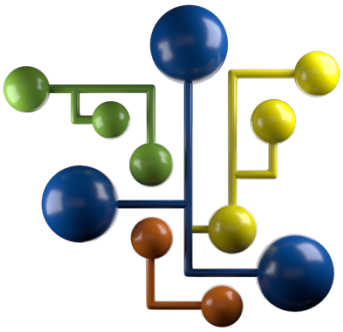


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Sistemas de Tradução Automática e Tradução Estatística

A tradução automática é a tradução de texto de uma língua natural (a fonte) para outra (o destino). Foi uma das primeiras áreas de aplicação previstas para computadores, mas somente na década passada a tecnologia teve seu uso disseminado de forma ampla. Aqui está uma passagem em inglês:

AI is one of the newest fields in science and engineering. Work started in earnest soon after World War II, and the name itself was coined in 1956. Along with molecular biology, AI is regularly cited as the "field I would most like to be in" by scientists in other disciplines.

Tradução para o português:

A IA é um dos campos mais recentes da ciência e da engenharia. O trabalho começou para valer logo após a Segunda Guerra Mundial, e o próprio nome foi cunhado em 1956. Junto com a biologia molecular, a IA é regularmente citada como o "campo onde eu mais gostaria de estar" por cientistas de outras disciplinas.

E aqui está a tradução de inglês para dinamarquês por uma ferramenta on-line, o Google Translate:

AI er en af de nyeste områder inden for videnskab og teknik. Arbejde startede for alvor lige efter Anden Verdenskrig, og navnet i sig selv var opfundet i 1956. Sammen med molekylær biologi, er AI jævnligt nævnt som "feltet Jeg ville de fleste gerne være i" af forskere i andre discipliner.

Para quem não entende dinamarquês, aqui está o dinamarquês traduzido de volta para o inglês. As palavras que apareceram diferente estão em itálico:

AI is one of the newest fields *of* science and engineering. Work *began* in earnest *just* after the *Second* World War, and the name itself was *invented* in 1956. *Together* with molecular biology, AI is *frequently mentioned* as -- "field I would most like to be in" by *researchers* in other disciplines.

As diferenças são todas paráfrases razoáveis, tal como *frequently mentioned* em vez de *regularly cited*. O único erro real é a omissão do artigo the, indicado pelo símbolo -- . Isso é precisão típica: das duas sentenças, uma tem um erro que não teria sido feito por um falante nativo, mas o significado foi transmitido claramente.

Historicamente, houve três principais aplicações da tradução automática. A tradução tosca, como a prevista por serviços on-line gratuitos, fornece a "essência" de uma sentença ou documento estrangeiro, mas contém erros. A tradução pré-editada é utilizada por empresas para publicar sua documentação e materiais de vendas em vários idiomas. O texto-fonte original está escrito em uma linguagem limitada que é mais fácil de traduzir automaticamente, e os resultados são editados geralmente por um ser humano para corrigir erros eventuais. A

tradução com restrição de origem funciona de forma totalmente automática, mas apenas em linguagem altamente estereotipada, como um boletim meteorológico.

