Robustness in Statistical Language Modelling: Review And Perspectives

presentation for
T-61.182 -
Robustness in Language and Speech Processing

Based on the article of the same title by
J. Bellegarda

presented by Matti Aksela

Introduction – 1

- A Statistical Language Model (SLM)
  - "A language model tries to encapsulate as much as possible of the syntactic, semantic and pragmatic characteristics for the task considered"
  - SLM : Statistical in nature, for example n-grams, Stochastic finite state automata etc.
  - In this presentation, consideration is based on the n-gram paradigm

- SLM robustness
  - The effectiveness of a SLM is directly related to its ability to discriminate between strings of words
  - This is influenced by two related issues, convergence and estimation
  - Coverage refers to the underlying vocabulary while estimation to the length of the string of words evaluated (n in n-grams)
  - The effect of training data cannot be over-emphasized
  - Constraining the speech naturally helps recognition, but effects generalization

Introduction – 2

- How to optimize performance despite mismatches between training and testing conditions

1. Coverage optimization – lexical coverage and model coverage - unseen elements cause problems!
2. Robust estimation – less than perfect coverage leads to unobserved strings, which must be handled somehow
3. Information aggregation – words behaving “like” each other provide information on one another
4. Span extension – extend/complement n-grams with larger-span information
5. Language model adaptation – use information from the task at hand in conjunction with an underlying model

Coverage Optimization – Lexical coverage

- Lexical coverage problem
  - Unknown, or out-of-vocabulary (OOV), words
  - OOV almost surely generates a substitution error
  - This may also cause the next word to be misrecognized ("ripple effect" of OOV words)
- General principles for vocabulary optimization
  - Inherently task-dependent
  - Coverage is strongly effected by the amount of training data used
  - Source and recency of the training data is very important
  - Trade-off: OOV rate vs. acoustic confusability
- Example: NAB (North American Business business publication news collection)
  - training data amount has effect until 30-50 mill.
  - optimal vocabulary size between 55 000 and 80 000
  - each OOV results in an average of 1.2 errors

Coverage Optimization – n-gram coverage

- Lexical coverage is a subproblem of n-gram coverage ($n = 1$)

- Frequency of the grams decreases rapidly as n increases
  - The amount of training data needed for reliable estimation is huge (100-200 million words for bi-grams)

- Language evolution effects n-gram coverage
  - Acquiring data takes time, during which the language patterns may shift...

- Also highly language-dependent
  - Compounds, inflection, tense, . . .

Robust Estimation

- Due to suboptimal n-gram coverage, some strings are never observed and many very infrequently
- Classical smoothing
  - The discounting and redistribution paradigm : a portion of the probability mass corresponding to frequent items is redistributed across infrequent and never observed ones
  - how to define how much of the probability mass to redistribute and how to redistribute it?
  - Approaches for discounting: Linear discounting, absolute discounting, floor discounting, Good-Turing discounting
  - Approaches for redistribution: Interpolation, back-off
- Robustness can also be sought through the maximum entropy criterion, leading to minimum discrimination information (MDI) estimation
  - Knowledge sources are introduced in terms of constraints that the underlying distribution should satisfy

Information Aggregation – Class Models

- Information from similar, rare, events may be aggregated
  - Class models to take advantage of words that behave "like" each other in the given context
  - Makes frequency counts more reliable

  $Pr(w_q|H^n_{q-1}) =$
  $\sum_{\{C_q\}} \sum_{\{C^n_{q-1}\}} Pr(w_q|C_q) Pr(C_q|C^n_{q-1}) Pr(C^n_{q-1}|H^n_{q-1}),$
  where $\{C_q\}$ is the set of possible classes and $\{C^n_{q-1}\}$ the set of possible class histories.

- Several class model approaches
  - Grammatical units such as part-of-speech or morphological units
  - Divisive clustering to maximize average mutual information of adjacent classes
  - Divisive clustering based on *a posterior* distributions on word co-occurrences

Information Aggregation – Mixture Models

- Information may also be aggregated across several domains
  - Combine models trained on K different corpora

$$Pr(w_q|H_{q-1}^n) = \sum_{k=1}^{K} \lambda_k(H_{q-1}^n) Pr_k(w_q|H_{q-1}^n),$$

- Interpolation coefficients can be estimated using the EM algorithm on a comparatively small amount of data closely related to the task at hand

Span extension – 1

- Related words may be far from another:
  stocks, as a result of the announcement, sharply fell

- Variable length models
  - Include frequent word compounds
  - Several approaches; join word pairs with high MI, decision trees to determine class equivalence
  - May expand span, but not by much

- Use of structure
  - Structural information may be added if a good parser is available
  - One approach is to take into account the hierarchical nature of language; determine headwords and use n-gram models on them
  - Performance highly dependent on the parser

Span extension – 2

- Topics
  - Use a large set of topics T_k,

$$Pr(w_q|H_{q-1}^n) = \sum_{k=1}^{K} Pr(w_q|T_k)Pr(T_k|H_{q-1}^n)$$

  - The main uncertainty is the topic clustering
  - Even knowledge of the correct topic may not help

- Word trigger pairs
  - Word pairs showing significant correlation in the training corpus may be used to trigger words
  - The first encountered part of the pair increase the others probability
  - In practice search for word pairs of high mutual information inside fixed length windows
  - Problems, as different pairs may have markedly different behavior

- Latent Semantic Analysis (LSA) may be used for trigger pair selection
  - Can find words that tend to appear in similar documents and documents that tend to convey the same semantic meaning

Language Model Adaptation

- Cache models
  - Short-term features are collected to a cache model, which is then combined (for example linearly) with a static underlying model
- Adaptive mixture models
  - Adaptive mixture SLMs estimate the interpolation coefficients from the history for the word under consideration
- If a dynamic model and a underlying static model are used, EM can be used to determine the weighting (or the robust smoothing techniques presented previously)

Conclusions

- The main problem is to overcome the potential weaknesses of the training data, limitations of the used paradigm and a possible mismatch between training and testing conditions.
  - Coverage optimization and robust estimation attempt to relieve problems caused by training data insufficient for common estimation methods
  - Information aggregation seeks to reduce the number of parameters needed to evaluate through equivalence classes
  - Span extension aims at encapsulating higher-level knowledge into the SLM
  - Language model adaptation seeks to update the SLM with task-specific information
- None of the approaches are mutually exclusive
- The first approaches seek to relieve the lack of data problem
- The latter two seek to incorporate more information, which may be more profitable in the future