Speech translation by statistical methods

Traduction automatique de la parole par méthodes automatiques

Daniel Déchelotte

LIMSI-CNRS

December 17, 2007

Introduction

- Speech-to-speech translation: a humanist's dream
- 50 years of progress in Automatic Speech Recognition (ASR) and Machine Translation (MT)
- Speech translation: more recent research topic
- Applications:
  - tourism, media monitoring, parliamentary proceedings, ...

Objectives of this thesis

1. Develop a translation system
2. Focus on translating speech

Translation tasks

- TC-STAR project: translation of the European Parliament Plenary Sessions (EPPS)
- 2006 and 2007 international evaluation campaigns
- English–Spanish, both ways
- Testing material: verbatim and automatic transcriptions
- Training material: proceedings published on the web

Sample Verbatim sentence

I take these allegations very very seriously indeed which are being made in order to undermine my integrity and my reputation .

Sample training sentence

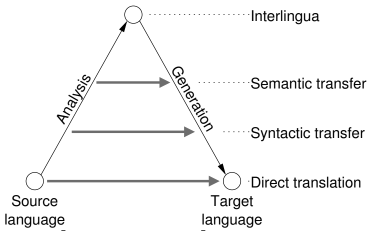
I take these allegations, which are aimed at undermining my integrity and reputation, very seriously indeed.

Outline of the defence

1. Models and algorithms for machine translation
   - Introduction to machine translation
   - A word-based translation system
   - A phrase-based translation system
   - Phrase-table discriminative training
2. Specifics of speech translation
   - Motivation
   - Translation of a stream of words
   - Integration with speech recognition

Approaches to machine translation

- Rule-based approaches
  - Expert and semi-automatic rule acquisition
- Interlingua-based approaches
  - Translation replaced by two monolingual processes
- Data-driven, or corpus-based, approaches
  - Learn from translated examples
  - Example-based MT
  - Statistical MT



[Vauquois, 68]

Statistical machine translation

- Translating from **f** (French) to **e** (English):

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) \qquad \text{[Brown et al., 90]}$$

- Bayes rule:

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{e})$$

- Model weighting:

$$\mathbf{e}^* \approx \operatorname*{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})^{\lambda_1} p(\mathbf{e})^{\lambda_2}$$

- (Log-)linear combination of features:

$$\mathbf{e}^* \approx \operatorname*{argmax}_{\mathbf{e}} \sum_i \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

where, e.g., $h_1(\mathbf{f}, \mathbf{e}) = \log p(\mathbf{f}|\mathbf{e})$, $h_2(\mathbf{f}, \mathbf{e}) = \log p(\mathbf{e})$, etc

BLEU : an automatic evaluation of translation quality

- Evaluating a translation is a problem in itself
- Subjective metrics, objective metrics
- Introducing BLEU...
- Measure similarity with reference translations
- Geometric mean of n-gram precisions

Computing n-gram precisions for BLEU

I am feeling good

Ref1: I am happy

Ref2: I am feeling very good

$p_1 = 1 \quad p_2 = \frac{2}{3} \quad p_3 = \frac{1}{2} \quad p_4 = \frac{0}{1}$

Outline

1. Models and algorithms for machine translation
   - Introduction to machine translation
   - A word-based translation system
   - A phrase-based translation system
   - Phrase-table discriminative training
2. Specifics of speech translation
   - Motivation
   - Translation of a stream of words
   - Integration with speech recognition

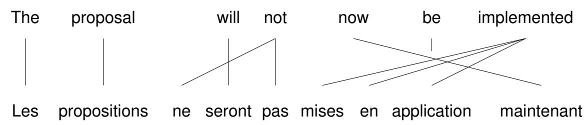
A word-based translation system

- Statistical MT equation:

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{e})$$

- $\Pr(\mathbf{e})$: target language model
- $\Pr(\mathbf{f}|\mathbf{e})$: use "IBM-4" translation model (TM)
- $\underset{\mathbf{e}}{\operatorname{argmax}}$ operation: own decoder developed

IBM-4: a word-based translation model [Brown 93]



4 sub-models:

- A fertility model: $n(\phi|e)$ (number of produced words)
- A lexical model: $t(f|e)$ (what words are produced)
- A distorsion model: $d(\Delta_j|\ldots)$ (where those words are placed)
- A parameter p_0 for the spontaneous production of words
- Alignment is not symmetric
- Parameters iteratively trained (Expectation-Maximization algorithm)

Decoder highlights

- Supports IBM-4 TM, with word classes
- Supports 2-, 3- and 4-gram language models (LM)
- Outputs search space as a word lattice
- A* decoding, with admissible heuristics
- Several configurable prunings
- Groups hypotheses in stacks

