

9 Months Progress Report

Louis Onrust

CLS, Radboud University ESAT-PSI, KU Leuven

September 29, 2014

Overview

- ▶ Scope of the project
- ▶ Bayesian language models
- ▶ Results
- ▶ Research plan
- ▶ Side projects
- ▶ Reflections
- ▶ Formalities

Scope of the project

Scope

- ▶ Language models
- ▶ Latent variable models
- ▶ Domain-dependence of LVLM
- ▶ Intrinsic & extrinsic evaluation

Goal

Bring back language modelling in Bayesian language modelling

Bayesian Language Model

- ▶ The goal is to derive the partition underlying the data
- ▶ But we only have the word counts

Clustering

- ▶ Each n -gram is a cluster
- ▶ Each n is a layer
- ▶ Each history is in a cluster at the $(n - 1)$ th layer

Hierarchical Pitman-Yor Chinese Restaurant Process

- ▶ CRP and DPCRP give logarithmic growth
- ▶ Language manifests typically in power law growth
- ▶ PYCRP as generalisation of CRP and DPCRP

CRP No parameters

DPCRP Concentration parameter α

PYCRP Concentration parameter α and discount
parameter γ

Bayesian Language Model: The Implementation

Implementation

We use the following software:

- `cypyp` an existing C++ framework on BNP with PYP priors
- `colibri` an existing C++ pattern model framework

Advantages

- ▶ We can now handle many patterns such as n -grams, skipgrams and flexgrams
- ▶ Thresholding patterns on many levels
- ▶ Efficient storage of patterns

Results

Data sets

- ▶ JRC English
- ▶ Google 1 billion words
- ▶ EMEA English

Backoff methods

- ▶ n -gram backoff
- ▶ Limited recursive backoff
- ▶ Full recursive backoff

Evaluation measure Intrinsic evaluation with perplexity

Summary

- ▶ Within domain evaluation yields best performance
- ▶ Adding skipgrams increases performance on cross domain evaluation
- ▶ For generic corpora, limited recursive backoff performs best
- ▶ Seems to outperform Generalised Language Model
- ▶ If significant, perhaps not enough for extrinsic evaluation

Results: Within domain versus cross domain

Training with only n -grams

| | jrc | 1bw | emea |
|------|-----|------|------|
| jrc | 13 | 1195 | 961 |
| 1bw | 768 | 158 | 945 |
| emea | 600 | 1143 | 4 |

Training with both n -grams and skipgrams

| | jrc | 1bw | emea |
|------|-----|------|------|
| jrc | 13 | 1162 | 939 |
| 1bw | 751 | 162 | 921 |
| emea | 581 | 1155 | 4 |

Relative differences

| | jrc | 1bw | emea |
|------|------|------|------|
| jrc | 2.0 | -2.8 | -2.3 |
| 1bw | -2.2 | 2.4 | -2.6 |
| emea | -3.2 | 1.1 | 0.7 |

Results: Effects of different backoff methods

| | | <i>n</i> -grams | | | Skipgrams | | |
|------|---------|-----------------|------|------|-----------|------|------|
| | | jrc | 1bw | emea | jrc | 1bw | emea |
| jrc | ngram | 13 | 1510 | 1081 | 13 | 1843 | 1295 |
| | limited | 14 | 1477 | 1122 | 13 | 1542 | 1149 |
| | full | 69 | 1195 | 961 | 65 | 1195 | 939 |
| 1bws | ngram | 768 | 158 | 946 | 879 | 163 | 1105 |
| | limited | 815 | 185 | 1025 | 751 | 162 | 921 |
| | full | 801 | 264 | 1039 | 768 | 252 | 988 |
| emea | ngram | 769 | 1552 | 4 | 969 | 2089 | 4 |
| | limited | 779 | 1385 | 4 | 838 | 1655 | 4 |
| | full | 600 | 1143 | 32 | 581 | 1155 | 32 |

Research Plan

Focus

- ▶ Cross domain language modelling with skipgrams

Experiments

- ▶ Validate significance by testing multiple languages
- ▶ Investigate influence skipgrams with qualitative analysis
- ▶ When we find a more substantial drop in perplexity:
 - ▶ Machine translation experiments
 - ▶ Automated speech recognition experiments
- ▶ Investigate multi-domain language models

Writing in progress

- ▶ TACL journal paper on our findings
 - ▶ ACL, EMNLP, ICASSP, ...
- ▶ Background/Methodology section of PhD thesis

Side Projects

Parsimonious Language Models

The goal is to model the differences between corpora

- ▶ Only store salient differences:
 - ▶ document-specific terms and patterns
 - ▶ domain-specific terms and patterns

Realistic Motif Detection

The goal is to find motifs in folk tales at a sentential level

- ▶ Take order of motifs in consideration
- ▶ Sentences can take any number of motifs
- ▶ Un-, semi-, and supervised learning
- ▶ Incorporation of domain and genre knowledge

Reflections

Struggling with reproducing results

- ▶ No data or code provisional
- ▶ Instructions unclear and fuzzy
- ▶ Fast pacing and non-dedicated research lines

Missed the boat

- ▶ Good ideas, but obviated by other publications
 - ▶ HPYLM with $n \rightarrow \infty$: Stochastic Memoiser
 - ▶ Bayesian PLM

Little help from outside, but learned anyway

- ▶ A lot of literature, but confusing or contradicting
- ▶ Still a relative small research community
- ▶ Good foundation for further work

Formalities

Teaching and Supervision

- ▶ Supervision of master students in a competition on sentiment analysis
- ▶ Supervision of a master student for a task to predict reduction in speech

Training and Education

Participated

- ▶ Academic writing
- ▶ Research methods and methodology
- ▶ Applied Bayesian statistics school on Bayesian non-parametrics

To participate in

- ▶ Mathematical methods
- ▶ Presentation skills
- ▶ Any relevant event