WiTi-Lex: Uma ferramenta para extração de informações sobre pessoas e análise de sentimentos de opiniões

Felipe de Morais[1], **Gabriel Souto Fischer**[1], **Rodrigo Smiderle**[1], **Sandro José Rigo**[1]

[1]Programa de Pós-Graduação em Computação Aplicada (PPGCA)
Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos, 950 – Bairro Cristo Rei – CEP 93022-000 – São Leopoldo – RS – Brazil

{felipmorais, gsfischer, rsmiderle}@edu.unisinos.br, rigo@unisinos.br

Abstract. *The extraction of information from unstructured documents is a resource of Natural Language Processing (NLP). Several applications can use these results once they have already been extracted and classified. Some NLP features provide information on structured data, such as the polarity of a word, which denotes the feeling associated with it. However, no studies integrating these two aspects are observed. This paper describes the development of the WiTi-Lex tool. WiTi-Lex is a web application for extracting information about people and identifying other people's opinions about the person being searched. In order to achieve it, the information about people is extracted from the database of Wikipedia and categorized with the Brazilian Academy of Letters (Academia Brasileira de Letras). Other people's opinions are extracted from Twitter and classified with SentiLex. Furthermore, this paper describes an ontology on grammar class extraction and word polarity, called WiTi-Onto. Finally, a set of evaluation procedures was performed with the tool. The evaluation have shown promising results when testing the WiTi-Lex tool against a public dataset.*

Resumo. *A extração de informações de documentos não estruturados é um dos recursos da área de Processamento de Linguagem Natural (PLN). Várias aplicações podem utilizar esses resultados, uma vez que já tenham sido extraídos e classificados. Alguns recursos de PLN fornecem informações sobre dados já estruturados como, por exemplo, a polaridade de uma palavra, que denota o sentimento associado com a mesma. Entretanto, não são observados trabalhos integrando estes dois aspectos. Este artigo descreve a implementação da ferramenta WiTi-Lex. WiTi-Lex é uma aplicação web para extração de informações sobre pessoas e identificação da opinião de outras pessoas em relação a pessoa pesquisada. Para isso, as informações sobre pessoas são extraídas da base de dados do Wikipédia e categorizadas com a Academia Brasileira de Letras. Além disso, as opiniões das outras pessoas são extraídas do Twitter e classificadas com o SentiLex. Neste contexto, este artigo também descreve uma ontologia sobre extração de classe gramatical e de polaridade de palavras, chamada WiTi-Onto. Finalmente, foi realizado um conjunto de procedimentos de avaliação com a ferramenta. A avaliação permitiu demonstrar resultados promissores ao testar a ferramenta WiTi-Lex uma base de dados pública.*

1. Introdução

O Processamento de Linguagem Natural (PLN) é uma das áreas de pesquisa da Inteligência Artificial, que pode ser dividida em duas subáreas de trabalho: *interpretação* e *geração* de linguagem natural. O conceito de *interpretação* baseia-se em mecanismos que buscam compreender frases em alguma linguagem natural e traduzir essas frases para uma representação estruturada que possa ser compreendida e utilizada por um sistema computacional. Por outro lado, o conceito de *geração* refere-se ao oposto, ou seja, o sistema traduz uma representação computacional de um conteúdo semântico para texto em linguagem natural [Allen 1995].

Para a interpretação e geração de linguagem, primeiramente, é necessário um processo de extração de informações. Para Kushmerick [Kushmerik 1999], extração de informação é "a tarefa de identificar os fragmentos específicos de um documento que constituem o núcleo de seu conteúdo semântico". Assim, a extração de informação encarrega-se de analisar e transformar o formato de apresentação da informação contida em diferentes tipos de documentos, exibindo-as em um formato desejável. Por exemplo, pode-se analisar uma página da internet e como saída ter um documento com nome, telefone e e-mail encontrados nessa página, em um formato de texto desejado [Cabral 2009].

Um possível tratamento dos dados extraídos de um determinado contexto é conhecido como análise de sentimentos. A análise de sentimentos tem por objetivo identificar e extrair, de forma automática, as opiniões, sentimentos e emoções expressadas em um documento de entrada. Existem diversos métodos de análise de sentimentos na literatura, que se diferenciam por suas técnicas de predição de sentimentos, utilizando as abordagens de aprendizado de máquina, dicionários léxicos e processamento de linguagem natural [Reis et al. 2015]. Uma das formas de se fazer isso é através da predição da orientação do sentimento, popularmente conhecida como polaridade, que pode ser positiva, negativa ou neutra [Narayanan et al. 2009].

