Improvements in DP Beam Search for Phrase-based SMT

Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany
Overview

1. Introduction & related work
2. Search for phrase-based MT
3. Experimental results
4. Summary & conclusions
Contributions

• clear & precise description of phrase-based search
• analysis of important aspects
  – rest score estimation
  – lexical vs. coverage hypotheses
  – beam search including cube pruning
on a large data task
Related Work

- based on
  - [Zens & Och+ 02]: phrase-based model
  - [Och 02]: rest score estimation (for AT)
  - [Tillmann & Ney 03]: search for SWB models

- other related work:
  - Pharaoh [Koehn 03], Moses [Koehn & Hoang+ 07]
  - many others, e.g. [Tillmann 06], [Moore & Quirk 07], ...
System Architecture

Source Language Text

Preprocessing

Global Search

\[ \hat{E} = \arg\max_{E} \{ p(E|F) \} \]
\[ = \arg\max_{E} \{ \sum_m \lambda_m h_m(E, F) \} \]

Postprocessing

Target Language Text

Models

Language Models

Phrase Models

Word Models

Reordering Models

. . .
interdependencies:
• find phrase boundaries
• reordering in target language
• find most ‘plausible’ sentence

costants:
• no gaps
• no overlaps
Search

• **goal:** \( \arg\max_{E} \left\{ \max_{S} \sum_{m=1}^{M} \lambda_m h_m(E, S; F) \right\} \)

  with target sentence \( E \), segmentation \( S \), source sentence \( F \), models \( h(\cdot) \), weights \( \lambda \)

• **models:**
  
  – within phrase models:
    phrase lexica, word lexica, word penalty, phrase penalty
  
  – \( n \)-gram backing-off language model
  
  – distortion penalty
Search Space

• **source sentence**  \( F = f_1, \ldots, f_J \)

• **states** \((C, \tilde{e}, j)\)
  
  – **coverage**  \( C \subseteq \{1, \ldots, J\} \): translated input positions
  
  – **LM history** \( \tilde{e} \) to predict the next target word
  
  – **source position** \( j \) for the distortion model

• **edges** \((\tilde{e}, j, j')\)
  
  – generate target phrase \( \tilde{e} \)
  
  – which covers the source sentence words \( f_j, \ldots, f_{j'} \)

• **expanding** \((C, \tilde{e}, j)\) with \((\tilde{e}', j'', j')\) results in state

\[
(C \cup \{j'', \ldots, j'\}, \tilde{e} \oplus \tilde{e}', j')
\]
Lexical vs. Coverage Hypotheses

• (partial) hypothesis: path to state \((C, \tilde{e}, j)\)

• for each cardinality \(c = |C|:\)
  we have a list of coverage hypotheses \(C\)

• for each coverage \(C:\)
  we have a list of lexical hypotheses \((\tilde{e}, j)\)

• beam search: limit the list sizes
Search Illustration

Legend:
- Coverage Hypothesis
- Lexical Hypothesis

Zens: Improvements in DP Beam Search for Phrase-based SMT
IWSLT Oct, 2008
Algorithm Details

- DP beam search
  - generate hypotheses with increasing cardinality by expanding hypotheses with lower cardinality
  - recombine hypotheses with same state
  - expand only promising hypotheses

- share computations between expansions, e.g. check for overlap, rest score computation, ...

- early pruning
  - stop expansion as soon as possible

- expand most promising candidates first
Rest Score Estimation

- estimated score of hypothesis completion (inspired by A*)

- previous work:
  - [Och 02, Och & Ney 04]
    TM & LM per source position, distortion
  - [Koehn 03]
    TM & LM per source sequence, no distortion

- here: comparison of
  - computation per position and per sequence
  - models: TM only; TM & LM; TM, LM & distortion
Experimental Results

- NIST Chinese-English large data task
- **TM:**
  - training data: 8 M sentence pairs, 250 M words
  - phrase-based, word-based lexica, word / phrase penalty
- **LM:**
  - 4-gram, trained on 650 M words, SRILM [Stolcke 02]
- reordering:
  - distortion penalty, reordering window: 10
  - lexicalized reordering model [Zens & Ney 06]
- evaluation:
  - case-insensitive Bleu score (mt-eval) on NIST 2002 test set
Effect of Search Errors

Translate test set with various pruning parameters settings.
Model score averaged over whole test set (878 sentences).
Rest Score Estimation

![Graph showing BLEU score improvement with varying maximum number of hypotheses per source word.](image)

- **None**
- **per Position:**
  - TM
  - +LM
  - +Dist
- **per Sequence:**
  - TM
  - +LM
  - +Dist
Lexical vs. Coverage Hypotheses

![Graph showing BLEU performance with different maximum coverage hypotheses and maximum number of lexical hypotheses per coverage hypothesis.](image)

- **BLEU [%]**: Accuracy metric for machine translation models.
- **Max. Number of Lex. Hyps per Cov. Hyp.**: Maximum number of lexical hypotheses per coverage hypothesis.
- **Max. Cov. Hyps**: Maximum number of coverage hypotheses.

- **Zens: Improvements in DP Beam Search for Phrase-based SMT IWSLT Oct, 2008**
Effect of Cube Pruning

Numbers averaged over whole test set; vary beam sizes.
Lexicalized reordering not used, just distortion penalty.
Comparison with Moses

Same TM, LM, etc.; vary beam setting
Lexicalized reordering not used, just distortion penalty.
Summary & Conclusions

• Summary
  – detailed problem description
  – efficient solution
  – in-depth analysis

• Conclusions
  – search important for good translation quality
  – rest score estimation allows for small beam sizes
  – distinction lexical vs. coverage hypothesis important
  – additional cube pruning not necessary
  – significantly faster than Moses
THANK YOU FOR YOUR ATTENTION!
References


