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

1. Introduction
2. Our approach
3. OOV words retrieval
4. Conclusion

Introduction

Automatic speech recognition

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

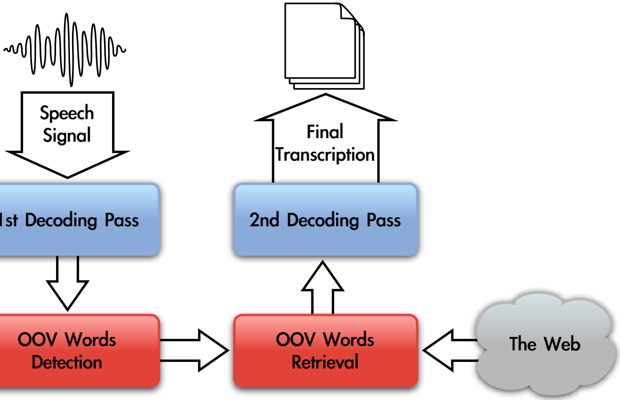
Problem

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



Experimental framework

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**

n-gram	**1**	**2**	**3**	**4**	**5**
Recall	100.00 %	88.22 %	50.54 %	27.29 %	16.12 %

TAB.: *n*-grams containing OOV words on Google depending on the size *n*.

N-gram Strategy

The gaol

- Retrieve words which occurs 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

n-gram	2	3	4	5
Recall	13.9 %	18.1 %	16.4 %	13.8 %
Set size	145	49	13	4

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

Pattern Strategy

The gaol

- Retrieve words which occurs 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

n-gram	**2**	**3**	**4**	**5**
Recall	20.0 %	20.3 %	17.5 %	12.2 %
Set size	411	139	34	15

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

Semantics Driven N-gram Strategy

The gaol

- 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

n/m	2/0	2/1	2/2	3/0	3/1	3/2
Recall	13.9 %	24.0 %	26.0 %	18.1 %	19.1 %	15.0 %
Set size	145	268	789	49	16	15

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