Uma forma que tem sido utilizada para armazenar conhecimento e relações entre informações são as ontologias. As ontologias possibilitam representar o conhecimento semântico de várias situações do mundo real. As relações existentes no mundo real são de difícil representação. Assim, as ontologias normalmente são estruturadas de tal maneira em que se pretende representar apenas determinada parte desse mundo, focando em domínios específicos onde apenas uma parte é formalizada e representada. Segundo Vasconcelos [Vasconcelos 2003], "As ontologias se propõem a capturar domínios de conhecimento de forma genérica, para fornecer um entendimento semântico desses domínios que poderá ser utilizado e compartilhado por diversas comunidades e aplicações".

Este artigo descreve o desenvolvimento da ferramenta web WiTi-Lex. WiTi-Lex tem o objetivo de possibilitar a busca nos sites do Twitter e da Wikipédia para extrair informações pertinentes, com base em um termo de pesquisa. A busca no Twitter tem como objetivo identificar as publicações mais recentes no site, com base no termo de pesquisa e extrair os sentimentos com relação a estas publicações baseado na polaridade das palavras. Por outro lado, a busca na Wikipédia tem como objetivo a coleta de informações pertinentes sobre o termo de pesquisa. Este termo ficou restrito aos nomes de empresas e de pessoas. Assim, a aplicação desenvolvida é capaz de extrair e apresentar estas informações ao usuário de forma estruturada. Ao mesmo tempo, a ferramenta é capaz de capturar os sentimentos de outras pessoas com relação ao termo pesquisado. Além disso,

as informações estruturadas, geradas pela ferramenta, são representadas e disponibilizadas através de uma ontologia, proporcionando maior flexibilidade ao trabalho.

2. Trabalhos Relacionados

Com o objetivo de identificar o estado da arte das tecnologias existentes na área de PLN e reconhecimento de emoções em textos, foi realizada uma busca de trabalhos relacionados com a ferramenta desenvolvida. Dos trabalhos analisados, foram selecionados cinco para uma análise comparativa. Os trabalhos avaliados são apresentados na Tabela 1, trazendo informações sobre o problema abordado, técnica de solução proposta, base de dados de origem para extração das informações para análise, qual a precisão obtida pela ferramenta desenvolvida e, por fim, as limitações destes trabalhos.

Tabela 1. Tabela comparativa

Trabalho	Problema	Técnica Proposta	Base de Dados	Precisão	Limitações
[Wu et al. 2016]	Melhorar precisão	Léxico de sentimentos específico para micro-blog	Micro-blog chinês	84,3%	Apenas para o micro-blog chinês
[Pênalver-Martinez et al. 2014]	Classificação de polaridade	Ontologia na fase de extração de recursos	Twitter	82,9%	Funciona apenas neste domínio
[Kumar and Bala 2016]	Análise de sentimentos com Big-Data	Análise de sentimentos em um ambiente cloud usando Hadoop	Twitter	70%	Analisa apenas opiniões com Hashtags
[Bhatia et al. 2015]	Melhorar precisão	Análise de sentimentos utilizando RST	Duas bases de dados de filmes	85,5%	Atuação em análises léxicas de sentimentos
[Rigo et al. 2013]	Melhorar precisão	Análise de emoção em AVAs	Moodle	91,9%	Baseado em regras cadastradas

Através destes trabalhos relacionados, percebe-se que a análise de sentimentos em textos de redes sociais e ambientes de aprendizagem é uma tendência. Entretanto, muitos desses sistemas ficam limitados a domínios específicos de linguagem, bases de dados muitos específicas ou a regras pré-cadastradas. Além disso, geralmente faltam a esses sistemas a capacidade de gerar textos em linguagem natural para os dados de entrada, em que foi realizada a análise de sentimentos. Outro ponto é que, em alguns dos casos, os testes se basearam em realizar consultas específicas, não permitindo liberdade para a realização da análise de sentimentos para diferentes dados de entrada.

3. Fontes de dados

A Wikipédia[1] é uma enciclopédia livre online que está em processo constante de construção e atualização por milhares de colaboradores de todas as partes do mundo. O sistema é baseado no conceito de Wiki, ou seja, é uma coleção de muitas páginas interligadas, onde cada uma delas pode ser acessada e alterada por qualquer pessoa. Em outras palavras, a Wikipédia é uma enciclopédia colaborativa totalmente online e livre. Seu principal objetivo é a definição de termos e conceitos, de forma informativa, sem apresentar opiniões ou sentimentos.

