Desafios na Tradução da Informação Espacial do EN para o PT-br

Challenges in Translating Spatial Information from EN to PT-br

Rafael Fernandes
Universidade de São Paulo
rafael.macario@usp.br

Rodrigo Souza
Universidade de São Paulo
rodrigo.aparecido.souza@usp.br

Marcos Lopes
Universidade de São Paulo
marcoslopes@usp.br

Resumo

A Tradução Automática Neural (TAN), atualmente a abordagem mais utilizada, ainda enfrentam desafios ao lidar com a tradução do conhecimento espacial. Neste estudo, utilizamos o Raciocínio Espacial Qualitativo (REQ) para representar informações espaciais em traduções automáticas do inglês para o português. Traduzimos 145 frases dos corpora CAM e COCA, utilizando Google Translate e DeepL, e identificamos as causas das traduções não naturais. Com o uso do REQ, mapeamos logicamente as diferenças de significado. Nossos resultados indicam que, apesar de um bom desempenho no geral, a TAN apresenta dificuldades com significados espaciais específicos, resultando em $10,6\%$ de erros semânticos e $12,0\%$ de erros de projeção sintática. Este trabalho explora os desafios práticos e teóricos da tradução automática.

Keywords

Tradução Automática Neural; Tradução Automática Inglês-Português; Raciocínio Espacial Qualitativo; Google Translate; DeepL.

Abstract

Keywords

Open-source LLMs, Neural Machine Translation, Spatial Semantics, Polysemy, Language Typology

1 Introdução

A Tradução Automática Neural (TAN) tornou-se o paradigma dominante na área de Tradução Automática, tanto em estudos acadêmicos quanto em aplicações práticas (**?**) (**??**). Esse avanço se deve, em grande parte, à capacidade aprimorada dos modelos de aprendizado profundo de captar dependências longas nas frases.

No entanto, apesar dos avanços, alguns tradutores automáticos ainda enfrentam desafios ao lidar com as nuances da linguagem espacial, como a polissemia das preposições e a projeção idiossincrática da maneira de movimento em inglês diretamente para verbos em português (). Um exemplo disso pode ser visto no Exemplo (1), retirado do Cambridge Online Dictionary (CAM), onde a tradução do inglês (EN) para o português (PT) foi realizada com o Google Translate (GT) e o DeepL (DL).

(1) He swam *across* the river. (CAM)
   a. ? Ele nadou do outro lado do rio. (GT)
      3SG.M swam from-the other side of-the river
   b. Ele atravessou o rio a nado. (DL)
      3SG.M crossed the river by swimming

A traducao do Exemplo **??** feita pelo modelo GT, embora gramaticalmente correta, erra ao não capturar a expressao mais natural em PT para a sentença em EN. O DL, por outro lado, acerta em cheio.

A razao por tras dessa traducao errada esta na polissemiada preposicao across, que pode significar tanto uma localizacao oposta fixa ao ponto de referencia quanto movimento de um lado de um espaco para o outro. Neste caso em particular, o significado pretendido e claramente o ultimo. Para ilustrar isso, vamos considerar as Figuras 1 e 2.

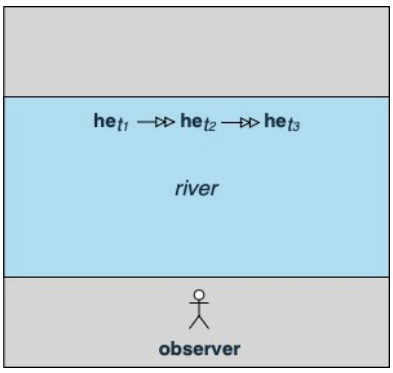

Figura 1: Diagrama semantico de (1)-a.

