Current Work 2013/4

Transcription

- transcribed and annotated ~7k words from pilot dataset (16 candidates: B1x5, B2x6, C1x5);
- comparison of crowd-sourced and gold-standard transcriptions: crowd-sourced transcriptions found to be 81% accurate (mean);
- common errors include spelling, omission of false starts and filled pauses, misinterpretation and transcribing what’s expected rather than what’s said (e.g. ‘helped’ instead of ‘help’, ‘two thousand and six’ instead of ‘two thousand six’).

Segmentation

- segmentation exercise: prosodically-motivated (automatic) where pauses are >=300ms; syntactically-motivated (manual) according to clause/phrase-structure;
- syntactic segmentation leads to longer segments (in words) and improved parse likelihoods (see Fig.4)
Annotation

- annotation of errors (in style of CLC, Nicholls 2003) and disfluencies (filled pauses, false starts, repetition, etc);
- e.g. for a small business <NS type="UD"><i>the</i></NS> word of mouth is the best advertising
- e.g. there is someone working with me <false>that is</false> that has been working at the office for five years
- disfluencies occur approx 12/100 tokens; formal errors 6/100; idiomatic errors 4/100

## tagclass tokens_affected per100words

<table>
<thead>
<tr>
<th></th>
<th>disfluency</th>
<th>846</th>
<th>12.082</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>formal</td>
<td>409</td>
<td>5.841</td>
</tr>
<tr>
<td></td>
<td>idiomatic</td>
<td>264</td>
<td>3.770</td>
</tr>
</tbody>
</table>

Parsing

- parsing experiments using RASP: 4 transcriptions modes varying in degree of transcription ‘clean-up’: A = ‘as-is’, B = ‘less-disfluency’, C = ‘less-formal-error’, D = ‘less-idiomatic-error’;
- each segment passed to RASP in each mode and in combination of modes (i.e. B+C, B+D, C+D, B+C+D);
- parse tree log likelihoods, normalized for segment length, taken as indicative of ‘canonical-ness’ of utterance (closer to zero = more likely);
- gives us an idea of distance between learner language and a target – i.e. what do we need to correct in order to improve the parse?
• greatest single improvement from disfluency correction; greatest combined improvement from BCD (disfluency + formal error + idiomatic error correction)

Figure 4: RASP parse log likelihoods (normalized) for (parallel) prosodic and syntactic segments in the BULATS corpus

<table>
<thead>
<tr>
<th>mode</th>
<th>µ</th>
<th>delta</th>
<th>segs</th>
<th>¬T/frag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A -2.429</td>
<td>0.0000</td>
<td>457</td>
<td>0.5646</td>
</tr>
<tr>
<td>2</td>
<td>B -1.894</td>
<td>0.5354</td>
<td>457</td>
<td>0.6980</td>
</tr>
<tr>
<td>3</td>
<td>C -2.409</td>
<td>0.0202</td>
<td>456</td>
<td>0.5921</td>
</tr>
<tr>
<td>4</td>
<td>D -2.396</td>
<td>0.0334</td>
<td>457</td>
<td>0.5689</td>
</tr>
<tr>
<td>5</td>
<td>BC -1.852</td>
<td>0.5770</td>
<td>457</td>
<td>0.7484</td>
</tr>
<tr>
<td>6</td>
<td>BD -1.865</td>
<td>0.5642</td>
<td>457</td>
<td>0.7155</td>
</tr>
<tr>
<td>7</td>
<td>CD -2.374</td>
<td>0.0550</td>
<td>457</td>
<td>0.5996</td>
</tr>
<tr>
<td>8</td>
<td>BCD -1.817</td>
<td>0.6121</td>
<td>457</td>
<td>0.7659</td>
</tr>
</tbody>
</table>

• in the above table: transcription ‘mode’ from A ‘as-is’ to BCD entirely corrected; mean RASP parse log likelihoods (µ); difference to baseline mode A (delta); number of segments parsed; proportion of non-fragmentary parse trees (¬T/frag)
Planned Work 2014/5

Transcription

• switch to ‘man’ multiply-transcribed dataset;
• new guidelines for transcribers? (transcribe what you hear, including disfluencies etc)

Segmentation

• how to automate syntactic-type segmentation?
• experiment with different algorithms;

Annotation

• further error annotation (crowdsourcing?);
• automatic disfluency detection and accompanying public data release;
• XML and annotation scheme documentation;

Parsing

• treebank analysis;
• subcategorization frame analysis;

New lines of work

• proficiency classification / clustering (based on ‘eval1’ dataset);
• semantic annotation;
• native speaker (and/or C2 non-native speaker) data collection on comparable tasks;
• prosody and pronunciation (c/o Calbert): sound segments and syllable structure, vowel formants and voiced onset time (VOT), acoustic (phonetic) cues [duration, intensity, pitch], accent distribution & intonation contours, rhythm metrics.