International Workshop on Spoken Language Translation
Kyoto, Japan
November 27-28, 2006

Automatic Sentence Segmentation and Punctuation Prediction
for Spoken Language Translation

Evgeny Matusov, Arne Mauser, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University
D-52056 Aachen
Content

1. motivation
2. related work
3. handling speech input in phrase-based MT
4. methods for predicting punctuation
5. automatic sentence segmentation algorithm
6. experimental results
7. summary
Motivation

ASR systems perform only acoustic segmentation
segments may be too short/long for machine translation systems
⇒ a novel approach to sentence segmentation
   – using a log-linear combination of lexical and prosodic features
   – using an explicit sentence length model and length contraints

MT users expect to see sentences with proper punctuation
⇒ three different strategies for punctuation prediction:
   – in the source language (before translation)
   – in the target language (after translation)
   – implicitly (using translation models)
Related work

• annotation of the ASR output as the end product delivered to the user:
  – combine lexical and prosodic cues to improve sentence boundary prediction (Liu et al., 2004)
  – maximum entropy model (Huang and Zweig, 2002)
    decision trees (Kim and Woodland, 2001)

• comma restoration for ASR output translation (Lee et al., 2006)
  – segments are already known and are assumed to end with a period
  – thresholds for bigram/trigram probabilities for commas
Phrase-based MT for Speech Translation

- translation of sentence units
- word and phrase reordering limit the maximum length of a SU to about 50-60 words
- punctuation is treated as words when training the MT system
- models: phrase-based and word-based lexica, language model, etc.
- loglinear model combination
MT Setups for Punctuation:

three different possible setups:

- MT without punctuation, prediction in the target language
- MT with punctuation, prediction in the source language
- implicit generation of punctuation marks in MT

MT without punctuation marks

* remove punctuation from corpus if present
* insert punctuation marks into the MT output based
+ improved phrase coverage
– errors in translation hypotheses reduce prediction quality
– only language information available
Various Punctuation Prediction Strategies

MT with punctuation marks

- keep all punctuation marks in training
+ prosodic features can be used for prediction
- incorrect prediction affects MT
- difficult to used with ASR lattice translation
- punctuation is different in languages
Various Punctuation Prediction Strategies

Implicit prediction of punctuation in the MT process

- remove punctuation marks only from the source
- punctuation marks will be “inserted” through phrasal translations
- example of alternative phrasal translations:

  * did you say
  * sagten Sie => , did you say
  * did you say ,
  * , did you say ,

+ SMT features help to select best translation and punctuation variant
+ MET of MT system also optimizes punctuation prediction
+ easy to use, applicable to translation of word lattices
Automatic Sentence Segmentation: Idea

Motivation:
- limit the maximum/minimum length of automatically determined SUs
- maximum: most MT algorithms work well up to 40-50 words
- minimum: context information is lost with too short SUs

Idea:
- utilize lexical and prosodic features (language model, pause duration)
- explicit optimization over segment length
- score of hypothesized segment boundary based on optimal previous segment boundary
- minimum/maximum SU length can be parametrized
Automatic Sentence Segmentation: Theory

• given: unsegmented word sequence \( w^N_1 := w_1, w_2, \ldots, w_N \)
• find optimal segmentation \( i^K_1 := (i_1, i_2, \ldots, i_K = N) \)
• i.e. segmentation with the highest posterior probability

\[
\hat{i}_1^K = \arg\max_{K, i^K_1} \{ Pr(i^K_1 | w^N_1) \}
\]  

\[ (1) \]

• posterior probability modeled as log-linear combination of several features:

\[
Pr(i^K_1 | w^N_1) = \frac{\exp \left( \sum_{m=1}^{M} \lambda_m h_m(i^K_1, w^N_1) \right)}{\sum_{K', i'^K_1} \exp \left( \sum_{m=1}^{M} \lambda_m h_m(i'^K_1, w^N_1) \right)}
\]

\[ (2) \]

