Disfluencies

A flight to um Berlin I mean Munich on Tuesday

→ A flight to Munich on Tuesday
Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- ‘pilot’ dataset: 1000 candidates, 19085 recordings;
- L1 Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, etc;
- CEFR proficiency levels $A2 \Leftrightarrow C2$. 

Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- ‘pilot’ dataset: 1000 candidates, 19085 recordings;
- L1 Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, etc;
- CEFR proficiency levels A2 ⇔ C2.
Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- ‘pilot’ dataset: 1000 candidates, 19085 recordings;
- L1 Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, etc;
- CEFR proficiency levels A2 ⇔ C2.
**Spoken language: learners**

- BULATS learner English (<60s prompted test recordings);
- ‘pilot’ dataset: 1000 candidates, 19085 recordings;
- L1 Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, etc;
- CEFR proficiency levels $A2 \Leftrightarrow C2$. 
**Spoken language: learners**

- BULATS learner English (<60s prompted test recordings);
- ‘pilot’ dataset: 1000 candidates, 19085 recordings;
- L1 Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, etc;
- CEFR proficiency levels $A2 \leftrightarrow C2$. 
Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- crowdsourced transcriptions;
- gold-standard: hand-annotated 5667 token subset, errors & disfluencies;
- $B_1, B_2, C_1$: c.1800 tokens each.
Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- crowdsourced transcriptions;
- gold-standard: hand-annotated 5667 token subset, errors & disfluencies;
- \(B_1, B_2, C_1: \) c.1800 tokens each.
Spoken language: learners

- BULATS learner English (<60s prompted test recordings);
- crowdsourced transcriptions;
- gold-standard: hand-annotated 5667 token subset, errors & disfluencies;
- $B_1, B_2, C_1$: c.1800 tokens each.
Spoken language: native speaker

- Switchboard (SWB) U.S. English telephone conversations;
- Used in previous work –

<table>
<thead>
<tr>
<th>System</th>
<th>Disfl.$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zwarts &amp; Johnson 2011 <em>ACL</em></td>
<td>83.8</td>
</tr>
<tr>
<td>Qian &amp; Liu 2013 <em>NAACL</em></td>
<td>84.1</td>
</tr>
<tr>
<td>Rasooli &amp; Tetreault 2014 <em>EACL</em></td>
<td>82.6</td>
</tr>
<tr>
<td><strong>Honnibal &amp; Johnson 2014</strong> <em>(Redshift)</em> <em>TACL</em></td>
<td>84.1</td>
</tr>
</tbody>
</table>
Transition-based dependency parsing

Dependency parser: we have a buffer (the sentence to parse), a stack (starts empty), and we want to build a set of arc relations. We use 4(+1) operations: –

- **SHIFT** – moves first word from buffer to stack
- **REDUCE** – pops stack
- **RIGHT ARC** – draws arc from stack to buffer, shifts buffer
- **LEFT ARC** – draws arc from buffer to stack, pops stack
- (EDIT ...)

10/41
Transition-based dependency parsing

his company went broke

S
N
A

Shift LeftArc Reduce RightArc
his company went broke

transition-based dependency parsing
Transition-based dependency parsing

S

N

A

company
went
broke
ROOT

his ← company

LeftArc
A company went broke. 

\[ S \quad N \quad A \]

- \text{company}
- \text{went broke ROOT}
- \text{his } \leftarrow \text{company}

\text{Shift}
Transition-based dependency parsing

\[ S \]

\[ N \]

\[ A \]

\begin{align*}
\text{went} & \quad \text{broke} \\
\text{ROOT} & \\
\end{align*}

\begin{align*}
\text{his} & \leftarrow \text{company} \\
\text{company} & \leftarrow \text{went} \\
\end{align*}
Transition-based dependency parsing

\[
\begin{align*}
S & \quad \text{went} \\
N & \quad \text{broke} \\
A & \quad \text{his} \leftarrow \text{company} \\
\end{align*}
\]

\[
\begin{align*}
\text{Shift} & \\
\text{broke} & \quad \text{ROOT} \\
\text{company} & \leftarrow \text{went} \\
\end{align*}
\]
Transition-based dependency parsing

\[ S \]

- broke
- went

\[ N \]

- ROOT

\[ A \]

- his ← company
- company ← went
- went → broke

RightArc
Transition-based dependency parsing

\[ S \]

\[ N \]

\[ A \]

\text{went}

\text{ROOT}

\text{his} \leftarrow \text{company}
\text{company} \leftarrow \text{went}
\text{went} \rightarrow \text{broke}

Reduce
Transition-based dependency parsing

\[ S \quad N \quad A \]

- \textit{his} \leftarrow \textit{company}
- \textit{company} \leftarrow \textit{went}
- \textit{went} \rightarrow \textit{broke}
- \textit{went} \leftarrow \text{ROOT}

LeftArc
Transition-based dependency parsing

ROOT

ROOT₀  his₁  company₂  went₃  broke₄
Transition-based dependency parsing

How are actions selected?

- selecting next action is a classification task,
- averaged perceptron to score action sequences,
- uses beam search to select next best step(s).

How do we score each possible action?

