Universitat Politècnica de València

Master en Inteligencia Artificial, Renoconocimiento de Formas e Imagen Digital

Curs 2007-2008

# MACHINE TRANSLATION

## STATISTICAL MACHINE TRANSLATION

# Statistical Framework to Machine Translation

Francisco Casacuberta

fcn@iti.upv.es

January 28, 2008





## Index







## Index







## Notation and Basic Concepts

- $x$ and $y$ will generally denote *source* and *target* texts, respectively
- CONDITIONAL AND UNCONDITIONAL PROBABILITIES:
  $\Pr(X = x \mid Y = y) \equiv \Pr(x \mid y), \quad \Pr(X = x) \equiv \Pr(x)$
- BAYES' RULE: $\Pr(x \mid y) \cdot \Pr(y) = \Pr(y \mid x) \cdot \Pr(x)$
- JOINT PROBABILITY: $\Pr(x, y) = \Pr(x) \cdot \Pr(y \mid x)$
- $\Pr(x_1, x_2, \ldots, x_I) = \Pr(x_1) \cdot \Pr(x_2 \mid x_1) \cdots \Pr(x_I \mid x_1, \ldots, x_{I-1})$
  $(\Pr(x_1^I) = \Pr(x_1) \cdot \Pr(x_2 \mid x_1) \cdots \Pr(x_I \mid x_1^{I-1}))$
- $\Pr(x) = \sum_y \Pr(x, y)$
- $\max_x \Pr(x) = \Pr(\operatorname{argmax}_x \Pr(x))$
- $\sum_x \Pr(x) \approx \max_x \Pr(x)$

## Index







## General Framework

- Every sentence y in a *target* language is considered as a possible translation of any other sentence x in another *source* language.
- For each possible pair of sentences y, x, there is a probability $\Pr(\mathsf{y} \mid \mathsf{x})$.
- $\Pr(\mathsf{y} \mid \mathsf{x})$ should be *low* for pairs $(\mathsf{y}, \mathsf{x})$ such as:
  ( *una habitación con vistas al mar , are all expenses included in the bill ?* )
- $\Pr(\mathsf{y} \mid \mathsf{x})$ should be *high* for pairs such as:
  ( *¿ hay alguna habitación tranquila libre ? , is there a quiet room available ?* )



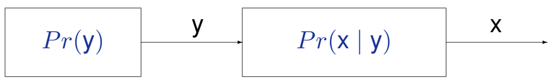


## An inverse approach

Decompose $\Pr(\mathsf{y} \mid \mathsf{x})$ using Bayes' rule:

$$\hat{\mathsf{y}} = \operatorname*{argmax}_{\mathsf{y}} \Pr(\mathsf{y} \mid \mathsf{x}) = \operatorname*{argmax}_{\mathsf{y}} \frac{\Pr(\mathsf{x} \mid \mathsf{y}) \cdot \Pr(\mathsf{y})}{\Pr(\mathsf{x})} = \operatorname*{argmax}_{\mathsf{y}} \Pr(\mathsf{x} \mid \mathsf{y}) \cdot \Pr(\mathsf{y})$$

A "distorted (noisy) channel model"




***Need:*** ***a target-language model + alignment and lexicon models***



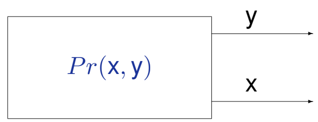


## A finite-state approach

The direct probability can be decomposed in a different way:

$$\hat{\mathsf{y}} = \operatorname*{argmax}_{\mathsf{y}} \Pr(\mathsf{y} \mid \mathsf{x}) = \operatorname*{argmax}_{\mathsf{y}} \frac{\Pr(\mathsf{x}, \mathsf{y})}{\Pr(\mathsf{x})} = \operatorname*{argmax}_{\mathsf{y}} \Pr(\mathsf{x}, \mathsf{y})$$

A "joint" model



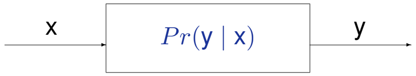

***A stochastic finite-state transducer can model the joint distribution***

## A direct approach

Search for a target sentence with maximum *posterior* probability:

$$\hat{y} = \operatorname*{argmax}_{y} \Pr(y \mid x)$$

A “direct model”




***Log-linear combination of models***





## A log-linear approach

Search for a target sentence with maximum *posterior* probability:

$$\hat{y} = \operatorname*{argmax}_{y} \Pr(y \mid x)$$

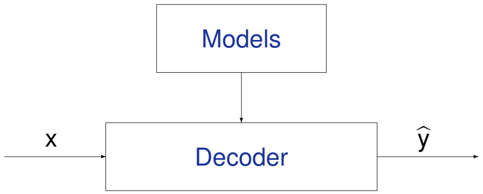
$$\hat{y} = \operatorname*{argmax}_{y} \frac{\exp\left(\sum_{k=1}^{K} \lambda_k h_k(x, y)\right)}{\sum_{y'} \exp\left(\sum_{k=1}^{K} \lambda_k h_k(x, y')\right)} = \operatorname*{argmax}_{y} \sum_{k=1}^{K} \lambda_k h_k(x, y)$$

- $h_1(x, y) = \log Pr(y)$, a language model
- $h_2(x, y) = \log Pr(y \mid x)$, a translation model model
- $h_3(x, y) = \log Pr(x \mid y)$, an inverse translation model model
- ...