A Figura **??**, representando a saida GT, indica movimento dentro de um local especifico (uma

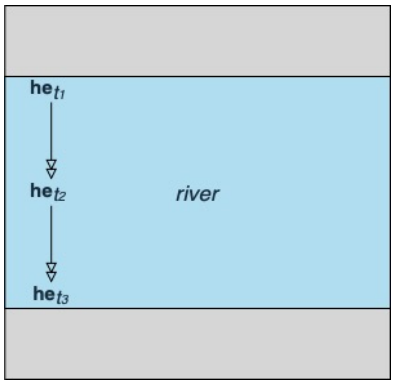

Figura 2: Diagrama semantico de (1)-b.

margem oposta do rio). No entanto, a Figura 2, representando a saida DL, transmite o significado de cruzar de uma margem do rio para a outra, capturando assim a natureza dinamica implicita na frase original EN.

Com isso em mente, este artigo explora a tradução automática de frases em EN que envolvem informações espaciais (topologia ou movimento) para PT, utilizando GT e DL. Nosso objetivo é duplo: primeiro, baseados nos trabalhos de Spranger et al. (2016), Freksa e Kreutzmann (2016) e Randell et al. (1992), formalizamos amostras de frases nas línguas de origem e destino. Em seguida, categorizamos as traduções para identificar erros comuns cometidos por ferramentas de TAN. Em vez de focar no processo de TAN em si, discutiremos os significados espaciais que essas ferramentas têm dificuldade em capturar, iluminando práticas e direções teóricas para pesquisa em linguagem espacial e TA. Nossos resultados mostram que, apesar do bom desempenho geral, os motores de TA ainda cometem erros sistemáticos em algumas categorias ao traduzir textos de EN para PT.

1.1 Desafios na Tradução da Espacialidade

A formatação ao longo do documento é a normal em documentos LaTeX, sem grandes alterações. No entanto, algumas sugestões:

- Para dar *ênfase* use sempre que possível o comando `\emph`;
- Para citar poderá usar o comando `\citep` que cria referências entre parêntesis (**?**). Para citar um **?**, use o comando `\citet`;
- Citações seguidas devem reaproveitar o comando de citação. Caso necessite de indicar a página a que se refere a citação, use (**?**, p. 40).
- Ao criar entradas bibliográficas assegure-se da correção do seu conteúdo. Não abrevie nomes de autores. Não coloque os nomes dos editores de livros de atas. Não se esqueça dos números das páginas do documento.
- Sempre que usar endereços web e outros tipos de URI, coloque-os com o comando `\url` e, sempre que possível, em nota de fim de página.[1]
- Nas notas de fim de página que sejam anexadas a palavras seguidas de pontuação, devem ser colocadas após a pontuação, como exemplificado no item anterior.
- Tenha em atenção a diferença entre -, – e —. O primeiro será usado entre palavras, como em curto-circuito, o segundo em intervalos, como 10–20 ou PT–EN e o terceiro — este — para introduzir pequenos comentários.
- As figuras devem ser legendadas e a legenda deve terminar com um sinal de pontuação.
- As referências a figuras, tabelas ou secções devem ser criadas usando as ferramentas do LaTeX.
- Sempre que possivel garanta a qualidade das imagens importadas, usando PDF ou PNG.
- Ao criar tabelas (Tabela **??**) tente diminuir a quantidade de traços usada. Grande parte das tabelas são legíveis apenas com um par de linhas como demonstrado.

	Homens	**Mulheres**
Crianças	10 032	32 341
Adultos	23 431	9 443

Tabela 1: Exemplo de tabela com poucos traços.

Agradecimientos

Os agradecimentos devem ser colocados sempre numa secção final, sem número, tal como neste exemplo. Sempre que o autor assim o entender, deverá agradecer aos revisores.

Referências

Dabre, Raj, Chenhui Chu & Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)* 53(5). 1–38.

[1] Assim. http://www.linguamatica.com

McCleary, Leland & Evani Viotti. 2004. Representação do espaço em inglês e português brasileiro: observações iniciais. *Revista da Anpoll* 1. doi:10.18309/anp.v1i16.552. https://revistadaanpoll.emnuvens.com.br/revista/article/view/552.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Yang, Shuoheng, Yuxin Wang & Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526* .