Acoustic Feature based Mixture Model

Original language model ranks documents according to

(1)

where V is the lexicon.

Query expansion: find the terms semantically relating to the query Q from N documents with highest S(d,Q) (top N documents).

1. **Word-based Mixture Model**

Top N documents, denoted as D, are used for estimating a query-related term distribution θR (unigram language model). For each word w in document d (d belongs to D), it is generated from either query-related model θR with probability αd or background model θB with probability 1-αd. αd is document dependent.

Likelihood Function:

Find θR and a set of αd for each document d in D maximizing F(θR, αd).

EM:

Estimate αd for each document d in D and θR by EM:

**E step:**

P(θR|w,d) is posterior probability of w generating from θR.

**M step:**

Testing Phase:

P(w|θR) is used to replace P(w|Q) in (1).

1. **Acoustic Feature based Mixture Model**

Here I only demonstrate the most naïve version.

Each spoken document is considered as a sequence of acoustic features instead of a sequence of word (from transcriptions)\*.

Each acoustic feature x in document d is generated from one of the Gaussian {θw , w in V’}\*. Each θw represents the acoustic feature distribution of a word. Then as word-based mixture model, the document d using θw for feature generation with probability .

\*It is preferred that each feature corresponds to a word.

\*\*Since sufficient data is necessary to estimate a reliable feature distribution for a word, I consider that V’ is composed of the word occurring frequently in D plus a word “other” as remaining words without sufficient data.

Likelihood Function:

Find θR, a set of αd for each document d in D, and θw for each word w in the lexicon V’ maximizing F(θR, αd, θw).

is the likelihood for θw generating x.

EM:

Estimate θR, a set of αd for each document d in D, and θw for each word w in the lexicon V’ by EM.

**E step:**

Given the acoustic feature x of d, P(w, θR|x,d) is the probability that x is generated from θw, andw is generated from θR.

**M step:**

M is for normalization, which is the number of features x in document d.

Mean of θw : mw

Variance of θw : Cw

Testing Phase:

Plan A. P(w|θR) is used to replace P(w|Q) in (1).

Plan B. Re-rank documents by S(d)