Trabalho

Objetivos

Neste trabalho, nosso objetivo é que o aluno faça uma prática com treinamento de modelos de linguagem usando ferramentas já existentes. A ideia é que seja realizado um autoaprendizado, fomentando a discussão e o aprendizado através da experimentação.

Organização

Você irá trabalhar com o SRILM (http://www.speech.sri.com/projects/srilm/). Você deve fazer um relatório explicando os detalhes dos seus experimentos, com passo a passo de cada comando utilizado e análises de cada resultado, além de breve explicação dos métodos utilizados (métodos de descontos – smoothing –, backoffs, etc).

Metodologia

Para os experimentos, caberá a você determinar sua base de treinamento e sua base de testes. Recomenda-se o uso de bases grandes o suficiente para obter resultados satisfatórios. Existem bases de treinamento disponíveis na web (Google ngram, Yahoo Labs ngram, bases como o Bosque, Floresta Sintactica e Amazonia para o português ou até mesmo textos da wikipedia que podem ser coletados via Dbpedia).

Experimentos a serem realizados

Neste trabalho, espera-se que o aluno se debruce sobre as formas de treinamento, problemas que podem surgir e analisar os dados que estão sendo gerados (nem sempre números maiores/menores são bons, certo?).

Minimamente, alguns tipos de análise devem ser feitas:

- Usar métodos intrínsecos de avaliação: (1) melhorar perplexidade, adotar normalizações no texto, inserção ou remoção de textos na base de treinamento, etc para alcançar melhores valores de perplexidade; (2) contagem de ngrams singletons: quantidade de ngrams com contagem igual a 1 (quanto mais, tende a ser pior); (3) quantidade de OOVs.
- Avaliação empírica I: no seu conjunto de testes, coloque frases que estão corretamente escritas e frases que não fazem sentido. O objetivo do teu LM é ser capaz de classificar corretamente frases bem escritas e frases mal escritas. Calcular acurácia do teu LM nessa base de testes.
- Avaliação empírica II: para avaliar o quão especializado teu LM está, utilize o método de visualização de Shannon para gerar frases com teu LM. Não sei se existe uma ferramenta própria pra isso. Imagino que você terá que criar um algoritmo para gerar essa visualização do teu LM (mas o método é bem simples). Permita que a frase seja iniciada aleatoriamente ou através de um input de texto (primeiras palavras da frase).
- Realizar experimentos com teu modelo alterando as formas de generalização do modelo. Testar, necessariamente: (1) interpolação; (2) backoff; (3) diferentes métodos de suavização. Não se preocupe, essas ferramentas já lhe oferecem os métodos prontos. Mas é necessário entendê-los. Tente usar um stupid backoff, se possível. Explore os métodos de desconto disponíveis (K-ney, WB, etc).
- **Extra:** um bom LM é aquele onde o seu P(W) é o mais alto para uma frase W que faça sentido para o LM. Como você poderia usar um LM como classificador? Crie um pequeno script que permita classificar uma frase em, digamos, esporte x música ou poesia x prosa, etc.

Referências

SRILM: http://www.speech.sri.com/projects/srilm/
Documentação do SRILM: http://www.speech.sri.com/projects/srilm/manpages/

OBS: Atualmente o KenLM (https://kheafield.com/code/kenlm/) é a ferramenta que substituiu o SRILM. Contudo, para fins didáticos, o SRILM é mais interessante, pois possui vários méodos implementados de suavização de backoff, enquanto o KenLM oferece apenas um.