UNIVERSIDADE FEDERAL DE ALAGOAS – UFAL
CAMPUS DE ARAPIRACA
CIÊNCIA DA COMPUTAÇÃO – BACHARELADO

ROQUE BARBOSA DA COSTA JÚNIOR

IDENTIFICAÇÁO E RESOLUÇÃO DE AMBIGUIDADES SEMÂNTICAS EM LÍNGUA PORTUGUESA UTILIZANDO ESTATÍSTICA

ARAPIRACA
2022

Roque Barbosa da Costa Júnior

Identificação e resolução de ambiguidades semânticas em língua portuguesa utilizando estatística

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal de Alagoas – UFAL, Campus de Arapiraca.

Orientador: Prof. Dr. Patrick Henrique da Silva Brito

Arapiraca

2022

Universidade Federal de Alagoas – UFAL
Campus Arapiraca
Biblioteca *Campus* Arapiraca - BCA

C837i Costa Júnior, Roque Barbosa da
Identificação e resolução de ambiguidades semânticas em língua portuguesa utilizando estatística / Roque Barbosa da Costa Júnior. – Arapiraca, 2022.
17 f.: il.

Orientador: Prof. Dr. Patrick Henrique da Silva Brito.
Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Federal de Alagoas, *Campus* Arapiraca, Arapiraca, 2022.
Disponível em: Universidade Digital (UD) – UFAL (*Campus* Arapiraca).
Referências: f. 17.

1. Ambiguidade semântica – Língua portuguesa. 2. Língua Brasileira de Sinais (LIBRAS). 3. Tradução automática. 4. Sistema FALIBRAS. I. Brito, Patrick Henrique da Silva. II. Título.

CDU 004

Bibliotecário responsável: Nestor Antonio Alves Junior
CRB - 4 / 1557

Roque Barbosa da Costa Júnior

Identificação e resolução de ambiguidades semânticas em língua portuguesa utilizando estatística

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal de Alagoas – UFAL, Campus de Arapiraca.

Data de Aprovação: 22/02/2022.

Banca Examinadora

Prof. Patrick Henrique da Silva Brito
Universidade Federal de Alagoas – UFAL
Campus Arapiraca
(Orientador)

Prof. Alexandre Paes dos Santos
Universidade Federal de Alagoas – UFAL
Campus Arapiraca
(Examinador)

Prof. Tércio de Morais Sampaio Silva
Universidade Federal de Alagoas – UFAL
Campus Arapiraca
(Examinador)

AGRADECIMENTOS

À minha família, que me deu a chance de estudar o que eu amo mesmo em meio às dificuldades da vida.

A todos os membros do projeto FALIBRAS que se tornou mais do que uma atividade extracurricular, sendo um ambiente de aprendizado permanente cheio de pessoas maravilhosas.

Aos professores que passaram pela minha vida, mostrando-me o caminho do aprendizado da melhor forma que lhes era possível.

Ao meu orientador, professor Patrick Henrique da Silva Brito, uma pessoa maravilhosa, capaz de trazer o melhor das capacidades de seus alunos mantendo uma personalidade amável e uma profundidade de conhecimento invejável.