Por outro lado, o Twitter[2] é um repositório de informações sobre o que está acon-

[1]http://pt.wikipedia.org/wiki/Wikipédia:Sobre_a_Wikipédia
[2]http://twitter.com/

tecendo no mundo agora. Mais especificamente, as pessoas utilizam o Twitter para compartilhar suas opiniões sobre algum determinado assunto com o resto do mundo através da internet. O Twitter possui mais de 300 milhões de usuários ativos, com uma média de 6 mil tweets por segundo, equivalente a 500 milhões de tweets por dia ou cerca de 200 bilhões de tweets por ano [ILS 2017].

Sendo assim, a ferramenta desenvolvida tem como objetivo capturar os dados não estruturados da Wikipédia, baseado em um termo de pesquisa que normalmente é um nome de pessoa. Além disso, com o uso das postagens dos usuários do Twitter, é possível capturar os sentimentos de várias pessoas sobre o termo de pesquisa, no momento exato da busca e apresentar de forma resumida ao usuário.

4. Ferramentas e Recursos Utilizados

Para obter os textos a serem analisados, WiTi-Lex acessa a *Application Programming Interface* (API) da Wikipédia, que se baseia em um Web-Service, fornecendo acesso através do protocolo HTTP. Os clientes requisitam ações particulares através de um parâmetro na URL para receber a informação desejada[3]. Assim, é possível para o WiTi-Lex capturar as informações contidas nas páginas da Wikipédia.

Para identificar e tratar as palavras recebidas através da API da Wikipédia, o WiTi-Lex acessa o site da Academia Brasileira de Letras (ABL)[4] e busca na sua base de dados as informações referentes as palavras a serem analisadas. A ABL possui um sistema de busca do Vocabulário Ortográfico da Língua Portuguesa que contém 381.000 verbetes, com as respectivas classificações gramaticais e outras informações, conforme descrito no Novo Acordo Ortográfico. Assim, o WiTi-Lex faz uma requisição de cada palavra extraída da Wikipédia para a ferramenta de busca da ABL e recebe um documento semi-estruturado no formato JSON.

Para obter os tweets mais recentes, baseado em um termo de pesquisa, WiTi-Lex acessa a API do Twitter. A API do Twitter[5] provê acesso para ler e gravar dados do Twitter. O principal interesse do WiTi-Lex é capturar os 100 tweets mais recentes, baseado em um termo de pesquisa. Para realizar tal pesquisa, WiTi-Lex possui credenciais de acesso que são validadas pela API do Twitter a cada requisição realizada. As respostas são fornecidas pela API no formato JSON.

Para analisar o sentimento ou opinião das pessoas, baseado no termo de pesquisa informado no WiTi-Lex e nos textos obtidos do Twitter, foi utilizado o SentiLex. SentiLex é um léxico de sentimento voltado para identificação da opinião das pessoas em textos publicados na mídia social [Silva et al. 2012]. Ele foi considerado um recurso pioneiro para a língua portuguesa, constituído por 7.014 lemas e 82.347 formas flexionadas. Mais especificamente são 16.863 adjetivos, 1.280 nomes, 29.504 verbos e 34.700 expressões idiomáticas [Carvalho and Silva 2015]. A informação da polaridade associada às entradas foi anotada manualmente, na maioria dos casos, porém algumas foram automaticamente atribuídas por uma ferramenta (JALC) desenvolvida pela equipe do SentiLex.

Para representar o conhecimento gerado pela WiTi-Lex, foi desenvolvida uma on-

[3]https://www.mediawiki.org/wiki/API:Main_page
[4]http://www.academia.org.br
[5]https://dev.twitter.com/rest/public

tologia. No desenvolvimento da ontologia, chamada WiTi-Onto, foi utilizado a ferramenta Protégé[6], que é um editor de ontologias livre e um *framework* para construção de sistemas inteligentes. Na implementação da ontologia proposta, foram utilizadas as linguagens SPARQL[7], uma linguagem de consulta para o padrão RDF, e a DL Query, ambas para a realização de consultas nas instâncias armazenadas na ontologia. Além disso, foi utilizada a linguagem SWRL, pertencente ao *plugin* OWL, para realizar a inferência de conhecimentos, baseado em regras previamente definidas.

