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), apesar de ser a abordagem dominante, ainda enfrenta desafios significativos ao traduzir conhecimento espacial. Neste estudo, utilizamos o Raciocínio Espacial Qualitativo (REQ) para representar e analisar informações espaciais em traduções automáticas do inglês para o português. Foram traduzidas 145 frases dos corpora CAM e COCA, utilizando os sistemas Google Translate e DeepL. Ao mapear logicamente as diferenças de significado entre as traduções e os originais, utilizando o REQ, identificamos que a TAN comete em média 10,6% de erros semânticos e 12% de erros de projeção sintática em traduções envolvendo conceitos espaciais. Nossos resultados evidenciam a necessidade de aprimorar os modelos de TAN para lidar com as nuances da linguagem espacial, contribuindo para o avanço da pesquisa em tradução automática.

brazil

tradução automática neural; semântica espacial; raciocínio espacial qualitativo; tradução automática inglês-português; polissemia; tipologia linguística

Abstract

Neural Machine Translation (NMT), currently the leading approach, still faces challenges in translating spatial knowledge. In this study, we used Qualitative Spatial Reasoning (QSR) to represent spatial information in automatic translations from English to Portuguese. We translated 145 sentences from the CAM and COCA corpora using Google Translate and DeepL, and identified the causes of unnatural translations. Using QSR, we logically mapped the differences in meaning. Our results indicate that, despite generally good performance, NMT struggles with specific spatial meanings, resulting in 10.6% semantic errors and 12.0% syntactic projection errors. This work explores the practical and theoretical challenges of machine translation.

Keywords

neural machine translation; spatial semantics; qualitative spatial reasoning; english-portuguese machine translation; 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, principalmente, à capacidade aprimorada dos modelos de aprendizado profundo em capturar dependências de longo alcance 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.
      3SG.M swam from-the other side of-the river
      (GT)

   b. Ele atravessou o rio a nado. (DL)
      3SG.M crossed the river by swimming

A tradução 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.

A razão por tras dessa traducao errada está na polissemia da 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.

A Figura **??** representa o GT, que indica mo-

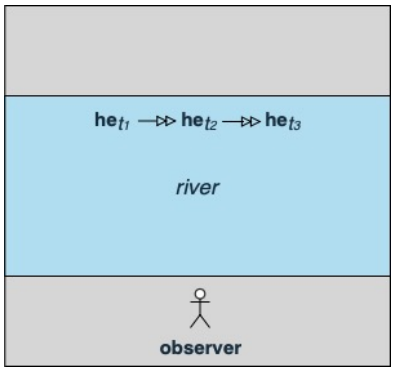

Figura 1: Diagrama semantico de (1)-a.

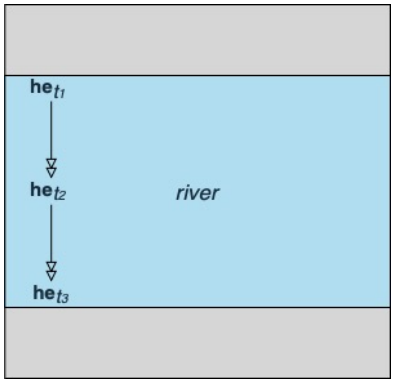

Figura 2: Diagrama semantico de (1)-b.

vimento dentro de um local especifico (uma margem oposta do rio). No entanto, a Figura **??**, 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

2 Metodologia

Nesta seção, apresentamos as etapas metodológicas do nosso trabalho, composta pela coleta dos dados, pela classificação das preposições, pelo processo de tradução, pelas formalizações das informações espaciais e pela categorização das traduções.

2.1 Coleta dos Dados

Para compor nosso corpus, compilamos 145 sentenças contendo cinco preposições em EN que transmitem conhecimento espacial: *across*, *into*, *to*, *through* e *via*. As sentenças foram obtidas do Cambridge Online Dictionary (CAM)[1] e do Corpus of Contemporary American English (COCA)[2]. Manualmente, anotamos cada sentença de acordo com o significado ligado às preposições, significado espacial e também atribuímos um identificador para o número da sentença. No Exemplo **??**, apresentamos uma das sentenças do corpus (amostra `Through-CAM-1-2`).

