A comparison of linguistically and statistically enhanced models for speech-to-speech machine translation

Alicia Pérez\textsuperscript{1} Víctor Guijarrubia\textsuperscript{1} Raquel Justo\textsuperscript{1} 
M. Inés Torres\textsuperscript{1} Francisco Casacuberta\textsuperscript{2}

\textsuperscript{1}University of the Basque Country
ines.torres@ehu.es

\textsuperscript{2}Technical University of Valencia
fcn@iti.upv.es

IWSLT (Trento, 2007)
Outline

1. Source words driven finite-state transducers
2. Category-based finite-state transducers
   - Architecture
   - Categorization techniques
3. Phrase-based finite-state transducers
   - Architecture
   - Segmentation techniques
4. Experiments
   - Task and corpus
   - Evaluation and confidence
5. Concluding remarks and further work
1. Source words driven finite-state transducers

2. Category-based finite-state transducers
   - Architecture
   - Categorization techniques

3. Phrase-based finite-state transducers
   - Architecture
   - Segmentation techniques

4. Experiments
   - Task and corpus
   - Evaluation and confidence

5. Concluding remarks and further work
Source words driven finite-state transducers

Statistical speech translation:

Notation:
\( x \): speech signal in the source language
\( t \): a string in the *target* language
\( s \): a string in the *source* language

\[ P(t, s) \sim P_{T_0}(t, s) \text{ being } T_0 \text{ WB-SFST} \]

\[ \hat{t} = \arg \max_t P(t|x) \]
\[ = \arg \max_t \sum_s P(t, s|x) \]
\[ = \arg \max_t \sum_s P(t, s)P(x|t, s) \]
\[ \approx \arg \max_t \sum_s P(t, s)P(x|s) \]
Source words driven finite-state transducers

Statistical speech translation:

Notation:

- $x$: speech signal in the source language
- $t$: a string in the target language
- $s$: a string in the source language

\[ P(t, s) \sim P_{\mathcal{T}_0}(t, s) \text{ being } \mathcal{T}_0 \text{ WB-SFST} \]

\[ \hat{t} = \arg \max_t P(t|x) \]
\[ = \arg \max_t \sum_s P(t, s|x) \]
\[ = \arg \max_t \sum_s P(t, s)P(x|t, s) \]
\[ \simeq \arg \max_t \sum_s P(t, s)P(x|s) \]
Source words driven finite-state transducers

Statistical speech translation:

Notation:

\( x \): speech signal in the source language

\( t \): a string in the target language

\( s \): a string in the source language

\( P(t, s) \sim P_{\mathcal{T}_0}(t, s) \) being \( \mathcal{T}_0 \) WB-SFST

\[
\hat{t} = \arg \max_t P(t|x)
\]

\[
= \arg \max_t \sum_s P(t, s|x)
\]

\[
= \arg \max_t \sum_s P(t, s)P(x|t, s)
\]

\[
\sim \arg \max_t \sum_s P(t, s)P(x|s)
\]
Source words driven finite-state transducers

Statistical speech translation:

Notation:
- \( x \): speech signal in the source language
- \( t \): a string in the target language
- \( s \): a string in the source language

\( P(t, s) \approx P_{\mathcal{T}_0}(t, s) \) being \( \mathcal{T}_0 \) WB-SFST

\[
\hat{t} = \arg \max_t P(t|x)
\]

\[
= \arg \max_t \sum_s P(t, s|x)
\]

\[
= \arg \max_t \sum_s P(t, s)P(x|t, s)
\]

\[
\approx \arg \max_t \sum_s P(t, s)P(x|s)
\]
Source words driven finite-state transducers

Integrated architecture for speech translation

\[
\text{arg max}_{t} \max_{s} P(t, s) P(x|s)
\]

- **bilingual corpus**
- **acoustic database**
- **grammatical inference**
- **training**

\[
P(\tilde{t}_i, s_i) \leftrightarrow x_i
\]

