

Avaliação de qualidade de traduções automáticas usando redes complexas

Diego R. Amancio¹, Maria G.V. Nunes², Thiago A.S.Pardo², Osvaldo N.O. Jr.¹

¹ Instituto de Física de São Carlos - Universidade de São Paulo (USP)
Caixa Postal 13560-970, São Carlos, SP, Brasil

²Instituto de Ciências Matemáticas e de Computação- Universidade de São Paulo (USP)
Caixa Postal 13560-970, São Carlos, SP, Brasil

diego.amancio@usp.br, {gracan,taspardo}@icmc.usp.br, chu@ifsc.usp.br

Abstract: *Establishing metrics to assess machine translation (MT) systems automatically is now crucial owing to the widespread use of MT over the web, and this can be done by representing the text as a complex network. In this study we extend our previous work by employing additional metrics of complex networks, whose results were used as input for machine learning methods and allowed MT texts of distinct qualities to be distinguished. This applied for Spanish-Portuguese and English-Portuguese pairs. The results presented here amount to a proof-of-principle that the possible capturing of a wider context with the hierarchical levels may be combined with machine learning methods to yield an approach for assessing the quality of MT systems.*

Resumo: *O estabelecimento de métricas para avaliação automática da qualidade dos sistemas de tradução automática (TA) é crucial devido ao amplo uso da TA na web, e isto pode ser feito representando textos como redes complexas. Aplicamos métricas de redes complexas, cujos resultados foram utilizados como entrada para métodos de aprendizado de máquina e permitiram que textos traduzidos automaticamente e manualmente fossem distinguidos. Tal método foi aplicado para o par de línguas Inglês-Português e Espanhol-Português. Os resultados apresentados aqui mostram que é possível capturar um contexto mais amplo com a utilização de níveis hierárquicos mais profundos em conjunto com os métodos de aprendizado de máquina como abordagem de avaliação de sistemas de TA.*

1. Introdução

Os conceitos e metodologias de redes complexas vêm sendo usados numa enorme variedade de áreas [Costa et al. 2008], incluindo a análise automática de textos na área de processamento de línguas naturais (PLN). Exemplos de aplicações de PLN são os tradutores automáticos, revisores gramaticais e sumarizadores automáticos. A adequação de redes complexas a esse tipo de análise foi demonstrada em várias instâncias, a partir da comprovação de que um texto pode ser representado por uma rede livre de escala [Cancho e Solé 2001], isto é, uma rede com poucos vértices fortemente conectados e muitos vértices fracamente conectados. Resultados consolidados incluem a análise automática de qualidade de textos, a avaliação da qualidade de sumarizadores automáticos, e a avaliação da qualidade de tradução automática [Costa et al. 2009].

Neste trabalho, aplicou-se a modelagem de redes complexas para o desenvolvimento de novas técnicas de avaliação de qualidade de traduções automáticas, mostrando que métricas baseadas na análise da vizinhança dos nós, ou seja, hierárquicas [Costa e Silva 2006], é capaz de distinguir os produtos de diferentes tradutores automáticos, ou seja, diferenciar a origem da tradução.

2. Modelando textos como redes complexas

Neste trabalho utilizou-se a modelagem proposta por Antiqueira et al. (2005), cuja rede resultante da modelagem dos textos é construída segundo relações de adjacência entre palavras. Inicialmente, em uma etapa de pré-processamento, o texto é lematizado e tem as *stopwords* removidas. A seguir, cada palavra é representada como um vértice na rede, sendo que palavras iguais são representadas por um mesmo vértice. As arestas são obtidas associando-se vértices cujas palavras correspondentes são imediatamente adjacentes (não se consideram limites de sentenças ou parágrafos). Caso uma associação seja repetida, incrementa-se em 1 o peso da aresta. Assim, obtém-se um grafo ponderado representado por uma matriz de adjacência quadrada W , em que cada $W(j,i)$ representa o número de associações $i \rightarrow j$ encontradas no texto, em que i e j são pares de palavras imediatamente adjacentes.