(2) He struggled through the crowd till he reached the front. (CAM)

a. ? Ele lutou no meio da multidão
3SG.M fought in-the middle of-the crowd
até chegar à frente. (GT)
until reach to-the front

b. ? Ele se debateu entre a multidão
3SG.M REFL struggled amongst the crowd
até chegar à frente. (DL)
until reach to-the front

2.2 Classificação das Sentenças

Sistematicamente, categorizamos cada sentença com base nos significados espaciais alinhados às entradas encontradas no CAM para cada preposição, como mostra a Tabela **??**.

[1] https://dictionary.cambridge.org/
[2] https://www.english-corpora.org/coca/

Preposição (EN)	Significado(s) Espaciais(s)
Across	(1) posição perpendicular (2) movimento de atravessar (3) localização oposta (4) em todas as partes de
Into	(1) movimento para um ponto não especificado de uma área ou contêiner (2) movimento até um ponto de contato com um obstáculo
Onto	(1) movimento sobre uma superfície sair de uma área delimitada
Through	(1) movimento de atravessar uma área de uma extremidade a outra (2) movimento de passar ou penetrar uma barreira
Via	(1) parte de uma rota

Tabela 1: Categorização das preposições *across*, *into*, *onto*, *through* e *via* baseada nas definições do CAM.

2.3 Tradução das Sentenças

As sentenças foram traduzidas com o Google Translate (GT)[3] e com o DeepL (DL)[4], utilizando suas versões publicamente disponíveis on-line em agosto-setembro de 2023. Além disso, para facilitar a comparação entre as traduções, fornecemos referências profissionais traduzidas por humanos para todas as sentenças.

2.4 Formalização das Sentenças

A partir do trabalho de **?**, definimos cada intervalo de tempo t como um conjunto de pontos e utilizamos o predicado *occurs_ in(θ, t)* para denotar que um evento θ ocorre durante um intervalo de tempo t. Com base em **?**, definimos os eventos θ por meio do cojunto de treze relações qualitativas espaço-temporais apresentadas na Figura **??**.

A Figura **??** apresenta as treze relações conjuntamente exaustivas e em pares disjuntivos baseadas no Cálculo de Intervalos de Allen (**?**). Essas relações podem ser descritas pelo seguinte cojunto: {*before, after, equal, meets, met by, overlaps, overlapped by, during, contains, starts, started by, finishes, finished by*}. Com esse cojunto de relações, podemos representar transições relacionadas ao movimento de objetos que fazem parte de um evento.

Por *default*, assumimos um espaço 3D para todos os objetos em nossas representações de movimento nas cenas. Para representar as informações espaciais em sentenças como as do Exemplo

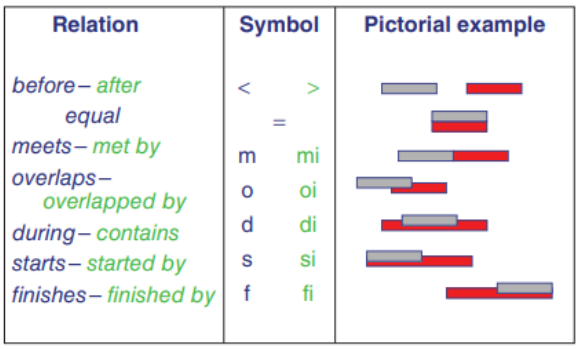

Figura 3: Treze relações qualitativas entre dois objectos lineares estendidos sobre uma reta orientada (**?**).

(1), em que a preposição "across" denota o movimento de atravessar uma superfície, definimos uma função *surface(r)*. Essa função mapeia relações como *during* ou *contains*, projetando um objeto em uma superfície 2D.

Para modelar relações mereotopológicas, nos baseamos no RCC-8 (**?**): $\{dc,\ ec,\ po,\ eq,\ tpp,\ ntpp,\ tpp^{-1},\ ntpp^{-1}\}$. Para formalizar uma sentença como a gerada pela tradução do GT no Example **??**, nós definimos uma Região de Referência (RR), que é uma parte de uma região R, ou Fundo, localizada fora da região onde a ação executada pelo objeto F, a Figura, acontece. A Região de Referência é separada do restante de R por uma linha transversal (chamada por nós de *meridiano*) que liga com R em dois pontos distantes (não consecutivos) e não toca $F : R_{op} = ntpp(F, R)$.