- **translation model**
- **lexical model**
- **acoustic model**

**acoustic representation of speech signal**

**translated and recognised sentence**

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta

Linguistically and Statistically Enhanced Models
Source words driven finite-state transducers

Grammar Inference and Alignments for Transducers Inference

- **Training corpus:**
  
  \[ s_1 s_2 s_3 \leftrightarrow t_1 t_2 t_3 \quad \text{and} \quad s_1 s_2 s_4 \leftrightarrow t_1 t_2 t_4 \]

- **Alignments:**

  \[
  \begin{align*}
  &s_1 \\
  &\downarrow \quad \downarrow \\
  &t_1 \quad t_2 \\
  &s_2 \\
  &\downarrow \quad \downarrow \\
  &t_3 \quad t_4 \\
  &s_3
  \end{align*}
  \]

- **Monotonic segmentation:**

  \[
  (s_1, t_1)(s_2, \lambda)(s_3, t_2 t_3) \quad \text{and} \quad (s_1, t_1)(s_2, \lambda)(s_4, t_2 t_4)
  \]

- **Infer a regular grammar:**

  \[
  \begin{align*}
  &q_0 \quad s_1 | t_1 \\
  &\downarrow \quad 1 \\
  &q_1 \\
  &s_2 | \lambda \\
  &\downarrow \quad 1 \\
  &q_2 \\
  &s_3 | t_2 t_3 \\
  &\downarrow \quad 0.5 \\
  &q_3 \quad s_4 | t_2 t_4 \\
  &\downarrow \quad 0.5
  \end{align*}
  \]
Category-based finite-state transducers

Architecture (CB model): decoupled categorization

\[
\hat{c} = \arg \max_c P_{\tau_1}(s, c)
\]
\[
\hat{t} = \arg \max_t P_{\tau_2}(c, t)
\]

Notation:
- \(s\): a string in the source language
- \(t\): a string in the target language
- \(c\): categorized string
Category-based finite-state transducers

Categorization techniques:

- Linguistically motivated categories: gather all the words sharing the same *lemma* within an equivalence-class. 1,135 running words $\leftrightarrow$ 561 classes.

- Statistically motivated categories: automatically obtained by means of *mkcls*. For comparison purposes 561 classes were selected.

Example

class-1: orduetara, \ldots, orduetarako, ordutan
class-2: arinduko, bihurtuko, \ldots, pasatuko
class-3: goradakada, igoera.
Categorization techniques:

- **Linguistically motivated categories**: gather all the words sharing the same *lemma* within an equivalence-class. 1,135 running words $\leftrightarrow$ 561 classes.

- **Statistically motivated categories**: automatically obtained by means of *mkcls*. For comparison purposes 561 classes were selected.

**Example**

class-1: orduetara, . . . , orduetarako, ordutan
class-2: arinduko, bihurtuko, . . . , pasatuko
class-3: goradakada, igoera.
Category-based finite-state transducers

Categorization techniques:

- **Linguistically motivated categories**: gather all the words sharing the same *lemma* within an equivalence-class. 1,135 running words $\leftrightarrow$ 561 classes.

- **Statistically motivated categories**: automatically obtained by means of *mkcls*. For comparison purposes 561 classes were selected.

Example

- class-1: orduetara, ..., orduetarako, ordutan
- class-2: arinduko, bihurtuko, ..., pasatuko
- class-3: goradakada, igoera.
Outline

1. Source words driven finite-state transducers
2. Category-based finite-state transducers
   - Architecture
   - Categorization techniques
3. Phrase-based finite-state transducers
   - Architecture
   - Segmentation techniques
4. Experiments
   - Task and corpus
   - Evaluation and confidence
5. Concluding remarks and further work
Phrase-based finite-state transducers

Architecture (PB model):

- Given a *segmented* corpus infer the SFST.

- At decoding time phrases are expanded into words.