Sample « A* » decoding, step by step (1/3)

The idea: extend the most promising partial hypothesis

- We wish to translate « une maison bleue »
- Start with ○○○ « »
- Extend it (also produces partial scores):

●○○ « a »	○●○ « house »	○○● « blue »
●○○ « one »	○●○ « home »	○○● « red »
●○○ « the »	○●○ « company »	○○● « bruise »

- Sort those partial translations
- And so on: extend the most promising hypothesis

Sample « A* » decoding, step by step (2/3)

The idea: extend the most promising partial hypothesis

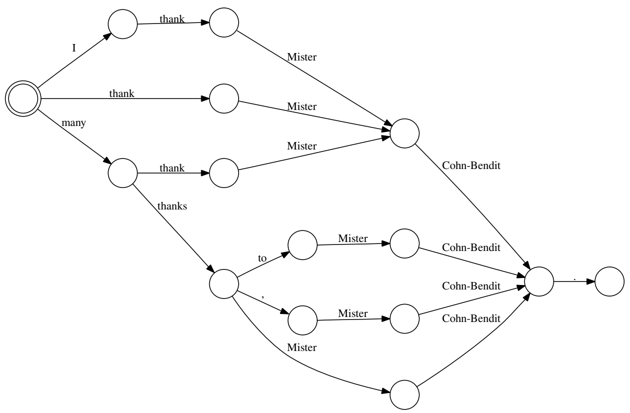
- Extend an hypothesis = translate one more word
- ●○○ « a » produces:
  - ●●○ « a house »
  - ●●○ « a home »
  - ●●○ « a company »
  - ●○● « a blue »
  - ●○● « a red »
  - ●○● « a bruise »

Sample « A* » decoding, step by step (3/3)

Bis repetita placent

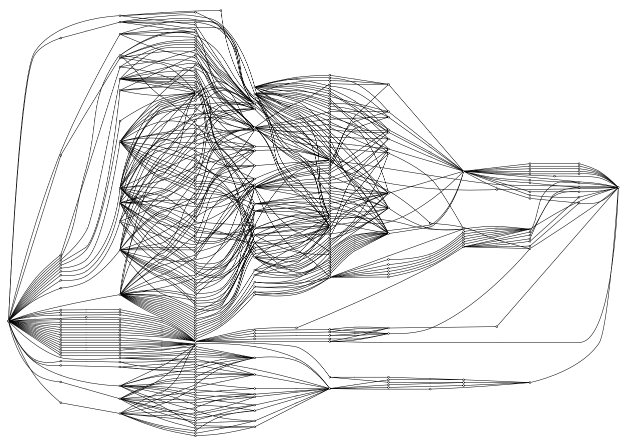
- New most promising hypothesis: ●○○ « one »
- It produces:
  - ●●○ « one house »
  - ●●○ « one home »
  - ●●○ « one company »
  - ●○● « one blue »
  - ●○● « one red »
  - ●○● « one bruise »
- Language model will penalize expansions of ●●○ « a house » (like ●●● « a house blue »)
- Repeat, until the most promising translation is a complete translation

Sample output lattices (1/2)



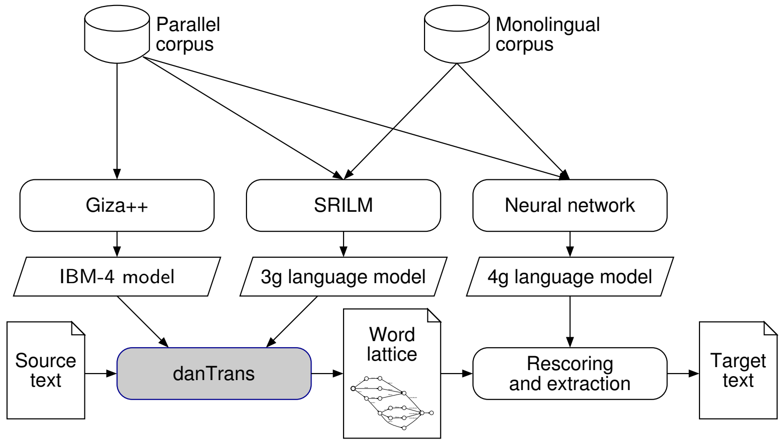
Pruned translation lattice for "muchas gracias señor Cohn-Bendit ."

Sample output lattices (2/2)



Full translation lattice for "muchas gracias señor Cohn-Bendit ."