De modo a representar a relação entre o predicado *occurs_ in(θ, t)* e as relações qualitativas apresentadas na Figura **??**, nós utilizamos o conectivo , que denota uma implicação revogável, isto é, uma forma de raciocíno que é racionalmente convincente, mas carece de validade dedutiva. Nesse contexto, as premissas do argumento oferecem suporte racional para a conclusão, mas há a possibilidade de as premissas serem verdadeiras e a ser falsa. Em resumo, a conexão entre as premissas e a conclusão são provisórias e podem ser anuladas por informações suplementares.

2.5 Categorização das Traduções

Categorizamos as 145 sentenças traduzidas pelo GT e pelo DL (ou seja, 290 no total), comparando as traduções das preposições com seus respectivos significados apresentados na Tabela **??**. Para tanto, utilizamos as seguintes categorias: tradução (C)orreta, tradução equivocada de

[3] https://translate.google.com
[4] https://www.deepl.com/translator

(S)ignificado e erro de (P)rojeção na tradução. A última envolve principalmente a incorporação inadequada da maneira do movimento no verbo das sentenças traduzidas para o português, em vez de representar essa informação com adjuntos (veja, por exemplo, **??**).

3 Resultados e Discussão

Inicialmente, para sintetizar os resultados obtidos por meio da nossa análise formal, discutiremos as formalizações das sentenças nos Examplos **??** and **??**. As fórmulas apresentadas a seguir descrevem relações espaciais qualitativas entre as sentenças originais e suas respectivas traduções automáticas. Em ambas as formalizações, os intervalos de tempo foram representados por t_1, t_2, t_3, onde t_1 e t_3 correspondem aos intervalos inicial e final, respectivamente. Já t_2, por sua vez, representa um intervalo de tempo entre t_1 e t_3. A tabela **??** mostra as fórmulas que representam o Exemplo **??**.

Original text: He swam across the river.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(moves\_across(he, river), t)$ $river' = surface(river) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$
GT: Ele nadou do outro lado do rio.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(moves\_on\_opposite\_side(he, river_{op}), t)$ $river' = surface(river_{op}) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$
DL: Ele atravessou o rio a nado.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(moves\_across(he, river_{op}), t)$ $river' = surface(river_{op}) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$

Tabela 2: Formalizações das sentenças do Exemplo **??**.

A formalização na Tabela **??** nos permite representar a diferença lexical mencionada na Seção **??**. Na sentença original, a preposição *across* é categorizada como sentido (2) de acordo com CAM (Tabela **??**). Entretanto, a tradução da GT opta pelo sentido (3). A expressão "do outro lado" transmite o significado de que a ação executada pelo indivíduo, *ele*, ocorreu em uma porção do *rio* que é separada da *RR*, diferindo da região onde a ação ocorreu na sentença original, e alinhando-se com a tradução da DL.

Na formalização das sentenças do Exemplo **??**, apresentada na TabelaTable **??**, também é possível perceber diferenças qualitativas relacionadas a informação espacial. No exemplo, *through* é empregado no sentido (1) (da Tabela **??**).

Original text: He struggled through the crowd till he reached the front.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(arduously(moves\_through(he, crowd), t))$ $starts(he, crowd, t_1) \wedge$ $during(he, crowd, t_2) \wedge$ $finishes(he, crowd, t_3)$
GT: Ele lutou no meio da multidão até chegar à frente.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(fights(he, crowd) \wedge moves\_to(he, crowd), t)$ $starts(he, crowd, t_1) \wedge$ $during(he, crowd, t_2) \wedge$ $finishes(he, crowd, t_3)$
DL: Ele se debateu entre a multidão até chegar à frente.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs\_in(flounder(he, crowd) \wedge moves\_to(he, crowd), t)$ $starts(he, crowd, t_1) \wedge$ $during(he, crowd, t_2) \wedge$ $finishes(he, crowd, t_3)$

Tabela 3: Formalizações para as sentenças do Exemplo **??**.

Como podemos observar, as três relações $\langle starts, during, finishes \rangle$ foram aplicadas em todas as sentenças na Tabela **??**. Essa escolha reflete que a ação começou em um ponto de entrada da *multidão* e terminou em um ponto de saída. As distinções entre as frases do evento estão na maneira como a ação ocorreu.

