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

Alicia Pérez[1] Víctor Guijarrubia[1] Raquel Justo[1]
M. Inés Torres[1] Francisco Casacuberta[2]

[1]University of the Basque Country
ines.torres@ehu.es

[2]Technical University of Valencia
fcn@iti.upv.es

IWSLT (Trento, 2007)

Outline

Outline

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

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

$$\begin{aligned}
\widehat{\mathbf{t}} &= \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}) P(\mathbf{x}|\mathbf{t}, \mathbf{s}) \\
&\simeq \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}) P(\mathbf{x}|\mathbf{s})
\end{aligned}$$

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

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

$$
\begin{aligned}
\widehat{\mathbf{t}} &= \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{t},\mathbf{s}) \\
&\simeq \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s})
\end{aligned}
$$

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

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

$$
\begin{aligned}
\widehat{\mathbf{t}} &= \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{t},\mathbf{s}) \\
&\simeq \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s})
\end{aligned}
$$

Source words driven finite-state transducers

Statistical speech translation:

Notation:

x: speech signal in the source language

t: a string in the *target* language

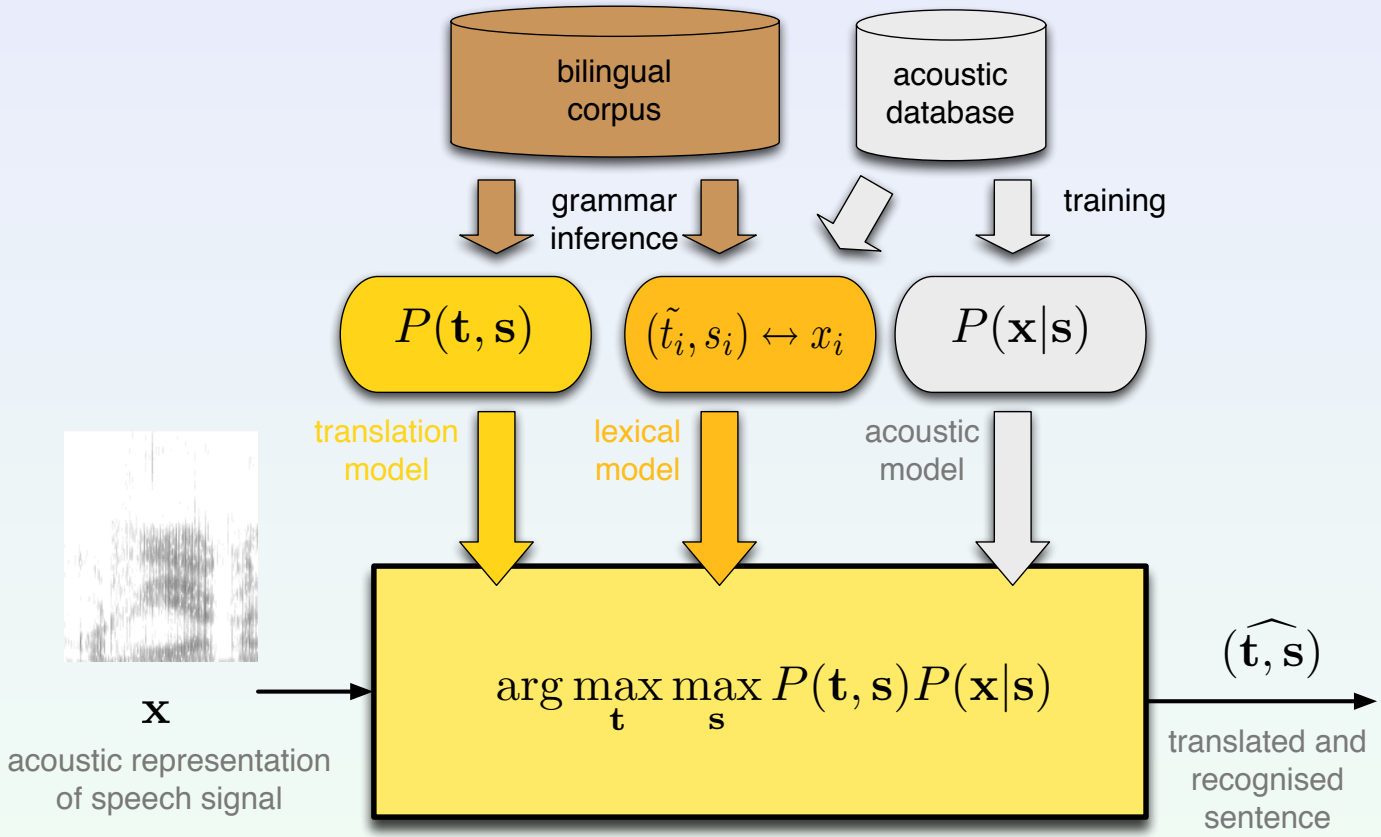
s: a string in the *source* language

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

$$
\begin{aligned}
\widehat{\mathbf{t}} &= \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}|\mathbf{x}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{t},\mathbf{s}) \\
&\simeq \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s})
\end{aligned}
$$

Source words driven finite-state transducers

Integrated architecture for speech translation



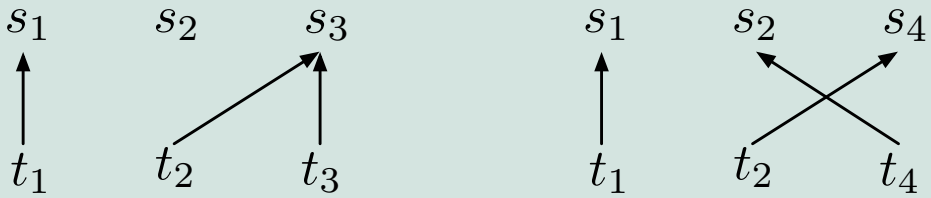
Source words driven finite-state transducers

Grammar Inference and Alignments for Transducers Inference

- Training corpus:

$s_1 s_2 s_3 \leftrightarrow t_1 t_2 t_3$ $\qquad$ $s_1 s_2 s_4 \leftrightarrow t_1 t_2 t_4$

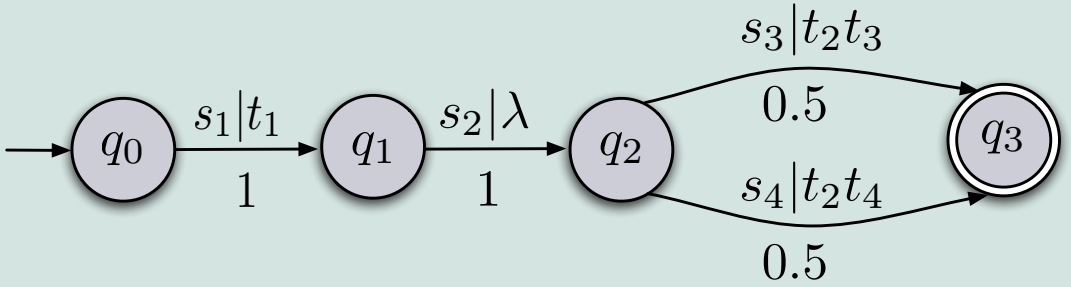
- Alignments:



- Monotonic segmentation:

$(s_1, t_1)(s_2, \lambda)(s_3, t_2 t_3)$ $\qquad$ $(s_1, t_1)(s_2, \lambda)(s_4, t_2 t_4)$

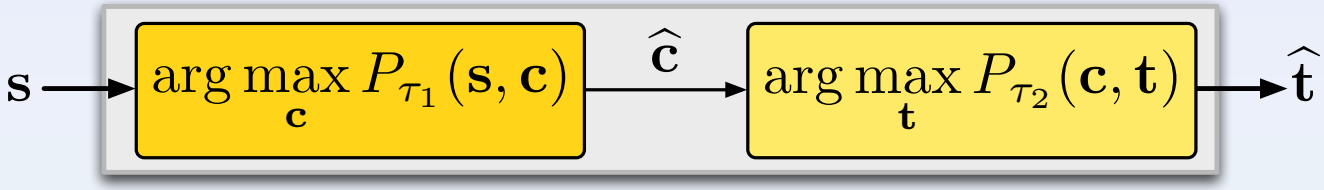
- Infer a regular grammar:



Outline

Category-based finite-state transducers

Architecture (CB model): decoupled categorization



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 $\longleftrightarrow$ 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 $\longleftrightarrow$ 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 $\longleftrightarrow$ 561 classes.

