Comentários do artigo

Abstract

O artigo se propõe a estudar a validade da lei de Zipf para unidades linguísticas elementares, a saber, “word forms” e “lemmas”. Os autores analisam vários textos literários em 4 línguas. Em todos os casos, a lei de Zipf é válida para word forms e lemmas. Os autores analisam então os expoentes e o limiar de baixa frequência. Os expoentes são similares, mas os limiares de baixa frequência são menos estáveis.

Introdução

Em vez de:

~~n(r) proporcional a 1 / r~~~~α~~

Usar:

**f(n) é proporcional a 1 / nγ, onde γ = 1 + 1 / α**

α costuma ser próximo de 1 e γ costuma ser próximo de 2.

f(n) é a probabilidade de uma palavra ter uma frequência n.

Perguntas:

A distribuição de frequências entre palavras e lemas é diferente? Duas subperguntas:

A lei de Zipf vale para lemas?

O expoente para word forms é diferente do de lemas?

Método:

10 textos literários longos de autor único em 4 línguas (Inglês, Espanhol, Francês, Finlandês)

Usam “type frequencies” em vez de rankings porque rankings têm propriedades matemáticas indesejadas que disfarçam diferenças entre textos reais e experimentos de digitação aleatória

Definições

Um texto é uma sequência de símbolos. Cada ocorrência de um símbolo é um token. A soma dos tokens é L, que é o tamanho do texto. O símbolo em si é um tipo e o número de tipos é o tamanho V do vocabulário, onde V <= L.

Lemma é a raiz da palavra, eu acho. E eu acho que word forms são cada token. Então houses, housed, housing são word forms diferentes do mesmo lemma, que é house.

Procedimento:

f(n) = C / nγ para valores de n >= a

Essa fórmula vai ser válida se a lei de potência for válida para frequências a partir de um limiar a que fiquem a até duas décadas até a frequência do tipo mais comum.

O teste: “fitar” o valor de γ para um a fixo e verificar se a lei de potência ainda é válida a partir de a. Repetir para vários valores de a até encontrar o mais satisfatório. O resultado vai ser γ +- σ, onde σ é o desvio-padrão de γ. Mais detalhes na referência 39. É o método da referência 38 adaptado para o caso discreto. Um resumo está disponível na seção Materiais e métodos.

Resultados

As palavras foram “lemmatizadas” com ferramentas computacionais de modo que cada token pertence a um lemma. => eles fazem posteriormente um comentário sobre possíveis erros de lematização que poderiam produzir tokens sem nenhum lemma associado.

A lei de Zipf é válida para word forms e lemmas

Tabela 2.

Consistência dos expoentes entre word forms e lemmas

Embora os expoentes sejam próximos entre si, é difícil quantificar a diferença entre os dois porque o desvio-padrão da diferença entre eles é difícil de calcular sem a suposição de que as variáveis são independentes (o que elas não são), então esses dados (da semelhança entre os expoentes) deve ser lido com cautela. É impossível afirmar de modo conclusivo que a lematização das palavras não afeta o valor do expoente na lei de Zipf.

Entretanto, é possível comparar pares de expoentes e fazer um teste t-Student. Esse teste pressupõe que o desvio-padrão entre cada par de expoentes vai ser igual, o que também não é verdade. Esse teste informa que: considerando todos os livros, podemos rejeitar a hipótese de que não haja diferenças significativas entre γw e γl. Entretanto, o resultado muda se desconsiderarmos um outlier.

O cálculo da correlação simples mostra uma correlação entre γw e γl considerando todos os livros. Os autores então formulam 3 hipóteses:

1. A distribuição de γw e γl é a mesma para um texto qualquer (mas não necessariamente para textos diferentes) => rejeitada em um teste que não entendi, mas não é rejeitada se tirarmos o outlier.
2. γw é dependente de γl. => não rejeitada
3. γl é dependente de γw. => não rejeitada

Mudar a escala dos gráficos auxilia na visualização e pode corroborar a ideia de que a lei de Zipf é invariante à lematização.

Consistência de lower cut-offs de frequência para word forms e lemmas

aw e al não são independentes e al tende a ser maior que aw. Testes estatísticos parecidos com os da seção anterior apoiam essa conclusão.

Discussão

Conclusões

Materiais e seleção

Seleção do corpus

Lemmatization

Procedimentos estatísticos