagens:

- Não possui suporte reconhecimento da língua portuguesa
- Não possui fórum de debate sobre a plataforma

Tabela 1. Comparativa entre as Ferramentas de PLN

	TABELA COMPARATIVA							
Ferramenta / Característica	Open Source	Manual da Ferramenta	Fórum	Suporte Português	Plataforma	Análise Semântica	Análise Sintática	Análise Morfológico
OpenNLP	X	X			JAVA	X	X	X
UIMA	X	X		X	JAVA	X	X	X
NLTK	X	X	X	X	PYTHON	X	X	
GATE	X		X		C++	X	X	X
Lucene	X	X			JAVA	X	X	X
Stanford CoreNLP	X				JAVA	X	X	X
CoGrOO	X			X	C++		X	
LX-Center	X	X			------	X	X	X
COGCOMP	X	X			JAVA	X	X	X
Freeling	X	X			C++			X
WordNET	X	X			-------			
LingPipe		X		X	JAVA			
ARK Syntactic & SemanticParsing	X				-------	X	X	X
FrameNET	X	X			-------			

CONCLUSÃO

PLN é uma área de pesquisa que mesmo não sendo estuda há anos, possui muitos desafios em aberto não resolvidos pela comunidade científica. A complexidade no tratamento da linguagem de comunicação humana faz com que as máquinas ainda tenham problemas no que concerne a precisão da interpretação da linguagem natural.

Nesse sentido, diversas ferramentas vêm sendo desenvolvidas visando melhorar a produção de novos softwares que façam uso de PLN. Por isso, neste artigo fizemos a análise de diversas

ferramentas que facilitam a aplicação de PLN em diversos cenários, evidenciado características, pontos positivos e negativos. Espera-se que esse levantamento do estado da arte das ferramentas de PLN facilite a busca de qual ferramenta se adequa melhor de acordo com a demanda de um desenvolvedor.

Uma das ferramentas que mais se destacou na pesquisa realizada foi a NTLK, pois esta ferramenta é uma das poucas que possui suporte para reconhecimento da língua portuguesa e contém uma linguagem de programação de fácil aprendizado.

Como trabalhos futuros pretendemos analisar a performance das ferramentas citadas em diferentes cenários, visando identificar o grau de acerto de cada ferramenta, tempo de resposta e a sua facilidade de uso

REFERÊNCIAS

FINATTO,M,J,B; LOPES, L; CIULLA, A. Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: parcerias que já dão (muito) certo. **Domínios de Linguagem**, v. 9, n. 5, p. 41-59, 2015.

GONZALEZ,M; LIMA, V, L, S. Recuperação de informação e processamento da linguagem natural. In: XXIII Congresso da Sociedade Brasileira de Computaçao. p. 347-395, 2003.

ROSA, Ademar Evandro et al. RECONHECIMENTO DE FALA E PROCESSAMENTO DA LINGUAGEM NATURAL

OpenNLP – Site Oficial: https://opennlp.apache.org/. Acesso em 09 de outubro de 2016.

Java- Site oficial: <https://www.java.com/pt_BR/>. Acesso em 09 de outubro de 2016.

IBM- Site oficial:< https://www.ibm.com/br-pt/>. Acesso em 09 de outubro de 2016.

Python- Site oficial:<https://www.python.org/>. Acesso em 09 de outubro de 2016.

C++- Site Wikipédia:< https://pt.wikipedia.org/wiki/C_(linguagem_de_programação)>. Acesso em 09 de outubro de 2016.

Apache-Site Oficial: < https://www.apache.org/> . Acesso em 09 de outubro de 2016.

UIMA – Site Oficial: https://uima.apache.org/. Acesso em 09 de outubro de 2016.

NLTK – Site oficial: http://www.nltk.org/. Acesso em 09 de outubro de 2016.

GATE – Site oficial: https://gate.ac.uk/. Acesso em 09 de outubro de 2016.

Lucene – Site Oficial: https://lucene.apache.org/core/. Acesso em 09 de outubro de 2016.

Core NLP – Site Oficial: http://stanfordnlp.github.io/CoreNLP/. Acesso em 09de outubro de 2016.

CoGrOO – Site: http://cogroo.sourceforge.net/. Acesso em 09 de outubro de 2016.

LX-Center – Site Oficial: http://lxcenter.di.fc.ul.pt/ . Acesso em 09 de outubro de 2016.

Cogcomp – Site Oficial: https://cogcomp.cs.illinois.edu/page/tools_view/13 . Acesso em 09 de outubro de 2016.

Freeling – Site Oficial: http://nlp.lsi.upc.edu/freeling/node/1 . Acesso em 09 de outubro de 2016.

Cognitive Computation Group– Site oficial <http://cogcomp.cs.illinois.edu/>. Acesso em 09 de outubro de 2016.

Wordnet – Site Oficial: http://ai.stanford.edu/~rion/swn/. Acesso em 09 de outubro de 2016.

LingPipe – Site Oficial: http://alias-i.com/lingpipe/. Acesso em 09 de outubro de 2016.

ARK Syntactic – Site Oficial: http://demo.ark.cs.cmu.edu/parse. Acesso em 09 de outubro de 2016.

FrameNet – Site Oficial: https://framenet.icsi.berkeley.edu/fndrupal/. Acesso em 09 de outubro de 2016.