Language Modeling: What Matters Most?

Nicolae Duta, Richard Schwartz

Overview

- **Using an out-of-domain corpus**
- **The importance of singletons**
- **What aspects of the language model are important?**

Using Out-of-Domain Corpus

- **When we don't have enough LM training data we try to compensate by using out-of-domain data.**
- **When will it help?**
- **We measure the "trigram hit rate": the percentage of trigrams in the test corpus that are explicitly in the language training.**

Arabic “Call-Home” Conversational Speech

- **Conversational training corpus: 150K words**
- **News corpus (mostly text): 300M words**
- **Development set: 32.5K words**

LM Training Data	Vocab Coverage	3-gram Hit Rate	Per-plexity	WER
Conversational Alone	90.6	20%	203	60.5
News	89.5	4%	5042	--
Conversational + News	96.5	21%	200	60.5

- **Even though the vocabulary coverage is high, the trigrams in the news are different so the 300M words of news do not help at all.**

English SWBD

- **Test set (Hub5 Eval 01): 70K words**

LM Training Data	3-gram Hit Rate	WER
SWBD (3.7M words)	70%	31.9%
Broadcast News (152M words)	82%	32.1%
News Text (377M words)	86%	32.5%
8 x SWBD + BN	85%	29.9%

same

- **High hit rate can make up for being out-of-domain**
- **But the statistics of the domain do matter**

Importance of Singleton Trigrams

- **Most 3-grams are seen very few times.**
  We usually discard trigrams seen less than 3 or 4 times and bigrams seen less than 2 or 3 times.
  Is this wise?

Condition	3-gram Hit Rate	WER
Discard singleton 3-grams from BN	80.5%	30.5%
No discarding	85.0%	29.9%

- **Discarding just singleton 3-grams from BN data (for Hub-5 test) affects hit rate and WER significantly.**
- **Even though these singletons are under-trained, and the model is 3 times larger, there is a significant gain for including them.**

Conclusions

- **Adding an out-of-domain corpus helps only if it has a high 3-gram hit rate.**
- **No trigrams should be discarded – even from an out-of-domain corpus.**
- **We need to understand the relative importance of trigram hit rate and the probabilities.**