## Translation search




$$\hat{y} = \operatorname*{argmax}_{y} Pr(y \mid x)$$

- Direct approach: log-linear combination of models (target-language model+translation models+...)
- Inverse approach: a target-language model + alignment and lexicon models
- Joint approach: stochastic finite-state transducer





## Index

# Speech-input translation

Given an input acoustic sequence v, search for a target sentence with maximum posterior probability:

$$\hat{y} = \operatorname*{argmax}_{y} \Pr(y \mid v)$$

But this can be seen as a *"two-step process"*:

$$v \longrightarrow x \longrightarrow y$$

where the *"hidden variable"* x accounts for all possible input decodings of v:

$$\hat{y} = \operatorname*{argmax}_{y} \sum_{x} \Pr(y, x \mid v) = \operatorname*{argmax}_{y} \sum_{x} \Pr(x, y) \cdot \Pr(v \mid x)$$

(with the assumption: $\Pr(v \mid x, y)$ does not depend on the target sentence y)





# Speech-input translation: An integrated approach

$$\boxed{\operatorname*{argmax}_{y} \Pr(y \mid v) \approx \operatorname*{argmax}_{y} \max_{x} (\Pr(x, y) \cdot \Pr(v \mid x))}$$

- $\Pr(v \mid x) \approx$ ACOUSTIC MODELS
- $\Pr(x, y) \approx$ STOCHASTIC FINITE-STATE TRANSDUCERS





# Speech-input translation: A serial approach

$$\operatorname*{argmax}_{y} \max_{x} (\Pr(x, y) \cdot \Pr(v \mid x)) = \operatorname*{argmax}_{y} \max_{x} (\Pr(x) \cdot \Pr(y \mid x) \cdot \Pr(v \mid x))$$

$$\boxed{\begin{aligned} \hat{x} &\approx \operatorname*{argmax}_{x} \Pr(x) \cdot \Pr(v \mid x) \\ \hat{y} &\approx \operatorname*{argmax}_{y} \Pr(y \mid \hat{x}) \end{aligned}}$$

- $\Pr(v \mid x) \approx$ ACOUSTIC MODELS
- $\Pr(y \mid x) \approx$ ALIGNMENT, LEXICON MODELS AND TARGET LANGUAGE MODELS OR STOCHASTIC FINITE-STATE TRANSDUCERS
- $\Pr(x) \approx$ SOURCE LANGUAGE MODELS





# Speech-input translation: Another serial approach

$$\operatorname*{argmax}_{y} \max_{x} (\Pr(x, y) \cdot \Pr(v \mid x)) = \operatorname*{argmax}_{y} \max_{x} (\Pr(y) \cdot \Pr(x \mid y) \cdot \Pr(v \mid x))$$

$$\boxed{\begin{aligned} \hat{x} &\approx \operatorname*{argmax}_{x} \Pr_{y}(x) \cdot \Pr(v \mid x) \\ \hat{y} &\approx \operatorname*{argmax}_{y} \Pr(y) \cdot \Pr(\hat{x} \mid y) \end{aligned}}$$

- $\Pr(v \mid x) \approx$ ACOUSTIC MODELS
- $\Pr_{y}(x) \approx$ ADAPTATIVE SOURCE LANGUAGE MODELS
- $\Pr(y \mid x) \approx$ ALIGNMENT AND LEXICON MODELS
- $\Pr(y) \approx$ TARGET LANGUAGE MODELS

## Index







## Text prediction for Computer-Assisted Translation (CAT)

Given a source text x and a "correct" *prefix* $y_p$ of the target text, search for a *suffix* $\hat{y}_s$, that maximizes the posterior probability over all possible sufixes:

$$\hat{y}_s = \operatorname*{argmax}_{y_s} \Pr(y_s \mid x, y_p)$$

Taking into account that $\Pr(x, y_p)$ does not depend on $y_s$, we can write:

$$\begin{aligned} \hat{y}_s &= \operatorname*{argmax}_{y_s} \Pr(x, y_p y_s) \\ &= \operatorname*{argmax}_{y_s} \Pr(x \mid y_p y_s) \cdot \Pr(y_p y_s) \end{aligned}$$

Main difference with text-input machine translation: search over the set of suffixes.





