Detecção automática de fatores (potenciais) no texto de origem levam a preconceito de gênero em traduções de maquinas.

Resumo

Este projeto de pesquisa mira em desenvolver uma metodologia compreensiva para ajuda fazer sistemas de Maquinas de Tradução (MT) mais inclusivas com gêneros para a sociedade. O objetivo é a criação de um sistema de detecção, um modelo de "aprendizado de máquina" (ML) treinado em anotações manuais, que podem automática analisar a fonte de origem e detectar e destaca palavras e frases que influenciam a inflexão o preconceito de gênero em alvos de tradução.

1 - Créditos

Este projeto é uma estratégia básica de pesquisa de PHD totalmente financiada pela The Research Foundation - Flanders (FWO) pelo intervalo de tempo de 4 anos, de 1/11/2023 até 31/10/2027, e liderada com o Time de Tecnologia de Linguagem e Tradução (LT3) na Universidade de Ghent.

2 - Introdução

Com o aumento do uso e interesse pelo desenvolvimento de Máquinas de Tradução (MT) e o crescimento da demanda pela inclusividade de genero na sociedade, as pesquisas sobre preconceito de genero em MTs está aumentando (Savoldi et al., 2021). Um sistema de MT é considerado preconceituoso se "sistematicamente e injustamente discriminar contra certos indivíduos ou grupos em favores de outros" (Savoldi et al., 2021, p. 846), perpetuando "imprecisão e potencial de discriminação de estereótipo" na sociedade (Vanmassenhove, 2024, p. 3).

Os estudos ao níveis de palavras mostram que a incorporação de palavras utilizadas na formação em MT são altamente flexionados em função do gênero, onde a incorporação de palavras de diferentes partes do discurso (POS) se agrupam fortemente com base no seu gênero em domínios variados (por exemplo, nos esportes, na cozinha, na grande tecnologia, em profanações sexuais) (Caliskan et al., 2022). No entanto, a investigação sobre os agrupamentos de palavras limita-se à análise ao nível das palavras e ainda não foi alargada ao

contexto completo das frases de origem em língua natural e às associações sistemáticas de gênero daí resultantes nas traduções automáticas.

3 Descrição do Projeto

Neste projeto de pesquisa, aplicamos uma nova abordagem para analisar caraterísticas linguísticas e orfológicas alinhadas, com enfoque no gênero, em textos de partida e de chegada, e aplicamos metodologias de aprendizagem automática. O inglês é escolhido como língua de partida, onde os nomes dos papéis não são geralmente marcados com um gênero (por exemplo, professor) e as traduções alvo são analisadas em alemão e espanhol, línguas de gênero gramatical (Savoldi et al., 2021), onde o gênero é claramente marcado (por exemplo, Lehrer/Lehrerin). O objetivo é identificar e classificar automaticamente as “palavras-gatilho” num contexto de origem que influenciam a inflexão gramatical de gênero na tradução de destino (ou seja, se uma MT traduz uma pessoa como feminina ou masculina, ou se, em vez disso, opta por neutralizar ou reformular a palavra). O projeto consiste nos seguintes resultados principais que diferem da investigação anterior: (1) um conjunto de dados anotados manualmente de associações de gênero humano em contextos de frases, (2) uma taxonomia baseada nestas anotações, (3) uma análise comparativa das associações de gênero humano vs. as inflexões de gênero da MT, e (4) um modelo linguístico de grande dimensão (LLM) aperfeiçoado que destaca as palavras desencadeadoras de gênero num texto de partida.

3.1 Coleta de Dados

A primeira etapa é a coleta de uma lista de palavras candidatas (nomes de cargos, por exemplo, *amigo*) incluindo inflexão de gênero, como amostra das suas palavras em estudos anteriores. Essas palavras são usadas para filtrar dados monolinguísticos em Inglês de diferentes domínios, ligeiramente assemelhando o metodologia possuída por Ondoño-Soler e Forcada (2022).
Seguindo o filtro automático, as sentenças em inglês são filtradas manualmente em um nível monolinguístico para selecionar casos de ambiguidade de gênero de termos da palavra singularmente candidata. Nós miramos por 2000 á 5000 frases para otimizar o modelo.

Próximo, os dados filtrados serão maquinalmente traduzidos para alemão e espanhol com kits MT publicamente disponíveis. Nós documentamos e comparamos entre qual gênero o Sistema MT traduz cada palavra candidata, primeiro em um nível de palavra (Exemplo: O termo individual *friend*) e depois em um contexto de frase (Exemplo: Depois um amigo sugeriu que ela experimentasse, Ann disse, "Claro!"). As duplas bilíngues (EN-DE, EN-ES) serão alinhados e enriquecidos ao nível da palavra pela informação morfossintáctica.

3.2 Análise e Anotação de Dados

No próximo passo, os dados ambíguos em inglês serão manualmente anotados para analisar como o contexto influencia em associações de gênero. Dessas anotações, nós podemos comparar até que ponto as associações de gênero humanas se sobrepõem à escolha de gênero gramatical de um sistema de MT em uma linguagem alvo. Para cada sentença, anotadores serão solicitados para anotar que contexto influencia suas associações de genero. Especialmente, eles serão solicitados para anotar qualquer palavras ou frases chaves (por exemplo, uma referência, localização ou qualquer POS) que considerem pessoalmente influenciar a inflexão de gênero de uma palavra candidata nessa frase. As anotações serão verificadas e classificadas e a partir disso, a taxonomia irá ser criada. Um caso com 22 anotadores de diferentes gêneros já estava realizado para avaliar as orientações de anotação, a concordância dos anotadores e a influencia do gênero. (Hackenbuchner et al., sobre revisão).

Em seguida, nós vamos analisar funcionalidades morfossintáticas de dados que estão usando ferramentas automatizadas. A combinação de toda informação morfossintática irá revelar padrões entre palavras chaves e palavras candidatas em questão. Baseado nesta análise, nós queremos excluir palavras chaves "irrelevantes", nos permitindo focar em palavras "relevantes" ou frases que tem a maior influência no gênero do candidato.

3.3 Otimização e avaliação do Modelo

No nosso último passo nós iremos aplicar o aprendizado de máquina (ML) ajustando um LLM baseado em nossos dados, anotados e verificados, e pela informação extraída da análise morfossintática. Procurando padrões de estudos anteriores que conduzem ao preconceito de gênero na MT, com este modelo otimizado nós pretendemos detectar automaticamente palavras ou frases chaves de gênero em um texto fonte e destacá-las.

4 Projetos alinhados

Em conjunto com a pesquisa deste projeto, os colegas do PHD, primeiro autor deste papel, é um co-fundador e membro do DeBiasByUs e co-organizador dos dois WorkShops internacionais sobre Tecnologias de Tradução Inclusivas em termos de gênero.

5 Conclusão

Nosso Benchmark pode sensibilizar os utilizadores de MT para as inflexões de gênero dos textos de partida que são traduzidos automaticamente, apoiar tecnologicamente os tradutores na pós-edição dos resultados da MT, orientar desenvolvedores dos sistemas de

MT para questões persistentes de preconceitos de gênero, e ajudar os criadores de conteúdos a identificar potenciais chaves no texto que possam levar a traduções com preconceitos de gênero.