A Systematic Analysis of Translation Model Search Spaces

Michael Auli Adam Lopez Hieu Hoang Philipp Koehn

Presented by Kenneth Heafield at MT Seminar

March 3, 2010

Error Categories

Output $\neq$ reference because:

Induction	**Not in search space**
Parametrization	Something else scores higher
Search	Reachable but pruned *at decoding time*

Justification for Reference Reachability

- Would prefer to ask if search
- Loose correlation with automatic metrics and accuracy
- Find systematic holes in the search space

Methodology for Reference Reachability

Moses

- Large stack size: 10^5
- No beam threshold
- Remove non-matching hypotheses

Hiero

- Insanely high rule and entry limits
- Remove non-matching hypotheses

Rule Pruning

Common Heuristic

Limit rules to *tol* translations per source

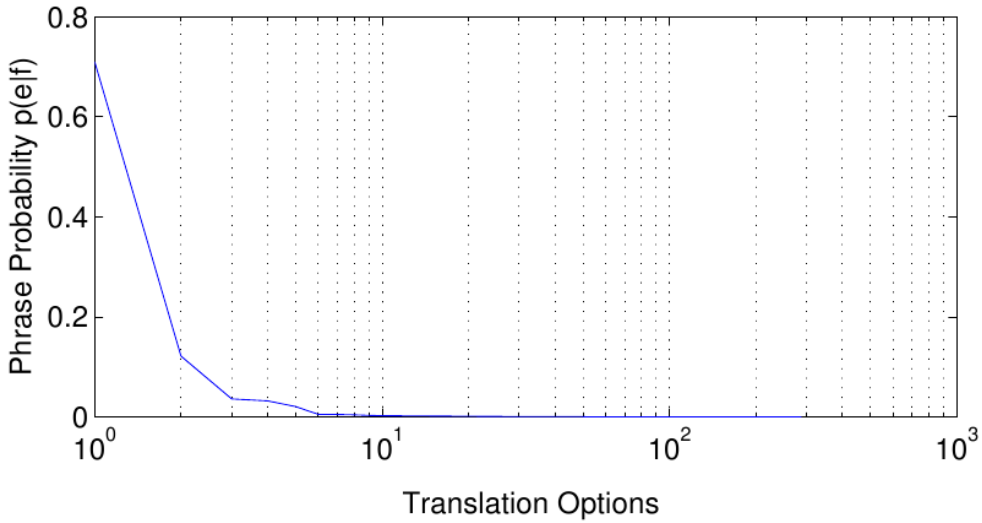

Figure: $p(e|f)$ for translations of problème

Hypothesis

Reference reachability improves with large *tol*

Rule Pruning Statistics

Statistics

- 20 rules/span
- 6.2% of rules are pruned
- 0.1% of French spans have > 20 rules

Cherrypicked Example

Rank 105: problème $\mapsto$ dilemma

Impact of Rule Pruning

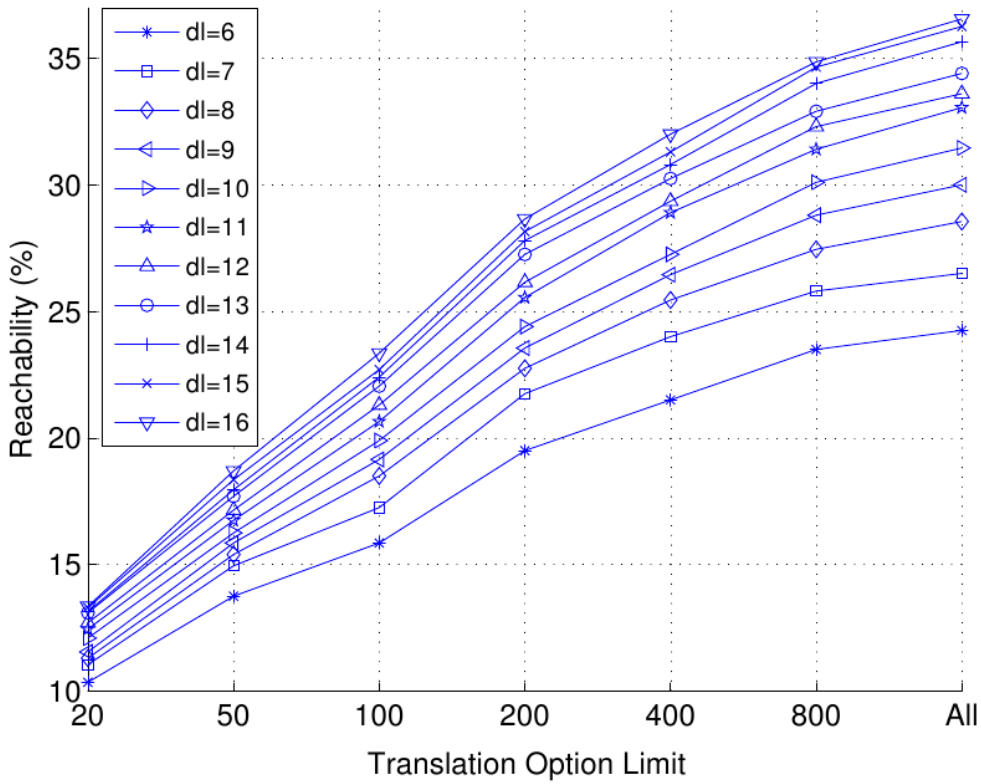

Figure: Reference reachability in Moses based by phrase pruning and distortion limit

Analysis of *tol* impact

Worry

More translations of "," spuriously increase reachability

Qualitative Analysis

Half of added reachability comes from legitimate translations

Skeptic

This analysis is conditioned on reaching a meaningful reference

Mutual Reachability

Definition

- Can Hiero produce Moses's 1-best?
- Can Moses produce Hiero's 1-best?

Answer

Mostly (93%) the same, but Hiero usually does more

What's Missing?

Moses

- Drop words in a rule: 12/14 times
- Generalize some phrase pairs

Hiero

- Fine-grained reordering

Conclusions

Pruning

Common rule pruning reduces reachability

Small reachability difference

Performance difference is model error, not induction error

The elephant

Increasing the search space adds a lot of junk