RESUMO

Há no mundo mais de 1 bilhão de pessoas com algum tipo de deficiência. No Brasil, essa realidade corresponde a cerca de 23,9% dos 190 milhões de brasileiros; entre estes, 9.6 milhões têm alguma deficiência auditiva. A surdez dificulta consideravelmente a interação social, uma vez que inibe o indivíduo de se comunicar através da via oral-auditiva. Esses problemas de comunicação costumam prejudicar consideravelmente a interação dos alunos surdos com colegas ouvintes, prejudicando o processo de integração social. Para facilitar a comunicação entre pessoas surdas e ouvintes, ferramentas de tradução automática Português-Libras podem ser utilizadas. Porém, de acordo com relatos na literatura, cerca de 75% da comunidade surda se sente insatisfeita com a tradução e relata como as principais causas dessa insatisfação: (1) uso de sinais inadequados para palavras com ambiguidade semântica (e.g., direito, público); (2) a ausência de expressão facial; (3) não utilização de modificadores da Libras, prejudicando a naturalidade da sinalização; e (4) utilização de sinais não usuais em determinada região. Um dos principais obstáculos na tradução eficiente de falas se encontra na ambiguidade, já que a linguagem natural é, por natureza, ambígua e muitas vezes depende de contexto para ser entendida, processos rígidos de tradução que focam apenas no contexto léxico se tornam extremamente limitados, e a estatística entra como uma alternativa para tornar tais métodos capazes de lidar com o contexto além do nível léxico. Neste trabalho é abordado o Problema 1; para isso, é proposto o aperfeiçoamento dos módulos de tradução do Sistema Falibras através da identificação de ambiguidades através de estatística. Foi percebido um impacto direto na melhoria da qualidade da tradução no tocante às críticas observadas na literatura. Os principais objetivos do projeto proposto são: (1) conhecer o estado da arte do processo de língua natural para resolução automática de ambiguidades na língua portuguesa; (2) aperfeiçoar o módulo semântico do tradutor para reconhecimento de ambiguidades em língua portuguesa; e (3) avaliar as atividades através de um experimento piloto. Devido às dificuldades enfrentadas decorrentes da pandemia, não foi possível avaliar a solução conforme previsto originalmente, isto é, envolvendo a participação direta de voluntários intérpretes. A avaliação da solução foi realizada de maneira preliminar, pela própria equipe do projeto. Por essa razão, os artefatos de software desenvolvidos ainda não foram totalmente validados e carecem de ajustes importantes antes da sua disponibilização gratuita à comunidade.

Palavras-chave: resolução de ambiguidade semântica em língua portuguesa; língua brasileira de sinais; tradução automática; falibras.

ABSTRACT

There are more than 1 billion people in the world with some form of disability. In Brazil, this reality corresponds to about 23.9% of the 190 million Brazilians; among these, 9.6 million have hearing impairment. Deafness considerably hinders social interaction, as it inhibits the individual from communicating through the oral-auditory route. These communication problems tend to considerably impair the interaction of deaf students with hearing colleagues, impairing the process of social integration. To facilitate communication between deaf and hearing people, Portuguese-Libras machine translation tools can be used. However, according to reports in the literature, about 75% of the deaf community feel dissatisfied with the translation and report as the main causes of this dissatisfaction: (1) use of inappropriate signs for words with semantic ambiguity (eg, law, public) ; (2) the absence of facial expression; (3) non-use of Libras modifiers, harming the naturalness of the signal; and (4) use of unusual signals in a given region. One of the main obstacles in the efficient translation of speeches is found in ambiguity, since natural language is, by nature, ambiguous and often depends on context to be understood, rigid translation processes that focus only on the lexical context become extremely limited, and statistics enters as an alternative to make such methods capable of dealing with the context beyond the lexical level. In this work, Problem 1 is addressed; for this, it is proposed to improve the translation modules of the Falibras System through the identification of ambiguities through statistics. A direct impact on improving the quality of the translation was perceived in terms of the criticisms observed in the literature. The main objectives of the proposed project are: (1) to know the state of the art of the natural language process for automatic resolution of ambiguities in the Portuguese language; (2) improve the translator's semantic module to recognize ambiguities in Portuguese; and (3) evaluate the activities through a pilot experiment. Due to the difficulties faced due to the pandemic, it was not possible to evaluate the solution as originally planned, that is, involving the direct participation of volunteer interpreters. The evaluation of the solution was carried out in a preliminary way, by the project team itself. For this reason, the developed software artifacts have not yet been fully validated and need important adjustments before being freely available to the community.

