UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE CIÊNCIAS MATEMÁTICAS E CIÊNCIAS COMPUTACIONAIS

SCC0633/SCC5908 - Processamento de Linguagem Natural, 2024

- **Nome dos integrantes do grupo:**
  - Arthur Santorum Lorenzetto, NUSP: 12559465
  - Bruno Berndt Lima, NUSP: 12542550
  - Eduardo Z. Monteiro, NUSP: 12559490
  - Pedro Henrique de Freitas Maçonetto, NUSP: 12675419
  - Vinicius Kazuo Fujikawa Noguti, NUSP: 11803121

- **Tópico do trabalho:**
  - Tradução Automática.

- **Lista de córpus escolhidos para o trabalho:**

  - https://www.kaggle.com/datasets/ramakrishnan1984/785-million-language-translation-database-ai-ml

Proposta:

1. **Introdução e Motivação**

Com a invenção da internet, as pessoas de hoje em dia têm acesso a conteúdos produzidos em todos os locais do mundo, mas muitas vezes pode não ser fácil entender o que se está querendo comunicar, já que pode haver uma barreira linguística entre o autor e o consumidor desses conteúdos.

Muito antes disso, os pesquisadores de PLN já buscavam um sistema capaz de fazer tradução entre duas línguas, com o objetivo de baratear e acelerar o processo de tradução, mas com a internet, hoje em dia é ainda mais interessante o desenvolvimento de um sistema do tipo, dado que ele pode ser utilizado diariamente para compreender conteúdos encontrados em línguas que o indivíduo não entende.

A princípio o tradutor proposto aqui é um tradutor da língua portuguesa para a língua inglesa, essa decisão foi tomada pois inglês é a língua que os autores têm mais familiaridade, e portanto é mais fácil avaliar as traduções geradas pelo sistema.

2. **Corpus Utilizado**

Inicialmente será utilizado apenas um corpus, pois atende as demandas necessárias do grupo, sendo que futuramente um segundo ou terceiro corpus pode ser utilizado.

O corpus escolhido trata-se de traduções de mais de 700 milhões de palavras do inglês para mais de 500 línguas diferentes. Vale ressaltar que utilizaremos

apenas a tradução Inglês-Português, portanto a gama de palavras traduzidas válidas pode ser menor que a mencionada, já que algumas palavras são exclusivas de uma determinada língua, não possuindo tradução direta.

Alguns exemplos de tradução do corpus:

Inglês	Português
happiness	felicidade
plant hormone	fitormônio
noodle	talharim
international air transport association airport code	código aeroportuário IATA

Aqui temos 4 diferentes exemplos de tradução, sendo:

1. Uma tradução simples e correta.
2. Uma conversão do termo “plant” para o “fito” e a aglutinação das palavras.
3. Uma tradução não convencional para o termo, o usual seria “Macarrão”.
4. Uma abreviação de um termo.

Portanto é evidente que as traduções podem não ser as ideias, porém serão o suficiente para o projeto em questão.

3. O Sistema Simbólico

O sistema simbólico proposto pelo grupo consiste em grande parte do uso de um conjunto de dados de dicionário, que contém uma palavra em português e sua(s) equivalente(s) em inglês. Em posse dos dados, o modelo pretende realizar a tradução do texto entrada pelo usuário ‘palavra por palavra’, ou seja, traduzir as palavras individualmente. Em seguida, o texto formado por palavras em inglês passará por uma série de regras baseadas na gramática da língua inglesa, essas regras serão aplicadas a cada frase individualmente, esse processo tem como objetivo traduzir as especificidades do português para as do inglês, um exemplo típico é a regra do uso de adjetivos, que em português são tipicamente encontrados após o substantivo, enquanto que em inglês eles tendem a aparecer antes. As regras utilizadas são descritas na seção 5.

4. Classificação Gramatical e Tempos verbais

A identificação dos tipos morfológicos das palavras será feita utilizando um tagger. Já a identificação das pessoas e tempo verbais será feita por um lematizador verbal, ambos da CINTIL-UPos corpus.

https://portulanclarin.net/workbench/lx-parser/ - Tagger indicado pelo Germano

- O LX-Parser é um analisador sintático de constituição para o Português baseado numa abordagem estatística.

https://portulanclarin.net/workbench/lx-lemmatizer/ - Lematizador usado pelas regras simbolicas

- LX-Lemmatizer é um serviço online disponível gratuitamente para lematização completa de verbos portugueses.
- O LX-Lemmatizer assume uma forma verbal em português e entrega todos os lemas correspondentes (formas infinitivas) juntamente com os valores dos traços flexionais. Os lemas menos prováveis, mas ainda ortograficamente possíveis, são agrupados em uma última seção sob o título "Outros lemas possíveis".

5. Regras

Aqui serão apresentadas as regras utilizadas no modelo de tradução após a aplicação da tradução inicial, é evidente que essas regras causarão erros na tradução. Entretanto, elas foram escolhidas com o propósito de melhorar a tradução na maioria dos casos em que aplicadas. Dada a quantidade de erros possíveis, não é possível pensar ou listar todos, dito isso, os principais erros identificados estão documentados.

