

COMP90042 Web Search & Text Analysis

Workshop Week 12

Zenan Zhai

May 28, 2019

University of Melbourne

Machine Translation

- Statistical MT - IBM I
- Phrase Translation
- Neural MT
- Evaluation

$$P(f, a|e) = \frac{\epsilon}{(1 + I)^J} \prod_{j=1}^J t(f_j|e_{a_j})$$

Components in IBM I

- E : Original language
- F : Target language
- A : Alignments from F to E

Why normalize by $(1 + I)^J$?

What is e_{a_j} representing ?

Expectation Maximisation

- Widely used in unsupervised learning
- Topic modeling (Latent Dirichlet Allocation)
- Clustering (Gaussian Mixture Model)

Intuition

- E-Step: Estimate distribution based on current parameters θ .
- M-Step: Update parameters θ based on the estimated distribution.

How is it applied in IBM I?

So how do we estimate the alignments?

- What is not observed? (So we need to estimate them)
- What are the variables θ in this case?

$$P(a|f, e) = \frac{P(f, e, a)}{P(f, e)} = \frac{P(f, a|e)P(e)}{P(f|e)P(e)} = \frac{P(f, a|e)}{P(f|e)}$$

Looks familiar?

$$P(a_j|f, e) = \frac{t(f_j|e_{a_j})}{\sum_j t(f_j|e_{a_j})}$$

Now, use the expectation of alignments to train our parameters θ .

Recall training for generative models (N-grams, HMM, etc).

$$t(f_j|e_{a_j}) = \frac{\text{count}(f_j, e_{a_j})}{\text{count}(e_{a_j})}$$

$$\text{count}(f_j, e_{a_j}) = \text{count}(f_j, e_{a_j}) + P(a_j|f, e)$$

What is the next step?

Phrase Translation - Key points

Steps for Phrase Translation

1. Learn word alignments (e.g. A in IBM I)
2. Extract phrase from the co-occurrence matrix
3. Train language model with phrase structures

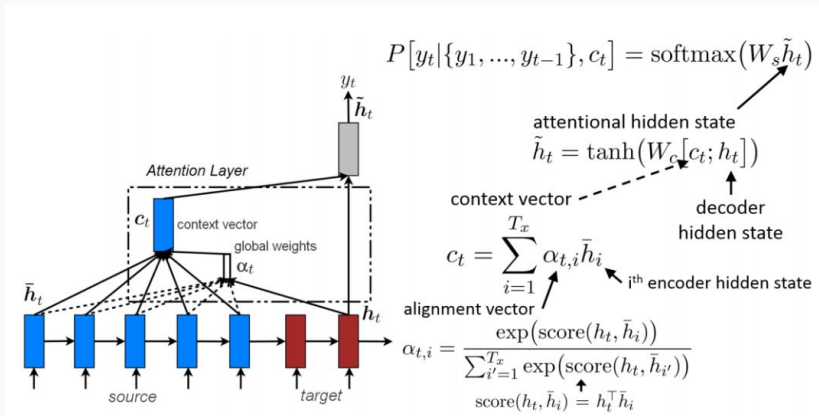
Inference

- Beam search.
- Dynamic programming - e.g. the Viterbi algorithm

Neural LM

RNN-Based seq2seq models, using encoder-decoder structure.

- Use the final representation from encoder as first input to the decoder.
- Global attention.



Transformers (Vaswani et al., 2017)

Evaluation - BLEU

Widely used evaluation method for MT, Image Captioning, Summarisation.

Measuring discrepancy between human/machine outputs.

$$BLEU = bp \times \exp\left(\frac{1}{N} \sum_{n=1}^N \log(p_n)\right)$$

- Weighted average of 1-4 grams precision.
- $bp = \min\left(1, \frac{\text{len}(\text{machine})}{\text{len}(\text{reference})}\right)$
- $p_n = \frac{\min(\text{correct}, \text{len}(\text{reference}))}{\text{len}(\text{machine})}$

Criticism: *BLEU at your own risk*

Indexing

- Data Structure
 - Document-Term Matrix
 - Inverted Index
- Compression
 - Variable Byte Compression
 - OptPFor Delta Compression
- Index Construction
 - Invert Batch Indexing
 - Auxiliary Indexing
 - Logarithmic Indexing

Search

- Vector Space Models
 - TF-IDF
 - BM25
- Efficient Query Processing
 - Operation GEQ
 - WAND
- Query Completion
 - Prefix Trie
 - Range Maximum Query
- Query Expansion
 - Relevance Feedback
 - Semantic-Based Methods
- Phrase Search
 - Inverted Index + Positional Information
 - Suffix Array
- Evaluation and Re-rank

Features

1. Word Semantics
 - Lexicon semantics
 - Distributional semantics
2. Sequence Labeling
 - Part-of-speech tagging
 - Named entity recognition
3. Parsing
 - Dependency parsing
 - Phrase-structure parsing

Applications

1. Text classification
2. Question answering
3. Discourse tasks
4. Machine translation
5. Summarization
- ...

Concept questions

- Concrete concept of any given topic, one sentence.

Method questions

- Describe general approach to given task.

Algorithm questions

- Methods in action. What we know how to train and infer.

Essay questions

- Can be boarder topics, application of NLP.