System architecture



Performance of the word-based translation system

		3g LM	4g LM	4g NNLM
En→Sp	Dev06	39.82	40.58	41.41
	Eval07	37.96	38.34	39.52
Sp→En	Dev06	37.86	38.36	39.04
	Eval07	39.31	39.48	40.39

- BLEU scores (%), the higher the better
- 4-gram LM (back-off): improves over 3-gram, not by much
- Neural network 4-gram LM: excellent generalization behavior
- Language model more important when translating to Spanish

Outline

1 Models and algorithms for machine translation
- Introduction to machine translation
- A word-based translation system
- A phrase-based translation system
- Phrase-table discriminative training

2 Specifics of speech translation
- Motivation
- Translation of a stream of words
- Integration with speech recognition

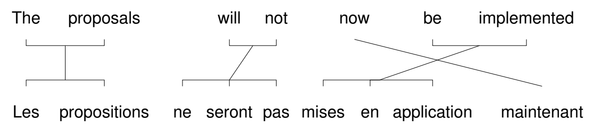
A phrase-based translation system

- Statistical MT equation:

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{e})$$

- $\Pr(\mathbf{e})$: target language model
- $\Pr(\mathbf{f}|\mathbf{e})$: use a phrase-based model (phrase = group of words)
- $\operatorname*{argmax}_{\mathbf{e}}$ operation: Moses [Koehn et al., ACL'07]

A typical phrase-based model [Koehn et al., 03]

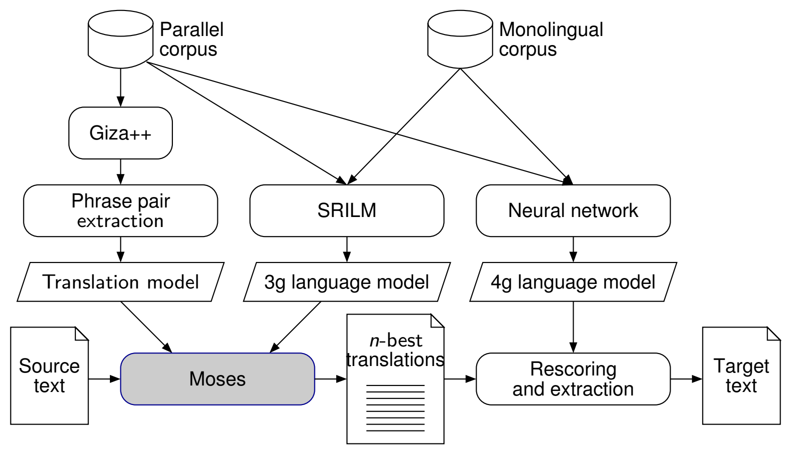


- A phrase-table: $t(\tilde{f}|\tilde{e})$ (how to translate *phrases*)
- A distortion model, for instance $d(\Delta_j| \ldots)$

A phrase-table is:

$\langle \tilde{e} , \tilde{f} \rangle$	Score
⟨`want a` , `veut`⟩	0.12
⟨`want a` , `veux une`⟩	0.15
⟨`want as` , `exigera`⟩	0.003
... , ...	...

System architecture



Performance of the phrase-based translation system

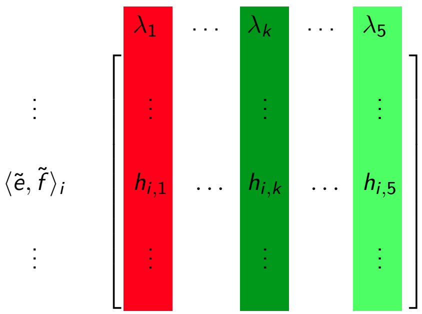
		Phrase-based	Word-based
En→Sp	Dev06	50.03	41.41
	Eval07	50.91	39.52
Sp→En	Dev06	47.93	39.04
	Eval07	48.93	40.39

- BLEU scores (%), the higher the better
- Results with the 4g NNLM
- Impact of better LM similar to with word-based system
- Phrase model ≈ 10 BLEU points better than word-based one

Outline

1. Models and algorithms for machine translation
   - Introduction to machine translation
   - A word-based translation system
   - A phrase-based translation system
   - Phrase-table discriminative training
2. Specifics of speech translation
   - Motivation
   - Translation of a stream of words
   - Integration with speech recognition

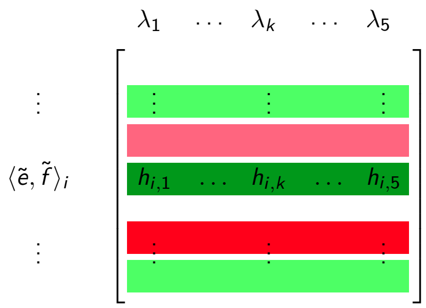
Current phrase-table training and tuning



- Millions of lines (phrase pairs)
- A few columns (scores)
- Score of each phrase pair: $\sum_k \lambda_k h_{i,k}$
- Tuning discriminates scores

Not shown here: LM score, distortion, ...

