

Desafios na Tradução de Informação Espacial do Inglês para o Português

Challenges in Translating Spatial Information from English to Portuguese

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

Os sistemas de Tradução Automática Neural (TAN), atualmente a abordagem mais utilizada na Tradução Automática, ainda enfrentam desafios ao lidar com a tradução da linguagem espacial. Neste estudo, utilizamos o Raciocínio Espacial Qualitativo (REQ) para representar informações espaciais nas 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 do bom desempenho geral, os motores TAN apresentam 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.

brazil

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 (TA), 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 de serem bastante eficien-

tes, alguns tradutores automáticos ainda enfrentam desafios ao lidar com as sutilezas da linguagem espacial, como a polissemia das preposições e a projeção idiossincrática da maneira do movimento em inglês diretamente em verbos no português (?). Um exemplo disso pode ser observado 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).

1.1 Desafios na Tradução da Espacialidade

A formatação ao longo do documento é a normal em documentos L^AT_EX, 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.

¹Assim. <http://www.linguamatica.com>

- 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 L^AT_EX.
- Sempre que possível garanta a qualidade das imagens importadas, usando PDF ou PNG.
- Ao criar tabelas (Tabela 1) 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.

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

2.2 Classificação das Preposições

2.3 Tradução das Sentenças

2.4 Formalização das Sentenças

A partir do trabalho de Spranger et al. (2016), definimos cada intervalo de tempo t como um conjunto de pontos e utilizamos o predicado $occurs_in(\theta, t)$ para denotar que um evento θ ocorre durante um intervalo de tempo t . Com base em Freksa & Kreutzmann (2016), definimos os eventos θ por meio do conjunto de treze relações qualitativas espaço-temporais apresentadas na Figura 1.

A Figura 1 apresenta as treze relações conjuntamente exaustivas e em pares disjuntivos baseadas no Cálculo de Intervalos de Allen (Allen, 1983). Essas relações podem ser descritas pelo seguinte conjunto: $\{before, after, equal, meets,$

Relation	Symbol	Pictorial example
<i>before</i> – <i>after</i>	< >	
<i>equal</i>	=	
<i>meets</i> – <i>met by</i>	m mi	
<i>overlaps</i> – <i>overlapped by</i>	o oi	
<i>during</i> – <i>contains</i>	d di	
<i>starts</i> – <i>started by</i>	s si	
<i>finishes</i> – <i>finished by</i>	f fi	

Figura 1: Treze relações qualitativas entre dois objectos lineares estendidos sobre uma reta orientada (Freksa & Kreutzmann, 2016).

met by, overlaps, overlapped by, during, contains, starts, started by, finishes, finished by}. Com esse conjunto 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 (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 (Randell et al., 1992): $\{dc, ec, po, eq, tpp, ntp, tpp^{-1}, ntp^{-1}\}$. 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 $F : R_{op} = ntp(F, R)$.

De modo a representar a relação entre o predicado $occurs_in(\theta, t)$ e as relações qualitativas apresentadas na Figura 1, nós utilizamos o conectivo \Leftarrow , que denota uma implicação revogável, isto é, uma forma de raciocínio 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

3 Resultados e Discussão

Original text: He swam <u>across</u> the river.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs_in(moves_across(he, river), t) \leftarrow$ $river' = surface(river) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$
GT: Ele nadou <u>do outro lado</u> do rio.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs_in(moves_on_opposite_side(he, river_{op}), t) \leftarrow$ $river' = surface(river_{op}) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$
DL: Ele <u>atravessou</u> o rio a nado.
$\forall t \in \{t_1, t_2, t_3\}, t_1 < t_2 < t_3$ $occurs_in(moves_across(he, river_{op}), t) \leftarrow$ $river' = surface(river_{op}) \wedge$ $starts(he, river', t_1) \wedge$ $during(he, river', t_2) \wedge$ $finishes(he, river', t_3)$

Tabela 2: Formalizations for sentences in Example ??.

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

Tabela 3: Formalizations for sentences in Example ??.

4 Conclusão

Agradecimentos

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

- 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*.