Enriching ASR Lattices with POS Tags for Dependency Parsing
Motivation

Parsing speech
- POS tags (or other labels) are helpful to downstream tasks
- Lattice-level tags allow for further task integration
Title: Enriching ASR Lattices with POS Tags for Dependency Parsing

Motivation

Parsing speech
- POS tags (or other labels) are helpful to downstream tasks
- Lattice-level tags allow for further task integration

First step

POS-enriched ASR word lattices
Title: Enriching ASR Lattices with POS Tags for Dependency Parsing

Motivation

Parsing speech
- POS tags (or other labels) are helpful to downstream tasks
- Lattice-level tags allow for further task integration

First step

POS-enriched ASR word lattices

A pipeline approach:
Title: Enriching ASR Lattices with POS Tags for Dependency Parsing

Motivation

Parsing speech
- POS tags (or other labels) are helpful to downstream tasks
- Lattice-level tags allow for further task integration

First step

POS-enriched ASR word lattices

Our approach:

n-best

ASR-lattice

This work: Joint ASR-POS

nsubj

nobj

ASR

POS tagger

DP

She likes CS

She|PRP likes|VBZ CS|NN

She PRP

likes VBZ

CS NN
Method

- Using the Kaldi ASR toolkit (Povey et al., 2011)
Method

- Using the Kaldi ASR toolkit (Povey et al., 2011)
Method

- Using the Kaldi ASR toolkit (Povey et al., 2011)
A word-POS paired lexicon

read
A word-POS paired lexicon

read

read reh d

read r iy d
A word-POS paired lexicon

Moritz Stiefel & Ngoc Thang Vu, Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart: Enriching ASR Lattices with POS Tags for Dependency Parsing
Data: Switchboard splits

- North-American English
- Treebank-3 transcription (not MS-State transcription!)

<table>
<thead>
<tr>
<th>Set</th>
<th>Conversations</th>
<th>Utterances</th>
<th>Tokens</th>
<th>Avg. tok./utt.</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>2xxx-3xxx</td>
<td>90823</td>
<td>677160</td>
<td>7.46</td>
<td>14759</td>
</tr>
<tr>
<td>dev</td>
<td>4519-4936</td>
<td>5697</td>
<td>50148</td>
<td>8.80</td>
<td>3761</td>
</tr>
<tr>
<td>eval</td>
<td>4004-4153</td>
<td>5822</td>
<td>48320</td>
<td>8.30</td>
<td>3695</td>
</tr>
<tr>
<td>lmdev</td>
<td>4154-4483</td>
<td>5949</td>
<td>50017</td>
<td>8.41</td>
<td>3742</td>
</tr>
</tbody>
</table>

Table: Summary of SWBD data splits
Data: Switchboard splits

- North-American English
- Treebank-3 transcription (not MS-State transcription!)

<table>
<thead>
<tr>
<th>Set</th>
<th>Conversations</th>
<th>Utterances</th>
<th>Tokens</th>
<th>Avg. tok./utt.</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>2xxx-3xxx</td>
<td>90823</td>
<td>677160</td>
<td>7.46</td>
<td>14759</td>
</tr>
<tr>
<td>dev</td>
<td>4519-4936</td>
<td>5697</td>
<td>50148</td>
<td>8.80</td>
<td>3761</td>
</tr>
<tr>
<td>eval</td>
<td>4004-4153</td>
<td>5822</td>
<td>48320</td>
<td>8.30</td>
<td>3695</td>
</tr>
<tr>
<td>lmdev</td>
<td>4154-4483</td>
<td>5949</td>
<td>50017</td>
<td>8.41</td>
<td>3742</td>
</tr>
</tbody>
</table>

**Table:** Summary of SWBD data splits

<table>
<thead>
<tr>
<th>LM</th>
<th>Baseline 2-gram</th>
<th>Baseline 3-gram</th>
<th>Joint 2-gram</th>
<th>Joint 3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPL</td>
<td>89.4</td>
<td>76.3</td>
<td>96.4</td>
<td>84.2</td>
</tr>
</tbody>
</table>

**Table:** Language model (LM) perplexities (PPL) on lmdev.
Data: Switchboard POS-enriched transcription

Processing at Mississippi State University ran a clean-up project which hand-checked and corrected the transcript of the 1126 Treebank conversations. They also produced word alignments, showing, for each transcript word, its start and end times in the audio file; word times were determined automatically, with partial manual corrections (see Deshmukh et al. 1998; Harkins 2003). We refer to the resulting time-aligned transcript as the MS-State transcript.