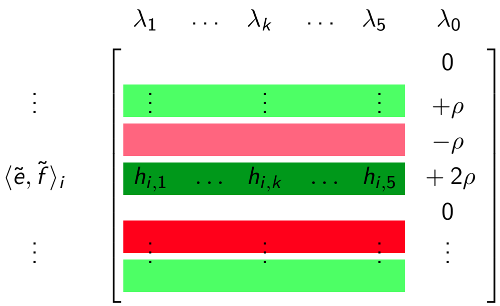
Proposed training



- Same phrase table
- Same weights λ
- Score of each phrase pair: $w_i = \sum_k \lambda_k h_{i,k}$
- Tuning discriminates phrase pairs

Not updated: λ for LM score, distortion, ...

Proposed training



- Start with optimized λ
- Translate corpus
- Adjust phrase pair scores accordingly
- Store updates in a new column

Score of each phrase pair: $w_i = \sum_k \lambda_k h_{i,k}$.

Example of Perceptron-inspired updates

- $\mathbf{f} = \texttt{le petit chat boit le lait}$
- $\mathbf{e_h} = \texttt{the | small | cat | drinks | the | milk}$
  - $\mathcal{C}(\texttt{le},\texttt{the}) = 2$, $\mathcal{C}(\texttt{petit},\texttt{small}) = 1$, $\mathcal{C}(\texttt{chat},\texttt{cat}) = 1$, $\mathcal{C}(\texttt{boit},\texttt{drinks}) = 1$ and $\mathcal{C}(\texttt{lait},\texttt{milk}) = 1$
- $\mathbf{e_d} = \texttt{the kitten | drinks | the | milk}$
  - $\mathcal{C}(\texttt{le petit chat},\texttt{the kitten}) = 1$, $\mathcal{C}(\texttt{boit},\texttt{drinks}) = 1$, $\mathcal{C}(\texttt{le},\texttt{the}) = 1$ and $\mathcal{C}(\texttt{lait},\texttt{milk}) = 1$

$$\begin{cases} w_i \leftarrow w_i + (1 - 2)\rho & \text{for pair } \langle \texttt{le},\texttt{the} \rangle \\ w_i \leftarrow w_i + (0 - 1)\rho & \text{for pair } \langle \texttt{petit},\texttt{small} \rangle \\ w_i \leftarrow w_i + (0 - 1)\rho & \text{for pair } \langle \texttt{chat},\texttt{cat} \rangle \\ w_i \leftarrow w_i + (1 - 0)\rho & \text{for } \langle \texttt{le petit chat},\texttt{the kitten} \rangle \\ w_i \text{ unchanged} & \text{for all other pairs} \end{cases}$$

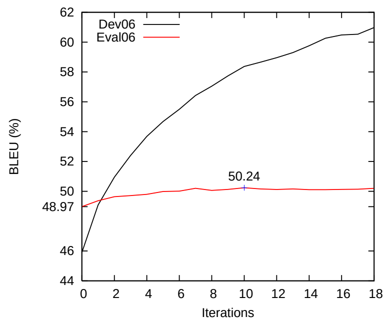
Update system inspired by the Perceptron

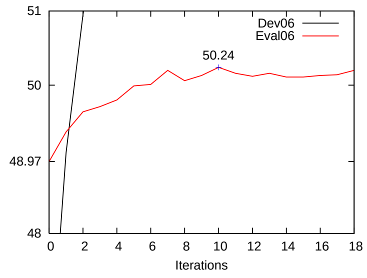
$$\left\{\begin{array}{ccccc} \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ w_i \leftarrow w_i & + & \rho\big(\mathcal{C}(\tilde{e}_{i,d}, \tilde{f}_i) & - & \mathcal{C}(\tilde{e}_{i,h}, \tilde{f}_i)\big) \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \end{array}\right.$$

- $\mathbf{f}$: sentence to translate
- $\mathbf{e_d}$: desired (expected) translation
- $\mathbf{e_h}$: hypothesized (produced) translation
- w_i: aggregated score of the i^{th} phrase pair
- $\mathcal{C}(\tilde{e}_i, \tilde{f}_i)$: how many times $\langle\tilde{e}_i, \tilde{f}_i\rangle$ is used to translate $\mathbf{f}$ into $\mathbf{e}$