• decision rule

\[
\hat{i}_1^K = \arg\max_{K, i^K_1} \left\{ \sum_{m=1}^{M} \lambda_m h_m(i^K_1, w^N_1) \right\}
\]

\[ (3) \]
Automatic Sentence Segmentation: Models Used - 1

features used in log-linear combination

- \( n \)-gram language model probability incl. hidden event for boundary
- explicit sentence length probability \( p(j - i) \) (lognormal distribution)
- normalized pause duration between \( w_i \) and \( w_{i+1} \)
- segment penalty
- possible extensions: prosodic features form as (pitch, energy, ...)

scaling factors tuned manually for precision/recall on development data
Automatic Sentence Segmentation: Search

process the word sequence $w_1^N$ from left to right

recursion

$$p(w^i_1) = \max_{j-l_{\text{max}} \leq i \leq j-l_{\text{min}}} p(w^i_1) \cdot p(w^j_{i+1})$$

$p(w^i_1)$ was determined in the previous recursion step

processing time: less than 2 seconds for segmentation of 30 000 words
Experimental results
Evaluation Methodology

sentence segmentation:
- precision and recall compared to manual segment boundaries

machine translation:
- objective measures WER, PER, BLEU, NIST with human references

evaluation of automatically segmented output:
- MT output automatically aligned to reference (Matusov et al. 2005)
Experiments

sentence segmentation:
• use language model and length model features only (IWSLT task)
+ add pause duration feature (TC-STAR task)

machine translation:
• test the influence of automatic segmentation
• three strategies for punctuation prediction
Quality of Sentence Segmentation (IWSLT 2006 task)

Compare with the standard approach: SRI hidden-ngram tool

<table>
<thead>
<tr>
<th>corpus</th>
<th>RWTH tool</th>
<th>hidden-ngram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>IWSLT test 2005</td>
<td>84.2</td>
<td>84.1</td>
</tr>
<tr>
<td>IWSLT dev 2006</td>
<td>59.5</td>
<td>64.6</td>
</tr>
<tr>
<td>IWSLT test 2006</td>
<td>56.4</td>
<td>61.0</td>
</tr>
<tr>
<td>IWSLT test 2006 (ASR)</td>
<td>56.0</td>
<td>55.2</td>
</tr>
</tbody>
</table>

- precision (P) and recall (R) in %
- 4-gram LM used in both tools
- minimum sentence length 3 words, maximum sentence length 30 words
- Chinese transcriptions
  - 2005 test set: 3208 words and 506 segments
  - 2006 test set: 5550 words and 500 segments
    (15.2% character error rate of the ASR output)
- 2005 test set is very similar to the language model training data
### Quality of Sentence Segmentation (TC-STAR task)

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>baseline (4-gram LM only)</td>
<td>54.2</td>
<td>52.1</td>
</tr>
<tr>
<td>+ length model</td>
<td>54.7</td>
<td>52.5</td>
</tr>
<tr>
<td>+ pause model</td>
<td>68.8</td>
<td>68.4</td>
</tr>
<tr>
<td>baseline + pause model</td>
<td>68.1</td>
<td>68.3</td>
</tr>
</tbody>
</table>

- precision (P) and recall (R) in %
- minimum sentence length 3 words, maximum sentence length 50 words
- English ASR output, word error rate of 6.9%, 28K words, 1155 reference segments
- pause duration feature is very important for good segmentation performance
## Translation Quality: IWSLT 2006 Chinese-to-English Task

