Tradução de Textos Baseado em estatística (N-grams)

A. Meirelles ra008072 C.E. Fernandes ra008275
R.D. de Castro ra009663

14 de junho de 2004

Resumo

A tradução de textos de maneira automática é um campo de estudos em atividade desde o início da própria computação. Dentro da inteligência artificial existem diversas ferramentas que tem sido usadas para se tentar produzir um mecanismo eficiente para este fim. A abordagem que vem tendo melhores resultados é a que se apoia nos mecanismos estatísticos bayesianos. Neste trabalho será abordada a técnica dos *N-grams*, que tenta simplificar a quantidade de elementos que necessitam de tratamento estatístico durante todo o processo.

1 Introdução

A tradução de textos automática é um problema que vem sendo pesquisado há várias décadas, mas por limitações computacionais, apenas nas últimas duas décadas o tema pôde avançar mais.

Mecanismos de tradução apresentam dificuldades profundas, pois as diferenças entre as várias línguas em uso no planeta não se restringem a grafia de palavras, mas também existe todo um conjunto de regras gramaticais, e o maior agravante, o fator "contexto"e "cultura". O contexto pode fazer com que uma palavra não se encaixe bem em uma frase embora ela seja uma tradução perfeita de uma outra palavra. Assim, seria interessante que a ferramenta de tradução pudesse tratar de alguma forma essa questão. O outro grande problema se refere a cultura de cada país. Existem expressões características de cada nação e ainda mais, algumas que são típicas de uma região, que carregam um significado próprio e que fica muito difícil reproduzir por um mecanismo automático. Nesta questão, mesmo para os seres humanos, não basta o conhecimento da língua, é necessário o conhecimento do ambiente.

Para essas questões é muito difícil que se chegue a uma solução automática. O que será tratado neste trabalho pode ser muito eficiente para textos menos personalizados, como produções científicas.

O método estatístico se baseia na premissa de que uma frase em uma língua é uma possível tradução de qualquer outra frase de outra língua. Assim, a cada par de frases (O, A) é designada uma probabilidade, $P(A|O)$, a probabilidade de um tradutor produzir a frase A na língua alvo a partir da frase O na língua de origem. A criação das relações entre as frases segue uma lógica inversa. Dada a

frase A, é feita uma busca pela frase O que produziu A. Dessa maneira é possível minimizar as margens de erro. A relação que se deve buscar maximizar é então $P(O|A)$. De acordo com o teorema de Bayes, então:

$$Pr(S|T) = \frac{Pr(S).Pr(T|S)}{Pr(T)}$$

Essa metodologia se baseia no fato de que a tradução de textos é inversível, ou seja, o caminho de A para O e O para A são equivalentes, valendo assim o teorema de Bayes.

Na equação, $Pr(S)$ se refere ao Modelo de Linguagem e $Pr(T|S)$ se refere ao Modelo de Tradução. Na seção 2 será discutido o Modelo de Linguagem, na seção 3 o Modelo de Tradução, na seção 4 é discutido o mode de busca para o Modelo de Linguagem, e por final, a seção 5 conclui o trabalho.

2 Modelo de Linguagem

O modelo de linguagem aqui descrito será o responsável pelo fator $Pr(S)$ do teorema de Bayes utilizado para modelar o tradutor. Este modelo, é utilizado em instituições como o *Wall Street Jornal* para modelar sequencias de 2 e 3 palavras, a fim de se obter as melhores sequencias de palavras. Isto significa, que o modelo de linguagem não se restringe apenas ao uso em traduções, porém é utilizado para análise de sentenças de uma determinada língua. Essa abordagem, será adotada para analisarmos o modelo, e após isso a aplicaremos à tradução.

No processo adotado para tradução, o Modelo de Linguagem é o responsável por aplicar a técnica de *N-grams*. Contudo, isso não exclue a estatística do processo, já que a mesma será utilizada em outras etapas.

A técnica de *N-grams* baseia-se no princípio de considerar os N-1 fatores que precedem um fator i, para calcular $Pr(s_i)$.

Seguindo o modelo proposto em nossa introdução, é preciso encontrar uma probabilidade $Pr(S)$ que maximize $Pr(S|T)$, ou seja, encontrar uma sentença de origem S que melhor se adeque a sentença traduzida T.