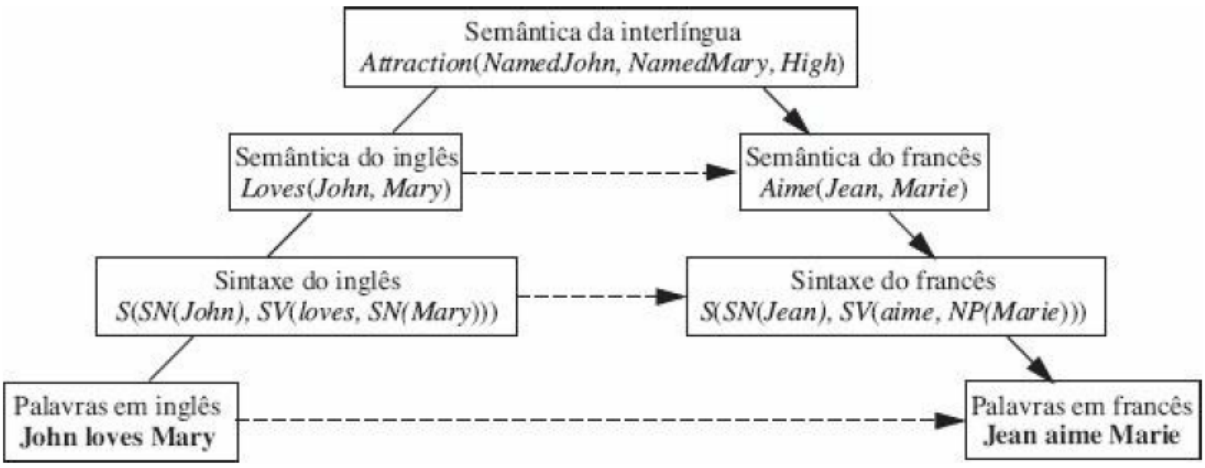
Tradução é difícil porque, no caso mais geral, requer conhecimeto profundo do texto. Isso é verdade mesmo para os textos muito simples – "textos" de uma palavra. Considere a palavra "Open" ("Aberta") em uma grande faixa com dizeres na parte de fora de uma loja construída recentemente. Significa que a loja está agora em operação diária, mas os leitores desse sinal não se sentiriam enganados se a loja fechasse durante a noite sem que a faixa fosse retirada. Os dois sinais utilizam a palavra idêntica para transmitir significados diferentes. Em alemão, o sinal na porta seria "Offen", enquanto na faixa seria lido "Neu Eröffnet".

O problema é que linguagens diferentes categorizam o mundo de forma diferente.

Por exemplo, a palavra francesa "doux" abrange ampla gama de significados, correspondendo aproximadamente às palavras em inglês "soft" ("suave"), "sweet" ("doce") e "gentle" ("gentil"). Da mesma forma, a palavra em inglês "hard" abrange praticamente todos os usos da palavra alemã "hart" (fisicamente recalcitrante, cruel) e alguns usos da palavra schwierig" (difícil). Portanto, representar o significado de uma sentença é mais difícil para a tradução do que é para a compreensão única do idioma. Um sistema de análise em inglês poderia usar predicados como Open(x), mas para a tradução a linguagem da representação teria de fazer mais distinções, talvez com Open1(x) representando o sentido de "Offen" e Open2(x) representando o sentido de "Neu Eröffnet". **Interlíngua** é a representação da linguagem que faz todas as distinções necessárias para um conjunto de línguas.

Um tradutor (humano ou máquina), muitas vezes precisa compreender a situação real descrita na origem, e não apenas as palavras individuais. Por exemplo, para a tradução da palavra inglesa "him" ("lhe", "a ele") em coreano a escolha deve ser feita entre a forma humilde e a honorífica, uma escolha que depende da relação social entre o falante e o referente de "him". Em japonês, os honoríficos são relativos, então a escolha depende das relações sociais entre o falante, o referente e o ouvinte. Os tradutores (tanto máquina como humanos), por vezes, acham difícil fazer essa escolha. Em outro exemplo, para traduzir "The baseball hit the window. It broke" para o francês, devemos escolher o feminino "elle"ou o masculino "il" para "it"(ele ou ela em inglês), por isso temos de decidir se "it" refere-se ao beisebol ou à janela. Para obter a tradução correta, é preciso entender a física, bem como a linguagem. Às vezes não há escolha que possa render uma tradução completamente satisfatória. Por exemplo, um poema de amor em italiano que utiliza o masculino "il sole" (sol) e o feminino "la luna" (lua) para simbolizar dois amantes terá de ser necessariamente alterado quando traduzido para o alemão, no qual os gêneros são invertidos, e ainda mais alterado quando traduzido para uma linguagem em que os gêneros são os mesmos.

Todos os sistemas de tradução devem modelar os idiomas de origem e de destino, mas os sistemas variam no tipo de modelos que utilizam. Alguns sistemas tentam analisar todo o texto do idioma de origem em uma representação do conhecimento interlíngua e geram sentenças na linguagem-alvo a partir dessa representação. Isso é difícil porque envolve três problemas não resolvidos: criação de uma representação do conhecimento completa de tudo; análise dessa representação; e geração de sentenças dessa representação. Outros sistemas são baseados em modelo de transferência. Eles mantêm um banco de dados de regras de tradução (ou exemplos) e, sempre que a regra (ou exemplo) combinam, traduzem diretamente. A transferência pode ocorrer no nível lexical, sintático ou semântico. Por exemplo, uma regra estritamente sintática faz o mapeamento do inglês [adjetivo substantivo] para o francês [substantivo adjetivo]. A mistura da regra sintática e lexical faz o mapeamento do francês [S1 “et puis” S2] para o inglês [S1 “and then” S2]. A abaixo faz um diagrama dos vários pontos de transferência.



