SVitchboard 1: Small Vocabulary Tasks from Switchboard 1

SVitchboard 1
Small Vocabulary Tasks from Switchboard 1

Simon King, Centre for Speech Technology Research, University of Edinburgh, UK
Chris Bartels, Jeff Bilmes,Dept. of Electrical Engineering,U Washington, USA

Simon.King@ed.ac.uk

Why?

Goal

Define an ASR task

- vocabulary is
  - small/medium sized
  - completely closed (0% OOV rate)
- conversational telephone speech (CST)

Motivation

- when developing **novel approaches** to ASR
  - want to move away from working with read-text corpora
  - cannot work with large corpora like Switchboard or Fisher
  - wish to **decouple** acoustic modelling from the problems of
    - constructing a lexicon, language modelling, decoding, dealing with words that are unseen in the training data
- lattice rescoring is not an ideal solution
  - a large lattice does not sufficiently limit computation
  - the low WER region of search space represented by a small lattice may not overlap with the low WER region of a novel model
- **ideal** solution is a corpus of spontaneous, conversational speech with a small, limited vocabulary; of course, **no such corpus actually exists**

How?

Overview

Requirements

- want to construct
  - a **set** of tasks
  - of **varying vocabulary sizes**
- for convenience, the vocabulary/data of any task will be a subset of the vocabulary/data for any larger task
- for any given task, all words should occur in train, validation and test sets
- minimal number of low frequency words (i.e. not a long-tailed Zipf distribution)
- maximise amount of speech data for each vocabulary size

Method

- start with a very small vocabulary
- grow vocabulary iteratively, one word at a time
- at each iteration, choose one new word to add to the vocabulary
  - *that maximises the available data*
- select utterances from Switchboard 1 that are within the current vocabulary
- can stop the algorithm at any desired vocabulary size

Algorithm

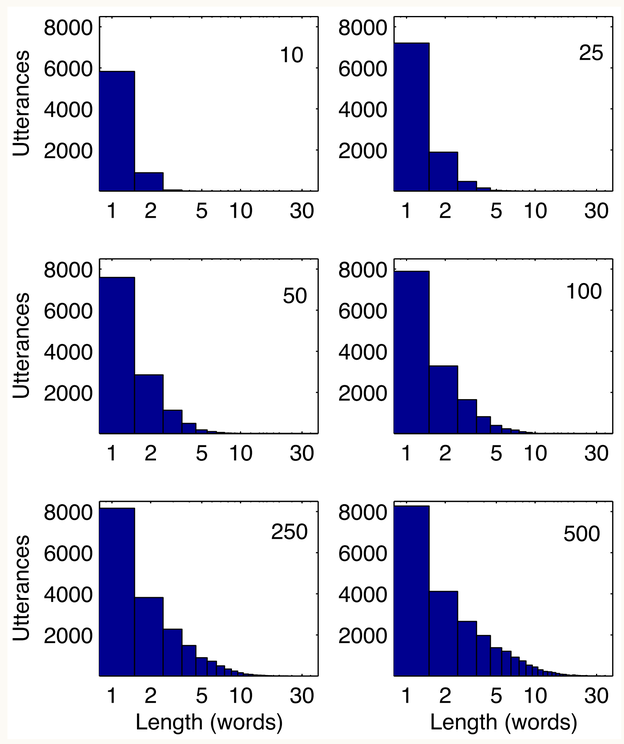
sv_vocabulary = 5 most common words in large corpus
oov_vocabulary = full_vocabulary $\setminus$ sv_vocabulary
while |sv_vocabulary| $<$ target number of words **do**
  for all $word \in$ oov_vocabulary **do**
    new_vocabulary = sv_vocabulary $\cup\ word$
    incoming_utterances = all utterances that only contain words in new_vocabulary
    count$_{word}$ = number of words in incoming_utterances
  end for
  $new\_word = \arg\max_{word}$ count$_{word}$
  sv_vocabulary = sv_vocabulary $\cup\ new\_word$
  oov_vocabulary = oov_vocabulary $\setminus\ new\_word$
end while

Practical details

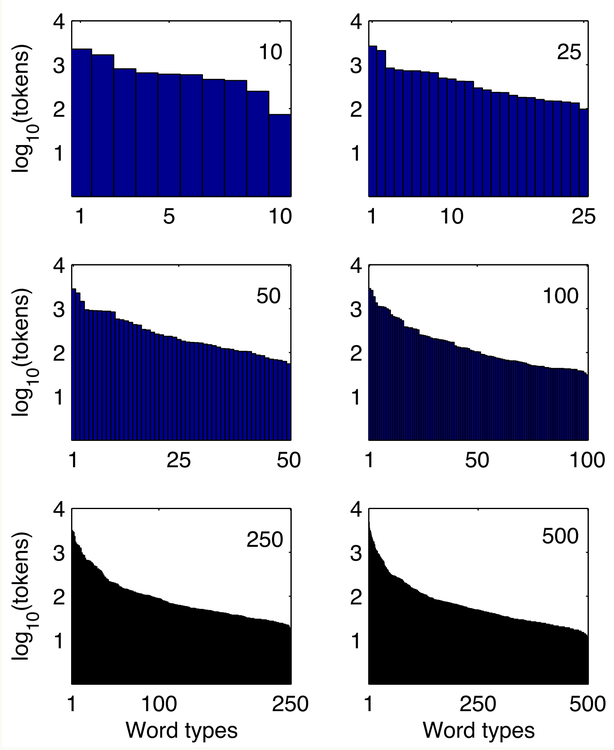
- remove all
  - disfluencies
  - word fragments
  - filled pauses
- each utterance must only contain in-vocabulary words
  - therefore, we “chunk” the data into small fragments by cutting at every silence longer than 500ms
- the algorithm is seeded with a vocabulary consisting of the 5 most common words in the corpus