It actually learns something

- TC-STAR task, Spanish to English
- Discriminative adaptation on dev06, calibration on eval06
- Blind evaluation on eval07: 48.67 (baseline: 47.81)



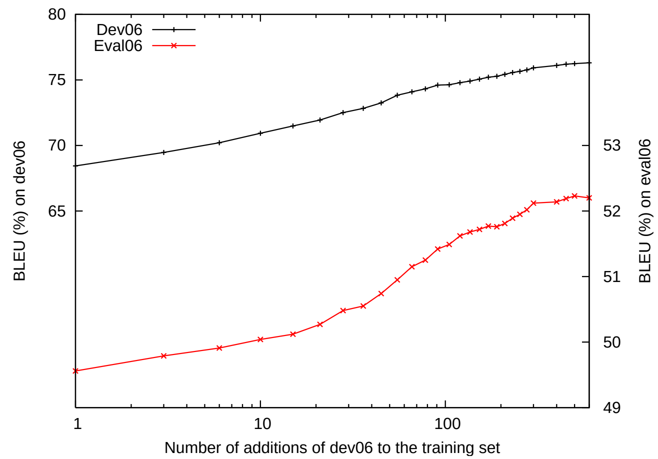


Alternative approaches

We should compare with other ways to include dev06 data:

- Simply add dev06 to the TM training data
  $\rightsquigarrow$ What relative weight? 1? 2?

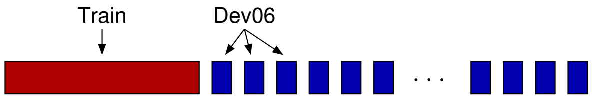
Adding dev06 data to the training data



Alternative approaches

We should compare with other ways to include dev06 data:

- Simply add dev06 to the TM training data
  $\rightsquigarrow$ What relative weight? ~~1? 2?~~ 600!



- TM trained on dev06 only

- Two TMs in parallel (on train and on dev06)

- LM adaptation
  Interpolation with an LM trained on dev06

Comparative results

- TC-STAR task, Spanish to English
- BLEU scores (%), on Eval07 set
- All weights λ_i retuned on Eval06

	BLEU	Δ Baseline
Baseline	48.22	0
Adapted LM	48.87	+0.65
Discriminative training of TM	48.90	+0.68
TM on train+600 dev	49.90	+1.68
TM on dev only	39.85	−8.37
TM train + TM dev	49.17	+0.95

Other results

- TC-STAR task, English to Spanish

	BLEU	Δ Baseline
Baseline	49.09	0
Discriminative training of TM	48.88	−0.21
TM on train+1 dev	48.84	−0.25
TM on train+300 dev	48.59	−0.50

- Also tried on training set
- Why doesn't it work?

Outline

1 Models and algorithms for machine translation
- Introduction to machine translation
- A word-based translation system
- A phrase-based translation system
- Phrase-table discriminative training

2 Specifics of speech translation
- Motivation
- Translation of a stream of words
- Integration with speech recognition

Specifics of speech translation

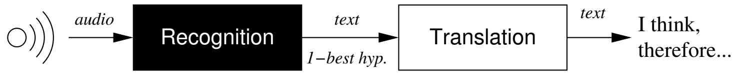
Translation of transcribed speech

- Spoken language (grammar? syntax?)
- Style, vocabulary, expressions
- Segmentation into sentences, punctuation

Translation of automatically transcribed speech

- Combination of two complex systems
- Towards a tighter integration

Speech translation



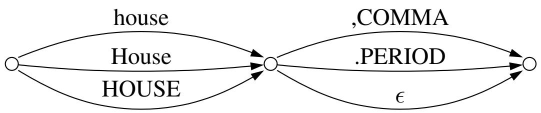
1. Translation of a word stream
2. Speech translation: theoretical motivation
3. Integration of recognition and translation
4. Tuning of recognition for translation

Case and punctuation restoration

Objective: Making ASR's output resemble MT's training data

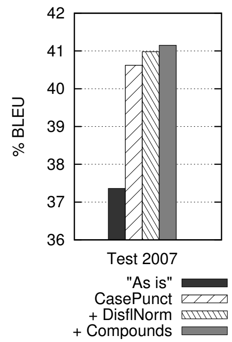
Example: Case and punctuation

- Input: CTM file (words and time information)
- Remove any punctuation and case
- Build a lattice for each word
- Tuning: Target 3.5% of periods and 5% of commas

