Punctuation Insertion for Real-time Spoken Language Translation

IWSLT 2015
Eunah Cho, Jan Niehues, Kevin Kilgour, Alex Waibel | December 4, 2015

Let’s eat, grandpa!
Let’s eat grandpa
OK but then after a while I realized this is my life this is six months of my life and this...

OK. But then, after a while, I realized this is my life. This is six months of my life. And this...

- Readability
- Performance of MT
### Punctuation and MT

- Impact of segmentation and punctuation marks on MT performance
- Oracle experiments
- German to English

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR output</td>
<td>20.70</td>
</tr>
<tr>
<td>+ Segmentation</td>
<td>21.42</td>
</tr>
<tr>
<td>+ Full stop</td>
<td>22.18</td>
</tr>
<tr>
<td>+ All punctuations</td>
<td>22.48</td>
</tr>
<tr>
<td>Transcript</td>
<td>27.99</td>
</tr>
</tbody>
</table>
Conventional strategies...

**LM and prosody based model**
- Incorporates acoustic features
- Low latency

**Monolingual translation system**
- MT that translates non-punctuated language into punctuated one
- Improve MT performance
- Overlapping window for segment insertion, longer context → latency
LM and prosody based model

- Consider two words prior and two after the possible punctuation mark.
- 4-gram LM trained on punctuated text
  - Score without an inserted punctuation mark: \( P(w_{i-1}, w_i, w_{i+1}, w_{i+2}) \)
  - Score with a comma: \( P(w_{i-1}, w_i, \text{COMMA}, w_{i+1}, w_{i+2}) \)
  - Score with a fullstop: \( P(w_{i-1}, w_i, \text{STOP}, w_{i+1}, w_{i+2}) \)
- Dynamic scaling factor to prevent very short or long sentences.
- Pauses over \( \theta \)s then force terminate segment.
Monolingual MT: Random segmentation for training data

- Original corpus
  
  this is Bill Lange.
  I am Dave Gallo.
  and we are going to tell you some stories from the sea here in video. we have got some of the most incredible video of Titanic that is...

- Randomly segmented corpus
  
  this is Bill Lange. I am
  Dave Gallo. and we are going to tell you some stories from the sea here in video. we have got some of the most incredible video of Titanic that is...
Monolingual MT: Sliding window for testing

- Sliding window to observe words in longer, various contexts
- Example segments: ... *der bildet die sogenannte konjunktive Normalform* wir haben gesehen dass wir diese ...

---

| der bildet die sogenannte konjunktive Normalform wir haben gesehen dass wir diese |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|

---

Table 6: Test set preparation for the monolingual translation system

Table 7: Translation using the monolingual translation system

5.2. Punctuation prediction criteria

A punctuation mark is chosen if the same punctuation mark is found same or more often than a given threshold. If more than one punctuation mark appears more than the threshold in the same word space, the most frequent one is chosen.

There are some cases where we have the same frequency for multiple punctuation marks; in this case we put a different priority on punctuation marks. For example, in this experiment we put higher priority for a period over a comma.

In this experiment, we evaluate the translation quality over a varying threshold, from 1 to 9. We exempt the case when the threshold is 10, the length of the sliding window. In this case, one punctuation mark has to appear all the 10 word spaces after a word in order to be inserted. This condition is so restrictive that only few full stops are generated, which causes unaffordable computational time consumption for the translation procedure.

In the same way as in the oracle experiment, we consider four punctuation marks here: period, comma, question mark, and exclamation mark. A new segment is introduced when either a period, question mark, or exclamation mark is predicted, in order to have congruence with the manual transcripts.

To make the hypotheses comparable with the oracle experiments, we considered three different hypotheses of reconstructing segmentation and punctuation:

- **MonoTrans-Segment**: monolingual translation system is used for segmentation prediction only.
- **MonoTrans-FullStop**: monolingual translation system is used for segmentation and full stop prediction.
- **MonoTrans-AllPunct**: monolingual translation system is used for segmentation and all punctuation marks prediction.

5.3. Results

In order to analyze the effect of the varying threshold for the monolingual translation system, first we use the same threshold value for all punctuation marks. The number of punctuation marks predicted using the same threshold are shown in Table 8. As shown in the table we could predict periods and commas, but we could not generate question marks and exclamation marks. A reason might be that question mark and exclamation mark are already rare in the manual transcript. In addition, we do not have many of them appearing in the training corpora, compared to the frequency of the other punctuation marks. The number of periods in Table 8, therefore, is the same as the number of segments predicted.

**Figure 1** presents the translation performance of the three hypotheses in BLEU over different threshold values. In this experiment as well, the same threshold value is used for all the different punctuation marks. Even though we ob...
Output of the Monolingual MT

- Punctuated words in bold letter
- An empirically chosen threshold

Table 6: Test set preparation for the monolingual translation system

<table>
<thead>
<tr>
<th>der bilde</th>
<th>bildet</th>
<th>die</th>
<th>sogenannte konjunktive Normalform.</th>
<th>konjunktive Normalform.</th>
<th>Wir haben gesehen, dass wir diese.</th>
</tr>
</thead>
<tbody>
<tr>
<td>bildet die sogenannte konjunktive Normalform.</td>
<td>Wir haben gesehen, dass wir diese.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sogenannte konjunktive Normalform.</td>
<td>Wir haben gesehen, dass wir diese.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ...der bildet die sogenannte konjunktive Normalform.
  Wir haben gesehen, dass wir diese...
Efficient punctuation insertion scheme

- For real-time SLT system
- Based on the monolingual translation system
- Designed to decrease the latency
- No loss in performance
Real-time spoken language translation

- Latency: the time between when a word is spoken and when its transcript and translation are displayed to the user