5. Descrição da Implementação

Segundo a arquitetura do WiTi-Lex, ilustrada na Figura 1, os usuários acessam a aplicação através de uma página web. O sistema possui um *Web-Service* responsável por comunicar os módulos WiTi-Sentiment e WiTi-Extract com a base de dados. O WiTi-Lex é composto por três módulos. O primeiro é o módulo de extração de informação, chamado WiTi-Extract, que usa o Wikipédia para trazer informações relevantes sobre o objeto de pesquisa. O segundo é o módulo de reconhecimento de sentimentos, chamado WiTi-Sentiment, que apresenta a polaridade sobre os textos escritos no Twitter contendo o nome pesquisado. O terceiro é o módulo Witi-Ontology, responsável por descrever a ontologia criada para a aplicação.

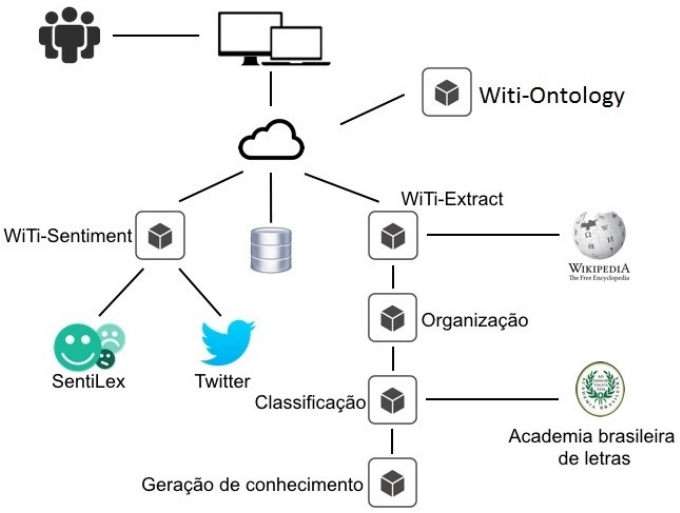

Figura 1. Arquitetura do WiTi-Lex
Fonte: Elaborada pelos autores

5.1. Módulo WiTi-Extract

O módulo de extração de informação, após alimentado pela entrada fornecida pelo usuário, faz uma requisição para API do Wikipédia para a procura de informações. Em caso de falha na busca, o módulo retorna sem informação. Em caso de sucesso, a informação recuperada é passada para os submódulos, sendo eles o de organização, classificação e o de geração de conhecimento. Tal estratégia foi adotada para que os dados possam ser usadas posteriormente pelos serviços desenvolvidos e por serviços futuros.

[6]http://protege.stanford.edu/
[7]https://www.w3.org/TR/rdf-sparql-query/

O primeiro submódulo de extração é o de organização. Ele é responsável por organizar o texto extraído, separando-o por frases, palavras e pontuação. Esse processo facilita a classificação das mesmas, possibilitando automatizar esse processo, além de facilitar o reconhecimento de padrões no processo de geração de conhecimento.

O segundo submódulo é o de classificação, o qual possui duas funções. A primeira é a de verificar se uma palavra tem classificação gramatical na base de dados local. Caso não encontre a classificação, ele procura informações sobre a palavra em fontes externas. No caso da aplicação desenvolvida, essa procura é feita na Academia Brasileira de Letras. A segunda função desse submódulo é, dentre as diversas classificações possíveis para uma determinada palavra, escolher qual é a classificação mais adequada a partir dos padrões já conhecidos.

O terceiro submódulo é o de geração de conhecimento. Ele utiliza as informações pré-processadas no submódulo de organização e classificadas no submódulo de classificação. Esse submódulo, através de padrões pré-estabelecidos, extrai informações específicas. As informações que o sistema se propõem a extrair são as seguintes:

- Se o termo de pesquisa é uma pessoa ou não;
- Data de nascimento;
- Lugar de nascimento;
- Data de falecimento;
- Lugar de falecimento;
- Uma breve descrição de quem é a pessoa.

