Pre- and Postprocessing Techniques for SMT Output Quality Improvement

Mark Fishel (fishel@ut.ee)

26. september 2006
Summary:

- Some words about SMT
- Some existing output improvement techniques
- Suggested new output improvement techniques
Statistical Machine Translation
Statistical Machine Translation

Translating from “foreign” into “English”:

- $e$ – English sentence
- $f$ – foreign sentence
- $p(e)$ – probability of $e$ being a correct sentence
- $p(e|f)$ – probability of $e$ being the translation of $f$

\[
\hat{e} = \arg \max_{e} p(e|f) = \arg \max_{e} \frac{p(e)p(f|e)}{p(f)} = \arg \max_{e} p(e)p(f|e)
\]
Statistical Machine Translation

\[ \hat{e} = \arg \max_e p(e)p(f|e) \]

- \( p(e) \) – language model
- \( p(f|e) \) – translation model
- \( \arg \max \) – decoder
Decoder Output

maison rouge $\rightarrow$ LM $\rightarrow$ $p(\text{house red}) = 0.6$

$\rightarrow$ TM $\rightarrow$ $p(\text{red house}) = 0.2$

$\ldots$
Output Correction

Via postprocessing:

\[
\text{maison rouge} \rightarrow \text{LM} \rightarrow \text{TM} \rightarrow \quad \begin{array}{l}
p(\text{house red}) = 0.6(0.3) \\
p(\text{red house}) = 0.2(0.8) \\
\ldots
\end{array}
\]

Via preprocessing:

\[
\text{rouge maison} \rightarrow \text{LM} \rightarrow \text{TM} \rightarrow \quad \begin{array}{l}
p(\text{red house}) = 0.7 \\
p(\text{house red}) = 0.1 \\
\ldots
\end{array}
\]
Existing output improvement techniques
Existing techniques: 1

- Replace inflected word with stem and morphological description
  - useful if translating from highly agglutinative to less agglutinative language (e.g. Estonian/German to English)
  - example:
    poiss – the boy
    poisiga → poiss ga – with the boy
    poisile → poiss le – to the boy
Existing techniques: 2

• Put source language words in the target language order
  ■ ich mache die Tür zu →
  ich zu mache die Tür →
  i close the door
Suggested new output improvement techniques
Idea-1

- replace words in monolingual corpus with parts-of-speech (PoS)
- train a LM on that corpus
- use it together with (or instead of) the usual LM
  - more general, assistive in case of sparse data
Idea-2

- replace words in bilingual corpus with PoS
- train 2 translators: for words and for PoS
- translate in parallel the sentence and the PoS sequence
- choose the word sequence in the N-best list which matches best the translated PoS sequence
  - translating parts-of-speech should work well because of the limited lexicon
Idea-2.1

- translate separately stemmed words and their morphological descriptions
- use morphological synthesis to obtain the actual translation
Idea-3

- let $f$ be the input and $e_i$ the output sentences from the N-best list;
- use one of ML techniques (NN, TBL, SVM, ...) to tell whether the pair $< f, e_i >$ is correct ($e_i$ is the translation of $f$)
Idea-3

• main problem: training/testing data
  ▪ is it possible to solve

\[ F(e_i, e_{orig}) = e_i \text{ “distance” from } e_{orig} \]

\[ (F \text{ decides whether } e_{i_0} \text{ is the closest to } e_{orig} \text{ for all possible } i) \]

• if possible, each N-best list provides us with 1 positive and N − 1 negative pairs
Thank You!