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

Example

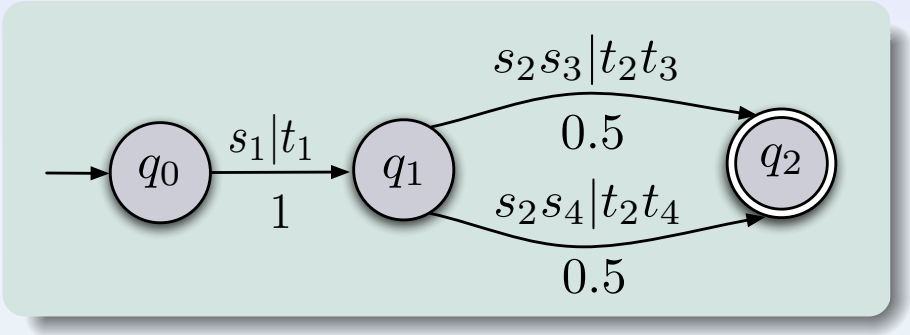
class-1: orduetara, . . . , orduetarako, ordutan
class-2: arinduko, bihurtuko, . . . , pasatuko
class-3: goradakada, igoera.

Outline

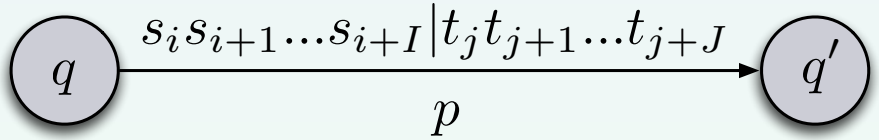
Phrase-based finite-state transducers

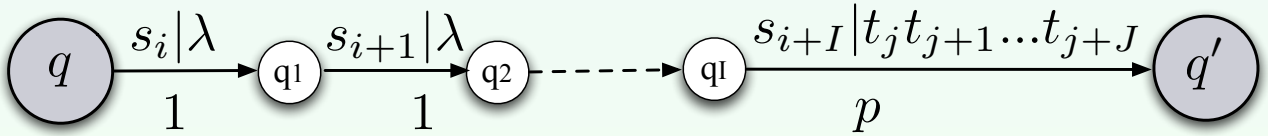
Architecture (PB model):

- Given a *segmented* corpus infer the SFST.



- At decoding time phrases are expanded into words.





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.

	size of the vocabulary		
	source	target	extended
Running words	702	1,135	8,789
Linguistic phrases	2,427	2,519	12,108
Statistical segments	1,085	1,499	13,243