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta
Linguistically and Statistically Enhanced Models
Phrase-based finite-state transducers

Segmentation techniques:

- **Linguistically motivated segments**: syntactic parsing joins words sharing the same syntactic function.

- **Statistically motivated segments**: the most frequent n-grams in the training set.

<table>
<thead>
<tr>
<th></th>
<th>size of the vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>source</td>
</tr>
<tr>
<td>Running words</td>
<td>702</td>
</tr>
<tr>
<td>Linguistic phrases</td>
<td>2,427</td>
</tr>
<tr>
<td>Statistical segments</td>
<td>1,085</td>
</tr>
</tbody>
</table>
Outline

1. Source words driven finite-state transducers
2. Category-based finite-state transducers
   - Architecture
   - Categorization techniques
3. Phrase-based finite-state transducers
   - Architecture
   - Segmentation techniques
4. Experiments
   - Task and corpus
   - Evaluation and confidence
5. Concluding remarks and further work

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta  Linguistically and Statistically Enhanced Models
## Experiments

### Task and corpus:

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test-1</th>
<th>Test-2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pair of sentences</strong></td>
<td>14,615</td>
<td>1,500</td>
<td>1,800</td>
</tr>
<tr>
<td><strong>Different pairs</strong></td>
<td>8,445</td>
<td>1,173</td>
<td>500</td>
</tr>
<tr>
<td><strong>Running words</strong></td>
<td>191,156</td>
<td>12.6</td>
<td>17.4</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>702</td>
<td>12.4</td>
<td>16.5</td>
</tr>
<tr>
<td><strong>Singletons</strong></td>
<td>162</td>
<td>3.6</td>
<td>4.8</td>
</tr>
<tr>
<td><strong>Average length</strong></td>
<td>13.1</td>
<td>4.3</td>
<td>6.7</td>
</tr>
<tr>
<td><strong>Perplexity (3-grams)</strong></td>
<td>3.6</td>
<td>4.3</td>
<td>6.7</td>
</tr>
</tbody>
</table>

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta

Linguistically and Statistically Enhanced Models
Experiments

Evaluation and confidence:
Mean value ($\mu$) and 95% confidence interval ($2\sigma$) of BLEU, NIST, WER and PER scores over 1,000 bootstrap test sets\(^1\).

\(^1\)Given the test-set $D$, consisting of $N$ sentences, a **bootstrap test set** $D^*$, is a set created by randomly selecting with replacement $N$ sentences from $D$. 
### Experiments

<table>
<thead>
<tr>
<th></th>
<th>Word Category</th>
<th>Phrase Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>µ</td>
</tr>
<tr>
<td></td>
<td>ling</td>
<td>stat</td>
</tr>
<tr>
<td></td>
<td>µ</td>
<td>2σ</td>
</tr>
<tr>
<td>BLEU</td>
<td>57.9</td>
<td>1.7</td>
</tr>
<tr>
<td>NIST</td>
<td>7.4</td>
<td>0.1</td>
</tr>
<tr>
<td>WER</td>
<td>32.8</td>
<td>1.5</td>
</tr>
<tr>
<td>PER</td>
<td>27.7</td>
<td>1.3</td>
</tr>
<tr>
<td>BLEU</td>
<td>41.1</td>
<td>1.3</td>
</tr>
<tr>
<td>NIST</td>
<td>6.0</td>
<td>0.1</td>
</tr>
<tr>
<td>WER</td>
<td>47.5</td>
<td>1.2</td>
</tr>
<tr>
<td>PER</td>
<td>39.4</td>
<td>1.1</td>
</tr>
<tr>
<td>BLEU</td>
<td>38.5</td>
<td>1.2</td>
</tr>
<tr>
<td>NIST</td>
<td>5.7</td>
<td>0.1</td>
</tr>
<tr>
<td>WER</td>
<td>51.3</td>
<td>1.3</td>
</tr>
<tr>
<td>PER</td>
<td>42.5</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Is there any significant difference in performance?
Is there any significant difference in performance?
Experiments

Comparing systems: instead of absolute scores, measure the discrepancy ($\Delta \text{Score}_{(\text{sys}_1, \text{sys}_2)}$) over a big number ($B$) of bootstrap test sets, and hence, the relative number of times that one system outperforms the other.