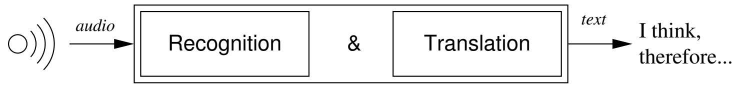


Making ASR's output resemble MT's training data

- Punctuation restoration is crucial for our system
- Additional gains with "easy" renormalizations
  - Greater improvements observed with other systems
- Small extra gains by recreating compounds



Speech translation



1. Translation of a word stream
2. Speech translation: theoretical motivation
3. Integration of recognition and translation
4. Tuning of recognition for translation

Theoretical motivation [Ney, ICASSP'99]

$\mathbf{X}$ is the audio in **f** rench, which we want to translate into **e** nglish

$$
\begin{aligned}
\mathbf{e}^* &= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{X}) \\
&= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{X}|\mathbf{e}) \\
&= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \sum_{\mathbf{f}} \Pr(\mathbf{X}, \mathbf{f}|\mathbf{e}) \\
&= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \sum_{\mathbf{f}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{X}|\mathbf{f}, \mathbf{e}) \\
&= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \sum_{\mathbf{f}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{X}|\mathbf{f})
\end{aligned}
$$

Theoretical motivation [Ney, ICASSP'99]

$\mathbf{X}$ is the audio in **f** rench, which we want to translate into **e** nglish

$$
\begin{aligned}
\mathbf{e}^* &= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \sum_{\mathbf{f}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{X}|\mathbf{f}) \\
&\approx \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \max_{\mathbf{f}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{X}|\mathbf{f})
\end{aligned}
$$

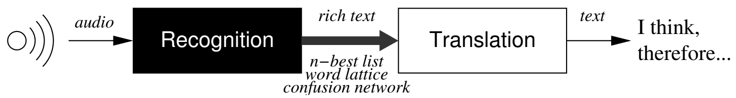
Target language model

(Reverse) translation model

Acoustic model

- Determination of **f** not necessary (hidden variable)
- Source language model not necessary
- Speech recognition formula: $\mathbf{f}^* = \operatorname{argmax}_{\mathbf{f}} \Pr(\mathbf{f}) \Pr(\mathbf{X}|\mathbf{f})$

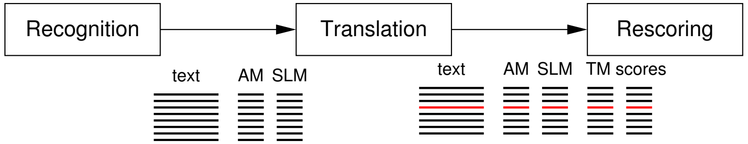
Speech translation



1. Translation of a word stream
2. Speech translation: theoretical motivation
3. Integration of recognition and translation
4. Tuning of recognition for translation

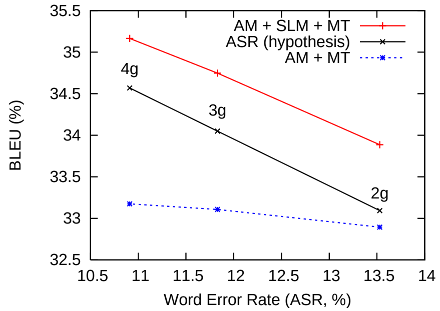
Translation of ASR's ambiguous output (1/2)

- Translation of ASR's *n*-best lists



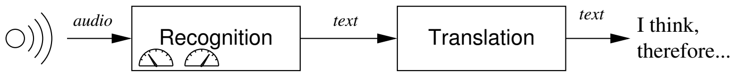
- AM: score from the Acoustic Model
- SLM: score from the Source Language Model
- TM: scores from the Translation Model
- 3 ASR systems: same acoustic model, different source language models

Translation of ASR's ambiguous output (2/2)



- Spanish to English
- Source language model useful indeed
- Would use confusion networks or word lattices nowadays

Speech translation



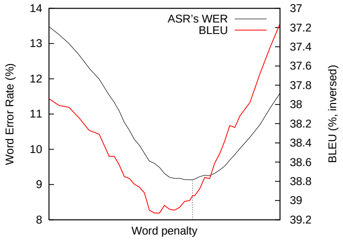
1. Translation of a word stream
2. Speech translation: theoretical motivation
3. Integration of recognition and translation
4. Tuning of recognition for translation

Tuning ASR to improve ASR+MT performance (1/2)

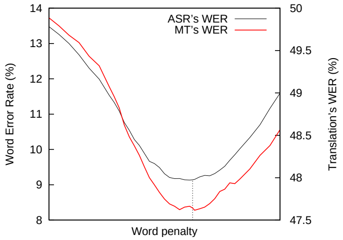
- ASR parameters tuned to minimize expected WER
- Rather, tune them to maximize ASR+MT performance
- Possible experiments: adjust word penalty, SLM weight, disable consensus decoding, ...
- Observe impact on several automatic measures

