DSP HW3 Report

Environment:

1. macosx 10.14 Mojave
2. CSIE workstation

How to compile my program?

macosx:

```
$ make MACHINE_TYPE=macosx SRIPATH=/your_path/srilm-1.5.10 all
```

CSIE workstation:

```
$ make MACHINE_TYPE=i686-m64 SRIPATH=/your_path/srilm-1.5.10 all
```

How to execute my program?

```
$ ./mydisambig -text <text_file> -map <mapping_file> -lm <language_model> -order <order>
```

macosx:

Create mapping file:

```
$ make map
```

Bigram:

```
$ make MACHINE_TYPE=macosx SRIPATH=/your_path/srilm-1.5.10 LM=bigram.lm run
```

Trigram:

```
$ make MACHINE_TYPE=macosx SRIPATH=/your_path/srilm-1.5.10 LM=trigram.lm run_trigram
```

(The result of trigram will be in “result3” directory.)

CSIE workstation:

Create mapping file:

```
$ make map
```

Bigram:

```
$ make MACHINE_TYPE=i686-m64 SRIPATH=/your_path/srilm-1.5.10 LM=bigram.lm run
```

Trigram:

```
$ make MACHINE_TYPE=i686-m64 SRIPATH=/your_path/srilm-1.5.10 LM=trigram.lm run_trigram
```

(The result of trigram will be in “result3” directory.)

Implementation:

Bigram:

I implement mydisambig by Viterbi algorithm, every candidate of a word at time t remember the best candidate of word at time t-1, and use backtracking from the last word to get the best sequence.

Trigram:

The trigram version of mydisambig is implemented by Viterbi algorithm with two variables. I speed up the process by pruning the candidates, first I sort the unigram probability of each candidate of a given word, then I keep the largest 150 candidates and prune the others. This will speed up the process and maintain good accuracy.

Observations:

When I ran mydisambig on the test data, I found that some of my result will be slightly different from the result of disambig from SRILM. For example, "ㄑ心準ㄅ" in mydisambig will be translated as "全心準備", while disambig from SRILM will translate it into "其心準備". I checked the language model, and logP(全)+logP(心|全) is actually bigger than logP(其)+logP(心|其), so I still don't understand why disambig from SRILM translate this sequence into the result.

Another observation is that the result of trigram model is better than the bigram one, but in some cases the trigram model will take too much words into account and chose the wrong word.