A case study on using speech-to-translation alignments for language documentation

Antonis Anastasopoulos, David Chiang
Goal

• Collect data now; analyze later

• The data must be:
  • Sufficient
  • Interpretable
How much data?

- Qur'an
- New Testament
- Hebrew Bible
- All classical Hebrew
- All classical Latin
- All classical Greek

Millions of words / Hundreds of hours
The linguistic graveyard
Making an audio Rosetta Stone

- Aikuma: Android/web-based app
- Push-to-talk, push-to-translate
Interpreting the audio
Rosetta Stone
Interpreting the audio
Rosetta Stone

Now their farm will not stay behind forever.
Interpreting the audio
Rosetta Stone

Now their farm will not stay behind forever.

Gila abur-u-n ferma hamišaluğ gügüna amuq’-da-č
Interpreting the audio
Rosetta Stone

now their farm will not stay behind forever
Background
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
Aligning Speech to Translation
Aligning Speech to Translation

[tanta plata]
Aligning Speech to Translation

[tanta plata]  [plata]
Aligning Speech to Translation

[tanta plata]  [plata]  [playa]
Aligning Speech to Translation

[tanta plata]  [plata]

[Mexico]  [playa]
Aligning Speech to Translation

[tanta plata]

[plata]

[Mexico]

[playa]
Aligning Speech to Translation

[tanta plata] [plata]

Mexico

[Mexico] [playa]
Example output

Valeria legge il giornale

[Valeria reads the newspaper]
Valeria legge il giornale

[Valeria reads the newspaper]
Example output

é Valeria meletá o’ giornále

Valeria legge il giornale

[Valeria reads the newspaper]
User Study
Griko
User Interface

Translation: non deve mangiare la sera

Transcription:
Results

error rates averaged across 6 Italian-speaking participants

---

- **None**: 27
- **Auto**: 25.7
- **Gold**: 24.5

Phone error rate (lower is better)
Results

error rates averaged across all participants
(6 Italian, 3 Spanish, 3 English)
## Consensus transcription

<table>
<thead>
<tr>
<th>user</th>
<th>transcription</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>it1</td>
<td>o ladro i so ndze mia butto</td>
<td>5</td>
</tr>
<tr>
<td>it2</td>
<td>o ladro isodZenti dabol tu</td>
<td>6</td>
</tr>
<tr>
<td>it3</td>
<td>o ladro isodzeem biabiddu</td>
<td>5</td>
</tr>
<tr>
<td>correct</td>
<td>o ladro isodZe embi apo ttu</td>
<td></td>
</tr>
</tbody>
</table>

Can we do better?

**Combine them!**
String averaging
String averaging

oladroisondzemiambuttuo

oladroisodzentidaboltu

oladroisodzeembiabiddu
String averaging

\[ o l a d r o i s o n d z e m i a b u t t u \]

\[ o l a d r o i s o d Z e n t i d a b o l t u \]

\[ o l a d r o i s o d z e e m b i a b i d d u \]
String averaging

oladroisondzemialbuttu

oladroisodZentidaboltu

oladroisodzeembia biddu
String averaging

oladrosondzemiabuttu

oladrosodZentidaboltu

oladrosodzeembiabiddu
String averaging

```
0 1 1 ...
1

oladroisondzemibuttu

oladroisodZentidaboltu

oladroisodzeembiabiddu

```

...
String averaging

Average:
String averaging

\[
\begin{array}{cccccc}
| 0 | 1 | 1 | \ldots | 1 | 0 |
\end{array}
\]

\[
\text{oladroisondzemiambuttu}
\]

\[
\text{oladroisodZentidaboltu}
\]

\[
\text{oladroisodzembiabiddu}
\]

Average: \[
\text{oladroisodzembiabuttu}
\]
String averaging

Average: $\text{oladrosondzemiabuttu}$
Results

consensus transcription

phone error rate (lower is better)

subset for averaging

Eng  Spa+Eng  Spa  Ita+Eng  all  Ita  Ita+Spa

32.1  27.7  26.9  24.7  24  23.9  23.2
Conclusions
Conclusions

• Alignment of speech with its translation is possible even in small corpora
Conclusions

• Alignment of speech with its translation is possible even in small corpora
• And *possibly* helpful for manual transcription (crowdsourced?)
Conclusions

- Alignment of speech with its translation is possible even in small corpora
- And *possibly* helpful for manual transcription (crowdsourced?)
- What are the needs for interface design?
Conclusions

- Alignment of speech with its translation is possible even in small corpora
- And *possibly* helpful for manual transcription (crowdsourced?)
- What are the needs for interface design?
- Next: large scale case-study with collected data
Conclusions

• Alignment of speech with its translation is possible even in small corpora
• And *possibly* helpful for manual transcription (crowdsourced?)
• What are the needs for interface design?
• Next: large scale case-study with collected data

Talk to me if you want to share comments/ideas/data!