1 Language Model Domain Adaptation

The general objective of this project is to explore methods for adapting a language model trained on a source domain for which relatively large corpora are available, to a somewhat similar target domain with limited data available. The idea is to come up with the best adaptation technique. The following can serve as a guide to explore some techniques, but there are many other possible methods that you can also consider.

1.1 Domain similarity

Download the data from

http://cs.nyu.edu/~mohri/asr09/data.tar.gz

This includes 4 files, each corresponding to a different domain. Create a Katz Back-off model for each of the 4 domains. Use the $L_2$ distance between these probability distributions to determine which two are the closest domains:

$$L_2(p,q) = \sqrt{\sum_{x \in \Sigma^*} (p(x) - q(x))^2}.$$  \hspace{1cm} (1)

Indicate how the distances can be computed using weighted automata algorithms. The books and kitchen domains are expected to be the most dissimilar ones.

1.2 Single-source adaptation

Consider the problem of domain adaptation from the books domain to the kitchen domain. Assume that only the first 200 lines of the kitchen corpus are available. The rest is used for testing.

1. What is the perplexity of the books bigram model measured on the test sample?

2. Explore different adaptation methods for improving that perplexity.
(a) Mixture of \( n \)-gram models: define model as

\[
p = \alpha p_S + (1 - \alpha)p_T
\]

(2)

where \( 0 \leq \alpha \leq 1 \), \( p_S \) is an \( n \)-gram model for the source, and \( p_T \) one for the target derived from the small amount of data. Try to find ways to determine the mixing parameter \( \alpha \) automatically, for example by using the \( L_2 \) distance. Explain how the weighted automaton for \( p \) can be constructed from those for \( p_S \) and \( p_T \).

(b) Mixture of counts: use similarly a mixture of the counts instead.

(c) Maximum entropy models: use the books bigram or trigram model as a prior and learn a maxent model for the target.

1.3 Multiple-source adaptation

As in the previous section, the objective is to create an accurate model for the target domain, by here by making use of all other three possible source domains.

1. What is the perplexity of each of the source bigram model measured on the test sample?

2. Explore different adaptation methods for improving that perplexity.

(a) Mixture of the three source \( n \)-gram models:

\[
p = \alpha_1 p_{S_1} + \alpha_2 p_{S_2} + \alpha_3 p_{S_3} + \alpha_4 p_T,
\]

(3)

where \( 0 \leq \alpha_i \leq 1 \), \( \sum_{i=1}^4 \alpha_i = 1 \), \( p_{S_i} \) is an \( n \)-gram model for source \( i \), and \( p_T \) one for the target derived from the small amount of data. Try to find ways to determine \( \alpha_i \)'s automatically, for example by using the \( L_2 \) distance. Explain how the weighted automaton for \( p \) can be constructed.

(b) Mixture of counts: use similarly a mixture of the counts instead.

(c) Maximum entropy models: use the books bigram or trigram model as a prior and learn a maxent model for the target.

Organization

You can work on the project as part of a team. But, teams should have no more than two members and the work expected is proportional to the size.