Como já dissemos, esse método não é utilizado apenas para tradução e para exemplificar sua atuação, faremos uma análise Inglês|Inglês por exemplo, utilizando o conceito de *bag translation*.

Por esse conceito, devemos pegar a sentença S, dividí-la em palavras e colocar as palavras em um *bag*. Então utilizamos *N-grams* para ranquear diferentes arranjos das palavras a fim de reconstruir a frase inicial S.

Portanto, dada uma frase inicial $S = s_1 s_2 s_3 \dots s_n$, pode-se definir:

$$Pr(s_1, s_2, ..., s_n) = Pr(s_1)Pr(s_2|s_1)...Pr(s_n|s_1 s_2 ... s_{n-1})$$

Para evitar calcular essa relação para todas as palavras, o modelo de *N-grams* considera apenas as n-1 palavras precedentes da palavra S_i.

Logo, um arranjo S pode ser entendido como melhor que outro S' se $Pr(S) > Pr(S')$.

Segundo Brown utilizando N=3, a reconstrução de sentenças de uma mesma linguagem é feita com sucesso em 63% dos casos, e, se levar-se em consideração

frases com o mesmo sentido porém com construção diferente, este índice aumenta para 84%.

Para o caso de realizarmos uma tradução, teremos o mesmo objetivo: Obter uma sentença S que maximize o teorema de Bayes dado uma sentença de saída T.

Deste modo, para realizar a tradução realizamos os procedimentos acima citados, de onde obtermos $Pr(S)$ para ser utilizado no teorema de Bayes.

3 Modelo de Tradução

O objetivo do Modelo de Tradução é fornecer a probabilidade $Pr(T|S)$ do teorema de Bayes da forma que foi enunciado em 1. Sendo assim, esse modelo deve fornecer um mecanismo para a identificação dos pares de palavras que tem a maior probabilidade de se corresponderem.

Aqui depende muito de quais são as línguas **Alvo** e **Origem**, pois baseado nas mesmas é que se pode construir um modelo que tente mapear as frases de uma na outra com a respectiva probabilidade de ocorrência.

Um modelo proposto que foi pesquisado envolve as línguas Francesa e Inglesa. O pilar dessa proposta está na idéia de que as palavras de uma língua em geral correspondem a palavras em outra, de uma maneira que essas podem ser alinhadas. Frases simples e curtas em geral fornecem um alinhamento perfeito, como no par (Il fait froid | It is cold). A realidade é, infelizmente, bem mais complicada que isso, como no seguinte par (Je n'a pas assez d'argent | I don't have all the money). Nesse exemplo **Je** alinha-se com **I**, **n'** e **pas** alinham-se com **don't** e **'a** com **have**. A construção verbal de negativo e possessão são exemplos de dificuldade de tradução. Dificuldade maior é imposta no caso de palavras que traduzidas para outra língua geram múltiplas palavras, o que define a fertilidade de uma palavra. Existem casos, também, de palavras que não tem correspondente direto na outra língua, mas são necessárias para a construção gramatical correta, como é o caso do **pas** na composição do negativo em francês.

Existem também dificuldades com respeito à ordem com que algumas palavras aparecem em relação a outras dentro da frase. Por exemplo, em inglês os adjetivos precedem o substantivo, mas no francês geralmente é o contrário. Casos assim também ocorrem com frequência com os advérbios, mas neste caso a distância geralmente é maior. O nome dado a essa diferença de posicionamento é distorção.

Usando esses conceitos, a forma que se encontrou para uma melhor representação dos alinhamentos de um par de frases segue o seguinte exemplo:
(Que faut-il faire si nous voulons aller au match? | What(1) should(2,3) we do(4) if(5) we(6) want(7) to watch(8) the game(9)?).
Aqui,cada palavra da língua Alvo é mapeada pela posição - entre parênteses - que ocupa(m) cada uma da(s) palavra(s) que se alinha(m) a ela. Como pode ser visto, existem palavras da língua *alvo* que não são mapeadas em nenhuma palavra da língua *origem* e o inverso também pode ocorrer.

