Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff

Matthias Sperber, Mirjam Simantzik, Graham Neubig, Satoshi Nakamura, Alex Waibel

Karlsruhe Institute of Technology (Germany)
Mobile Technologies (Germany)
Nara Institute of Science and Technology (Japan)

Transactions of the Association for Computational Linguistics (TACL), April 2014
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold *(they)* in *(apron)*, and *(a)* clocks were striking thirteen.*
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold *(they) in *(apron), and *(a) clocks were striking thirteen.*
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: It was a bright cold day in April, and the clocks were striking thirteen.

Speech recognizer:

*It was a bright cold* (they) *in* (apron), and (a) *clocks* were striking thirteen.

Correcting sentences [Sperber+2013]:

wastes effort!
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold (they) in (apron), and (a) clocks were striking thirteen.*

Correcting sentences

[Sperber+2013]: wastes effort!

*It was a bright cold (they) in (apron), and (a) clocks were striking thirteen.*
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold *(they)* in *(apron)*, and *(a)* clocks were striking thirteen.

Correcting sentences
[Sperber+2013]: wastes effort!

Correcting individual words
[Sanchez-Cortina+2012]: frequent context switches \(\rightarrow\) high cognitive overhead!
User Study: Supervision efficiency drops for short segments!

Average supervision time per token [sec]

- Transcription task
- Word segmentation task
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold* *(they)* *in* *(apron)* *, and* *(a)* *clocks were striking thirteen.*

Correcting sentences
[Sperber+2013]: wastes effort!

Correcting individual words
[Sanchez-Cortina+2012]:

frequent context switches → high cognitive overhead!
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold (the) in (apron), and (a) clocks were striking thirteen.*

Correcting sentences
[Spéder+2013]: wastes effort!

Correcting individual words
[Sanchez-Cortina+2012]: frequent context switches → high cognitive overhead!
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold (they) in (apron), and (a) clocks were striking thirteen.*

Better: moderate-sized segments with high percentage of errors!

Correcting sentences

[Sperber+2013]: wastes effort!

Correcting individual words

[Sanchez-Cortina+2012]: frequent context switches → high cognitive overhead!
Supervision Strategies

Example: Let’s manually correct an automatic speech transcript…

Reference: *It was a bright cold day in April, and the clocks were striking thirteen.*

Speech recognizer:

*It was a bright cold (they) in (apron), and (a) clocks were striking thirteen.*

Previous approaches:

- Sentence-based baseline
- Word-based baseline

Proposed method:

*It was a bright cold (they) in (apron), and (a) clocks were striking thirteen.*
Roadmap

1. Segmentation Approach
2. Predictive User Models
3. User Study
Our approach to efficient supervised language annotation

Given some language corpus…

1. **Predict utility and cost** of supervising any given segment
   - according to supervision modes (e.g. typing, respeaking, skipping)

2. Define **constrained optimization goal**
   - E.g. minimize annotation time to remove at least 100 errors

3. **Choose segments** (*locations/lengths & supervision modes*), such that human supervision yields maximum efficiency according to (1) and (2)

**User Study:**
- Post-editing for speech transcription
- Active learning for word segmentation
Optimization Framework

1 (at) 2 (what’s) 3 a 4 bright 5 cold 6 ...

Institute for Anthropomatics, Department of Computer Science

Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff
Optimization Framework

Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff
Optimization Framework

1 (at) 2 (what’s) 3 a 4 bright 5 cold 6 ...

[SKIP] [TYPE] [SKIP] [TYPE] [TYPE] [SKIP] [SKIP]
Optimization Framework

1 (at) 2 (what's) 3 a 4 bright 5 cold 6 ...

- [RESPEAK]
- [SKIP]
- [TYPE]
- [TYPE]
- [TYPE]
Optimization Framework

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

"typing corrects ~2 words, needs ~5 seconds"
Optimization Framework

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

1. (at) 2. (what’s) 3. a 4. bright 5. cold 6. ...