Keywords: semantic ambiguity resolution in portuguese; brazilian sign language; machine translation; falibras.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

GQM	Goal-Question-Metric
UFAL	Universidade Federal de Alagoas
LaTAD	Laboratório de Tecnologias Assistivas Digitais

SUMÁRIO

1 INTRODUÇÃO

Com o objetivo de facilitar a comunicação entre ouvintes e surdos e apoiar a formação de intérpretes, surgiu em 2001 o projeto Falibras (CORADINE *et al*., 2001). Desde seu início, o projeto tem como um de seus principais objetivos traduzir frases em português nas respectivas animações da Libras. Além do Falibras, uma ampla gama de novas ferramentas de tradução surgiram, tais como HandTalk, Prodeaf e VLibras. Porém, relatos recentes da literatura (ROCHA; MELGAÇO, 2018) mostram que cerca de 75% da comunidade surda se sente muito insatisfeita com a tradução e relata como sendo as principais causas dessa insatisfação: (1) uso de sinais inadequados para palavras com ambiguidade semântica (e.g., direito, público); (2) a ausência de expressão facial; (3) não utilização de modificadores da Libras, prejudicando a naturalidade da sinalização; e (4) utilização de sinais não usuais em determinada região. Porém, o mesmo estudo ao tempo que relata a falta de estímulo para utilizar os tradutores automáticos, devido às limitações apontadas, também relata a indisponibilidade de intérpretes humanos para intermediar a comunicação em situações cotidianas, o que aumenta a dependência das pessoas surdas em relação aos tradutores automáticos.

O objetivo principal deste trabalho foi, a partir do estudo das técnicas de processamento de língua natural para resolução automática de ambiguidades, aperfeiçoar o módulo semântico do tradutor Falibras de modo a resolver ambiguidades semânticas em frases na língua portuguesa. Para isso, foram executadas atividades que foram principalmente os objetivos específicos: Conhecer o estado da arte do processo de língua natural para resolução automática de ambiguidades na língua portuguesa; aperfeiçoar o módulo semântico do tradutor para reconhecimento de ambiguidades em língua portuguesa e avaliar as atividades através de um experimento piloto do projeto.

2 METODOLOGIA

Devido à pandemia, o projeto não pôde ser desenvolvido nas dependências do LaTAD, localizado no Campus Arapiraca da UFAL. O projeto foi desenvolvido de forma remota, utilizando os recursos computacionais dos próprios pesquisadores. A ferramenta utilizada para gerenciar a execução das atividades foi o Trello. O Google Meet foi a ferramenta escolhida para as reuniões síncronas do projeto, que ocorreram semanalmente. A linguagem de programação utilizada foi Java e o ambiente de desenvolvimento escolhido foi a IDE Eclipse.

A primeira atividade realizada durante o projeto foi a revisão da sistemática da literatura; essa revisão não foi feita de forma isolada já que todos os alunos participantes do projeto participaram promovendo seminários abordando assuntos relevantes ao projeto, desde a história do Falibras até assuntos específicos como métodos para tradução automática. Tal revisão sistemática seguiu o modelo proposto por Kitchenham (2012). Tal método foi adotado por ser bem aceito perante a comunidade acadêmica que utiliza pesquisas experimentais na área da computação, em especial na comunidade de engenharia de software experimental.

No tocante ao acompanhamento e execução das atividades de desenvolvimento de software, será adotado um método com características de desenvolvimento ágil (MARTIN, 2003), combinando atividades do processo SCRUM e Kamban (KNIBERG, 2010). A ferramenta que será utilizada para o desenvolvimento será ontologias, além da linguagem de programação Java e suas bibliotecas.

No tocante à avaliação dos resultados da pesquisa, era prevista inicialmente uma avaliação dos resultados de forma quanti-qualitativa, através de duas avaliações complementares, sendo a primeira de caráter qualitativo e a segunda de caráter quantitativo. A avaliação quantitativa teve seu planejamento iniciado, utilizando a técnica GQM (BASILI 1992), com a previsão de se definir questões de pesquisa com caráter comparativo. As respostas a tais questões deveriam se dar por meio da comparação dos resultados obtidos entre as principais soluções existentes: Falibras (versão atual), Falibras (nova versão), Handtalk, Prodeaf e Vlibras. Porém, apesar dessa intenção inicial, devido às dificuldades enfrentadas decorrentes da pandemia, não foi possível avaliar a solução conforme previsto originalmente, isto é, envolvendo a participação direta de voluntários intérpretes. A avaliação da solução foi realizada de maneira preliminar, pela própria equipe do projeto.