Este é o triângulo de Vauquois: diagrama semântico esquemático das escolhas de um sistema de tradução automático (Vauquois, 1968). Inicia-se com um texto em inglês no topo. Um sistema baseado em interlíngua segue as linhas contínuas, analisando primeiro o inglês em uma forma sintática, em seguida em uma representação semântica e em uma representação interlíngua, e depois através da geração de uma forma semântica, sintática e lexical em francês. Um sistema baseado em transferência utiliza as linhas tracejadas como atalho. Sistemas diferentes fazem a transferência em pontos diferentes: alguns o fazem em vários pontos.

Agora que vimos como pode ser complexa a tarefa de tradução, não deve ser surpresa que os sistemas de tradução automática de maior sucesso são construídos pelo treinamento de um modelo probabilístico utilizando estatísticas recolhidas a partir de um corpus de texto grande. Essa abordagem não precisa de uma ontologia complexa de conceitos interlíngua nem precisa de gramáticas artesanais das linguagens de origem e destino, nem de uma floresta sintática rotulada à mão. ***Tudo o que precisa são de dados*** − amostras de traduções de onde um modelo de tradução pode ser aprendido. Para traduzir uma sentença, digamos, do inglês (e) para o francês (f), encontramos a cadeia de palavras f* que maximiza

$$f^* = \underset{f}{\text{argmax}}\, P(f \mid e) = \text{argmax}\, P(e \mid f)\, P(f)$$

Aqui o fator P(f) é o modelo do idioma alvo para o francês; informa a probabilidade de dada sentença em francês. P(e | f) é o modelo de tradução, que informa a probabilidade de que uma sentença em inglês seja a tradução de determinada sentença em francês. Da mesma forma, P(f | e) é um modelo de tradução do inglês para o francês. Devemos trabalhar diretamente em P(f | e) ou aplicar a regra de Bayes e trabalhar em P(e | f) P(f)?

Em aplicações de diagnóstico como a medicina, é mais fácil modelar o domínio na direção causal: P(sintomas | doença), em vez de P(doença | sintomas). Mas, na tradução, as duas direções são igualmente fáceis. Os primeiros trabalhos de tradução automática estatística aplicaram a **Regra de Bayes** — em parte porque os pesquisadores tinham um modelo bom de linguagem, P(f), e queriam fazer uso dele, em parte porque eles vinham de um conhecimento em reconhecimento de voz, que é um problema de diagnóstico. Seguimos a sua orientação neste capítulo, mas notamos que os trabalhos recentes em tradução automática estatística, muitas vezes otimizam o P(f | e) diretamente, utilizando um modelo mais sofisticado que leva em conta muitas das características do modelo de linguagem. O modelo de linguagem, P(f), poderia abordar qualquer nível no lado direito da figura com o digrama anterior, mas a abordagem mais fácil e mais comum é a de construir um modelo de n-grama a partir de um corpus em francês, como vimos antes. Isso apreende apenas uma ideia local parcial de sentenças em francês; no entanto, muitas vezes é suficiente para traduções toscas.