- [RESPEAK: 1.5/2]
- [SKIP: 0/0]
- [TYPE: 1/4] [TYPE: 1/4]
- [TYPE: 2/5]
- [SKIP: 0/0]
- [TYPE: 0/6]

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

Constrained shortest path problem

1 (at) 2 (what’s) 3 a 4 bright 5 cold 6 ...

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

Constrained shortest path problem

“typing corrects ~2 words, needs ~5 seconds”
Optimization Framework

Constrained shortest path problem

Define as **integer linear program (ILP)**

Use **off-the-shelf ILP solver** to find near-optimal solution
Roadmap

1. Segmentation Approach

2. Predictive User Models

3. User Study
Roadmap

1. Segmentation Approach

2. Predictive User Models
   - Here:
     - Cost model:
       - Supervision time
     - Utility models:
       - # removed errors (post-editing setting)
       - Model improvement (active-learning setting)

3. User Study
Predicting supervision time

Input features:
- Segment length
- Segment confidence
- Audio duration

Regressor:
- Trained via annotator enrollment
- predict

Supervision time
Post editing utility: # removed errors

# errors in automatic output

\[ \sum_{t \in \text{tokens}} \text{uncertainty}(t) \]

Reliability of human annotator

Fixed human error rate (e.g. from enrollment)

# errors removed by human
Active learning utility: model improvement

Token uncertainties

Scaled token uncertainties, want:
util(tok) ∝ Model improvement
Roadmap

1. Segmentation Approach
2. Predictive User Models
3. User Study
Roadmap

1. Segmentation Approach

2. Predictive User Models

3. User Study
   - Post-editing task for speech transcription
   - Active-learning task for word segmentation
Experiment: Post Editing for Speech Transcription

- Optimization goal: given a faulty transcript, improve from 20% to 15% WER as fast as possible
- 2 supervision modes: **TYPE, SKIP**
- 3 participants corrected 2 TED transcripts (24 minutes of audio)
- Preparative step: annotator enrollment (200 random segments)
Experiment: Post Editing for Speech Transcription

- **Sentence-based**, confidence-order baseline [Sperber+2013]:
  - *Fixed* segmentation into sentence-like units
    - average length: 8.6 words
  - Order segments by average word confidence
  - Choose top-\(n\) segments, such that user model predicts 15% WER after supervision

- Proposed method:
  - Directly **optimize segmentation** according to optimization goal

- Participants transcribe 2 talks twice (once baseline, once proposed method)
  - Switch order of methods after 1\(^{st}\) talk to minimize learning bias
User Interface

Mobile Technologies Transcription Tool

42  like so this is a transaction through

43  blood in the top left hand corner

44  here you've got this yellow green area the yellow green are is the fluids of blood which is mostly water but it's

45  also anybody shivers hormones that kind of thing and the

46  red region is a slice in the red blood cell

47  and red molecules a

48  actually read that's what gives blood its color and hemoglobin

49

50  I was very much inspired by this image many

Playback: Segment 47  02:35 / 09:08
Simulation (w/ perfect time predictions & annotator accuracy)

Target WER after post-editing [%]

Post-editing time [minutes]

Sentence-based:
- Linear baseline
- Confidence-order baseline

Explicitly segmented:
- Proposed method
- Proposed w/ oracle utilities
Simulation \((w/ \text{perfect time predictions \& annotator accuracy})\)

Target WER after post-editing [%]

- **Sentence-based:**
  - Linear baseline
  - Confidence-order baseline

- **Explicitly segmented:**
  - Proposed method
  - Proposed w/ oracle utilities

Post-editing time [minutes]
Simulation (w/ perfect time predictions & annotator accuracy)

Target WER after post-editing [%]

Post-editing time [minutes]