A probabilidade de todo o par, no exemplo acima é calculado da seguinte maneira: multiplicar a probabilidade de **What** ter fertilidade 1 pela probabilidade $Pr(Que|What)$, multiplicado pela probabilidade de **should** ter fertilidade 2 multiplicado por $Pr(fault|should).Pr(il|should)$.... e assim por diante. As palavras em frances que não se alinham com nenhuma do inglês, por convenção,

são alinhadas com a palavra especial <*Nulo*>.

O que se percebe é que o valor final $Pr(T|S)$ a que se quer chegar, é uma combinação das probabilidades das fertilidades e dos pares entre cada palavra do francês e do inglês ($P(f|i)$). É possível ainda usar alguma relação sobre a distorção, que envolva a posição da palavra e o tamanho da frase, uma vez que quanto maior a frase, maior a chance de alguma palavra aparecer em um lugar trocado.

4 Search

Nesta etapa, deve ser implementado o *decoding* que é o real responsável por encontrar a sentença S que maximiza $Pr(S)Pr(T|S)$.

Para isso, precisamos fazer uma busca dentre todas as sentenças possíveis. Esse procedimento é extremamente caro pois existem infinitas possibilidade de sentenças, o que demanda a aplicação de heurísticas.

O mecanismo escolhido para implementar nossa busca será a *stack search* que é um processo iterativo, e que nos possibilitará traduzir tanto um texto quanto uma simples sentença.

Apesar de caro, o algoritmo para se descobrir S é simples.

Primeiro criamos uma pilha (*stack*).

A partir de agora chamaremos a senteça S de E (English), para identificarmos melhor as sentenças em francês e inglês.

Com a pilha criada, pegaremos uma sentença F para analisar.

Pegamos uma palavra f_i de F. Inicia-se o processo com f_0.

Para a palavra f_i, que pegamos no passo acima, iremos analisar sua fertilidade e seu histórico *trigram* ($N = 3$). Para a palavra f_i que tiver o maior *score* considerando fertilidade e histórico, tomaremos sua equivalente e_i e colocaremona na pilha.

O passo acima deve ser repitido até o fim da sentença F, para cada palavra f_i da sentença. Deste modo, estaremos empilhando as palavras em inglês com maior probabilidade de terem gerado as palavras em francês.

Como estaremos formando uma nova sentença E, podemos calcular os valores parciais de $Pr(S|T)$, no caso $Pr(E|F)$. Para analisar a variação de $Pr(S|T)$, podemos trocar a palavra e_j com maior *score* pela segunda opção e verificar o resultado da alteração. Deve-se repetir esse procedimento para todas as palavras da pilha, até que a sentença acabe ou a $Pr(S|T)$ atinja o valor máximo.

Caso desejemos traduzir um texto, devemos repetir estes procedimentos para todas as k sentenças do mesmo. Deste modo, teremos k valores de $Pr(S|T)$.

Pode ocorrer que a sentença S' obtida é diferente da sentença S que o tradutor utilizou. Neste caso temos algumas situações a analisar:

- $Pr(S')Pr(T|S') > Pr(S)Pr(T|S)$

  Neste caso caímos nos modelos de linguagem e tradução, logo temos uma sentença S' que maximiza nosso modelo.

- $Pr(S')Pr(T|S') < Pr(S)Pr(T|S)$

  Neste caso, nossa busca por um S que maximizaria o modelo falhou.

5 Conclusão

O índice de acertos da tradução é suficiente para se fornecer meios de compreensão superficiais, e até para auxiliar estudantes e pesquisadores em geral na leitura de textos de línguas desconhecidas. Os resultados não são suficientes para automaticamente lançar versões em várias línguas de um artigo para publicação ou de um livro.

As dificuldades impostas pelas diferenças linguísticas são muitas, e variam de acordo com as escolhas **Alvo** e **Origem**. O tratamento estatístico ainda é o melhor meio atual, mas a maior dificuldade, como foi exposto, é retratar questões culturais e contextuais presentes na escrita.

Referências

[1] L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179-190, March 1983

[2] J.D. Ferguson. Hidden Markov analysis: an introduction. In J.D. Ferguson, editor, *Hidden Markov Models for Speech*, October 1980

[3] Peter F. Brown, John Cocke, Stephen A.D. Pietra, Vincent A.D. Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, *A Statistical Approach to Machine Translation*

[4] William W. Cohen *Fast Effective Rule Induction*