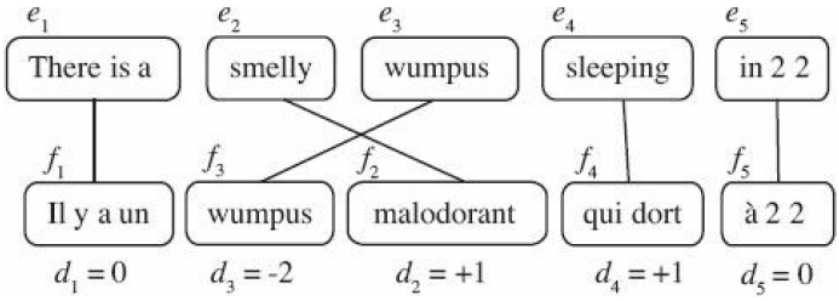
O modelo de tradução é aprendido a partir de um corpus bilíngue — uma coleção de textos paralelos, cada par inglês/francês. Agora, se tivéssemos um corpus infinitamente grande, traduzir uma sentença seria apenas uma tarefa de pesquisa: teríamos visto a sentença em inglês antes no corpus, assim poderíamos apenas retornar a sentença francesa paralela. Mas é claro que nossos recursos são finitos, e a maioria das sentenças a serem solicitadas para traduzir será recente. No entanto, serão compostas de sintagmas que já vimos antes (mesmo que alguns sintagmas sejam tão curtos como uma palavra). Por exemplo, sintagmas comuns incluem “neste exercício vamos”, “tamanho do espaço de estados”, “em função do” e “notas no final do capítulo”. Se solicitado para traduzir a sentença nova “Neste exercício, vamos calcular o tamanho do espaço de estados como uma função do número de ações” para o francês, devemos ser capazes de quebrar a sentença em sintagmas, encontrar os sintagmas no corpus em inglês, encontrar os sintagmas franceses correspondentes (a partir da tradução francesa) e depois remontar os sintagmas franceses em uma ordem que faça sentido em francês. Em outras palavras, dada uma sentença cuja fonte é inglês e, encontrar um tradução francesa f é uma questão de três etapas:

1. Quebrar a sentença em inglês em sintagmas e1,…, en.

2. Para cada sintagma ei, escolher um sintagma em francês correspondente fi. Usamos a notação P(fi | ei) para a probabilidade frasal de que fi seja uma tradução de ei.

3. Escolher uma permutação da frases f1…, fn. Vamos especificar essa permutação de uma forma que parece um pouco complicada, mas é projetada para ter uma distribuição simples de probabilidade: para cada fi, escolhemos uma distorção di, que é o número de palavras que o sintagma fi moveu com relação a fi-1; positivo se moveu para a direita, negativo se moveu para a esquerda e zero se fi seguiu imediatamente fi-1.

A figura abaixo mostra um exemplo do processo. No topo, a sentença "**Há um wumpus fedorento dormindo em 2 2**" é dividida em cinco sintagmas, e1,…, e5. Cada um deles será traduzido em um sintagma f1 correspondente, e então eles serão permutados pela ordem f1, f3, f4, f2, f5. Especificaremos a permutação em termos de distorções di de cada sintagma em francês, definido como



$$d_i = \text{INÍCIO}(f_i) - \text{FIM}(f_{i-1}) - 1$$

onde INÍCIO(fi) é o número ordinal da primeira palavra do sintagma fi na sentença em francês e FIM(fi−1) é o número ordinal da última palavra do sintagma fi-1. Na figura observamos que f5, "à 2 2" segue imediatamente f4, "qui dort"" e, assim, d5 = 0. Entretanto, o sintagma f2 moveu uma palavra à direita de f1, assim d2 = 1. Como caso especial temos d1 = 0, porque f1 inicia na posição 1 e FIM(f0) é definido como 0 (embora f0 não exista).

Agora que definimos a distorção, di, podemos definir a distribuição de probabilidade para a distorção, P(di). Observe que, para sentenças limitadas pelo comprimento n temos |di| ≤ n e de modo que o total de distribuição de probabilidade P(di) tem apenas 2n + 1 elementos, muito menos números para aprender do que os números de permutações, n!. É por isso que definimos a permutação nesse caminho tortuoso. Claro, esse é um modelo bastante empobrecido de distorção. Não informa que os adjetivos são geralmente distorcidos para aparecer após o substantivo quando estamos traduzindo do inglês para o francês — esse fato é representado no modelo francês de língua, P(f). A probabilidade da distorção é completamente independente das palavras nos sintagmas — depende apenas do valor inteiro di. A distribuição de probabilidade fornece um resumo da volatilidade das permutações; qual a probabilidade de uma distorção de P(d = 2), em comparação com P(d = 0), por exemplo.