O acompanhamento das atividades ocorreu por meio de reuniões semanais para aferir o andamento do projeto. Tal acompanhamento aconteceu para sanar rapidamente eventuais

dúvidas e estreitou os laços dos integrantes do projeto, composto por alunos de mestrado e outros alunos de graduação.

3 RESOLUÇÃO DE AMBIGUIDADES

Ambiguidades são onipresentes no processamento de linguagem natural e ocorrem quando uma sentença ou uma expressão linguística apresentam mais de um significado possível, gerando problemas de interpretação. Assim sendo, elas podem ser divididas em dois tipos, as de caráter sintático, onde os diferentes significados podem ser encontrados ao alterar as posições dos lexemas dentro da oração. O outro tipo de ambiguidade, e o foco desse projeto, são as ambiguidades lexicais, onde a dificuldade de interpretação se dá pelo fato do lexema possuir mais de um significado dado um contexto específico.

Esse projeto então aborda as ambiguidades em um contexto ontológico, uma ontologia é uma especificação explícita de um determinado conceito, e o significado de uma expressão de linguagem natural no contexto de uma interpretação ontológica é o conceito ontológico que a expressão representa (UNGER; CIMIANO, 2011), ontologias linguísticas por sua vez armazenam conceitos lexicalizados, conceitos que são expressos por uma ou mais palavras de uma língua, um compartimento de significados, conceitos e sentidos compartilhados por um contexto linguístico, como uma nação ou uma comunidade na internet. Por isso, certas ontologias linguistas podem acabar não possuindo certos ‘significados’ já que em sua língua não existe o lexema que representa esse conceito, apesar de o mesmo ‘significado’ ainda poder ser representado por um conjunto de diferentes lexemas.

Antes da resolução do problema em si, se vê necessário a criação de um mecanismo facilitador, um banco de palavras e seus múltiplos significados para serem utilizados durante o projeto. Iniciou-se a criação de um corpus, tal mecanismo têm como objetivo a criação de um banco de informações que facilite o acesso e o processamento dos modelos, permitindo um local de consulta para informações semânticas com um determinado contexto.

Para solucionar esse problema é proposto o uso de redes Bayesianas, também conhecidas como redes causais, redes de opinião e gráficos de dependência probabilística, que consistem em um tipo de abordagem probabilística para tomada de decisões. Elas funcionam combinando conceitos de teoria dos Grafos, teoria de probabilidades, Ciências da Computação e Estatística e “podem ser consideradas como uma representação visual e informativa da tabela de probabilidade conjunta de todas as variáveis que envolvem o domínio do problema” (SOUZA, 2010, p. 22).

Dada a ocorrência de lexemas na oração, seus respectivos significados tem suas probabilidades dadas conforme o Teorema de Bayes, também chamado de probabilidade condicional. com isso, é criado um gráfico direcional com as probabilidades anteriores e

posteriores ao evento analisado, computando a distribuição de probabilidades posterior para um conjunto de variáveis de busca (os lexemas e o contexto), a partir daí é possível 'caminhar no grafo' para encontrar o mais provável significado dado o lexema, seu contexto, e a probabilidade de vezes que ele aparece em determinado contexto com determinado significado, então, a inferência em redes Bayesianas “trata-se de um mecanismo para cálculo da distribuição posteriori de probabilidade para um conjunto de variáveis, dado um conjunto de evidências, ou seja, variáveis aleatórias com valores instanciados” (SOUZA, 2010, p. 35).

Entre os pontos fortes que motivaram o uso das redes Bayesianas se encontram o fato delas realizarem o aprendizado de relações causais, o que permite mais interpretabilidade dos resultados, o que em um domínio como o processamento de linguagem natural é muito bem vindo, e assumindo que os valores das probabilidades utilizados são corretos, o valor de classificação é ótimo.

A seguir, são apresentados os resultados preliminares decorrentes da adaptação do módulo tradutor do Falibras, para resolver ambiguidades semânticas.