A partir da modelagem do texto descrita anteriormente, caracterizaram-se os textos com as medidas padrões de grafos (graus, coeficiente de aglomeração, entre outras) [Costa et al. 2007] em níveis hierárquicos. Basicamente, definem-se hierarquias como extensão do conceito de um vértice v para seu nível hierárquico i , ao aderir ao vértice v seus vizinhos de distância máxima i , sendo que sua nova aresta será representada pelas arestas dos vértices de distância i que ligam a vértices de distância k , com $k > i$. Tal definição torna-se interessante na rede de palavras por permitir que a topologia da rede seja quantificada de uma maneira mais compreensiva, já que se leva em consideração contextos mais amplos de cada palavra representada como nó.

3. Avaliação de qualidade de traduções automáticas

Nestes experimentos pretendeu-se extrair padrões de distinção de tradutores automáticos a fim de correlacionar os padrões encontrados com a qualidade do tradutor utilizado. O corpus utilizado era composto de 50 textos de 500 vértices aproximadamente cada. Os tradutores envolvidos destacam-se pela divergência de qualidade para o par português-espanhol: *Apertium*¹ (alta qualidade) e *Intertran*² (baixa qualidade) e tradutor do Google³ (qualidade média).

Para cada par texto_fonte – texto_alvo, foram extraídas as métricas hierárquicas referentes a cada vértice. Adicionalmente, o mapeamento entre vértices (palavras) do texto fonte e vértices do texto alvo foi realizado pelo alinhador lexical automático LIHLA [Caseli et al. 2005]. Em uma tradução de n s palavras numa sentença fonte para uma

¹ <http://www.apertium.org/?lang=pt>

² <http://www.tranexp.com:2000/Translate/result.shtml>

³ http://www.google.com.br/language_tools?hl=pt-BR

sentença alvo com n palavras, LIHLA tenta criar um alinhamento entre palavras desta sentença. Quando um alinhamento simples considera mais do que uma palavra na sentença fonte ou alvo, tais palavras são agrupadas em um único vértice da rede. Alguns vértices podem não pertencer a qualquer alinhamento. Isto ocorre quando um conjunto de palavras da sentença fonte não for traduzido diretamente em palavras da sentença alvo. Com a informação das métricas para cada vértice e do alinhamento lexical automático, define-se a taxa de acerto para um par texto_fonte – texto_alvo. Dada uma palavra fonte, seleciona-se a palavra alvo cujas métricas hierárquicas mais se aproximam das métricas da palavra fonte, de forma que se este mapeamento corresponder ao alinhamento automático, então um acerto é contabilizado. Por fim, toma-se a média dos acertos para representar a taxa de acerto do par texto_fonte – texto_alvo.

A Figura 1 ilustra as taxas de acerto em função do nível hierárquico para o corpus proposto, em traduções espanhol-português, feitas pelos tradutores *Apertium* e *Intertran*. Em concordância ao trabalho desenvolvido em Amancio et al. (2008) para análise de qualidade com métricas não hierárquicas, tal resultado mostra que o tradutor automático de alta qualidade apresenta taxas de acerto sempre maiores que o de baixa qualidade, expressando uma preservação na estrutura global da rede para traduções de alta qualidade. Para o inglês (tradutores *Google* e *Intertran*), esta distinção não é tão evidente (resultado não mostrado), no entanto, o desvio padrão das taxas de acertos em cada texto mostrou-se capaz de prover padrões de distinção.