5.2. Módulo WiTi-Sentiment

O módulo de reconhecimento de sentimentos, denominado WiTi-Sentiment, é responsável por reconhecer o sentimento nos textos publicados pelos usuários do Twitter sobre o termo pesquisado. Para isso, esse módulo utiliza duas ferramentas, o Twitter e o SentiLex. Ao entrar com o nome de uma pessoa no campo de pesquisa do sistema, WiTi-Sentiment executa uma pesquisa na base de dados do Twitter, utilizando esse nome como termo de busca. Essa pesquisa é realizada através de um *request* na API do Twitter. Foram elaboradas as seguintes definições para a obtenção dos textos: **a)** Apenas os 100 tweets mais recentes, que contenham o termo de busca, são selecionados; **b)** Apenas textos em português são pesquisados; **c)** Cada texto (tweet) é separado em um conjunto de palavras; **d)** Imagens, links e menções a outras pessoas são descartados; **e)** Hashtags são removidas, mas a palavra é utilizada (ex.: #chateado); **f)** Retweets são contabilizados como tweets normais, pois se uma pessoa compartilhou o tweet de alguém, chamado de retweet, significa que essa pessoa pensa da mesma forma ou concorda com o que foi postado.

Na implementação foi utilizada uma tabela hash onde a chave é cada palavra, de todos os 100 tweets, e o valor é o número de ocorrências dessa palavra. Ao analisar todas as palavras, é gerado uma lista ordenada pelas palavras com maior número de ocorrências, como se fosse um *ranking* de palavras mais utilizadas. Após obter essa lista, foi utilizado o léxico de sentimento SentiLex para verificar a polaridade dessas palavras.

Ao carregar o SentiLex na máquina do usuário, em formato JSON, o sistema gera uma outra tabela hash para armazenar a polaridade de cada palavra contida no SentiLex. Sendo assim, a chave de cada registro da tabela é a própria palavra e o valor é a polaridade

associada a esta palavra. Essa estratégia, de representar o SentiLex com tabelas hash, foi utilizada para reduzir a complexidade do sistema de O(n) para constante, onde n é o número de palavras contidas no SentiLex, diminuindo assim o tempo de execução.

Todas as palavras obtidas nos textos são pesquisadas na tabela hash que representa o SentiLex. Caso a busca retorne um valor, essa palavra é contabilizada de acordo com sua polaridade. O SentiLex utiliza três tipos de polaridade: positiva, neutra ou negativa. Assim, ao final da busca de todas as palavras na tabela hash do SentiLex, é obtida a quantidade de palavras associada a cada polaridade. A ideia em obter o número de palavras associadas com cada polaridade é poder dizer qual o sentimento associado com o termo de pesquisa obtido dos 100 tweets mais recentes. Foi utilizada uma barra de progresso na interface gráfica para exibir o resultado. Essa barra é dividida em três partes. O número de palavras com polaridade positiva, neutra e negativa é apresentado nas cores verde, amarelo e vermelho, respectivamente.

5.3. Módulo WiTi-Ontology

O Witi-Ontology é o módulo responsável por descrever a ontologia desenvolvida para o Witi-Lex, com o objetivo de facilitar a descoberta da classificação de palavras. Ela está divida em três classes responsáveis pela ligação e estruturação da ontologia, sendo elas: **(i) Classe**, que representa as classes gramaticais possíveis para uma palavra, **(ii) Polaridade**, que representa as polaridades possíveis para uma palavra e **(iii) Palavra**, que serve para representar cada palavra, tendo relações com as Classes e as Polaridades.

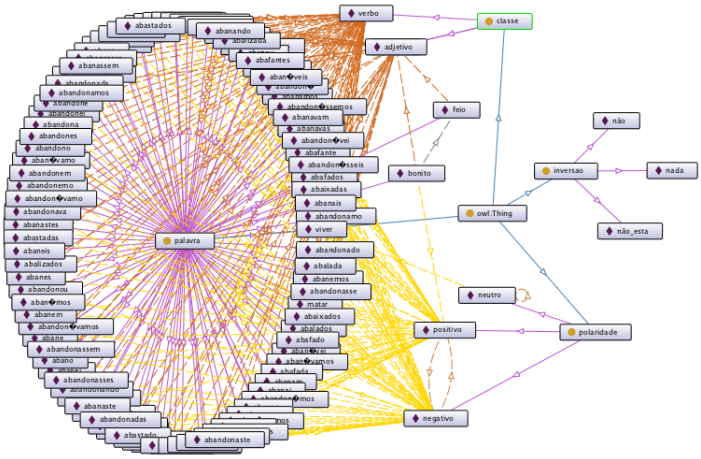

Figura 2. Grafo da Ontologia
Fonte: Imagem exportada da Ontologia no software Protégé 5.2