Since both the Treebank3 and MS-State transcripts had been enriched with distinct annotations, we included both transcripts separately in our corpus, using an NXT pointer to link equivalent words in the two versions. Section 5.1 describes the method used to create the alignment between the two transcriptions. We refer to the words from the Treebank3 transcript as words and the words from the MS-State transcript as phonwords, since the MS-State transcript words have start and end times in the audio file and hence are slightly more phonetically grounded. The double inclusion does result in redundancy, but has the advantage of retaining the internal consistency of prior annotations. For the most part, the MS-State transcription is more accurate than the Treebank3, so the other option would have been to attach all of the annotations that were derived from the Treebank transcription to the MS-State transcription and discard the original. However, attaching the Treebank annotations exactly as they are would have made the resource difficult for the end-user to interpret. For instance, where the MS-State transcription adds words to the original, the syntactic annotation would appear inconsistent. On the other hand, creating new annotations to cover the changed portions of the transcription would have been time-consuming for little gain and would have greatly complicated the relationship between the NXT-format data and the original.

Figure 1 shows our solution diagrammatically. As can be seen, where there are differences in the representation of a word in the two transcripts (e.g. in the treatment of contractions like doesn’t), one Treebank3 ‘word’ may link to more than one MS-State ‘phonword’, or vice versa.
Data: Switchboard POS-enriched transcription

- Orthography/tokenization and POS tags from the Treebank data (word)
- Timestamps from linked MS-State transcriptions (phonword)

MS-State vs Treebank-3 transcription, from Calhoun et al. (2010, p. 392)
Intermediate results: ASR

\textit{tri4}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{asr_diagram.png}
\caption{1-best ASR result: \textit{She likes CS}}
\end{figure}
**Intermediate results: ASR**

**tri4**

*1-best*  
She likes CS

---

**Joint-POS**

*1-best*  
She |PRP likes |VBZ CS |NN
Intermediate results: ASR

tri4

Joint-POS

1-best
She likes CS

1-best
She\text{PRP} likes\text{VBZ} CS\text{NN}

<table>
<thead>
<tr>
<th>Set</th>
<th>\textit{tri4}</th>
<th>Joint-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>28.75 (65.83)</td>
<td>28.93 (65.28)</td>
</tr>
<tr>
<td>test</td>
<td>29.41 (64.41)</td>
<td>29.26 (64.15)</td>
</tr>
</tbody>
</table>

\textbf{Table}: ASR results: numbers are WER (SER)
Intermediate results: POS

\[ tri4+AP \]

\[ \begin{align*}
\text{tri4} & \quad 1\text{-best} & \quad \text{AP tagger} \\
\text{She} & \quad \text{PRP} & \quad \text{likes} & \quad \text{VBZ} & \quad \text{CS} & \quad \text{NN}
\end{align*} \]
Intermediate results: POS

*tri4+AP*

```
tri4+ME.pre tri4+AP .pre tri4+spaCy.pre tri4+ME.70k tri4+AP
```

**Joint-POS**

```
She|PRP likes|VBZ CS|NN
```

```
Joint-POS 1-best

She|PRP likes|VBZ CS|NN
```

```
tri4 1-best

AP tagger

She|PRP likes|VBZ CS|NN
```
** Intermediate results: POS **