Outline

Experiments

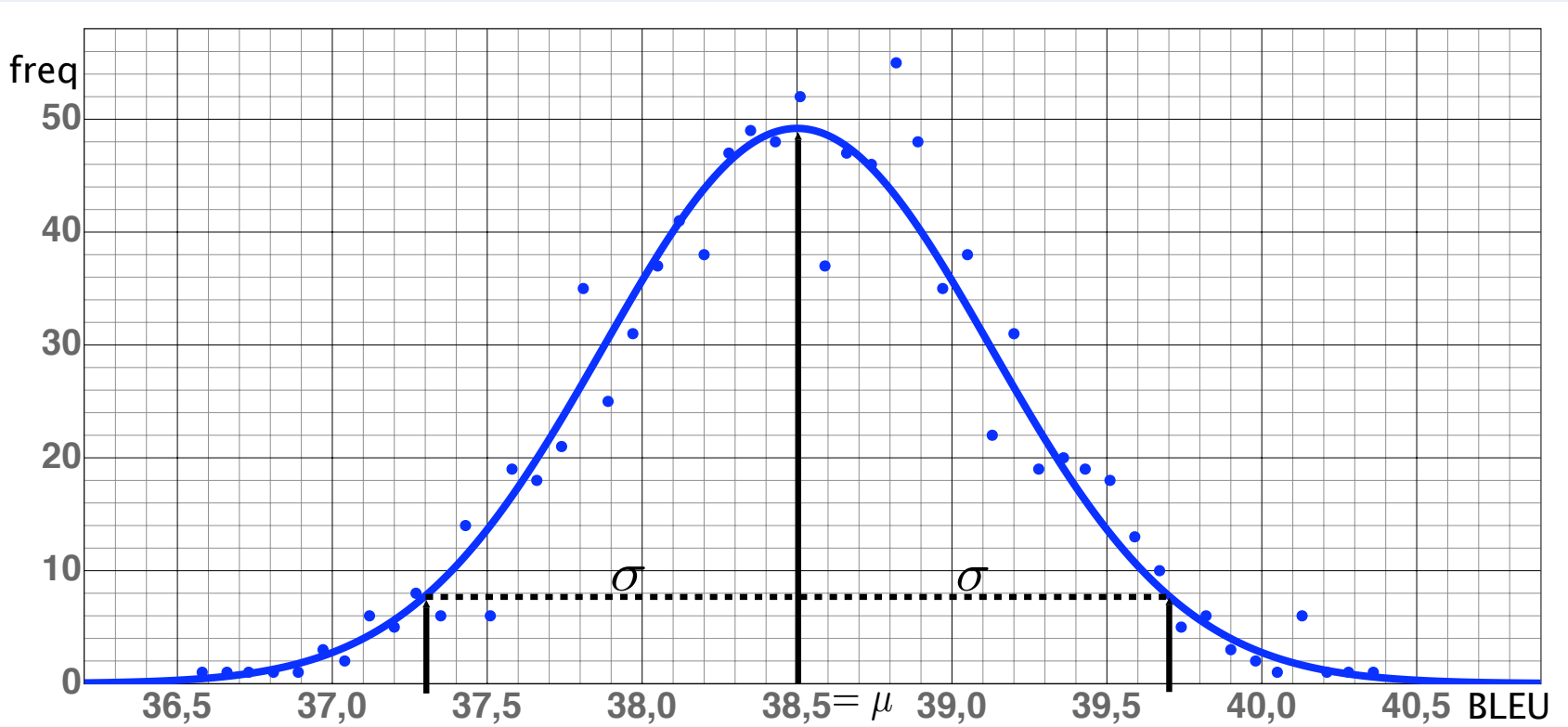
Task and corpus:

		Spanish	Basque
Training	Pair of sentences	14,615	
	Different pairs	8,445	
	Running words	191,156	187,195
	Vocabulary	702	1,135
	Singletons	162	302
	Average length	13.1	12.8
Test-1	Pair of sentences	1,500	
	Different pairs	1,173	
	Average length	12.6	12.4
	Perplexity (3-grams)	3.6	4.3
Test-2	Pair of sentences	1,800	
	Different pairs	500	
	Average length	17.4	16.5
	Perplexity (3-grams)	4.8	6.7

Experiments

Evaluation and confidence:

Mean value (μ) and 95% confidence interval (2σ) 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

		Word		Category				Phrase			
				ling		stat		ling		stat	
		μ	2σ	μ	2σ	μ	2σ	μ	2σ	μ	2σ
Test-1 text	BLEU	57.9	1.7	60.3	1.7	58.9	1.6	66.1	1.8	62.6	1.8
	NIST	7.4	0.1	7.6	0.1	7.5	0.1	8.1	0.1	7.8	0.2
	WER	32.8	1.5	31.4	1.6	32.3	1.7	27.6	1.7	29.9	1.6
	PER	27.7	1.3	26.6	1.3	27.1	1.5	22.3	1.4	24.3	1.3
Test-2 text	BLEU	41.1	1.3	41.6	1.2	42.0	1.2	43.6	1.2	41.4	1.2
	NIST	6.0	0.1	6.0	0.1	6.1	0.2	6.3	0.1	6.0	0.1
	WER	47.5	1.2	48.0	1.2	47.5	1.1	48.0	1.3	51.0	1.4
	PER	39.4	1.1	40.4	1.0	39.4	1.1	38.9	1.1	41.1	1.2
Test-2 speech	BLEU	38.5	1.2	38.9	1.2	38.8	1.2	40.2	1.2	40.0	1.4
	NIST	5.7	0.1	5.8	0.1	5.7	0.1	5.9	0.1	5.9	0.1
	WER	51.3	1.3	50.5	1.3	51.4	1.3	50.5	1.3	53.9	1.4
	PER	42.5	1.10	41.8	1.0	42.4	1.1	41.1	1.1	44.1	1.3

Is there any significant difference in performance?

