Improving Translation Model by Monolingual Data

Ondřej Bojar, Aleš Tamchyna
{bojar, tamchyna}@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague

Overview
Target-side monolingual data help in LM. Can we use it also in TM?
The Trick: Reverse self-training with back-off
Helps:
• in small data setting and
• into morphologically rich languages.

Reverse Self-training
- Monolingual Data
- Parallel Data
- Translate
- Phrase Table
- MERT
- Final Translation Model

Target-side monolingual data help in LM. Can we use it also in TM?
The Trick: Reverse self-training with back-off
Helps:
- in small data setting and
- into morphologically rich languages.

Learning Unseen Forms
Small Parallel Data:
Source English
a cat chased...
saw a cat
read about a dog
Target Czech
kočka honiša...
viděl jsem kočku
četl jsem o psovi
Czech Lemmatized
kočka honil...
vidět byt kočka
číst byt o pes

Large Monolingual Data:
? četl jsem o kočce číst byt o kočka

⇒ Learned a new phrase (o kočce) including a form never seen in parallel data (kočce).

Comparison Across Languages
- Absolute gain over the
  Mono LM baseline
- Parallel Data:
  90-125 thousand sentences
  Monolingual Data:
  0.6-0.9 million sentences

Case-insensitive BLEU scores
- German
  - No improvement in BLEU score.
  - Parallel data already sufficiently large
  - Not all available data used in reverse self-training.
- Czech
  - Achieved a small improvement.
  - Only 2010 and 2011 News data used in reverse self-training.

Impact of Data Size
Mono LM and TM
Mono LM
Parallel data
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
BLEU
0 1 2 3 4 5 6 7 8 9 10
Monolingual data (millions of sents.)
(Monolingual fixed: 5 million sents.)

German
14.8
14.8
Czech
15.7
15.9

German – constrained
Czech – constrained except LM

Our WMT11 System Submissions

Translation with Back-off
- Reverse translation must handle unknown forms (these will become the newly learned forms).
- Factored model with alternative decoding paths.
- The back-off factor (e.g. lemmas) unifies different word forms.

Problems
- MERT had to optimize many weights of two very similar models.
- Many derivations lead to the same hypothesis
- 100-best list contained only ~6 unique strings, compared to ~35 in the baseline setup
  ⇒ unstable,
  ⇒ diverging runs had to be repeated.
- Lattice MERT did not help.
- Possible solution: "Better Hypothesis Testing..." (Clark et al., 2011).

Conclusions
- Reverse self-training learns to produce forms not seen in parallel data.
- Greater effect for language pairs with very different vocabulary sizes.
- More monolingual data ⇒ greater effect.
- More parallel data ⇒ the effect diminishes.
- Good back-off: forms with last 3 characters removed.