5.1 Substituição de adjetivos

Sempre que for encontrado pelo tagger o par Substantivo - Adjetivo em ordem, será feita a inversão dessas palavras na tradução final. Essa regra serve para compreender a diferença gramatical entre o português o inglês, em que no português tipicamente os adjetivos aparecem depois dos substantivos, enquanto que no inglês eles ocorrem antes, a seguir um exemplo de substituição bem sucedida da regra:

O menino chutou a bola azul → The boy kicked the blue ball

5.2 Correção de estrutura de perguntas

É característica da língua inglesa a presença de verbos auxiliares no começo de perguntas, são eles: ‘do’, ‘did’, ‘does’. Já no português esses verbos não existem, portanto uma vez feita a tradução, será necessário fazer a adição no início da sentença, essa adição será feita sempre que for identificada frase terminada em

interrogação que não contém um dos seguintes verbos: ‘sou’, é’, ‘posso’, ‘pode’, ‘sei’, ‘sabe’, esses verbos serão tratados na próxima regra. A adição do verbo auxiliar ‘do’ será feita quando a pergunta original começar com uma das seguintes palavras: ‘Você’, ‘eles’, ‘elas’. Segue um exemplo:

Você joga basquete? → Do you play basketball?

Já o verbo auxiliar ‘does’ ocorrerá quando a pergunta começar com uma das palavras: ‘ele’, ‘ela’, ‘isso’, da mesma forma que a substituição acima:

Ele joga basquete? → Does he play basketball?

Finalmente, o auxiliar ‘did’ será adicionado quando for encontrada uma palavra na frase traduzida terminada em ‘ed’, nesse caso além da adição o sufixo será removido. Essa regra tomará prioridade sobre as 2 anteriores, retomando o exemplo acima:

Você jogava basquete? → Did you played basketball?

É importante notar que essa regra poderá causar problemas na tradução quando ocorrem palavras como: Fred, LED, ou outras terminadas em ‘ed’.

5.3 Correção de perguntas com ‘sou’, ‘é’, ‘posso’, ‘pode’, ‘sei’, ‘sabe’

Quando traduzindo perguntas do português para o inglês, é necessário aplicar uma correção especial para casos em que a pergunta contenha verbos como 'sou', 'é', 'posso', 'pode', 'sei' ou 'sabe'. Esses verbos não têm equivalentes diretos em inglês e, portanto, requerem uma abordagem específica para garantir uma tradução precisa e natural.

Exemplos:
Você sabe nadar? → Can you swim?

Posso pegar emprestado a caneta? → Can I borrow the pen?

Regra:
Quando uma pergunta em português contém um dos verbos especiais ('sou', 'é', 'posso', 'pode', 'sei', 'sabe'), a tradução para o inglês deve seguir as seguintes diretrizes:

Você sabe? → Can you?
Posso? → Can I?
Sou? → Am I?

5.4 Negação de frases afirmativas

Assim como na estrutura de perguntas, é característica do inglês usar verbos auxiliares (do, does, did) em frases de negação. E como já explicado, será necessário fazer a adição desses verbos na sentença. Essa adição será feita sempre que for encontrado um advérbio de negação (não → not), combinando o verbo auxiliar com a pessoa verbal. A negação do presente simples e do passado simples de todos os verbos principais (exceto 'be' e alguns usos de 'have' como verbos principais) é feita com o auxiliar 'do' + 'not', que pode ser abreviado para don't (do not), doesn't (does not) e didn't (did not). Segue exemplos:

Para 1°, 2° pessoa do singular, e 1°, 2°, 3° pessoa do plural, será utilizado 'do':

Eu não gosto de café → I do not like coffee → I don't like coffee

Para 3° pessoa do singular, será utilizado 'does':

Ela não quer ir → She does not want to go→ She doesn't want to go

Sempre que a frase estiver no passado, será utilizado 'did' para qualquer pessoa do singular ou plural:

Eles não gostaram do filme → They did not like the movie → They didn't like the movie

5.5 Frases com verbos no infinitivo:

A forma infinitiva é a forma mais básica de um verbo. Não tem tempo verbal e não está vinculado a nenhum sujeito de uma frase. Para identificar a forma infinitiva de um verbo conjugado no português, é preciso relacioná-lo com uma das três terminações dos verbos no infinitivo: -ar, -er, -ir (e.g. estudar, aprender, partir). Em inglês, o infinitivo do verbo é formado pela adição do marcador de infinitivo 'to' antes do verbo base. Portanto, ao ser observado um verbo no infinitivo no português, ao fazer a tradução para o inglês é necessário adicionar a partícula 'to' antes do verbo. Segue exemplos para os 3 tipos de terminações verbais:

O garoto adora jogar football → The boy loves to play football

Eu quero ser desenvolvedor de software → I want to be a software developer

Ele precisa dormir → He needs to sleep

5.6 Frases com sujeitos com nomes próprios:

O inglês e o português possuem diferenças marcantes no que diz respeito a nomes de pessoas e instituições. Sendo assim, a fim de realizar uma tradução mais fidedigna e que não comprometa o sentido semântico da sentença, todo e qualquer substantivo próprio (simples ou composto), ou seja, nomes que possuem letras maiúsculas, serão mantidos em sua forma original. Dessa forma, é necessário garantir que a regra dos nomes próprios seja aplicada antes da regra dos adjetivos (5.1). Seguem os exemplos abaixo:

Substantivos próprios compostos:

Ponte Alta possui 7000 habitantes → Ponte Alta has 7000 inhabitants.

Venha para a Lagoa Azul aproveitar esse feriado → Come to Lagoa Azul enjoy this holiday.

Substantivos próprios simples:

A notícia foi publicada em um jornal na Globo → The news was published in a newspaper on Globo