Experiments

		Word		Category				Phrase			
				ling		stat		ling		stat	
		μ	2σ	μ	2σ	μ	2σ	μ	2σ	μ	2σ
Test-1 text	BLEU	57.9	1.7	60.3	1.7	58.9	1.6	66.1	1.8	62.6	1.8
	NIST	7.4	0.1	7.6	0.1	7.5	0.1	8.1	0.1	7.8	0.2
	WER	32.8	1.5	31.4	1.6	32.3	1.7	27.6	1.7	29.9	1.6
	PER	27.7	1.3	26.6	1.3	27.1	1.5	22.3	1.4	24.3	1.3
Test-2 text	BLEU	41.1	1.3	41.6	1.2	42.0	1.2	43.6	1.2	41.4	1.2
	NIST	6.0	0.1	6.0	0.1	6.1	0.2	6.3	0.1	6.0	0.1
	WER	47.5	1.2	48.0	1.2	47.5	1.1	48.0	1.3	51.0	1.4
	PER	39.4	1.1	40.4	1.0	39.4	1.1	38.9	1.1	41.1	1.2
Test-2 speech	BLEU	38.5	1.2	38.9	1.2	38.8	1.2	40.2	1.2	40.0	1.4
	NIST	5.7	0.1	5.8	0.1	5.7	0.1	5.9	0.1	5.9	0.1
	WER	51.3	1.3	50.5	1.3	51.4	1.3	50.5	1.3	53.9	1.4
	PER	42.5	1.10	41.8	1.0	42.4	1.1	41.1	1.1	44.1	1.3

Is there any significant difference in performance?

Experiments

Comparing systems: instead of absolute scores, measure the discrepancy ($\Delta Score_{(sys_1, sys_2)}$) 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

$$poi(\Delta Score_{(sys_1, sys_2)}) = \lim_{B \to \infty} \left[\frac{1}{B} \sum_{i=1}^{B} \Theta(Score^{(i)}_{sys_1} - Score^{(i)}_{sys_2}) \right]$$

`if` *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:

poi		Score			
sys_1	**sys_2**	BLEU	NIST	WER	PER
CB-stat	WB	0.812*	0.641	0.385*	0.461
CB-ling	WB	0.844	0.955	0.982	0.958
CB-ling	CB-stat	0.667*	0.936	0.995*	0.973
PB-stat	WB	0.996*	0.999	0.000*	0.002
PB-ling	WB	0.999	1.000	0.934	0.997
PB-ling	PB-stat	0.612*	0.527	1.000*	1.000

- 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:

poi		Score			
sys_1	**sys_2**	BLEU	NIST	WER	PER
CB-stat	WB	0.812*	0.641	0.385*	0.461
CB-ling	WB	0.844	0.955	0.982	0.958
CB-ling	CB-stat	0.667⋆	0.936	0.995⋆	0.973
PB-stat	WB	0.996*	0.999	0.000*	0.002
PB-ling	WB	0.999	1.000	0.934	0.997
PB-ling	PB-stat	0.612⋆	0.527	1.000⋆	1.000

- 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:

poi		Score			
sys_1	**sys_2**	BLEU	NIST	WER	PER
CB-stat	WB	0.812*	0.641	0.385*	0.461
CB-ling	WB	0.844	0.955	0.982	0.958
CB-ling	CB-stat	0.667⋆	0.936	0.995⋆	0.973
PB-stat	WB	0.996*	0.999	0.000*	0.002
PB-ling	WB	0.999	1.000	0.934	0.997
PB-ling	PB-stat	0.612⋆	0.527	1.000⋆	1.000

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

Outline

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!

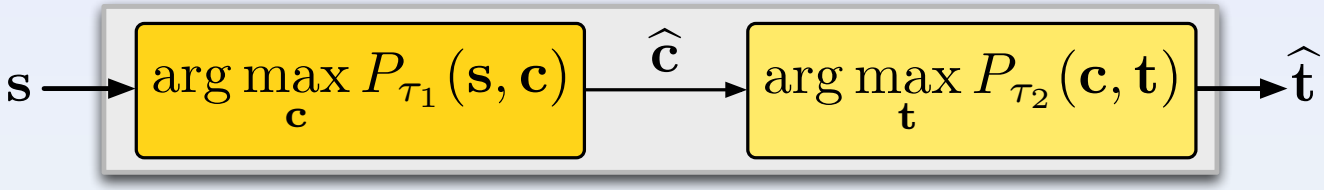
Thank you!

