Local methods for on-demand OOV word retrieval

Stanislas Oger, Georges Linarès, Frédéric Béchet

Laboratoire d’Informatique d’Avignon (LIA) - University of Avignon
339 ch. des Meinajaries, BP 1228
F-84911 Avignon Cedex 9 (France)

{stanislas.oger, georges.linares, frederic.bechet}@univ-avignon.fr
Automatic speech recognition

1. Speech signal $\rightarrow$ Lexicon $\rightarrow$ Transcription
2. All the words in the transcription are in the Lexicon
3. Word not in the lexicon = Transcription error

Problem

1. Finite lexicon size
2. Always Out-Of-Vocabulary (OOV) words
Plan

1. Introduction

2. Our approach
   - Overview
   - Experimental framework

3. OOV words retrieval

4. Conclusion
Overview of our approach

1st Decoding Pass

Speech Signal

2nd Decoding Pass

Final Transcription

OOV Words Detection

OOV Words Retrieval

The Web
The speech corpus
- 6 hours of French broadcast news from ESTER
- A 65k lexicon
- 1.03% of OOV words
- 73% named entities / 24% technical words

The Web corpus
- Google search engine
Plan

1. Introduction

2. Our approach

3. OOV words retrieval
   - Our approach
   - The Web as corpus
   - N-grams Strategy
   - Patterns Strategy
   - Semantics Driven N-gram Strategy

4. Conclusion
Our approach

We have

- OOV words identified in the transcription

We want

- Retrieve the OOV words

Our method

- The local context bring information on the OOV words
- Use this information to retrieve the OOV words on the Web
Using the Web

1. The Web considered as an unlimited source of words
2. Continuously updated

<table>
<thead>
<tr>
<th>n-gram</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>100.00 %</td>
<td>88.22 %</td>
<td>50.54 %</td>
<td>27.29 %</td>
<td>16.12 %</td>
</tr>
</tbody>
</table>

Tab.: $n$-grams containing OOV words on Google depending on the size $n$. 

S. Oger, G. Linarès, F. Béchet - University of Avignon
N-gram Strategy

The goal

- Retrieve words which occur in the same context

The method

- Search the N-grams with the same head
- Build requests and retrieve documents
- Search the pattern in the documents

Example

- “Les otages Christian chez nos et Georges [...]”
- “otages Christian * ”
Experimental results

<table>
<thead>
<tr>
<th>n-gram</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>13.9%</td>
<td>18.1%</td>
<td>16.4%</td>
<td>13.8%</td>
</tr>
<tr>
<td>Set size</td>
<td>145</td>
<td>49</td>
<td>13</td>
<td>4</td>
</tr>
</tbody>
</table>

**Tab.**: Recall and sets size of the $n$-grams strategy for OOV word retrieval using Google depending on the size $n$. 
Pattern Strategy

The goal

- Retrieve words which occur in about the same context

The method

- The same method that previously
- Relax constraints on stop-words
- Allow words insertion

Example

- “Les otages Christian chez nos et Georges [...]”
- “otages * Christian * ”
Experimental results

<table>
<thead>
<tr>
<th>n-gram</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>20.0 %</td>
<td>20.3 %</td>
<td>17.5 %</td>
<td>12.2 %</td>
</tr>
<tr>
<td>Set size</td>
<td>411</td>
<td>139</td>
<td>34</td>
<td>15</td>
</tr>
</tbody>
</table>

**Tab.**: Recall and sets size of the pattern strategy for OOV word retrieval using Google depending on the size $n$. 
The goal

- Allow the search engine to better rank documents

The method

- The same method that the n-gram strategy
- Add a relevant context words (Drive Words)

Example

- “Les otages Christian chez nos et Georges [...]”
- “otages Christian * ” +Georges
Experimental results

<table>
<thead>
<tr>
<th>n/m</th>
<th>2/0</th>
<th>2/1</th>
<th>2/2</th>
<th>3/0</th>
<th>3/1</th>
<th>3/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>13.9 %</td>
<td>24.0 %</td>
<td>26.0 %</td>
<td>18.1 %</td>
<td>19.1 %</td>
<td>15.0 %</td>
</tr>
<tr>
<td>Set size</td>
<td>145</td>
<td>268</td>
<td>789</td>
<td>49</td>
<td>16</td>
<td>15</td>
</tr>
</tbody>
</table>

Tab.: Recall and sets size of the semantics-driven n-gram strategy for OOV word retrieval using Google depending on the n-gram size $n$ and the number of drive-words $m$. 
Conclusion

Strong potential of the Web

- The web contains OOV words
- We can retrieve them

Local context brings information