- Features are defined by 73 templates from Zhang & Nivre (2011), plus extensions by H&J, which refer to 12 context tokens around word;
- e.g. $S_{0p}N_{0wp}$, POS-tag for first word on stack plus word and POS-tag for first word on buffer.
Introducing the EDIT action, to remove disfluent items:

- **Shift, Reduce, Right Arc, Left Arc**
- **Edit** – on detection of disfluency, removes words and their dependencies.
- use extra features introduced by H&J to detect disfluency.
His company went broke. I mean went bankrupt.
Transition-based dependency parsing

\[ S \quad N \quad A \quad D \]

his company went\textsubscript{3} broke i\_mean went\textsubscript{6} bankrupt ROOT

Shift \quad LeftArc \quad Reduce \quad RightArc \quad Edit
Transition-based dependency parsing

\[
\begin{align*}
S & : \text{went}_3 \\
N & : \text{bankrupt} \\
A & : \text{his } \leftarrow \text{company} \\
&D & : \text{went}_3 \rightarrow \text{broke} \\
&D & : \text{i_mean } \leftarrow \text{went}_6
\end{align*}
\]
Transition-based dependency parsing

S

company

N

gone

went₆

bankrupt

ROOT

A

his ← company

company ← went₃

went₃ → broke

i_mean ← went₆

D

went₃

broke

i_mean
Transition-based dependency parsing

\[
\begin{align*}
S & \quad N & \quad A & \quad D \\
\text{ROOT} & & \text{went} & \quad \text{went} \\
\text{his} & \quad \text{company} & \quad \text{went} & \quad \text{broke} \\
\text{company} & \quad \text{went} & \quad \text{bankrupt} & \quad \text{i_mean} \\
\text{went} & \quad \text{ROOT} & & \\
\end{align*}
\]
Transition-based dependency parsing

ROOT

poss

nsubj
dobj

ROOT_0 his_1 company_2 went_3 broke_4 i_mean_5 went_6 bankrupt_7
Experiment: tasks

Following H&J:

- **Task 1**: dependency relations
  - correctly identify head of each token
  - $\rightarrow$ **UAS**: unlabelled attachment score

- **Task 2**: disfluency detection
  - correctly apply EDIT transition to RM tokens
  - $\rightarrow$ **Disfl.F**: disfluency $F$-measure $= 2\left(\frac{p \times r}{p + r}\right)$
EXPERIMENT: TASKS

Following H&J:

▶ **Task 1: dependency relations**
  ▶ correctly identify head of each token
  ↔ **UAS**: unlabelled attachment score

▶ **Task 2: disfluency detection**
  ▶ correctly apply EDIT transition to RM tokens
  ↔ **Disfl.F**: disfluency $F$-measure $= 2\left(\frac{p \times r}{p + r}\right)$
Experiment: tasks

Following H&J:

- **Task 1: dependency relations**
  - correctly identify head of each token
  - $\rightarrow$ **UAS**: unlabelled attachment score

- **Task 2: disfluency detection**
  - correctly apply **EDIT** transition to RM tokens
  - $\rightarrow$ **Disfl.$F$**: disfluency $F$-measure $= 2\left(\frac{p \times r}{p + r}\right)$
Experiment: tasks

Following H&J:

- **Task 1:** dependency relations
  - correctly identify head of each token
  - $\rightarrow$ uAS: unlabelled attachment score

- **Task 2:** disfluency detection
  - correctly apply EDIT transition to RM tokens
  - $\rightarrow$ Disfl.$F$: disfluency $F$-measure $= 2\left(\frac{p \times r}{p + r}\right)$
Experiment: Tasks

Following H&J:

- **Task 1:** dependency relations
  - correctly identify head of each token
  - $\Rightarrow$ **UAS:** unlabelled attachment score

- **Task 2:** disfluency detection
  - correctly apply EDIT transition to RM tokens
  - $\Rightarrow$ **Disfl.F:** disfluency $F$-measure $= 2\left(\frac{p \times r}{p+r}\right)$
## Experiment: Results

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Tokens</th>
<th>UAS</th>
<th>Disfl.P/R</th>
<th>Disfl.F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWB: test</td>
<td>45,405</td>
<td>90.5</td>
<td>n/a</td>
<td>84.1</td>
</tr>
<tr>
<td>SWB: dev</td>
<td>45,381</td>
<td>90.9</td>
<td>92.3/76.5</td>
<td>83.7</td>
</tr>
<tr>
<td>BULATS</td>
<td>5667</td>
<td>91.1</td>
<td>82.6/33.6</td>
<td>47.8</td>
</tr>
<tr>
<td>B1</td>
<td>1895</td>
<td>88.9</td>
<td>85.3/31.4</td>
<td>45.9</td>
</tr>
<tr>
<td>B2</td>
<td>1879</td>
<td>91.2</td>
<td>79.2/33.2</td>
<td>46.8</td>
</tr>
<tr>
<td>C1</td>
<td>1893</td>
<td>93.0</td>
<td>83.8/37.3</td>
<td>51.6</td>
</tr>
</tbody>
</table>
EXPERIMENT: DISCUSSION

And the people from twenty five to fifty years old

they have the percentage of them

RM \quad RP

\rightarrow And the people from twenty five to fifty years old the percentage...
Experiment: Discussion

My colleagues will advise me

for the working stra they will guide for

RM

how to behave in the office

RP

→ My colleagues will advise me how to behave in the office
Experiment: discussion

That I want they maybe they maybe help for my

RM

they maybe took my advice

RP

→ That they maybe took my advice
Experiment: discussion

That I want they maybe they maybe help for my
ncsubj  ncsubj

they maybe took my advice
ncsubj

→ That they maybe took my advice
ncsubj