Improved Models of Distortion Cost for Statistical Machine Translation

Spence Green, Michel Galley, and Christopher D. Manning

Stanford University

June 4, 2010
Motivation

Why phrase-based MT?

- Fast, simple, and scalable
- Good performance for many language pairs
  (Zollmann et al., 2008; Lopez, 2008; etc.)

Reordering in (baseline) phrase-based decoders controlled by:

- A distortion cost model
- A distortion limit
Motivation

Why phrase-based MT?

- Fast, simple, and scalable
- Good performance for many language pairs
  (Zollmann et al., 2008; Lopez, 2008; etc.)

Reordering in (baseline) phrase-based decoders controlled by:

- A distortion cost model
- A distortion limit

Cost model is poor, so a low distortion limit is typically used
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis:
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all
Motivating Example

Gloss:  Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis:  Followers of all Christian
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all engaged
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all engaged them
Motivating Example

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all engaged them in waiting ...
Distortion Limit v. Distortion Cost

Cost is a *soft* constraint

- Does **not** prune the search space
- Feature in the log-linear decoder framework

Limit is a *hard* constraint

- Prunes translations from the search space
Distortion Limit v. Distortion Cost

Cost is a *soft* constraint

▶ Does **not** prune the search space
▶ Feature in the log-linear decoder framework

Limit is a *hard* constraint

▶ Prunes translations from the search space

For Moses, low(er) distortion limit improves translation quality!
Translation Quality Decreases at High Distortion Limits

Moses BLEU-4 Performance

Arabic-English

Chinese-English
Hard Constraints Reduce Reference Reachability

Reference Reachability (%)

Translation Option Limit

(Auli, Lopez, Hoang, and Koehn, 2009)
A New Distortion Cost Model

Guide search *without* hard constraints

- Maintain baseline performance at high distortion limits
- Solution: Improve heuristic search with future cost estimation (Moore and Quirk, 2007)

Encourage linguistically-appropriate reorderings

- Solution: Transition-based discriminative distortion model

Worst-case $O(n)$ cost computation

- Maintain linear running time of decoding!
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

NP-SBJ: في الانتظار
NP-OBJ: هم شاركوا الدراسات المشتركة
VP: والسيشية والطوائف كل أبناء

limit = 5

Hypothesis:
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

NP-SBJ \hspace{5cm} \text{limit} = 10

VB \hspace{1cm} \text{NP-OBJ}

الانتظار

و

السماوية

و

العثمانية

الشراك

هم

كل

ابناء

Hypothesis:
Search Errors at High Distortion Limits

Gloss:  Followers of all of the Christian and Islamic sects engaged in waiting for them

NP-SBJ  VB  NP-OBJ  PP

 Ebola  تَفَعَّل  البرية  و  الطوائف  المسيحية  والإسلامية  و  شاركوا  في  الانتظار  هم

Cost = 0

Hypothesis:  Followers of all
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all **Christian and Islamic**
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic engaged
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic engaged them
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic engaged them in waiting

NP-SBJ
Cost = 1

VB NP-OBJ PP
0 0 0

Cost = 0
Search Errors at High Distortion Limits

Gloss: Followers of all of the Christian and Islamic sectors engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic engaged them in waiting sectors
An Admissible Future Cost Heuristic

\[ s_j \leftarrow \text{First uncovered source position} \]
\[ s_{j'} \leftarrow \text{First source position of phrase } p \]
\[ C_j \leftarrow \text{Coverage set to the right of } s_j \]
\[ D(s_{j'}, s_j) \leftarrow \text{Linear distortion from } s_{j'} \text{ to } s_j \]

When \( j' > j \), the estimate is

\[ F = |C_j| + D(s_{j'}, s_j) \]

Update the estimate at each translation step \( n \)

\[ \Delta F = F_n - F_{n-1} \quad n > 0 \]
Linear Distortion with Future Cost

**Gloss:** Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic engaged
Linear Distortion with Future Cost

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic
Linear Distortion with Future Cost

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

Hypothesis: Followers of all Christian and Islamic

Linear Distortion with Future Cost

Gloss: Followers of all of the Christian and Islamic sects engaged in waiting for them

NP-SBJ Cost = 1

VB NP-OBJ PP

شُارَكَوْا هُمُ فِي الانتظار

Cost = 0

Cost = 4

Hypothesis: Followers of all Christian and Islamic sects

Green, Galley, and Manning  Improved Models of Distortion Cost for SMT 26 / 37
Transition-based Discriminative Distortion Cost

**Problem**: Cost model still penalizes all reorderings

- Consider a verb-final language like Japanese
- Skipping over the entire verb complement is *good*
- Model should prefer particular reorderings

A **transition-based** discriminative distortion model

- Idea: compute cost of word-to-word transitions
- Source side features
- Discretized transition classes
Procedure

1. Classify discretized transitions with a log-linear model
   \[
   p_\lambda (D_j, j' \mid \bar{s}, j, j') \propto \exp \left[ \bar{\lambda} \cdot \bar{h} (\bar{s}, j, j', D_j, j') \right]
   \]

2. Train with sorted word-to-word alignments
   - e.g. Arabic \(\rightarrow\) Arabic' (English word order)

3. Query model for transitions at each translation step
Features

This evaluation

▶ Words and POS tags
▶ Relative source sentence position (discretized)
▶ Source sentence length (discretized)

Future work

▶ POS tag chains (bigram, trigram, etc.)
▶ Agreement morphology
▶ Subject has been translated? (binary, global)
Further Motivation

Incremental processing like shift-reduce parsing

- Process source items in decoding order
- Classes are discrete jumps instead of “operations”
- Beam search via the decoder
Implementation: Constant-time During Decoding

Nine discrete distortion classes

- Same number of training examples per class

Separation into inbound/outbound models
(Al-Onaizan and Papineni, 2006)

- Simplifies caching during decoding
- Future work: Combine into a single model

Model has four decoder features

- Inbound and outbound scores
- Alignment penalty
- Future cost estimate
Evaluation

MT system is Phrasal (Cer et al., 2010)
  ▶ Baseline: Moses feature set
  ▶ Lexicalized reordering model of Galley and Manning (2008)

NIST MT09 Ar-En constrained track training data
  ▶ Removed UN and comparable data
    ▶ Same baseline, faster experiments
  ▶ 6.20M English and 5.73M Arabic tokens

Evaluated BLEU-4 on MT03/05/06/08 at dlimit = 15
High Distortion Limit (15)

- MT06
  - Baseline: 41.0
  - Lex: 41.9
  - FutureCost: 42.8
  - This Work: 43.7

- MT08
  - Baseline: 38.8
  - Lex: 39.4
  - FutureCost: 41.0
  - This Work: 41.7
Improvement Over the Low Distortion Limit Baseline

MT06

- d5 Baseline: 43.0
- d15 This Work: 43.7

MT08

- d5 Baseline: 41.3
- d15 This Work: 41.7
Conclusion

Contributions of this work

▶ Fixed search errors caused by linear distortion
▶ Added a distortion model with linguistic features
▶ Modest improvement over Moses at a high distortion limit

Software:

- Phrasal  http://nlp.stanford.edu/phrasal/
Thank You!

Thanks to Daniel Cer and Claude Reichard.
Distortion Cost Curve for the adjective *American*

![Graph showing the distortion cost curves for different quintiles of the adjective "American".](image)