Sentence-based:
- Linear baseline
- Confidence-order baseline

Explicitly segmented:
- Proposed method
- Proposed w/ oracle utilities
Simulation (w/ perfect time predictions & annotator accuracy)

Target WER after post-editing [%]

Post-editing time [minutes]

Sentence-based:
- Linear baseline
- Confidence-order baseline

Explicitly segmented:
- Proposed method
- Proposed w/ oracle utilities
Simulation (w/ perfect time predictions & annotator accuracy)

Target WER after post-editing [%]

Post-editing time [minutes]

- Linear baseline
- Confidence-order baseline
- Proposed method
- Proposed w/ oracle utilities

10% target WER
Simulation \(\text{w/ perfect time predictions & annotator accuracy}\)

Target WER after post-editing [%]

- Sentence-based:
  - Linear baseline
  - Confidence-order baseline

- Explicitly segmented:
  - Proposed method
  - Proposed w/ oracle utilities

- Sentence-based:
  - Linear baseline
  - Confidence-order baseline

- Explicitly segmented:
  - Proposed method
  - Proposed w/ oracle utilities

- Post-editing time [minutes]:
  - 7 min
  - 10 min

- 10% target WER
Simulation (w/ perfect time predictions & annotator accuracy)

Target WER after post-editing [%]

Sentence-based:
- Linear baseline
- Confidence-order baseline

Explicitly segmented:
- Proposed method
- Proposed w/ oracle utilities

Post-editing time [minutes]
# Real User Study

<table>
<thead>
<tr>
<th>Participant</th>
<th>Confidence baseline</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>Time</td>
</tr>
<tr>
<td>P₁</td>
<td>12.26</td>
<td>44:05</td>
</tr>
<tr>
<td>P₂</td>
<td>12.75</td>
<td>36:19</td>
</tr>
<tr>
<td>P₃</td>
<td>12.70</td>
<td>52:42</td>
</tr>
<tr>
<td>Avg.</td>
<td>12.57</td>
<td>44:22</td>
</tr>
</tbody>
</table>
## Real User Study

<table>
<thead>
<tr>
<th>Participant</th>
<th>Confidence baseline</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>Time</td>
</tr>
<tr>
<td>P₁</td>
<td>12.26</td>
<td>44:05</td>
</tr>
<tr>
<td>P₂</td>
<td>12.75</td>
<td>36:19</td>
</tr>
<tr>
<td>P₃</td>
<td>12.70</td>
<td>52:42</td>
</tr>
<tr>
<td>Avg.</td>
<td>12.57</td>
<td>44:22</td>
</tr>
</tbody>
</table>
Real User Study

<table>
<thead>
<tr>
<th>Participant</th>
<th>Confidence baseline</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>Time</td>
</tr>
<tr>
<td>P_1</td>
<td>12.26</td>
<td>44:05</td>
</tr>
<tr>
<td>P_2</td>
<td>12.75</td>
<td>36:19</td>
</tr>
<tr>
<td>P_3</td>
<td>12.70</td>
<td>52:42</td>
</tr>
<tr>
<td>Avg.</td>
<td>12.57</td>
<td>44:22</td>
</tr>
</tbody>
</table>

25.2% speed-up!

Problem w/ WER prediction: Reached ~12% WER, even though we wanted only 15%…
Word Segmentation Experiment: Summary

- Active learning scenario for Japanese word segmentation
  - Optimize model improvement, s.t. 30 minutes time budget
- Baseline: **word-based** uncertainty sampling [*Neubig+2011*]

![Graph showing classification accuracy over annotation time]

- Improvement for expert annotator over baseline
  - Time prediction in absolute terms was inaccurate: training data mismatch
- For non-experts (students), improvement less clear, due to high annotation error rate
Conclusion

Annotating **sentences** wastes effort:

- “It was a bright cold *(they)* in *(apron)*, and *(a)* clocks were striking thirteen”.