Outline

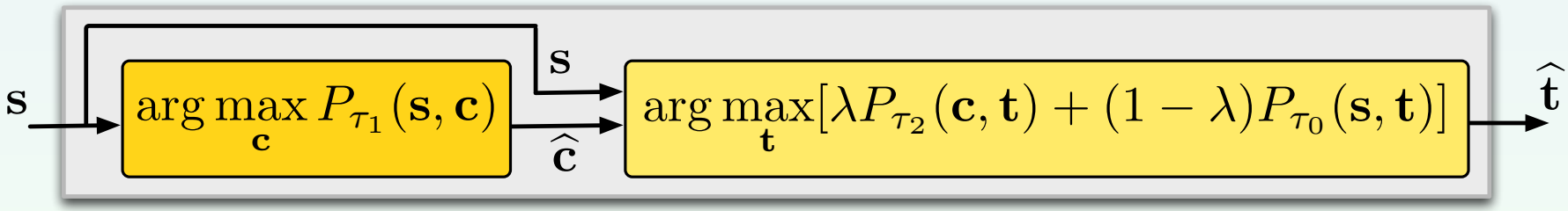
Category-based finite-state transducers

Architecture:

Decoupled categorization (CB model):



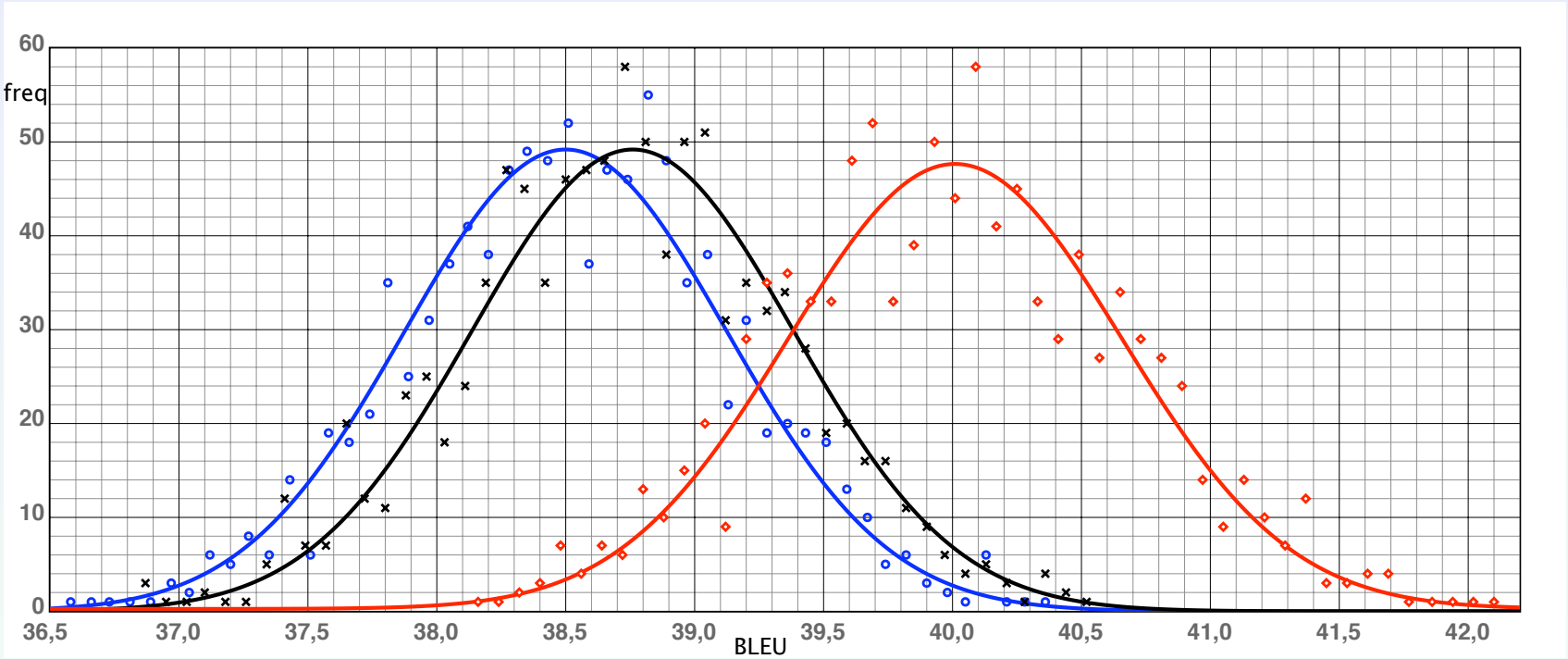
Interpolation between category and word-based models:



