Investigação de Modelos Estatísticos de Tradução Automática Baseados em Phrases

Wilker Ferreira Aziz, Thiago Alexandre Salgueiro Pardo, Ivandré Paraboni

Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

wilker@grad.icmc.usp.br, taspardo@icmc.usp.br, ivandre@usp.br

1. Introdução

A abordagem estatística de tradução automática (TA) tem apresentado baixo tempo de desenvolvimento e bom desempenho [1]; consiste em modelar a tarefa da tradução como uma função de probabilidades extraídas de exemplos de tradução, o córpus paralelo. O modelo de tradução é o componente que rege as escolhas do processo de tradução. Por muito tempo pesquisou-se modelos cuja unidade de aprendizado era a palavra; o estado da arte propõe a utilização de conjuntos de palavras em seqüência, as *phrases*, com o intuito de melhorar o desempenho do aprendizado.

2. Objetivos

Este trabalho tem por objetivo investigar modelos de tradução orientados a *phrases*. Modelos de tradução orientados a palavra apresentam bom desempenho para línguas com estrutura semelhante, como são o português e o espanhol [1], onde a tradução se dá muitas vezes palavra por palavra, porém são pouco eficientes para pares de línguas distantes. Esta pesquisa é motivada pelo baixo desempenho destes modelos na tradução entre as línguas portuguesa e inglesa, como demonstram os experimentos realizados com um tradutor baseado em palavras, aqui relatados.

3. Material e Método

Realizou-se experimentos em TA estatística utilizando modelos baseados em palavras entre os pares português-espanhol (PB-ES), espanhol-português (ES-PB), português-inglês (PB-IA) e inglês-português (IA-PB). A Tabela 1 apresenta a avaliação automática segundo a métrica BLEU [2].

Tabela 1: Score BLEU para os tradutores *word-based*

PB-ES	ES-PB	PB-IA	IA-PB
0,5673	0,5734	0,1892	0,2228

O experimento, realizado sobre o mesmo córpus de exemplos, demonstra que a qualidade da tradução cai drasticamente com a mudança do par utilizado.

A modelagem estatística parte do raciocínio intuitivo de lidar com a tarefa de traduzir uma sentença *f*, na língua-fonte *lf*, para uma sentença *a*, na língua-alvo *la*, como uma probabilidade condicional, expressa pela regra de Bayes apresentada na equação (1).

$$P(a \mid f) = \frac{P(a)P(f \mid a)}{P(f)} \qquad (1)$$

Em termos computacionais, traduzir *f* para *a* implica em encontrar *a* que maximiza (2) dado que o denominador de (1) é constante.

$$P(a \mid f) = P(a)P(f \mid a) \qquad (2)$$

Identificam-se dois componentes distintos em (2), o modelo de língua (LM) *P(a)*, responsável por selecionar as sentenças mais freqüentes da língua-alvo, e o modelo de tradução (TM) *P(f|a)*, responsável pelas escolhas do processo de tradução.

Assumindo a existência de um córpus bilíngüe que expresse todos os possíveis alinhamentos α entre elementos da *lf* e da *la*, o segundo fator de (2) pode ser expresso, pelo teorema da probabilidade total, como a contribuição de cada alinhamento possível.

$$P(f \mid a) = \sum_i P(\alpha_i, f \mid a) \qquad (3)$$

Esta última equação oferece meios de modelar o aprendizado das transformações entre a *lf* e a *la*. Considere u^L uma unidade (palavra ou *phrase*) na língua *L* e $t(u^{lf} \mid u^{la})$ a probabilidade de que a unidade u^{lf} seja a tradução da unidade u^{la}, pode-se aproximar o termo geral da soma em (3) como na equação (4), abaixo.

$$P(\alpha, f \mid a) = t(f \mid a) = \prod_i t(u_i^{lf} \mid u_i^{la}) \qquad (4)$$

A modelagem baseada em phrases assume que a unidade de aprendizado é a *phrase*. Trabalhos estimam que *phrases* de no máximo três palavras já produzam um bom aprendizado [3]. Tais *phrases* podem ser extraídas do próprio alinhamento lexical através de regras de boa formação. Um método consiste em alinhar o córpus paralelo em nível lexical nos dois sentidos possíveis (*lf-la* e *la-lf*), selecionar todos os alinhamentos contidos na intersecção dos conjuntos obtidos e extrair desta todos os pares de *phrases* nos quais as palavras que as constituem alinham-se somente entre si. A Figura 1 dá exemplos de construções bem e mal formadas, respectivamente.

$$p_1 : \{\alpha_a^{lf} \beta_b^{lf} \gamma_c^{lf}\} \Leftrightarrow \{a_\alpha^{la} c_\gamma^{la} b_\beta^{la}\}$$

$$p_3 : \{\alpha_a^{lf} \beta_t^{lf} \gamma_c^{lf}\} \Leftrightarrow \{a_\alpha^{la} c_\gamma^{la} b_\theta^{la}\}$$

Figura 1: Alinhamento interno das p*hrases*

Por uma abordagem *n-gram* conta-se a freqüência com que pares de *phrases* ocorrem no córpus paralelo, produzindo valores para o parâmetro *t* utilizado em (4).

4. Resultados

O modelo (4) em abordagem *unigram* foi avaliado pelo trabalho [3] e comparado com o modelo 1 da IBM (baseado em palavras), que também utiliza a probabilidade de tradução *t* como único parâmetro do modelo. Os resultados são reproduzidos na Tabela 2. O dicionário de tradução de N3 considera somente pares de *phrases* que ocorrem no mínimo três vezes no córpus. A Tabela 2 apresenta um ganho significativo em adotar-se *phrases* ao invés de palavras.

Tabela 2: Comparação entre modelos [3]

Modelo	Unidades obtidas	BLEU
IBM 1	1,23 M	$0{,}11 \pm 0{,}01$
N3	1,22 M	$0{,}18 \pm 0{,}01$

5. Conclusões

Na tentativa de em algum nível expressar a relação entre as palavras, abordagens baseadas em palavras carregam o modelo com uma maior quantidade de parâmetros, os quais se relacionam somente na medida em que devem ser analisados em conjunto para maximizar a função $P(\alpha, f \mid a)$; nesse sentido, *phrases* são mais eficientes, pois já no aprendizado consideram a dependência entre as palavras diminuindo a necessidade de parâmetros envolvidos no modelo.

6. Agradecimentos

Este trabalho conta com o apoio FAPESP.

7. Referências Bibliográficas

[1] Aziz, W. F.; Pardo, T. A. S.; Paraboni, I. An Experiment in Spanish-Portuguese Statistical Machine Translation. SBIA-2008.

[2] Papineni, K.; Roukos, S.; Ward, T. and Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of he 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318. (2002)

[3] Tillman, C.; Xia, F. A Phrase-Based Unigram Model for Statistical Machine Translation.