Exploiting noisy web data by OOV ranking for low-resource keyword search

Zhipeng Chen, Ji Wu
Department of Electronic Engineering, Tsinghua University, Beijing
zp-chen11@mails.tsinghua.edu.cn

Motivation

Web text is easy to collect and often used to expand the vocabulary and assist language model training. But it is not easy for web crawlers to fetch fully matching text. Therefore, there might be two kinds of noise in web text.

- **Out-of-language text**: many languages share a similar alphabet or borrow words directly from other languages. (e.g., Swahili & English)
- **Out-of-domain text**: web text is usually a mixture of various genres and domains.

Web text is quite beneficial to OOV keyword search. But our oracle experiments indicated that false alarm OOVs from web data hurt ASR and IV KWS.

This work tried to improve the overall ASR and KWS performance by filtering out false alarm OOVs or re-assigning probabilities to them.

Word Ranking Criteria

1. **Unigram probabilities in web text (Web 1gram)**
2. **Word perplexity in grapheme-based LM (Word PPL)**
3. **Average perplexity of related sentences (Sent PPL)**
4. **Approximate counts in target speech (Appr. Count)**

   via a first-pass decoding, indexing and searching
5. **Re-estimated counts in target speech (Re-est Count)**

   use “Web 1gram” as prior of new words. This criterion is actually the sum of “Web 1gram” and “Appr. Count” (on the log scale)
6. **Logistic regression (LR)**

   combination of 1-4, as well as average OOV rate of related sentences and maximum posterior score in first-pass KWS. Regression model is trained on a held-out tuning set.

Probabilities Re-assigning

The recall of OOVs are critical to OOV KWS performance and the overall WER. We keep all new words in LM but re-assign their unigram probabilities according to the ranking results. The well-known Zipf’s law is applied.

$$\hat{P}(w) = \alpha \left(\frac{R(w)}{|C|} \right)^{\beta}$$

where $R(w)$ is the order of w, $|C|$ is the number of words in the new word set C.

Experiments

IARPA Babel Program Swahili VLLP (3h, 5k word types)

Web data: officially provided in NIST OpenKWS15, with simple filtering (2.1M sentences, 169K new words)

Test set: 10h, 71.1% words not covered by VLLP

Tuning set: 3h, 61% words not covered by VLLP

Lexicon: graphemic, G2P is not needed.

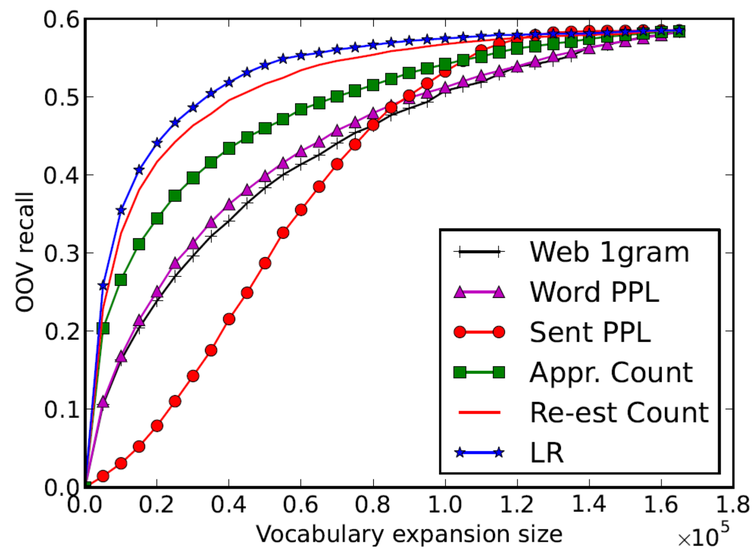
Keyword list: 756 IV, 512 OOV (454 in web data)

Top 4 words from LR:

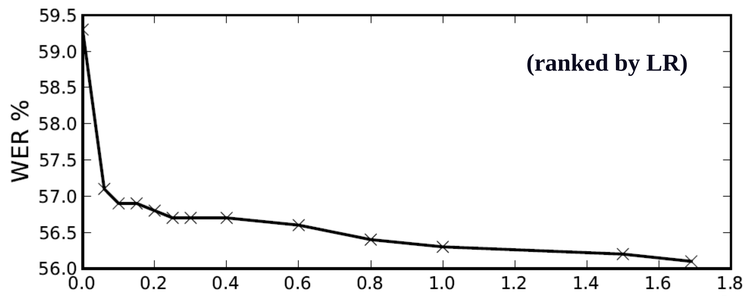
ujumbe (messages), *kuwasaidia* (help), *kuandika* (write), *mabadiliko* (changes)

Last 4 words from LR:

qqqq, *qqqkq*, *qqq*, *kqqq*



Most of the gain comes from the first 10K new words.



IV MTWV is improved by ~1% with only 1/8 vocabulary size and 50% time and space for WFST construction

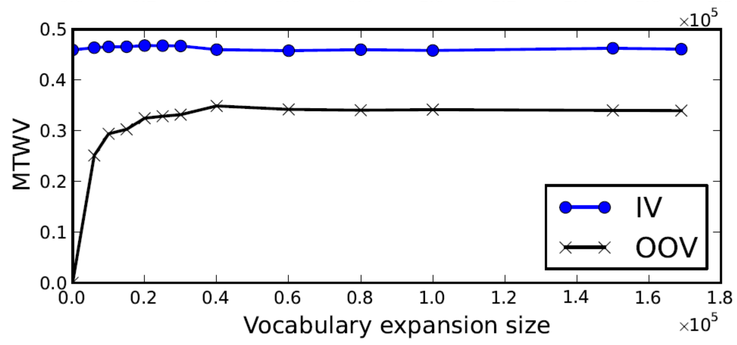


Table 2: *STT and KWS performance in WER and MTWV.*

LM	WER	IV	OOV	Overall
Baseline				
VLLP	60.5	0.4304	0	0.2566
PPL	57.0	0.4514	0.3239	0.3999
Diff. CE	56.2	**0.4626**	**0.3420**	**0.4139**
Full expansion	**56.1**	0.4608	0.3396	0.4119
Partial expansion (top 20k)				
Web 1gram	59.2	0.4334	0.1659	0.3254
Word PPL	57.7	0.4592	0.1299	0.3262
Sent PPL	59.9	0.4343	0.0348	0.2730
Appr. Count	57.3	0.4527	0.2243	0.3605
Re-est Count	58.8	0.4193	0.2973	0.3700
LR	**56.7**	**0.4679**	**0.3246**	**0.4100**
Probabilities re-assigning				
Web 1gram	**56.0**	0.4649	0.3434	0.4158
Word PPL	**56.0**	**0.4709**	0.3446	**0.4199**
Sent PPL	56.2	0.4682	0.3417	0.4171
Appr. Count	56.2	0.4624	0.3389	0.4125
Re-est Count	56.1	0.4649	0.3445	0.4163
LR	**56.0**	0.4641	**0.3448**	0.4159