Annotating **tokens** leads to frequent context switches:

- “It was a bright cold *(they)* in *(apron)*, and *(a)* clocks were striking thirteen”.

Instead, we propose to use compact **segments that yield optimal annotation efficiency** according to cost and utility predictions:

- *It was a bright cold *(they)* in *(apron)*, and *(a)* clocks were striking thirteen.*

• Some future work…
  • Larger scale user study
  • Improve user modeling
  • Investigate other tasks such as translation
  • …
Formal Problem Definition

- Corpus of words $w_1 \ldots w_n$ e.g. ASR transcript
- Set of “supervision modes” e.g. $K=\{\text{TYPE, RESPEAK, SKIP}\}$
- Set of optimization criteria e.g. $L=\{\text{supervision time, # of errors removed}\}$
- User models $u_{l,k}$ predict outcome of supervising any given segment w.r.t. to supervision mode $k$ and criterion $l$
  - e.g., user model would probably predict that:
    - Long segments have higher cost than short segments
    - TYPING is slow, but corrects many errors
    - RESPEAKING is fast, but does not correct all errors
    - SKIPPING a segment costs no time and corrects no errors
- Specify a constrained optimization problem
  - One criterion is the objective, others are constraints
  - e.g.: correct as many errors as possible, given a 2h time budget
→ Want to find the segmentation and supervision modes that optimize this objective!
Appendix: Integer Linear Program Formulation

\[
\begin{align*}
\min_{x} & \quad \sum_{i,j \in V} \sum_{k \in E_{ij}} x_{ijk} u_{l',k}(s_{i}^{j-1}) \\
\text{s.t.} & \quad \sum_{i,j \in V} \sum_{k \in E_{ij}} x_{ijk} u_{l,k}(s_{i}^{j-1}) \leq C_{l} \\
& \quad (\forall l \in L_{-V}) \\
& \quad \sum_{i \in V} \sum_{k \in E_{ij}} x_{ijk} = \sum_{i \in V} \sum_{k \in E_{ij}} x_{jik} \\
& \quad (\forall j \in V \setminus \{1, n\}) \\
& \quad \sum_{j \in V} \sum_{k \in E_{1j}} x_{1jk} = 1 \\
& \quad \sum_{i \in V} \sum_{k \in E_{in}} x_{ink} = 1 \\
x_{ijk} & \in \{0, 1\} \quad (\forall x_{ijk} \in x)
\end{align*}
\]
Appendix: Example Segments

(3) SKIP: “nineteen forty six until today you see the green”
(4) TYPE: <annotator types: “is the traditional”>
(5) SKIP: “Interstate conflict”
(6) TYPE: <annotator types: “the ones we used to”>
(7) SKIP: …
## Appendix: Word Segmentation Experiment - Details

<table>
<thead>
<tr>
<th>Participant</th>
<th>Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Acc.</td>
</tr>
<tr>
<td>Expert</td>
<td>25:50</td>
<td>96.17</td>
</tr>
<tr>
<td>NonExp&lt;sub&gt;1&lt;/sub&gt;</td>
<td>22:05</td>
<td>95.79</td>
</tr>
<tr>
<td>NonExp&lt;sub&gt;2&lt;/sub&gt;</td>
<td>23:37</td>
<td>96.15</td>
</tr>
<tr>
<td>NonExp&lt;sub&gt;3&lt;/sub&gt;</td>
<td>25:23</td>
<td>96.38</td>
</tr>
</tbody>
</table>

Table 2: Word segmentation task results, for our expert and 3 non-expert participants. For each participant, the resulting classifier accuracy [%] after supervision is shown, along with the time [min] they needed. The unsupervised accuracy was 95.14%.
Appendix: Possible Supervision Modes

- Input modalities [Suhm et al., 2001]
- Several annotators with different expertise and cost [Donmez & Carbonell, 2008]
- Correct (post-edit) vs. annotate from scratch in MT [Specia, 2011]