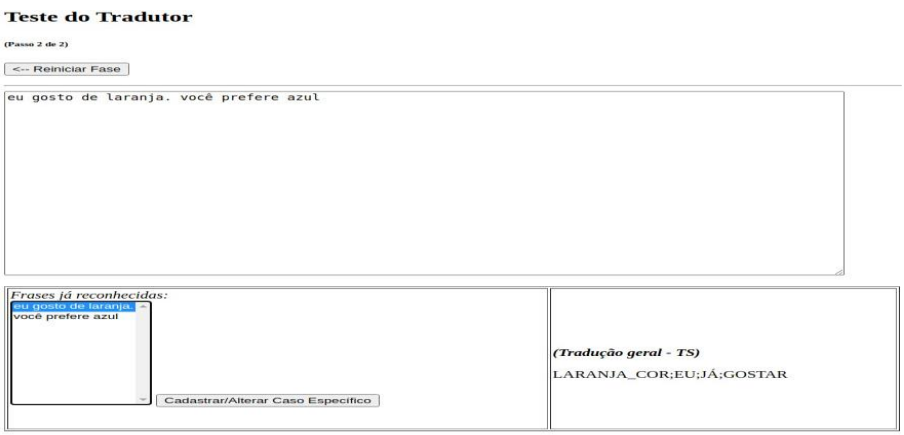
.

4 PROCESSO DE RESOLUÇÃO DO FALIBRAS

A versão atual do resolvedor de ambiguidades semânticas integrado ao Sistema Falibras utiliza uma classificação ontológica dos sinais, os agrupa de acordo com a semântica de utilização de cada um. Por exemplo, um sinal para representar a cor laranja, poderia fazer parte de grupos semânticos tais como “Cor” e “Arte”, enquanto que um sinal para representar a fruta laranja, poderia fazer parte de grupos semânticos tais como “Alimento” e “Perecível”. Quando uma palavra possui mais de um possível sinal, como é o caso da palavra “laranja”, a palavra é considerada ambígua. Nesses casos, o resolvedor de ambiguidade do Falibras identifica o sinal mais provável, a partir da proximidade e frequência de grupos semânticos na mesma frase e em frases próximas. Quanto mais próxima a localização do grupo semântico, maior a influência dele na resolução da ambiguidade.

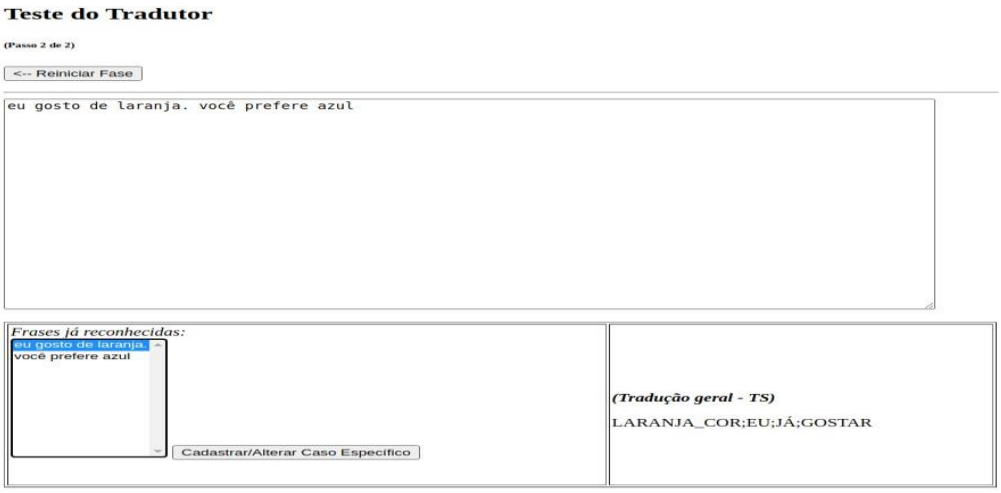
Para ilustrar melhor o funcionamento do resolvedor de ambiguidades desenvolvido e integrado ao Falibras, as Figuras 1 e 2 apresentam dois cenários de execução do tradutor, após a integração com o módulo de resolução de ambiguidades. Como pode ser observado, no primeiro exemplo (Figura 1), foi solicitada a tradução da sentença “eu gosto de laranja". Você prefere azul.” A ambiguidade da sentença está relacionada à palavra “Laranja”, que pode assumir o sentido de “Laranja-fruta” ou “Laranja-cor”. Devido à proximidade da palavra “Azul”, cujo sinal pertence à categoria semântica de “Cor”, o sinal escolhido para resolver a ambiguidade de “Laranja” foi “LARANJA-COR”. Já no segundo exemplo (Figura 2), foi solicitada a tradução da sentença “eu gosto de laranja. Você prefere maçã. Nesse caso, devido à proximidade da palavra “Maçã”, cujo sinal pertence à categoria semântica de “FRUTA”, o sinal escolhido para resolver a ambiguidade de “Laranja” foi “LARANJA-FRUTA”

Figura 1 – Tradução laranja-Cor.



