First Step Towards Enhancing Word Embeddings with Pitch Accents for DNN-based Slot Filling on Recognized Text

Sabrina Stehwien, Ngoc Thang Vu

IMS, University of Stuttgart

March 16, 2017
Slot Filling

- sequential labelling task to assign semantic labels to each word in an input sequence
- key query terms “fill” a semantic frame or *slot* e.g. locations, time periods
- benchmark corpus: Airline Travel Information Systems (ATIS)
- state-of-the-art DNN models yield around 95% F1-Score
- typical features: word embeddings (lexico-semantic representations)
- example:

  SHOW 0 FLIGHTS 0 FROM 0 BURBANK B-fromloc.city_name TO 0 MILWAUKEE B-toloc.city_name FOR 0 TODAY B-depart_date.today_relative
Motivation

- slot filling is a text-based task, however:
- spoken language understanding (SLU) involves automatic speech recognition (ASR) as first step
- realistic setting: apply and optimize on ASR output, taking recognition error into account
- related work shows that slot filling performance drops on recognized text
- additional information that is extracted from the speech signal and not present in text may help
- prosodic information, e.g. pitch accents
Pitch Accents in Slot Filling

- certain words are marked as salient to highlight important information (focus, contrast, information status)
- pitch accents are useful for various NLP and SLU tasks: named entity recognition, coreference resolution, dialog act segmentation, etc.
- human listeners may recover recognition errors using context information and prosodic cues
- content words with new information status are typically pitch accented
- e.g. *List FLIGHTS from DALLAS to HOUSTON*
- a previous study has shown that words with automatically predicted pitch accents account for 90% of the slots in a subset of ATIS (Stehwien & Vu, 2016)
Bidirectional Recurrent Neural Network with Ranking Loss (Vu et al. 2015)

- **bi-directionality:** combination of forward and backward hidden layer models past and future context
- **ranking loss function** maximizes distance between true label and best target
- **100-dimensional word embeddings**
- **95.56% F1-score on ATIS**
Bidirectional Sequential Convolutional Neural Network (Vu, 2016)

- combination of two CNNs that model past and future contexts respectively
- additional surrounding context gives current word more weight
- 50-dimensional word embeddings
- 95.61% F1-score
Word Embeddings with Pitch Accent Extensions

• word embeddings are vector representations of words based on their lexical and semantic context

• word embedding of $w$ concatenated with a binary flag indicating the absence or presence of a pitch accent on $w$:

$$embs(w) = [lexical\_embs(w), pitch\_accent\_flag(w)]$$  \hspace{1cm} (1)

• combines acoustic-prosodic information and lexico-semantic word embeddings
Method

- recognize ATIS corpus from audio signal with ASR (7% WER)
- obtain the word, syllable, and phone alignments
- pitch accent detector determines the binary label for each word
- the word embeddings are trained and concatenated with the binary pitch accent flag
- compare slot filling performance on original transcriptions and recognized version
Pitch Accents in ATIS

- analyze co-occurrence of (predicted) pitch accents and slots in ATIS
- compare on manual transcriptions and recognized test set
- almost 93% of slots are pitch accented in both versions

<table>
<thead>
<tr>
<th></th>
<th>manual</th>
<th>recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td># words</td>
<td>9551</td>
<td>9629</td>
</tr>
<tr>
<td># slots</td>
<td>3663</td>
<td>3560</td>
</tr>
<tr>
<td>pred. accents on slots</td>
<td>64.1%</td>
<td>64.0%</td>
</tr>
<tr>
<td>slots with pred. accent</td>
<td>92.7%</td>
<td>92.9%</td>
</tr>
</tbody>
</table>
Pitch Accents in Neural Models: Results

- results on ASR output are much worse than on manual transcriptions
- pitch accent extensions do not help on original text → context information suffices
- pitch accent extensions slightly improve F1-score on ASR output

<table>
<thead>
<tr>
<th></th>
<th>RNN</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptions (lexical word embeddings)</td>
<td>94.97</td>
<td>95.25</td>
</tr>
<tr>
<td>+ pitch accent extensions</td>
<td>94.98</td>
<td>95.25</td>
</tr>
<tr>
<td>ASR output (lexical word embeddings)</td>
<td>89.55</td>
<td>89.13</td>
</tr>
<tr>
<td>+ pitch accent extensions</td>
<td>90.04</td>
<td>89.57</td>
</tr>
</tbody>
</table>
Analysis

- *unknown* tokens replace words in the benchmark dataset that occur only once
- the ASR system also produces more unknown tokens due to recognition errors
- analysis of RNN results on unknown tokens, independent of slot type:
  - baseline: 43% correct
  - with pitch accent extensions: 51% correct
  - indicates that pitch accent information helped to localize a slot, even though the actual label may be incorrect
  - unknown tokens may still carry helpful information that is captured by this method
Examples

| reference                                      | I NEED THE FLIGHTS FROM **WASHINGTON** TO MONTREAL ON A SATURDAY |
| recognized                                    | I NEED THE FLIGHTS FROM `<UNK>` TO MONTREAL ON SATURDAY |
| ref. slots                                    | 0 0 0 0 0 B-fromloc.city_name O B-toloc.city_name O B-depart_date.day_name |
| with accents                                  | 0 0 0 0 0 B-fromloc.city_name O B-toloc.city_name O B-depart_date.day_name |
| baseline                                      | 0 0 0 0 0 O O B-toloc.city_name O B-depart_date.day_name |

→ unknown token is labelled correctly

| reference                                      | WHICH AIRLINES FLY BETWEEN **TORONTO** AND SAN DIEGO |
| recognized                                    | WHICH AIRLINES FLY BETWEEN **TO ROUND `<UNK>`** AND SAN DIEGO |
| ref. slots                                    | 0 0 0 0 0 O O O O B-toloc.city_name I-toloc.city_name |
| with accents                                  | 0 0 0 0 0 O O O O B-toloc.city_name I-toloc.city_name |
| baseline                                      | 0 0 0 0 B-fromloc.city_name B-round_trip I-round_trip O B-toloc.city_name ...

→ misrecognized words are labelled more appropriately
Conclusion

- we addressed the notion of overcoming the performance drop of state-of-the-art slot filling methods on speech recognition output
- extended word embedding vectors with pitch accent features
- small but positive effects were obtained on two models (RNN and CNN)
- limited and closed-domain nature of ATIS may be accountable for small differences
- evidence that pitch accent features may help in the case of misrecognized or unknown words
References

N.T. Vu et al. (2015)
Bi-directional Recurrent Neural Network with Ranking Loss for Spoken Language Understanding
*IEEE Transactions on Audio, Speech and Language Processing*

N. T. Vu (2016)
Sequential Convolutional Neural Networks for Slot Filling in Spoken Language Understanding
*Proceedings of Interspeech*

G. Mesnil et al. (2015)
Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding
*IEEE Transactions on Audio, Speech and Language Processing*

S. Stehwien and N. T. Vu (2016)
Exploring the Correlation of Pitch Accents and Semantic Slots for Spoken Language Understanding
*Proceedings of Interspeech*

A. Schweitzer (2010)
Production and Perception of Prosodic Events - Evidence from Corpus-based Experiments
*Ph.D. thesis, Universität Stuttgart*