The data selected

Utterance lengths



Word frequencies



Size

Data is divided into 5 partitions to allow a cross-validation scheme

Task	Partition	Utterances	Word tokens	Duration (hours) Total	Speech
10	A	1384	1617	0.67	0.20
	B	1275	1455	0.60	0.17
	C	1196	1389	0.56	0.16
	D	1446	1628	0.69	0.20
	E	1474	1703	0.70	0.20
	Total	6775	7792	3.22	0.93
25	A	1943	2698	0.95	0.29
	B	1887	2560	0.90	0.26
	C	1732	2359	0.83	0.25
	D	2078	2789	1.01	0.30
	E	2138	2918	1.04	0.31
	Total	9778	13324	4.74	1.42
50	A	2474	4228	1.24	0.39
	B	2392	3932	1.16	0.36
	C	2233	3789	1.10	0.34
	D	2594	4292	1.29	0.40
	E	2749	4673	1.37	0.43
	Total	12442	20914	6.16	1.93
100	A	2916	5814	1.51	0.51
	B	2794	5290	1.40	0.46
	C	2632	5237	1.34	0.45
	D	3059	5981	1.57	0.53
	E	3201	6289	1.64	0.55
	Total	14602	28611	7.47	2.48
250	A	3741	10400	2.10	0.81
	B	3681	10060	2.02	0.77
	C	3415	9336	1.88	0.71
	D	3927	10581	2.18	0.83
	E	4169	11573	2.32	0.89
	Total	18933	51950	10.50	4.01
500	A	4675	17948	2.92	1.30
	B	4673	17519	2.86	1.26
	C	4249	15857	2.60	1.13
	D	4871	18075	3.00	1.32
	E	5202	20021	3.23	1.43
	Total	23670	89420	14.62	6.44

No low-frequency words

In all but the 500 word task, every vocabulary word appears in every partition (A–E) and therefore in every train, validation and test set.

In the 500 word task, each partition has between 1 and 4 words fewer than 500, but the missing words are different for every partition so therefore the subtask training sets (ABC, BCD, etc) all contain every vocabulary word.

The lowest frequency word in each subtask (adding all 5 partitions together) has a frequency of 73, 97, 55, 31, 16, 10 in the 10, 25, 50, 100, 250 and 500 word tasks, respectively.

How to obtain and use SVitchboard 1

Download it for free (waveforms not included)

http://www.cstr.ed.ac.uk/research/projects/svitchboard

Using the corpus

- Each vocabulary size corresponds to a *task*
- Within each *task* there are five *subtasks*, numbered 1–5
- Each *subtask* specifies a particular arrangement of the five *partitions* (labelled A–E) into training, validation and testing sets according to this scheme:

Subtask	Train	Validate	Test
1	ABC	D	E
2	BCD	E	A
3	CDE	A	B
4	DEA	B	C
5	EAB	C	D

- Each subtask is to be performed independently of the others
- Where practical, it is preferable to perform a set of five independent experiments (as was done for the baseline results here). Report results by subtask and as an overall word error rate per task
- For systems employing a language model, this must also be trained on the training data specified above (the language model cannot be shared across all five sub-task experiments), or trained on other data.

See paper for full details.

Acknowledgements

The construction of SVitchboard 1 relies on the word alignments for Switchboard 1 provided by Mississippi State University (freely downloadable from their website).
Work partially carried out whilst the first author was hosted by the University of Washington.

Baseline whole-word HMMs + bigram

Results

Task	Subtask	Perplexity Val	Test	WER(%) Val	Test
10	1	3.1	3.2	20.2	20.8
	2	3.2	3.4	20.1	21.3
	3	3.4	3.4	21.6	21.0
	4	3.4	3.3	21.3	24.5
	5	3.3	3.1	23.3	20.9
	overall				21.6
25	1	5.0	5.2	35.6	34.9
	2	5.2	5.3	35.5	35.9
	3	5.2	5.4	35.0	37.2
	4	5.4	5.3	35.4	37.9
	5	5.2	4.9	37.6	35.4
	overall				36.2
50	1	7.6	8.1	48.4	48.4
	2	8.1	8.1	48.6	46.3
	3	8.1	8.4	46.1	49.3
	4	8.3	8.1	48.1	51.4
	5	8.2	7.6	50.8	48.3
	overall				48.7
100	1	11.4	11.5	57.9	56.8
	2	11.4	11.7	56.4	55.1
	3	11.7	11.7	55.1	58.8
	4	11.6	11.6	55.6	57.4
	5	11.8	11.5	57.5	57.2
	overall				57.0
250	1	21.7	23.2	65.8	66.2
	2	23.2	22.3	66.8	64.4
	3	22.2	23.5	63.9	67.1
	4	23.4	22.4	65.3	66.7
	5	22.6	21.7	66.7	64.5
	overall				65.7
500	1	38.4	39.5	69.8	70.8
	2	39.4	38.5	70.0	67.9
	3	38.1	39.7	67.6	70.0
	4	39.4	38.0	69.2	69.7
	5	38.2	37.9	70.1	68.9
	overall				69.5