Outline

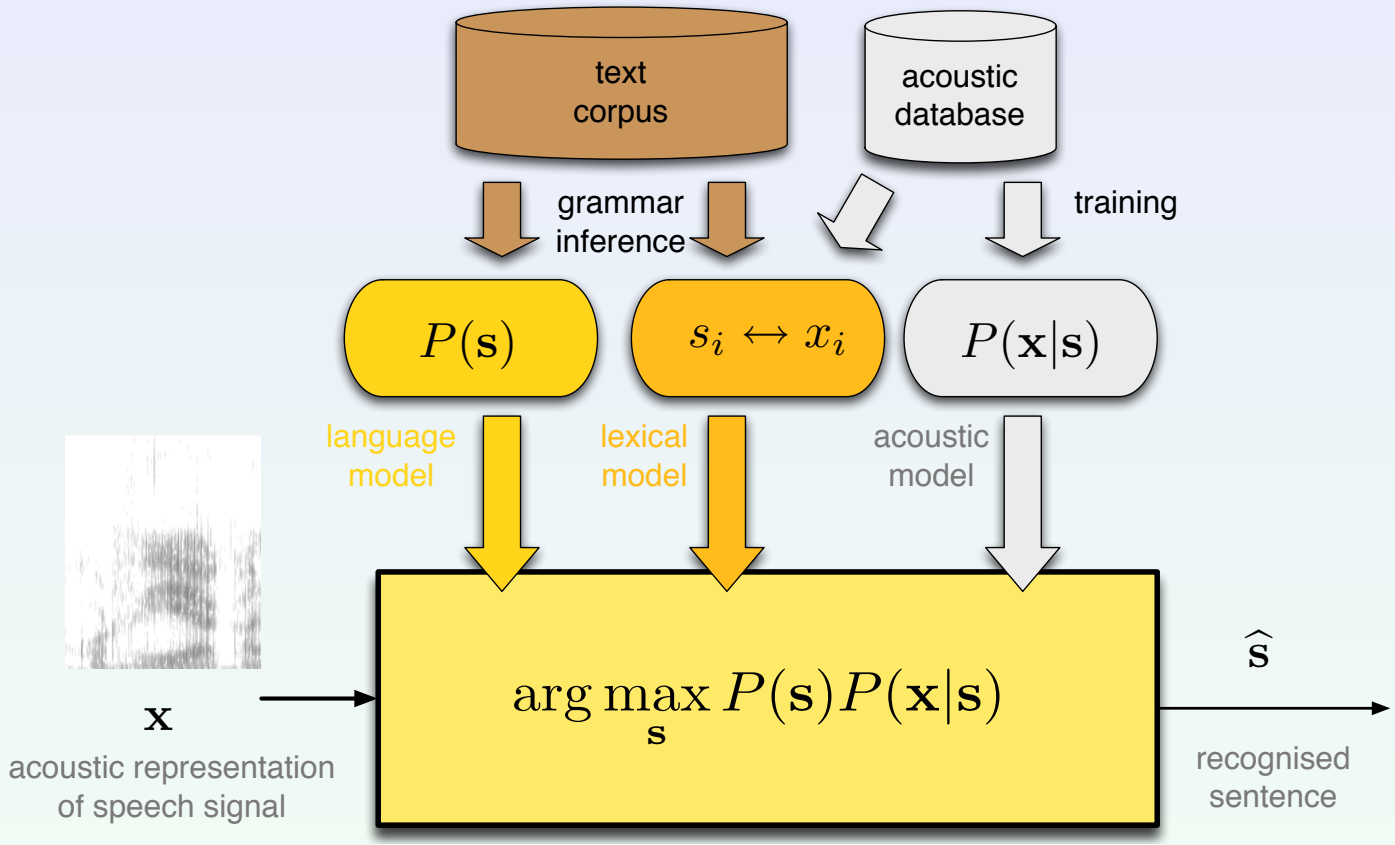
Experiments

- ○ Word-based SFST
- × Category-based SFST with statistical categories
- ◊ Phrase-based SFST with statistical segmentation