A Figura 2 ilustra o grafo que representa a arquitetura obtida através das interligações entre classes e instâncias na ontologia do Witi-Lex. Nesta figura é possível identificar alguns exemplos como “adjetivo” que é uma instância da classe Classe, “feio” que é uma instância da classe Palavra e que possui uma relação com “adjetivo” através da propriedade *hasClasse*, indicando que a palavra “feio” é um adjetivo. Outro exemplo é “bonito”, que também é uma instância da classe Palavra e que está relacionada com a instância “positivo” da classe Polaridade. Assim, há uma relação de *hasPolaridade* entre “bonito” e “positivo”, indicando que a palavra “bonito” tem polaridade positiva.

Nesta ontologia foram realizadas dois tipos de pesquisas, utilizando DL Query e SPARQL. DL Query proporciona uma interface amigável, onde é possível navegar facilmente pela ontologia, permitindo uma melhor visualização, através do uso de palavras-chave pré-definidas pela linguagem. Porém, DL Query é limitada quando comparada com SPARQL, que é muito mais poderosa, possuindo uma sintaxe similar ao SQL, permitindo recuperar e manipular dados da ontologia.

Em uma ontologia é possível inferir conhecimento através de propriedades e instâncias já conhecidas. No Protégé isso se dá através da linguagem SWRL. Assim, é possível adicionar novas propriedades e gerar novos conhecimentos sobre palavras, no caso da WiTi-Onto. Existe uma propriedade, chamada *hasInverter*, da classe polaridade que é responsável por relacionar positivo com negativo. Assim, a regra criada atribui aos antônimos das palavras a polaridade inversa da palavra original. Isso é bastante útil quando se deseja criar polaridade de várias palavras de forma automatizada. Para isso, é necessário existir uma relação de antônimos de cada uma das palavras.

5.4. Funcionamento

WiTi-Lex é uma ferramenta web, porém nesse momento está disponível somente em um servidor local, sem acesso externo. Ao acessar o sistema, o usuário pode entrar com um termo de pesquisa em um campo de texto livre. Inicialmente, espera-se que o usuário digite o nome de uma pessoa, embora o sistema também seja capaz de capturar informações de algumas empresas. Como o sistema é dividido em módulos diferentes, WiTi-Sentiment e WiTi-Extract, um módulo pode retornar sem resultados enquanto o outro pode retornar algum resultado.

Figura 3. Exemplos de pesquisas no WiTi-Lex
Fonte: Elaborada pelos autores

A Figura 3 ilustra um exemplo de execução do WiTi-Lex. Na parte **Results** são exibidas as informações que o sistema foi capaz de extrair da Wikipédia. Na parte **Palavras mais usadas nos últimos 100 tweets + SentiLex** é exibido uma barra de progresso divida em três partes, conforme descrito na Seção 5.2. Na interface há um botão que o usuário pode clicar para exibir mais informações sobre os resultados obtidos do módulo WiTi-Sentiment. Assim, o usuário pode visualizar: **(i)** as palavras identificadas pelo SentiLex, ordenadas pela quantidade, **(ii)** todas as palavras obtidas nos tweets, também ordenadas pela quantidade e, por fim, **(iii)** o texto de cada um dos tweets.

6. Avaliação e Resultados

A fim de testar o sistema, foi definida a utilização do mesmo para realizar a extração de informações de uma lista de nomes de pessoas famosas. Para tanto, resolveu-se utilizar a

lista das 50 celebridades mais bem pagas de 2016 da Revista Forbes[8], pois a mesma é uma revista conhecida e que pode fornecer uma lista adequada de nomes. Como na listagem existem nomes de bandas e grupos famosos também, retirou-se os mesmos do teste final, originando uma lista de 25 nomes, sendo estes utilizados para o experimento. Já que o sistema também possui a capacidade de identificar empresas, resolveu-se por realizar o mesmo processo em uma nova listagem com o nome das dez melhores multinacionais para se trabalhar no Brasil em 2017, conforme o Ranking *Great Places to Work*[9].

