Bayesian skipgram language model

In this work we add skipgrams to Bayesian language models based on the hierarchical Pitman-Yor processes language models (HPYPLM). HPYPLM follows the analogy of the Chinese restaurant process, where the meals represent the pattern types, and the occurrences the pattern tokens. We look at backoff strategies and the effects of domains, and report perplexities (lower is better).

Skipgrams are beneficial for language modelling

- Within-domain reduction of 10 (6%) perplexity
- Cross-domain reduction of 680 (39%) perplexity
- Within-domain n-gram performance converges to skipgram performance (with enough data)
- For generic corpora, use limited recursive backoff
- For cross-domain use full recursive backoff evaluation

Effects of skipping

- We skip meals (dishes) by thresholding on the unigram frequency (W), and on customers by thresholding on n-gram frequency (T). Conclusion: If you have enough data and sufficient power, don’t threshold on meals. It makes you hungry.

Contribution of skipgrams

- For similar domains, use limited recursive backoff evaluation
- For cross-domain use full recursive backoff evaluation

When do skipgrams work?

Most-helpful patterns: domain-specific patterns with no direct n-gram match if the \{1\} blood, : in \{1\} with, treatment \{1\} cervical cancer

Most-frequently helpful patterns: more generic, structural patterns \{2\} , \{2\} , see \{2\} , of \{2\} and Skipping over low-frequent words: generic patterns the \{1\} of the, the \{1\} of a, in the \{1\} of, to the \{1\} of

Backoff methods

- \texttt{ngram} — full recursive backoff to shorter n-grams
- \texttt{limited} — recursive backoff to all patterns \(\leq n\) until match
- \texttt{full} — recursive backoff to all patterns \(\leq n\)

Data and evaluation

- \texttt{emea}: 12M words of biomedical texts from the European Medicines Agency
- \texttt{jrc}: 31M words of legislative texts from the European Union (JRC-Acquis v3.0)
- \texttt{1bw}: 800M words generic texts from the Google 1 Billion Words language modelling benchmark
- \texttt{wps}: 70M Wikipedia texts (5% sample of November 2013’s snapshot)

Reproduce my results!

C++ toolkit for Bayesian language modelling with skipgrams: https://github.com/naiaden/cococyp

louis onrust
l.onrust@let.ru.nl

Antal van den Bosch
Hugo Van hamme

KU LEUVEN

Radboud University