**Probability of Improvement**

\[
\text{poi}(\Delta \text{Score}_{(\text{sys}_1, \text{sys}_2)}) = \lim_{B \to \infty} \left[ \frac{1}{B} \sum_{i=1}^{B} \Theta(\text{Score}_{\text{sys}_1}^{(i)} - \text{Score}_{\text{sys}_2}^{(i)}) \right]
\]

If $\text{Score}$ is an accuracy value

Then $\Theta(x) = H(x)$

Else $\Theta(x) = H(-x)$

where $H(x) = \begin{cases} 
0 & x \leq 0 \\
1 & x > 0 
\end{cases}$
Experiments

Test-2 speech translation results:

<table>
<thead>
<tr>
<th>poi</th>
<th>sys₁</th>
<th>sys₂</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>BLEU</td>
</tr>
<tr>
<td><strong>CB-stat</strong></td>
<td><strong>WB</strong></td>
<td></td>
<td>0.812</td>
</tr>
<tr>
<td><strong>CB-ling</strong></td>
<td><strong>WB</strong></td>
<td></td>
<td>0.844</td>
</tr>
<tr>
<td><strong>CB-ling</strong></td>
<td><strong>CB-stat</strong></td>
<td></td>
<td>0.667</td>
</tr>
<tr>
<td><strong>PB-stat</strong></td>
<td><strong>WB</strong></td>
<td></td>
<td>0.996</td>
</tr>
<tr>
<td><strong>PB-ling</strong></td>
<td><strong>WB</strong></td>
<td></td>
<td>0.999</td>
</tr>
<tr>
<td><strong>PB-ling</strong></td>
<td><strong>PB-stat</strong></td>
<td></td>
<td>0.612</td>
</tr>
</tbody>
</table>

- There is not a complete agreement between all the automatic evaluation scores.
- Linguistic approaches outperform statistical ones in this particular case.
Experiments

Test-2 speech translation results:

<table>
<thead>
<tr>
<th>poi</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
</tr>
<tr>
<td>sys₁</td>
<td></td>
</tr>
<tr>
<td>CB-stat</td>
<td>0.812*</td>
</tr>
<tr>
<td>CB-ling</td>
<td>0.844</td>
</tr>
<tr>
<td>CB-ling</td>
<td>0.667*</td>
</tr>
<tr>
<td>sys₂</td>
<td></td>
</tr>
<tr>
<td>CB-stat</td>
<td></td>
</tr>
<tr>
<td>CB-ling</td>
<td>0.667*</td>
</tr>
<tr>
<td>CB-ling</td>
<td></td>
</tr>
<tr>
<td>CB-ling</td>
<td></td>
</tr>
<tr>
<td>CB-ling</td>
<td></td>
</tr>
<tr>
<td>CB-ling</td>
<td></td>
</tr>
<tr>
<td>PB-stat</td>
<td>0.996*</td>
</tr>
<tr>
<td>PB-ling</td>
<td>0.999</td>
</tr>
<tr>
<td>PB-ling</td>
<td>0.612*</td>
</tr>
<tr>
<td>PB-stat</td>
<td></td>
</tr>
<tr>
<td>PB-ling</td>
<td></td>
</tr>
<tr>
<td>PB-ling</td>
<td></td>
</tr>
<tr>
<td>PB-ling</td>
<td></td>
</tr>
</tbody>
</table>

- There is not a complete agreement between all the automatic evaluation scores.
- Linguistic approaches outperform statistical ones in this particular case.
Experiments