**tri4+AP**

**Joint-POS**

<table>
<thead>
<tr>
<th>Set</th>
<th>tri4+ME.pre</th>
<th>tri4+AP.pre</th>
<th>tri4+spaCy.pre</th>
<th>tri4+ME.70k</th>
<th>tri4+AP</th>
<th>Joint-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>43.29 (94.23)</td>
<td>45.46 (95.84)</td>
<td>39.17 (82.38)</td>
<td>33.24 (68.18)</td>
<td>32.30 (67.67)</td>
<td><strong>32.05 (67.32)</strong></td>
</tr>
<tr>
<td>test</td>
<td>44.49 (94.19)</td>
<td>46.18 (95.74)</td>
<td>40.42 (81.86)</td>
<td>36.23 (67.26)</td>
<td>33.10 (66.85)</td>
<td><strong>32.52 (66.52)</strong></td>
</tr>
</tbody>
</table>

**Table:** POS tagging results: numbers are WER (SER)
DP results

- 1-best hypotheses of standard Kaldi \textit{tri4} setup plus AP tagger vs our Joint-POS
- Xiang Yu’s parser after (Weiss et al., 2015): greedy neural transition-based parser, uses word and POS features

<table>
<thead>
<tr>
<th></th>
<th>dev</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>utts</td>
<td>900</td>
<td>tokens</td>
<td>4881</td>
<td>94.30</td>
<td>92.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95.41</td>
<td>93.63</td>
</tr>
<tr>
<td>test</td>
<td>882</td>
<td></td>
<td>4827</td>
<td>94.68</td>
<td>93.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>94.92</td>
<td>93.52</td>
</tr>
</tbody>
</table>

Table: Parsing results for subsets of correct tokenizations. Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) given as percentages.
DP results

- 1-best hypotheses of standard Kaldi `tri4` setup plus AP tagger vs our Joint-POS
- Xiang Yu’s parser after (Weiss et al., 2015): greedy neural transition-based parser, uses word and POS features

<table>
<thead>
<tr>
<th>Set</th>
<th>#utts</th>
<th>#tokens</th>
<th><code>tri4+AP</code></th>
<th>Joint-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>dev</td>
<td>900</td>
<td>4881</td>
<td>94.30</td>
<td>92.71</td>
</tr>
<tr>
<td>test</td>
<td>882</td>
<td>4827</td>
<td>94.68</td>
<td>93.06</td>
</tr>
</tbody>
</table>

Table: Parsing results for subsets of correct tokenizations. Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) given as percentages.
DP results

- 1-best hypotheses of standard Kaldi \textit{tri4} setup plus AP tagger vs our Joint-POS
- Xiang Yu’s parser after (Weiss et al., 2015): greedy neural transition-based parser, uses word and POS features

<table>
<thead>
<tr>
<th>Set</th>
<th>#utts</th>
<th>#tokens</th>
<th>\textit{tri4+AP}</th>
<th>Joint-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>dev</td>
<td>900</td>
<td>4881</td>
<td>94.30</td>
<td>92.71</td>
</tr>
<tr>
<td>test</td>
<td>882</td>
<td>4827</td>
<td>94.68</td>
<td>93.06</td>
</tr>
</tbody>
</table>

\textbf{Table:} Parsing results for subsets of correct tokenizations. Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) given as percentages.

- High scores, but only on utterances with \textbf{correct} tokenizations
DP results extended

- Number of correctly tokenized utterances ≠ number of utterances
DP results extended

- Number of correctly tokenized utterances ≠ number of utterances
  - How can we evaluate incorrectly recognized utterances?
DP results extended

- Number of correctly tokenized utterances ≠ number of utterances
  - How can we evaluate incorrectly recognized utterances?
  - Our answer: fuzzy relation-based measure that ignores word position altogether
Initialize UAS, LAS, US and LS with zero count
For all reference utterances $R$ that have a hypothesis $H$
For all tokens $R_i$ in $R$

- Get dependent of $R_i$ in $H$
- Get dependent of $R_i$ in $R$
- Get head of $R_i$ in $H$
- Get head of $R_i$ in $R$