## Target language dictation in CAT

A *human* translator *dictates* the translation of a source text, x, producing a *target language* acoustic sequence v.

Given v and x, the system should search for a most likely decoding of v:

$$\hat{y} = \operatorname*{argmax}_{y} \Pr(y \mid x, v)$$

By the assumption that $\Pr(v \mid x, y)$ does not depend on x,

$$\hat{y} = \operatorname*{argmax}_{y} \Pr(v \mid y) \cdot \Pr(x \mid y) \cdot \Pr(y)$$

- $\Pr(v \mid y) \approx$ (TARGET LANGUAGE) ACOUSTIC MODELS
- $\Pr(x \mid y) \approx$ TRANSLATION MODEL
- $\Pr(y) \approx$ TARGET LANGUAGE MODEL

*Similar to plain speech decoding, where:* $\hat{y} = \operatorname{argmax}_{y} \Pr(v \mid y) \cdot \Pr(y)$





## Further use of speech recognition in CAT

Let x be the source text and $y_p$ a "correct" prefix of the target sentence. As in pure text CAT the system suggests an optimal suffix:

$$\hat{y}_s = \operatorname*{argmax}_{y_s} \Pr(y_s \mid x, y_p) \ . \tag{1}$$

The user is now allowed to *utter some words*, v, generally aimed at amending parts of $\hat{y}_s$ and the system has then to obtain a most probable decoding of v:

$$\hat{\mathbf{d}} = \operatorname*{argmax}_{\mathbf{d}} \Pr(\mathbf{d} \mid x, y_p, \hat{y}_s, v) \ . \tag{2}$$

Finally, the user can enter additional amendment keystrokes $k$, to produce a new consolidated prefix, $y_p$, based on the previous $y_p$, $\hat{\mathbf{d}}$, $k$ and parts of $\hat{y}_s$.

## Further use of speech recognition in CAT (cont.)

From Eq. (2):

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\text{argmax}} \Pr(\mathbf{d} \mid \mathsf{x}, \mathsf{y}_p, \hat{\mathsf{y}}_s) \cdot \Pr(\mathsf{v} \mid \mathsf{x}, \mathsf{y}_p, \hat{\mathsf{y}}_s, \mathbf{d})$$

and, by making the assumption that $\Pr(\mathsf{v} \mid \mathsf{x}, \mathsf{y}_p, \hat{\mathsf{y}}_s, \mathbf{d})$ only depends on **d**:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\text{argmax}} \Pr(\mathbf{d} \mid \mathsf{x}, \mathsf{y}_p, \hat{\mathsf{y}}_s) \cdot \Pr(\mathsf{v} \mid \mathbf{d})$$

- $\Pr(\mathsf{v} \mid \mathbf{d}) \approx$ (TARGET LANGUAGE) ACOUSTIC MODELS
- $\Pr(\mathbf{d} \mid \mathsf{x}, \mathsf{y}_p, \hat{\mathsf{y}}_s) \approx$ TARGET LANGUAGE MODEL CONSTRAINED BY THE SOURCE SENTENCE, THE PREFIX AND THE SUFFIX





## Index







## Bibliography

1. P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roosin: *A statistical approach to machine translation*. Computational Linguistics, 16:79–85, 1990.
2. P. Brown, S. Chen, S. Della Pietra, V. Della Pietra, A. Kehler, R. Mercer: *Automatic speech recognition in machine aided translation*. Computer Speech and Language, 8(3): 177–87. 1994.
3. H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, S. Vogel: *Algorithms for statistical translation of spoken language*. IEEE Transactions on Speech and Audio Processing, 8: 24–36, 2000.
4. F. Casacuberta, E. Vidal, J. Vilar: *Architectures for speech-to-speech translation using finite-state 5 models*. ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems, July 11, 2002.
5. F. Casacuberta, E. Vidal, A. Sanchis, and J. Vilar: *Pattern recognition approaches for speech-to-speech translation*. Cybernetic and Systems: an International Journal, 35(1):3-17, 2004.
6. P. Langlais, G. Foster, G. Lapalme; *Unit Completion for a Computer-aided Translation Typing System.* Machine Translation, 15(4): 267–294, 2000.
7. F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, Ch. Tillmann: *Statistical and finite-state approaches to speech-to-speech translation*. Computer Speech and Language, 18:25-47, 2004.
8. E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera and C. Martínez *Computer-assisted translation using speech recognition*. IEEE Transactions on Speech and Audio Processing, 14(3):941-951, 2006.