Tuning ASR to improve ASR+MT performance (2/2)

- BLEU: benefits slightly from higher insertion rates



- MT's WER (and others): optimized when ASR's WER minimized



Conclusion

- Fully developed a translation system
  - Decoder for IBM-4 model
  - Outputs search space as a word lattice
  - Neural language model brought significant improvements
- Experiments with phrase-based approach
  - Based on the open-source decoder Moses
  - Proposed a discriminative training algorithm for the phrase table
- Integration of ASR and MT
  - Efficient processings to translate a black-box ASR system
  - Source LM necessary, "despite theory"
  - Integration still not easy, subject to trade-offs
  - ASR's WER predicts well ASR+MT performance

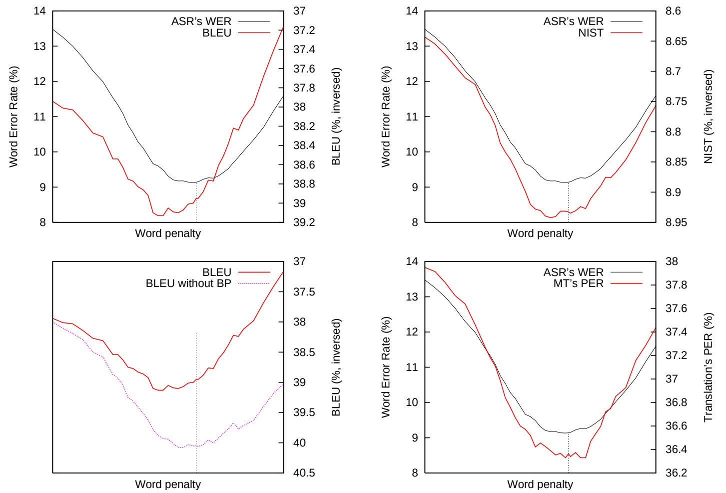
Perspectives

- Phrase-table parameter tying
- Phrase-table discriminative training
- Domain independence
- Or fast and automatic data acquisition

Thank you

Merci!

More on the effect of ASR's word penalty



Translating the output of different STT systems (1/2)

Motivation: "tune" ASR to improve ASR+MT performance

- Consensus decoding (CD) "break phrases"
- Rover combination even more so
- How to measure "phrase breakage"?
  - BLEU score of ASR's output against the manual transcription
  - Size of the filtered phrase table
- What impact?

Translating the output of different STT systems (2/2)

	ASR			MT	
	System	WER	BLEU	# phr.	BLEU
Dev06	Rover	7.18	70.22	2231k	43.58
	Limsi CD	9.14	63.98	2260k	42.95
	Limsi MAP	9.53	63.92	2264k	43.05
Eval07	Rover	7.08	67.92	2103k	41.15
	Limsi CD	9.33	61.29	2123k	40.30
	Limsi MAP	9.66	61.14	2130k	40.19

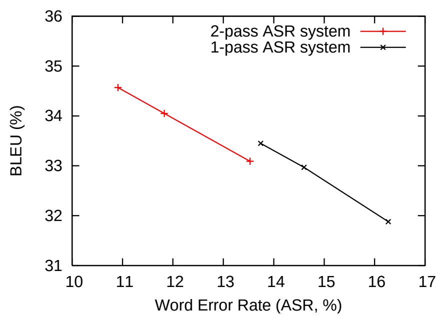
- Dev06: Limsi MAP slightly better translated than Limsi CD
- Results on Eval07 prevents any definitive conclusion

Impact of the SLM and the AM (1/2)

- Different ASR systems, of varying SLM and AM quality
- Impact on ASR+MT performance?
- Two acoustic models
  - "first-pass" model
  - "second-pass" model, after adaptation
- Three (source) language models
  - 2-gram (back-off)
  - 3-gram (back-off)
  - 4-gram (neural)

$\rightsquigarrow$ 6 different ASR systems

Impact of the SLM and the AM (2/2)



- Near-linear correlation between BLEU and ASR's WER
- Src language model at least as important as acoustic model

Current phrase-table training and tuning

EM algorithm	Learn IBM-4 model parameters $t(f\|e)$, $t(e\|f)$, $\phi(n\|e)$, ...
Best alignment	Compute best word-to-word alignments (source-to-target and target-to-source) according to the above models
Heuristics	Symmetrize word-to-word alignments Extracts phrase pairs
Relative frequencies + heuristics	Score phrase pairs 5 scores per phrase pair
Discriminative tuning	Tune feature function combination