Na extração de informações de pessoas, o sistema apresentou 92% de precisão ao identificar os nomes da lista como pessoas e 90,48% ao trazer informações extras sobre essas pessoas. O sistema apresentou algumas situações de falha, exemplificadas em duas situações identificadas. A primeira foi na identificação do termo "David Copperfield", quando o sistema trouxe a informação que não era uma pessoa e sim um romance de Charles Dickens. Embora o sistema tenha acertado, deveria também ter identificado a existência de uma pessoa. A segunda ocorreu com o termo "Diggy". Neste caso, o termo não se refere diretamente a um nome e sim a um apelido. Como este apelido foi utilizado na lista apresentada, o sistema não conseguiu identificar um nome válido e trazer informações relevantes.

Na extração de informações de empresas, o resultado obtido foi de 50% de precisão na identificação da lista e 100% nas informações trazidas pelo sistema. O alto nível de erro na identificação das empresas se deve principalmente pelo fato de que o sistema foi originalmente desenvolvido para identificar pessoas e não empresas. Porém, graças a arquitetura modular do sistema, ele também pode ser usados para buscar informações sobre qualquer assunto.

7. Conclusão

Este trabalho apresentou o sistema WiTi-Lex. A aplicação desenvolvida é capaz de extrair informações do Wikipédia, organizá-las e classificá-las. A ferramenta também é capaz de apresentar ao usuário essas informações de forma estruturada, com auxílio da base da dados de palavras da Academia Brasileira de Letras. Ao mesmo tempo, a ferramenta também é capaz de capturar os sentimentos de outras pessoas com relação ao termo pesquisado, obtidos através de informações extraídas do Twitter, com auxílio da base de dados do SentiLex. Além disso, a ontologia desenvolvida conseguiu atingir o objetivo para qual foi criada, sendo capaz de estruturar a classificação da polaridade e classe gramatical das palavras.

Como possibilidade de trabalhos de futuros, vislumbra-se uma melhoria do módulo de análise de sentimentos, de forma a permitir uma análise mais precisa dos sentimentos relacionados ao termo pesquisado. Além disso, outro objetivo é aumentar o número de informações extraídas da Wikipédia, bem como gerar uma melhor classificação das palavras extraídas de forma a se obter uma melhor geração de linguagem natural. Também, vislumbra-se aumentar o número de classes da ontologia, para atender às demais informações extraídas da aplicação. Por fim, espera-se disponibilizar publicamente a Witi-Onto integrada, em tempo real, ao WiTi-Lex. Atualmente ainda é necessário uma

[8]http://forbes.uol.com.br/listas/2016/07/50-celebridades-mais-bem-pagas-do-mundo-em-2016
[9]http://www.greatplacetowork.com.br/ranking/brasil-2017-medias-multinacionais.htm

intervenção humana na alimentação dos dados da ontologia pelos dados presentes na base de dados do Witi-Lex, por meio de um agente integrador desenvolvido.

Referências

Allen, J. (1995). *Natural Language Understanding (2Nd Ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.

Bhatia, P., Ji, Y., and Eisenstein, J. (2015). Better Document-level Sentiment Analysis from RST Discourse Parsing. *ArXiv e-prints*.

Cabral, L. S. (2009). Extração de informação usando integração de componentes de PLN através do framework GATE.

Carvalho, P. and Silva, M. J. (2015). Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*, 7(1).

ILS (2017). Twitter Usage Statistics. http://internetlivestats.com/twitter-statistics, Maio.

Kumar, M. and Bala, A. (2016). Analyzing twitter sentiments through big data. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 2628–2631.

Kushmerik, N. (1999). Gleaning the web. *IEEE Intelligent Systems and their Applications*, 14(2):20–22.

Narayanan, R., Liu, B., and Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Ángel Rodríguez-García, M., Moreno, V., Fraga, A., and Sánchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995 – 6008.

Reis, J. C., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015). Uma abordagem multilıngue para análise de sentimentos. In *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)*.

Rigo, S. J., Alves, I. M., Gazola, O., Belau, F., Barbosa, J. L. V., and Costa, C. (2013). Abordagem linguística para identificação da dimensão afetiva expressa em textos de Ambientes Virtuais de Aprendizagem – um Léxico da Emoção. *Simpósio Brasileiro de Informática na Educação - SBIE*, 24(1):738.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Vasconcelos, K. F. (2003). Ontoeditor: Um editor para manipular ontologias na web. Master's thesis, Universidade Federal de Campina Grande, Paraíba.

Wu, F., Huang, Y., Song, Y., and Liu, S. (2016). Towards building a high-quality microblog-specific chinese sentiment lexicon. *Decis. Support Syst.*, 87(C):39–49.