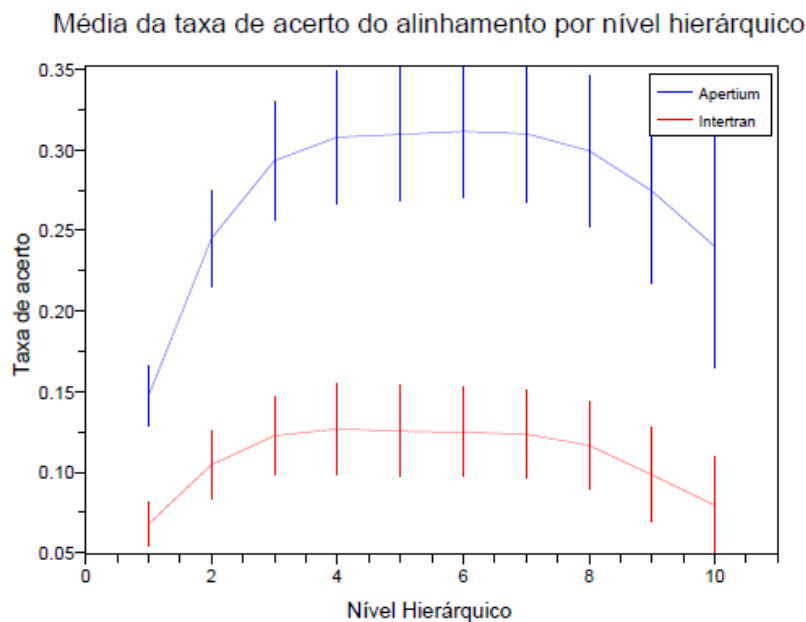


Figura 1: Dinâmica da taxa de acerto segundo o nível hierárquico – Espanhol-Português

A eficiência do uso de níveis hierárquicos para distinção de traduções automáticas é sumarizada na Tabela 1, que exhibe altas porcentagens de exatidão obtidas para distinção por aprendizado de máquina (AM), ao considerar como atributo a taxa de acerto em cada nível, e como classe, o tradutor utilizado. Desta forma, tal método mostra-se igualmente promissor quando comparado àqueles desenvolvidos em Amancio et al. (2008).

Tabela 1. Taxa de acerto para distinção de tradutores automáticos por AM.

Algoritmo	Acerto na distinção entre Apertium e Intertran	Acerto na distinção entre Google e Intertran
Regras	99 %	87 %
Árvore de decisão	98 %	84 %

4. Conclusão

Os experimentos confirmaram a eficácia da modelagem de textos como redes complexas em uma das aplicações mais tradicionais em PLN: a tradução automática. As redes foram usadas com sucesso para reconhecimento de padrões de qualidade por taxas de acerto definidas por métricas hierárquicas em redes complexas, refletindo que boas traduções (principalmente para o par espanhol-português) tendem a manter a estrutura textual, até mesmo em contextos mais amplos.

Referências

- Amancio, D.R.; Antikeira, L.; Pardo, T.A.S.; Costa, L.F.; Oliveira Jr., O.N.; Nunes, M.G.V (2008) Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(4):583-598.
- Antikeira, L; Nunes, M. G. V.; Oliveira Jr., O. N.; Costa, L.F (2007) Strong correlations between text quality and complex networks features. *Physica A*, 373:811-820.
- Cancho, R.F.; Solé, R.V (2001) The Small World of Human Language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268, 2261-2265.
- Caseli, H.M.; Nunes, M.G.V.; Forcada (2005) M.L. LIHLA: A Lexical Aligner Based on Language-Independent Heuristics. In *Proceedings of the V Encontro Nacional de Inteligência Artificial (ENIA 2005)*, p.641-650. São Leopoldo-RS, Brazil.
- Costa, L.F.; Rocha, L.E.C (2006) A generalized approach to complex networks. *European. Physical Journal B*, v. 50, p. 237-242.
- Costa, L. da F.; Silva, F. N. (2006) Hierarchical characterization of complex networks. *Journal of Statistical Physics*, v. 125, p. 845-876.
- Costa, L. F.; Oliveira Jr., O. N.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P. R.; Antikeira, L.; Viana, M. P.; Rocha, L. E. C (2008) Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *Physics and Society*.