Fonte: O autor (2021).

.Figura 2 – Tradução laranja-Fruta.



Fonte: O autor (2021).

5 CONSIDERAÇÕES FINAIS

Apesar das dificuldades impostas pela pandemia e principalmente pelo distanciamento social, o projeto atingiu resultados promissores. A inserção de um corpus próprio do projeto é de grande ajuda aos modelos estatísticos, além disso, o corpus pode ficar mais robusto com o tempo, com novas palavras, significados e contextos sendo adicionados a ele.

Quanto às questões de predição de significado e, por conseguinte, tradução , os resultados atuais também são satisfatórios apesar de começarem a se aproximar dos limites de modelos totalmente probabilísticos.

Entretanto, algumas das atividades não foram executadas, tais como o teste com pesquisadores voluntários e os usuários finais. Tal impossibilidade prejudicou não só a avaliação em si, mas também a condução da pesquisa, uma vez que a identificação das principais ambiguidades possuía como uma das etapas a realização de reuniões com brainstorm envolvendo pesquisadores convidados. Também houve prejuízos no tocante à análise qualitativa da solução proposta. A decorrência desse problema se deu principalmente pelas consequências da pandemia da Covid-19, que se instaurou ao longo do tempo de execução das atividades, o que impossibilitou a comunicação mais direta com pessoas fora do domínio do laboratório de desenvolvimento.

REFERÊNCIAS

BASILI, V. R**. Software modeling and measurement:** the goal/question/metric paradigma. College Park, Maryland, USA: University of Maryland, 1992.

CORADINE, L. C. *et al.* Levantamento de hipóteses para implementação do Projeto Falibras. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 7., 2001, Vitória, ES. **Anais** [...]. Vitória, ES: SBC, 2001. Pôster.

FREITAS, L.; VIEIRA, R. Ontologias e língua portuguesa. *In:* CÍRCULO DE ESTUDOS LINGUÍSTICOS DO SUL, 8., 2008, Porto Alegre, RS. **Anais** [...] Porto Alegre: UFRGS, 2008. p. [1-10].

KITCHENHAM, B. A. Systematic review in software engineering: where we are and where we should be going. *In*: INTERNATIONAL WORKSHOP ON EVIDENTIAL ASSESSMENT OF SOFTWARE TECHNOLOGIES, 2., 2012, New York, USA. **Proceedings** [...]. New York, USA: Association for Computing Machinery, 2012. p. 1– 2.

KNIBERG, H. **Kanban and Scrum:** making the most of both. Toronto, CA: C4Media Inc, 2010.

MARTIN, R. C. **Agile software development:** principles, patterns, and practices. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2003.

ROCHA, C; MELGAÇO, S. C. de. O uso de aplicativos para tradução de Libras. *In*: SIMPÓSIO INTERNACIONAL DE INOVAÇÃO EM MÍDIAS INTERATIVAS, 5., 2018, Goiânia, GO. **Anais** [...]. Goiânia, GO: UFG, 2018. p.36-47.

SOUZA, A. L. Ara. Redes Bayesianas: uma introdução aplicada a Credit Scoring. *In*: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 19., 2010, São Pedro, SP. **Anais** [...]. São Pedro, SP: ABE, 2010. p. 1-99. Relatório .

UNGER, C.; CIMIANO, P. Representing and resolving ambiguities in ontology-based question answering. *In*: WORKSHOP ON TEXTUAL ENTAILMENT, 1., 2011, Edinburgh, Scotland, UK. **Proceedings** [...]. Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011. p. 40–49.