- Each component adds to the latency
  - Computation time $\Rightarrow$ fast servers with multiple cores, parallelized computations, smaller faster models
  - Communication time $\Rightarrow$ fast connection, low overhead between components
  - Required (future) context, . . .
Resending of ASR

- Reduce the apparent latency
- ASR continually outputs its current best hypothesis e.g., once a second
- Updated by newer, possibly better, hypothesis
- Higher user acceptance than waiting for a complete, stable hypothesis
Example: resending of ASR

\[\ldots\text{in this planet you would have to prove}\ldots\]

\[\ldots\text{in this planet you would have to provide 36 million translation}\ldots\]

\[\ldots\text{many dialects it is of course a dog}\ldots\]

\[\ldots\text{many dialects it is of course a daunting challenge}\ldots\]
Segmentation for streamed input

- MT system requires the ASR output to be segmented
- The MT/ASR components constantly update its current hypothesis
  - The punctuation prediction component has to deal with possibly changing, frequently sent inputs fast
  - The MT component requires sentence boundary info as soon as possible

- Monolingual translation system with overlapping window
  - Long shifting window of 10 words, requiring a long future context ⇒ latency!
Model

- Finalized ASR output $w = \{w_l, \ldots, w_m\}$
- Flexible ASR output $v = \{v_{m+1}, \ldots, v_n\}$

*... would not exist in one hundred years one hundred years why because they look at the curb its why because they look at the curve and say if the curve and say if the population keeps growing at these keeps growing at this rate*

- Another stack for recognized words before $w$ for punctuation insertion setup:
- History stack $h = \{h_{l-c}, \ldots, h_{l-1}\}$, with context length $c$
city of New York would not exist in one hundred years. One

not exist in one hundred years why because they look at the curb its

not exist in one hundred years. Why? Because they look at the curb its
then, after a while, I realized this is my life. This is six months of my life. And this . . .
What about overlapping window?

and I said, “OK, it’s the huge file. OK, I said, “OK, it’s the huge file. OK, but said, “OK, it’s the huge file. OK, but then OK, it’s the huge file. OK, but then, after it’s the huge file. OK, but then, after a it’s the huge file. OK, but then, after a while, the huge file. OK, but then, after a while, I huge file. OK, but then, after a while, I realised file. OK, but then, after a while, I realised this OK, but then, after a while, I realized. this is but then, after a while, I realized. this is my then, after a while, I realized. this is my life. after a while, I realized. this is my life. this a while, I realized. this is my life. this is while I realized. this is my life. this is six I realized. this is my life. this is six months realized. this is my life. this is six months of this is my life. this is six months of my is my life. this is six months of my life. this is six months of my life, and life. this is six months of my life, and this this is six months of my life. and this fire is six months of my life. and this fire. so six months of my life. and this fire. so, I months of my life. and this fire. so I was of my life. and this fire. so I was a my life. and this fire. so I was a little life, and this fire. so I was a little bit and this fire. so I was a little bit skeptical this fire. so I was a little bit skeptical of

- With 10 words of overlapping window, a much longer future context is required
Experimental setup

- Punctuate English tst13
  - LM and prosody based segmentation
  - Monolingual translation system with an overlapping window 10
  - Monolingual translation system with streaming input*

- Translate it into German
  - Online setup
  - Not possible to prepare 100% fit phrase table for each test data
  - Phrase table is built based on training data vocabulary, filtered
Monolingual translation system

- Built on EPPS, NC, TED, and cleaned crawl corpus of English, 10.1 m words
- Randomized sentence boundaries, so that punctuation marks can be observed in all possible locations
- Removed .,?! from the corpus and lowercased all letters for the source side
- Phrase-based translation model
Results

<table>
<thead>
<tr>
<th>Punctuation</th>
<th>ASR Output</th>
<th>Manual Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM, Prosody</td>
<td>9.74</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>11.18</td>
<td>19.57</td>
</tr>
<tr>
<td>StreamingInput</td>
<td>11.55</td>
<td>19.41</td>
</tr>
</tbody>
</table>

Scores in BLEU

- Baseline: monolingual translation system with an overlapping window 10
- 1.8 BLEU points of improvements
I also ask myself does not really work can they really store all this information about us and every time I use my mobile phone, so I ask my phone company Deutsche Telekom which was at that time the largest phone company in Germany and they ask them please send me all the information you have started about me and there is some one thousand against and I got no real...
I also ask myself, does not really work? can they really store. all this information about us. and every time I use my mobile phone. so I ask my phone company, Deutsche Telekom, which was at that time the largest phone company in Germany, and they ask them, please send me all the information you have started about me. and there is some one thousand against, and I got no real, ...
LM and prosody based

... the city of New York would
not exist in one hundred years.
why because they look at the curve and say if the population keeps
growing at.
this rate to move the population of New York year round they would have needed.
 Streaming input segment

the city of New York would not exist in one hundred years. why? because they look at the curve, and say, if the population keeps growing at this rate to move the population of New York year round. they would have needed . . .
Conclusion

- Latency issue of real-time spoken language translation systems
- Performance of punctuation prediction

- Monolingual translation system: good performance but latency due to input in sliding window

- Using streamed input
- Based on monolingual translation system
- Comparable performance
Thank you!
Example 1

I also ask myself does not really work. can they really store all this information about us. and every time I use my mobile phone.
so I ask my phone company, Deutsche Telekom, which was at that time the largest phone company in Germany, and they ask them, please send me all the information you have started about me and there some one thousand against, and I got no real and . . .
Example 2

the city of New York would not exist in one hundred years. why? because they look at the curve, and say, if the population keeps growing at this rate to move the population of New York year round. they would have needed . . .