A sentença original do Exemplo **??** descreve um evento em que o movimento da figura acontece de um ponto dentro ou além da multidão para uma extremidade, em direção à frente da multidão e realizado comdificuldade. Essa dificuldade é incorporada ao verbo *to struggle* do EN. De modo semelhante, os tradutores automáticos, GT e DL, optaram por traduzir a mesma informação usando verbos do PT como "lutar" (*to fight*) e "se debater" (*to flounder*), respectivamente, produzindo traduções que soam execessivamente exageradas. Uma tradução mais precisa seria "Ele atravessou a multidão *com dificuldade* até chegar à frente." Nessa versão, o ato de atravessar é expresso por "atravessar", enquanto a dificuldade da ação é transmitida por "com dificuldade".

Para representar essa distinção nas formalizações, introduzimos um predicado de segunda ordem (*arduously*). Na Tabela **??**, o esforço envolvido na execução da ação vinculada ao verbo *lutar* é expresso pelo predicado *arduously*. Em contraste, as expressões "lutou no meio de" e "se debateu entre" denotam eventos individualizados.

Por fim, as formalizações na Tabela **??** apresentam os desafios que o GT e o DL enfrentaram ao traduzir as informações sobre a maneira dos eventos, resultando em erros de tradução. A Tabela **??** resume a avaliação de cada categoria en-

contrada em nossa análise: tradução (C)orreta, tradução equivocada de (S)ignificado e erro de (P)rojeção na tradução.

	Correta	**Significado**	**Projeção**
GT	106 (73.1%)	19 (13.1%)	20 (13.8%)
DL	118 (81.4%)	12 (8.3%)	15 (10.3%)
Total	**224 (77.2%)**	**31 (10.6%)**	**35 (12.0%)**

Tabela 4: Categorização da performance do GT e do DL.

A tabela **??** mostra que o DL superou o GT na geração de traduções corretas. O DL traduziu corretamente 118 sentenças (81,4%), enquanto o GT traduziu 106 (73,1%). O DL também apresentou menos erros de significado (8,3%) e erros de projeção (10,3%) em comparação com o GT, que teve 19 (13,1%) e 20 (13,8%), respectivamente. Os erros de significado referem-se a situações em que o tradutor automático gera uma frase gramaticalmente correta que não transmite o significado original. Por exemplo, ao traduzir *across*, independentemente de seu significado, o GT escolheu predominantemente a tradução "do outro lado". Os erros de projeção, por sua vez, são influenciados por padrões distintos de lexicalização de informações espaciais em EN e PT, conforme descrito em **??**.

4 Conclusão

Neste trabalho, analisamos 145 sentenças que descrevem relaçõs espaciais coletadas dos corpora CAM e COCA e traduzidas do EN para o PT pelo Google Translate e pelo DeepL. Utilizando métodos QSR, formalizamos as informações espaciais para destacar as diferenças nas relações qualitativas entre as frases de origem e de destino. Também analisamos as traduções de todas as sentenças e verificamos que os erros de sentido relacionados à polissemia e os erros de projeção sintática desafiam a tradução automática.

Para tornar nossos resultados mais consistentes, seria interessante (i) formalizar mais exemplos; (ii) testar computacionalmente as formalizações; e (iii) analisar traduções automáticas para outros idiomas de destino. Uma limitação óbvia dessas sugestões é que elas consomem muito tempo, pois todas as formalizações devem ser feitas manualmente. Além disso, reconhecemos a dificuldade de desenvolver métodos automáticos para representar logicamente a linguagem espacial e incorporar camadas formais aos modelos de tradução automática neural. De modo geral, esperamos que este trabalho destaque as questões práticas e teóricas da TAN.

5 Agradecimentos

Referências

Allen, James F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11). 832–843.

Freksa, Christian & Arne Kreutzmann. 2016. Neighborhood, conceptual. *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology* 1–12.

Randell, David A, Zhan Cui & Anthony G Cohn. 1992. A spatial logic based on regions and connection. Em *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, vol. 92, 165–176.

Spranger, Michael, Jakob Suchan & Mehul Bhatt. 2016. Robust natural language processing-combining reasoning, cognitive semantics and construction grammar for spatial language. *arXiv preprint arXiv:1607.05968* .