<table>
<thead>
<tr>
<th>transcription</th>
<th>segmentation</th>
<th>punctuation</th>
<th>BLEU [%]</th>
<th>WER [%]</th>
<th>PER [%]</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DEV 2006</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>correct</td>
<td>source</td>
<td>19.8</td>
<td>70.5</td>
<td>54.3</td>
<td>5.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>implicit</td>
<td>22.0</td>
<td>71.0</td>
<td>53.0</td>
<td>5.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>target</td>
<td>18.9</td>
<td>70.7</td>
<td>55.2</td>
<td>6.03</td>
</tr>
<tr>
<td>automatic</td>
<td>source</td>
<td>implicit</td>
<td>17.3</td>
<td>66.1</td>
<td>54.9</td>
<td>5.34</td>
</tr>
<tr>
<td></td>
<td>target</td>
<td>implicit</td>
<td>20.7</td>
<td>62.1</td>
<td>52.0</td>
<td>5.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>target</td>
<td>17.5</td>
<td>67.2</td>
<td>55.9</td>
<td>5.49</td>
</tr>
<tr>
<td>automatic</td>
<td>correct</td>
<td>source</td>
<td>15.9</td>
<td>73.9</td>
<td>58.5</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>implicit</td>
<td>19.0</td>
<td>69.1</td>
<td>56.7</td>
<td>5.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>target</td>
<td>15.4</td>
<td>73.2</td>
<td>58.2</td>
<td>5.37</td>
</tr>
<tr>
<td>automatic</td>
<td>source</td>
<td>implicit</td>
<td>14.4</td>
<td>68.4</td>
<td>58.2</td>
<td>4.51</td>
</tr>
<tr>
<td></td>
<td>target</td>
<td>implicit</td>
<td>17.1</td>
<td>64.8</td>
<td>55.2</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>target</td>
<td>13.8</td>
<td>69.0</td>
<td>59.1</td>
<td>4.60</td>
</tr>
</tbody>
</table>
Translation Quality on the IWSLT Task: Discussion

- errors introduced by ASR have a higher impact than errors introduced by automatic segmentation
- implicit prediction of PM in translation process performs best
  + phrasal translations are not being “broken” by incorrectly inserted PM on the source side
  + especially important for small vocabulary tasks, with limited phrase coverage
## Translation Quality: TC-STAR English-to-Spanish Task

<table>
<thead>
<tr>
<th>transcription</th>
<th>segmentation</th>
<th>punctuation prediction</th>
<th>BLEU [%]</th>
<th>WER [%]</th>
<th>PER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>correct</td>
<td>manual (source)</td>
<td>45.2</td>
<td>43.3</td>
<td>32.2</td>
</tr>
<tr>
<td>automatic</td>
<td>correct (aligned)</td>
<td>source</td>
<td>37.8</td>
<td>50.6</td>
<td>37.6</td>
</tr>
<tr>
<td>automatic</td>
<td>source</td>
<td></td>
<td>36.7</td>
<td>51.2</td>
<td>38.1</td>
</tr>
<tr>
<td>automatic</td>
<td>implicit</td>
<td></td>
<td>36.1</td>
<td>51.5</td>
<td>38.6</td>
</tr>
<tr>
<td>automatic</td>
<td>target</td>
<td></td>
<td>36.3</td>
<td>51.3</td>
<td>38.4</td>
</tr>
<tr>
<td>automatic</td>
<td>full stop only (source)</td>
<td></td>
<td>35.8</td>
<td>50.2</td>
<td>38.6</td>
</tr>
</tbody>
</table>

- only a small degradation in MT quality when automatic segmentation is used
- low recognition WER on this corpus makes punctuation prediction in the source language sufficiently reliable to serve as input to the MT system
Conclusions

• methods for automatic segmentation and punctuation prediction

• improved interface between ASR and MT

• sentence segmentation:
  – performs at least as well as existing state-of-the-art algorithm
  – sentence length can be explicitly adjusted
  – robust also in terms of machine translation quality

• three different approaches for punctuation prediction

• the three methods have advantages and disadvantages

• for small vocabulary tasks: implicit generation punctuation leads to higher translation quality
References

• Huang, J., and Zweig, G. “Maximum entropy model for punctuation annotation from speech”. In Proc. of ICSLP, pp. 917-920, 2002.