Estamos prontos agora para juntar tudo: podemos definir P(f, d | e), a probabilidade de que a sequência de sintagmas f com distorções d é uma tradução da sequência de sintagmas e. Fizemos a suposição de que cada tradução de sintagma e cada distorção é independente das outras, e, assim, podemos fatorar a expressão como

$$P(f, d \mid e) = \prod_i P(f_i \mid e_i) P(d_i)$$

Isso oferece uma maneira de calcular a probabilidade P(f, d | e) para uma tradução candidata f e uma distorção d. Mas, para encontrar os melhores f e d não podemos apenas enumerar sentenças; talvez com 100 sintagmas em francês para cada sintagma em inglês no corpus existam 1005 traduções diferentes de 5-sintagmas e 5! reordenações para cada uma delas. Teremos de procurar uma boa solução. Uma busca de feixe local com heurística que estima a probabilidade provou ser eficaz em encontrar a tradução mais provável.

Tudo o que resta é aprender as probabilidades frasais e de distorção. Esboçamos o procedimento abaixo:

1. Encontrar textos paralelos: em primeiro lugar, reúna um corpus paralelo bilíngue. Por exemplo, um Hansard9 é um registro do debate parlamentar. Canadá, Hong Kong e outros países produzem hansards bilíngues, a União Europeia publica seus documentos oficiais em 11 idiomas, e as Nações Unidas publicam documentos multilíngues. O texto bilíngue também está disponível on-line, alguns sites publicam conteúdo com URLs em paralelo, por exemplo, /en/ para a página em inglês e /fr/ para a página correspondente francesa. Os sistemas de tradução estatística principais experimentam em centenas de milhões de palavras de textos em paralelo e bilhões de palavras do texto monolíngue.

2. Segmento em sentenças: a unidade de tradução é uma sentença, então teremos de quebrar o corpus em sentenças. Períodos são fortes indicadores do final de uma sentença, mas considere “Dr. J. R. Smith de Rodeo Dr. paid $29.99 on 9.9.09.”; somente o ponto final termina a sentença. Uma maneira de decidir se um ponto termina uma sentença é analisar um modelo que tome como características as palavras ao redor e suas partes da fala. Essa abordagem atinge cerca de 98% de precisão.

3. Alinhar sentenças: determinar a quais sentenças corresponde na versão francesa cada sentença na versão em inglês. Geralmente, a próxima sentença em inglês corresponde à próxima sentença em francês em uma combinação de 1:1, mas às vezes há uma variação: uma sentença em um idioma será dividida em uma combinação de 2:1 ou a ordem de duas sentenças será trocada, resultando em uma combinação de 2:2. Ao olhar para o comprimento da sentença em si (isto é, sentenças curtas deveriam se alinhar com sentenças curtas), é possível alinhá-las (1:1, 1:2 ou 2:2 etc.) com precisão na faixa de 90-99% usando uma variação do algoritmo de Viterbi. Pode também ser conseguido um alinhamento melhor através de marcos que são comuns em ambos os idiomas, tais como números, datas, nomes próprios ou

palavras que, de um dicionário bilíngue, sabemos que têm tradução inequívoca. Por exemplo, se a terceira sentença em inglês e a quarta em francês contiverem a sequência “1989” e as sentenças vizinhas não tiverem, é uma boa evidência de que as sentenças deveriam ser alinhadas em conjunto.

4. Alinhar sintagmas: dentro de uma sentença, os sintagmas podem ser alinhados por um processo que é semelhante ao que é utilizado para o alinhamento de sentença, mas requer melhoria iterativa. Quando começamos, não temos nenhuma maneira de saber que “qui dort” alinha-se com “sleeping”, mas podemos chegar a esse alinhamento por um processo de agregação de evidências. Em todas as sentenças de exemplo que vimos, percebemos que “qui dort” e “sleeping” coocorrem com alta frequência e que no par de sentenças alinhadas nenhum sintagma diferente de “qui dort” coocorre com tanta frequência em outras frases com “sleeping”. Um alinhamento completo do sintagma sobre o nosso corpus nos dá as probabilidades frasais (após o alisamento adequado).

5. Extrato de distorções: uma vez que temos um alinhamento de sintagmas podemos definir as probabilidades de distorção. Basta contar quantas vezes a distorção ocorre no corpus para cada distância d = 0, ± 1, ± 2,… e aplicar o alisamento.

6. Melhorar as estimativas com EM: utilizar expectativa de maximização para melhorar asestimativas dos valores P(f | e) e P(d). Calculamos os melhores alinhamentos com os valores atuais desses parâmetros na etapa E, em seguida atualizamos as estimativas na etapa M e iteramos o processo até a convergência.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig