p(conclusions|Skipping {*2*})

Bayesian Language Modelling with Skipgrams

Radboud University

Bayesian Language Modelling with Skipgrams

Louis Onrust
Centre for Language Studies, Radboud University
Center for Processing Speech and Images, KU Leuven

l.onrust@let.ru.nl
github.com/naiaden

LaTeX Beamer template for RU corporate style
See my github page for example code:
github.com/naiaden/presentations/tree/master/ruhuisstijl

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
- Improve cross domain langauge modelling with skipgrams

Language Model

Traditional method

The process:

- Read n-gram p
- Increment frequency of p
- Repeat, preferably ad infinitum

n-gram probabilities are then determined by their MLE

Smoothed Traditional Language Model

What to do when the occurrence count of p is 0?

- Not assign 0 as probability $\rightarrow$ smoothing
- Fall back to the $(n-1)$ words of $p \rightarrow$ backoff

One of the best methods is still Modified Kneser-Ney: backoff and smoothing

Language Model

Bayesian method

- Assume texts are generated by some process
- Consider the texts to be a sample from the process
- Infer underlying process

Bayesian Language Model

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

Chinese Restaurant Process

- Clusters are tables, n-grams tokens are customers
- Initially tokens seat at the same table
- In the inference step, customers get to choose a new identity

Bayesian Language Model

Chinese Restaurant Process: Inference

When n are in the restaurant, people sit a table t with probability

- $\frac{1}{n}$ with another n-gram p, or $\frac{|t|}{n}$ at the same table as p
- $\frac{1}{n}$ at a new table

The number of tables grows logarithmically

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

- HPYCRP to model inherent hierarchical structure n-gram

Bayesian Language Model: The Implementation

Motivation

Existing Bayesian language models...

- are merely an algorithmic showcase without real language modelling aspirations
- cannot handle really big data sets

Implementation

We use the following software:

cpyp an existing C++ framework on BNP with PYP priors

colibri an existing C++ framework for pattern modelling

Advantages

- We can now handle many patterns such as *n*-grams, skipgrams, and flexgrams
- Tresholding patterns on many levels

Results: The Setup

Data Sets

- JRC English
- Google 1 billion words
- EMEA English

Backoff Methods

***n*-gram** full recursive backoff to shorter *n*-grams

Limited recursive backoff to all patterns $\leq n$ until match

Full recursive backoff to all patterns $\leq n$

Evaluation Measure

- Intrinsic evaluation with perplexity

Results: An Overview

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: Domains and Patterns

Observations

domains Within domain evaluation yields best performance

patterns Adding skipgrams increases performance on cross domain evaluation

Training with only *n*-grams

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

and with 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: Effect of Different Backoff Methods

Observations

backoff For generic corpora, limited recursive backoff performs best

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

Future Work

Experiments

- Validate significance by testing on 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 (DHPYPLM)
- Generalise skipgrams to flexgrams
- ...

http://dilbert.com/strips/comic/2009-09-17