Test-2 speech translation results:

<table>
<thead>
<tr>
<th>poi</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
</tr>
<tr>
<td>CB-stat</td>
<td>0.812</td>
</tr>
<tr>
<td>CB-ling</td>
<td>0.844</td>
</tr>
<tr>
<td>CB-ling</td>
<td>0.667</td>
</tr>
<tr>
<td>PB-stat</td>
<td>0.996</td>
</tr>
<tr>
<td>PB-ling</td>
<td>0.999</td>
</tr>
<tr>
<td>PB-ling</td>
<td>0.612</td>
</tr>
</tbody>
</table>

- There is not a complete agreement between all the automatic evaluation scores.
- Linguistic approaches outperform statistical ones in this particular case.
Outline

1. Source words driven finite-state transducers
2. Category-based finite-state transducers
   - Architecture
   - Categorization techniques
3. Phrase-based finite-state transducers
   - Architecture
   - Segmentation techniques
4. Experiments
   - Task and corpus
   - Evaluation and confidence
5. Concluding remarks and further work
Concluding remarks and further work

Concluding remarks

- GIATI approach has been explored with categories and phrases.
- With respect to the baseline, the improvements are slight with CB approach and significant with PB approach.
- Linguistic approaches outperform statistical ones.

Further work

- Explore these methods on wider tasks.
- Explore other kind of categorization techniques, such as interpolation or on-the-fly categorization.
Grazie mille!

Thank you!
Outline

6 Alternative category-based finite-state transducers

7 Confidence

8 Speech translation with SFSTs

9 Stochastic Finite-State Transducers
Category-based finite-state transducers

Architecture:
Decoupled categorization (CB model):

\[
\arg \max_c P_{\tau_1}(s, c) \rightarrow \hat{c} \rightarrow \arg \max_t P_{\tau_2}(c, t) \rightarrow \hat{t}
\]

Interpolation between category and word-based models:

\[
\arg \max_c P_{\tau_1}(s, c) \rightarrow \hat{c} \rightarrow \arg \max_t \left[ \lambda P_{\tau_2}(c, t) + (1 - \lambda) P_{\tau_0}(s, t) \right] \rightarrow \hat{t}
\]
Outline

6 Alternative category-based finite-state transducers

7 Confidence

8 Speech translation with SFSTs

9 Stochastic Finite-State Transducers
Experiments

- Word-based SFST
- Category-based SFST with statistical categories
- Phrase-based SFST with statistical segmentation

NO overlapping between the 95% confidence intervals ⇒ the performance of the systems differ significantly (95% certainty)
Speech recognition

\[ \text{arg max}_s P(s)P(x|s) \]

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta

Linguistically and Statistically Enhanced Models
Outline

6 Alternative category-based finite-state transducers

7 Confidence

8 Speech translation with SFSTs

9 Stochastic Finite-State Transducers
Decoupled architecture for speech translation

A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta

Linguistically and Statistically Enhanced Models
6 Alternative category-based finite-state transducers

7 Confidence

8 Speech translation with SFSTs

9 Stochastic Finite-State Transducers
### Definition

An **SFST** is a tuple $\mathcal{T} = \langle \Sigma, \Delta, Q, q_0, R, F, P \rangle$ where:

- $\Sigma$ is a finite set of input symbols (source words);
- $\Delta$ is a finite set of output symbols (target words);
- $Q$ is a finite set of states;
- $q_0 \in Q$ is the initial state;
- $R \subseteq Q \times \Sigma \times \Delta^* \times Q$ a set of transitions.
- $P : R \rightarrow [0, 1]$ transition probability;
- $F : Q \rightarrow [0, 1]$ final state probability;

The probability distributions satisfy the stochastic constraint:

$$\forall q \in Q \quad F(q) + \sum_{\forall s, \tilde{t}, q'} P(q, s, \tilde{t}, q') = 1$$