Machine Translation Models IBM 1 and IBM 2

First Author
Maartje de Jonge
0194107
maartjedejonge@gmail.com

Second Author
Lina Murady
xxx
lina.murady@gmail.com

Abstract

Scope: empirical evaluation of IBM 1 and IBM 2 models

Contributions: - performance of the different models on test data - analyses of the models and their performance

1 Introduction

- Statistical Machine Translation - Baysian split: $p(e|f) \propto p(e) * p(f|e)$ - We focus on translation model $p(f|e)$

- alignment model: word pairs (f, e) with the constraint that each french word matches exactly one english word. - Null word added to english sentence to align words in french that do not have an equivalent word in english (insertions)

- decomposition into: sentence length probability, alignment prob, translation prob (- alignment prob mixture component) - IBM 1: assume uniform alignment probability, - train with EM (explain why)

- Shortcomings of IBM 1 (assumption uniform alignments) - IBM 2 learn probabilities $p(i, j, I, J)$ - problem: too many parameters for small training sets - approach: jump probabilities (**?**), model probabilities as jumps from diagonal. - train with EM

- Problem with maximum likelihood estimaion, arguments for Bayesian approach - Problem with posterior inference which motivates variational inference - We use Dirichlet Prior and Variational Inference to meet these limitations

- In this report we compare alignment models IBM 1, IBM 2 and IB 1 with variational inference - We empirically evaluate how these models perform on a corpus and we discuss their differences

- Section 3.1 - Section 3.2 - Section 3.3 - Section 3.4

2 Models

2.1 IBM 1

- describe the model mathematically - mathematical assumptions - factorisation - parameterisation - limitations - parameter estimation: EM - inference techniques: viterbi alignment

- cite some literature

2.2 IBM 2

- describe the model mathematically - mathematical assumptions - factorisation - parameterisation - limitations - parameter estimation: EM - inference techniques: viterbi alignment

- cite some literature

2.3 IBM 1 with Variational Inference

- describe the model mathematically - mathematical assumptions - factorisation - parameterisation - limitations - parameter estimation: variational inference - inference techniques: viterbi alignment

- cite some literature

Jump Parameterization - why: lot of parameters for small data set - math: formula - intuition: diagonal - literature: vogel

3 Experiments

3.1 Experimental Setup

- datasets: training, validation, test - numbers, languages, where does the data come from?

- setup

- Viterbi Alignment: how do we deal with unknown words in the validation/test set

- metric: AER

- stop/convergence criteria: 1) based on training log likelihood Relative log-likelihood convergence: $\frac{ll_i - ll_{i-1}}{ll_{i-1}} < \epsilon$ why relative? what epsilon?

2) best AER on validation set Absolute criterion. why? what epsilon? $prevAER - AER < 0$

3.2 IBM 1 with Expectation Maximization

Training Conditions - uniform initialisation

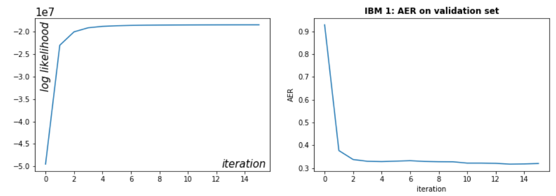

Figure 1: Evolution of training log likelihood and validation AER for IBM 1.

Results - Figure: training log-likelihood vs iteration

- Figure: validation AER vs iteration
- Figure/tabel: AER on test set using model selected based on AER and based on log-likelihood [remark: use official tool instead of python code]

3.3 IBM 1 with Variational Inference

Training Conditions - uniform initialization

- choice of hyper parameter

Results - Figure: training log-likelihood vs iteration

- Figure: validation AER vs iteration
- Figure/tabel: AER on test set using model selected based on AER and based on log-likelihood [remark: use official tool instead of python code]

3.4 IBM 2 with Expectation Maximization

Training Conditions - initialization Non-convex, thus local minimum, result depends on initialization 1) uniform 2) random 3 times 3) staged, use result of model 1 run

Results - Figure: training log-likelihood vs iteration using different initializations a) uniform b) random 3 times c) staged

- Figure: validation AER vs iteration using different initializations a) uniform b) random 3 times c) staged
- Figure/tabel: AER on test set using model selected based on AER and based on log-likelihood [remark: use official tool instead of python code] compare IBM 1 with IBM 2

3.5 Discussion

- (non)-convexity - stability - convergence.
- complexity - qualitative insight: i.e. distributions for rare words, frequent words and jump distribution

4 Conclusion and Future Work

- Future work: Dirichlet on IBM 2
- Future work: IBM 2 with jumps for other languages with completely different word order, i.e. not most probability mass on diagonal
- Comparison: Which model is the best, why?
- contributions
- limitations