Rewriting the score of translation hypothesis

$$\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_5 \\ \hdashline \lambda_6 \\ \vdots \\ \lambda_M \end{bmatrix}^{\mathsf{T}} \cdot \begin{bmatrix} \sum_p h_1(\tilde{e}_p, \tilde{f}_p) \\ \vdots \\ \sum_p h_5(\tilde{e}_p, \tilde{f}_p) \\ \hdashline h_6(\mathbf{e}, \mathbf{f}) \\ \vdots \\ h_M(\mathbf{e}, \mathbf{f}) \end{bmatrix} = \begin{bmatrix} \vdots \\ \vdots \\ \sum_{k=1}^{5} \lambda_k h_{i,k} \\ \vdots \\ \vdots \\ \hdashline \lambda_6 \\ \vdots \\ \lambda_M \end{bmatrix}^{\mathsf{T}} \cdot \begin{bmatrix} \vdots \\ \vdots \\ \mathcal{C}(\tilde{e}_i, \tilde{f}_i) \\ \vdots \\ \vdots \\ \hdashline h_6(\mathbf{e}, \mathbf{f}) \\ \vdots \\ h_M(\mathbf{e}, \mathbf{f}) \end{bmatrix}$$

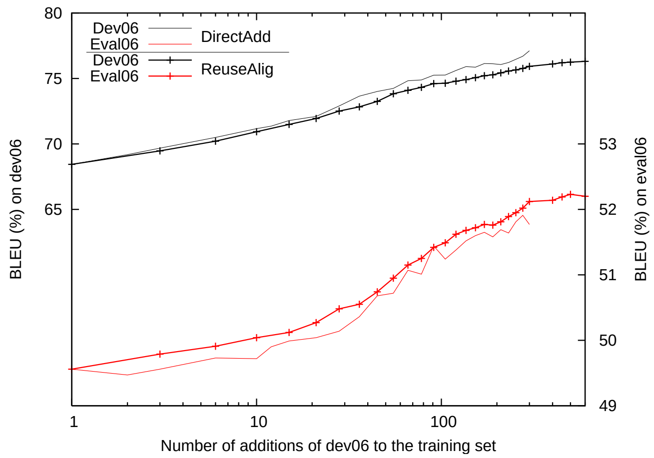
Discriminative training details

- How to determine the desired output $\mathbf{e_d}$?
  $\rightsquigarrow$ the n^{th}-best translation of highest *smoothed* bleu score

$$\text{BLEU}_{\text{smoothed}}(\mathbf{e}, \mathbf{e_r}) = \sum_{i=1}^{4} \frac{\text{BLEU}_i(\mathbf{e}, \mathbf{e_r})}{2^{5-i}}$$

- Inspired from [Liang et al., ACL'06]
- How to determine ρ?
  $\rightsquigarrow$ $\rho = 0.05$ seems to work well...
- What corpus?
  $\rightsquigarrow$ Discriminative training on *development* data

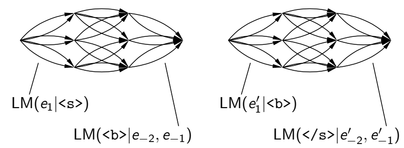
Adding dev06 data to the training data



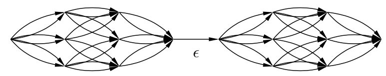
Efficient handling of long sentences

danTrans

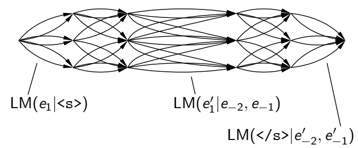
Independent translations of the two sentence fragments



Merge of the translation lattices



Lattice expansion with an *n*-gram language model



Speech translation: motivation

Coming up a twenty-seven year veteran of the FBI is arrested and charged with spying for the Russians	Monter des vingt-sept vétérans d'an du FBI est arrêté et chargé de l'espionnage pour les Russes
Well we are still spying on each other because things happen in governments that other governments think they need to know about	Bien nous remarquons toujours sur l'un l'autre parce que les choses se produisent dans les gouvernements que d'autres gouvernements pensent qu'ils doivent savoir
They didn't have any snow on the ground but boy it was eh	Ils n'ont eu aucune neige sur la terre mais le garçon qu'elle était hein
yeah but it's just that it's just that clear cold that's the way it is been here it's just been cold	ouais mais c'est juste qu'il fait juste ce froid clair qui est la manière qu'il est été ici il est juste été froid