Utilizing Source Context in Statistical Machine Translation

Aleš Tamchyna

June 5, 2013
Outline

- Introduction
  - Phrase-Based MT
  - Why Consider Wider Source Context?

- Related Work

- First Experiments
  - Context Similarity Feature
  - Beyond Phrase-Sense Disambiguation

- Conclusion
Phrase-based MT

- Word alignment: $P(e|f)$
  - Learned from sentence-aligned parallel data.
  - Example query:
    What is the probability of 'car' given German 'Auto'?

- Translation model (phrase table): $P(e|f)$
  - Trained heuristically based on the word alignment.
  - Example query:
    What is the probability of 'a fast car' given 'ein schnelles Auto'?

- Language model: $P(e)$
  - Trained from target-side monolingual data.
  - How probable are the words 'a fast car' in an English sentence?

- Feature weights
  - Model weights are tuned towards a metric of MT quality (e.g. BLEU, Papineni et al. 2002).
  - Minimum Error Rate Training (MERT, Och 2003).

- Decoding
  - Search in the hypothesis space for the most adequate translation.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

Long-distance agreement in Czech:
- Input Google Translate: Kids like to play football. Dˇ eti rády hrají fotbal.
- Kids mostly like to play football. Dˇ eti vˇ etˇ sinou rády hrají fotbal.

Lexical Selection:
- Shooting of the expensive film. Nat´ aˇ cen ´ı drah´e filmu.
- Shooting of the least expensive film. Stˇ relba z nejlevnˇej ´s ´ıch filmu.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

- Long-distance agreement in Czech:

  **Input**
  Kids like to play football.

  **Google Translate**
  Děti rády hrají fotbal.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

- Long-distance agreement in Czech:

  **Input**
  
  Kids like to play football.
  
  Kids mostly like to play football.

  **Google Translate**
  
  Děti rády hrají fotbal.
  
  Děti většinou rádi hrají fotbal.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

- **Long-distance agreement in Czech:**
  
  **Input**
  Kids like to play football.
  Kids mostly like to play football.
  
  **Google Translate**
  Děti rády hrají fotbal.
  Děti většinou rádi hrají fotbal.

- **Lexical Selection:**
  
  **Input**
  Shooting of the film.
  
  **Google Translate**
  Natáčení filmu.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

- **Long-distance agreement in Czech:**
  
  **Input**  
  Kids like to play football.  
  Kids mostly like to play football.

  **Google Translate**  
  Děti rády hrají fotbal.  
  Děti většinou rádi hrají fotbal.

- **Lexical Selection:**
  
  **Input**  
  Shooting of the film.  
  Shooting of the expensive film.

  **Google Translate**  
  Natáčení filmu.  
  Natáčení drahé filmu.
Why Consider Wider Source Context?

Phrase length and language model scope are always limited. Many language phenomena are local, but:

- Long-distance agreement in Czech:
  
  **Input**
  Kids like to play football.
  Kids mostly like to play football.

  **Google Translate**
  Děti rády hrají fotbal.
  Děti většinou rádi hrají fotbal.

- Lexical Selection:
  
  **Input**
  Shooting of the film.
  Shooting of the expensive film.
  Shooting of the least expensive film.

  **Google Translate**
  Natáčení filmu.
  Natáčení drahé filmu.
  Střelba z nejlevnějších filmu.
Related Work

Word Sense Disambiguation (WSD)

- Select correct word sense in a particular sentence.
- Possible senses = meanings in the target language (Vickrey et al., 2005)
  ⇒ essentially machine translation.
  ▶ Relies only on source-side features.
  ▶ Suitable setting for discriminative models.

Using Source Context in SMT

Related Work: Phrase Sense Disambiguation

Carpuat and Wu (2007)

- Disambiguate senses (=translations) of phrases.
- Use a state-of-the-art WSD system to provide $P(e|f, f_{context})$.
- Train one WSD model for each source phrase in vocabulary.
- Use the WSD score as a feature in the decoder.

Features:

*The DT dog NN saw VBD a DT cat NN with IN a DT telescope NN .*.

- Words around the phrase (fixed window length), e.g. “saw\_1”.
- POS tags around the phrase, e.g. “VBD\_1”.
- Syntactic features, e.g. predicate: “pred\_saw”.
- Local $n$-gram collocational features, e.g. “col\_dog\_saw\_”.

[9x251]Related Work: Phrase Sense Disambiguation

Carpuat and Wu (2007)

- Disambiguate senses (=translations) of phrases.
- Use a state-of-the-art WSD system to provide $P(e|f, f_{context})$.
- Train one WSD model for each source phrase in vocabulary.
- Use the WSD score as a feature in the decoder.

Features:

*The DT dog NN saw VBD a DT cat NN with IN a DT telescope NN .*.

- Words around the phrase (fixed window length), e.g. “saw\_1”.
- POS tags around the phrase, e.g. “VBD\_1”.
- Syntactic features, e.g. predicate: “pred\_saw”.
- Local $n$-gram collocational features, e.g. “col\_dog\_saw\_”.

Aleš Tamchyna (ÚFAL) Utilizing Source Context in SMT June 5, 2013 6 / 12
Related Work: Rich Features in the Decoder

Gimpel and Smith (2008)

- Similar idea and features, but very different architecture:

\[ P(e|f, f_{\text{context}}) = \frac{\text{count}(e|f, f_{\text{context}})}{\sum_{e'} \text{count}(e'|f, f_{\text{context}})} \]
Related Work: Global Lexicon Model

Mauser et al. (2009)

- Trained one classifier for each target word.
- Features: only source-side bag of words, extracted from the whole sentence.
- Binary classification: does the word belong in the sentence translation?
- In decoding: hypothesis score is the product of scores of individual target words.
Context Similarity Feature

- Phrase context: words in a window of fixed length, ignoring word positions.
- For each phrase pair \((e, f)\), remember all context words in training data (and their counts).
- In decoding, compute the cosine similarity between current context and the observed contexts of possible phrase translations.
- No improvements in BLEU so far:
  - Feature is unstable, small changes in phrase segmentation drastically impact similarity scores.
  - Function words have the same weight as content words.
- Possible solutions (future work):
  - Ignore function words.
  - Use word lemmas instead of surface forms.
Beyond PSD: Learning Setting

- Participated in a workshop where we integrated PSD in the Moses decoder
  ⇒ (relatively) easy to run PSD experiments.
- Still work in progress, no improvement in BLEU so far.
- Vowpal Wabbit
  ▶ Fast, scalable ML toolkit.
  ▶ Online learning algorithm.
  ▶ Linear: \( f(x) = \mathbf{w} \cdot \mathbf{x} \)
  ▶ Reductions for multi-class/multi-label classification.
- Our setting:
  ▶ Multi-label classification: each translation option has a loss.
    ★ 0 for correct, 1 for others.
    ★ 0 for correct, BLEU-approximation discounts for others.
  ▶ Two namespaces: Source, Target.
  ▶ Quadratic features: Cartesian product of S and T features.
  ▶ Reduction: cost-sensitive one-against-all.
  ▶ Label-dependent features.
Beyond PSD: Current Work

- Use 2 classifiers:
  - Predictor of content words (t-lemmas) based on wider context and lexical features (bag of source t-lemmas, ...).
  - Predictor of morphological categories (tags, +lemmas for prepositions) based on more local, morpho-syntactic features.

- Avoid 0/1 loss, use the BLEU approximation.

- Overall translation scheme: form|t-lemma|tag → form|t-lemma|tag.
References