- Head and dependent of $R_i$ in both $R$ and $H$ match in word and position? yes
  - Increase UAS count

- Relation label of $R_i$ exists for this relation in $H$? yes
  - Increase LS count

- Head and dependent of $R_i$ exist in a directed relation in $H$? yes
  - Increase US count

- Relation label of $R_i$ in $R$ and $H$ match? yes
  - Increase LAS count

- Head and dependent of $R_i$ in both $R$ and $H$ match in word and position? no
- Relation label of $R_i$ in $R$ and $H$ match? no
- Head and dependent of $R_i$ exist in a directed relation in $H$? no
- Relation label of $R_i$ exists for this relation in $H$? no
DP results extended

- Number of correctly tokenized utterances $\neq$ number of utterances
  - How can we evaluate incorrectly recognized utterances?
  - Our answer: fuzzy relation-based measure that ignores word position altogether
DP results extended

- Number of correctly tokenized utterances $\neq$ number of utterances
- How can we evaluate incorrectly recognized utterances?
- Our answer: fuzzy relation-based measure that ignores word position altogether

<table>
<thead>
<tr>
<th>Model</th>
<th>Set</th>
<th>UAS</th>
<th>LAS</th>
<th>US</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>tri4+AP</td>
<td>dev</td>
<td>32.20</td>
<td>31.20</td>
<td>52.02</td>
<td>49.40</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>31.21</td>
<td>30.29</td>
<td>50.72</td>
<td>48.33</td>
</tr>
<tr>
<td>Joint-POS</td>
<td>dev</td>
<td>32.41</td>
<td>31.43</td>
<td>52.21</td>
<td>49.71</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>31.56</td>
<td>30.73</td>
<td>51.21</td>
<td>48.99</td>
</tr>
</tbody>
</table>

Table: Parsing results on full dev and test sets. LAS and UAS given as percentages. LS (labeled score) and US (unlabeled score) are a fuzzy evaluation metric devised to be able to evaluate tokenization mismatches between the ASR hypotheses and the corresponding treebank data. LS and US are also given as percentages. The dev set has 3994 utterances with 44760 tokens and the test set has 3912 utterances with 43277 tokens. Best scores per set are bold-faced.
DP-based error analysis 1/3

- tri4 token incorrect, subsequent POS tag, too

**Figure:** Correct Joint-POS tree on the left, incorrect tri4 tree on the right.
DP-based error analysis 2/3

- *tri4* ASR deletion error

```
  do  you  like  rap  music
  VBP PRP VB NN NN

  you  like  rap  music
  PRP VB NN NN
```

**Figure**: Correct Joint-POS tree on the left, incorrect *tri4* tree on the right.
DP-based error analysis 3/3

- Joint-POS token sequence incorrect resulting in erroneous parse

Figure: Correct tri4 tree on the left, incorrect Joint-POS tree on the right.
Conclusions

- Successful joint ASR and POS tagging
  - Increased search space in the decoding graph
  - No performance loss compared to pipeline approach

⇒ POS tags in an ASR lattice structure

- Possible avenues of exploration:
  - Systematic error analysis
  - Use transcriptions tagged with a POS-tagger and compare results
  - Comparison against the approach of Velikovich (2016), who tag lattices
Moritz Stiefel & Ngoc Thang Vu
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart

eMail moritz.stiefel@ims.uni-stuttgart.de
Telefon +49-711-685 813 60
Fax +49-711-685 813 66
References


This work was funded by the German Research Foundation (DFG) through the Collaborative Research Center (SFB) 732, project A8, at the University of Stuttgart.
DP-based error analysis extra 1/2

- $tri4$ token incorrect, subsequent POS tag, too

Figure: Correct Joint-POS tree on the left, incorrect $tri4$ tree on the right.
• *tri4* with correct tokenization, but POS tagging error

**Figure:** Correct Joint-POS tree on the left, incorrect *tri4* tree on the right.