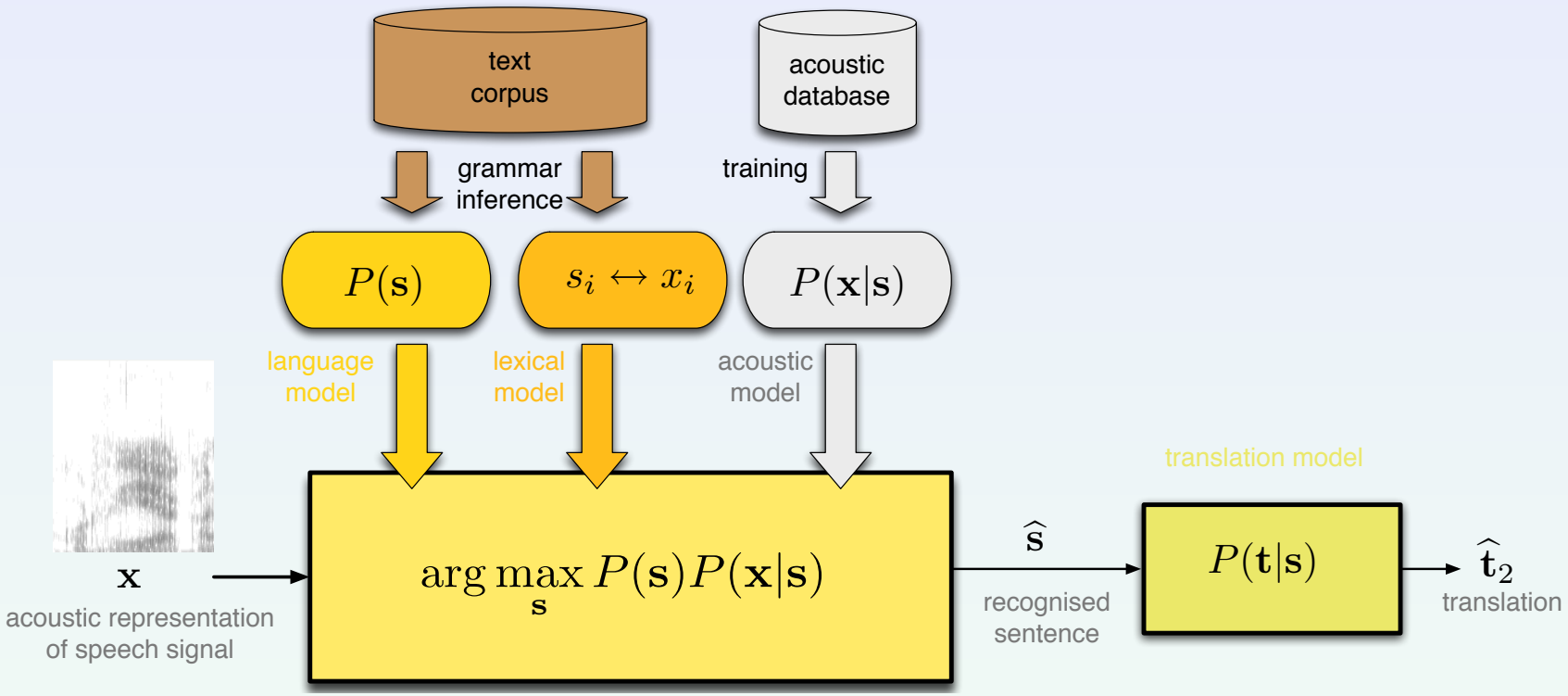
NO overlapping between the 95% confidence intervals $\Rightarrow$ the performance of the systems differ significantly (95% certainty)

Speech recognition



Outline

Decoupled architecture for speech translation



Outline

Stochastic Finite-State Transducers

Definition

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

Σ is a finite set of input symbols (source words);

Δ 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:

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

Bibliography

F. J. Och, “An efficient method for determining bilingual word classes,” in *EACL ’99: Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, Bergen, Norway, Jun. 1999, pp. 71–76.