Traduzione Automatica del Parlato

di

Sessioni del Parlamento Europeo

M. Federico, N. Bertoldi, R. Cattoni, M. Cettolo, B. Chen, D. Gupta

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica

38050 Povo (Trento), Italy

Statistical Machine Translation

SMT goal: given a source string f, find the best translation e:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{f})$$

Maximum Entropy framework for the word-alignment MT approach:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\}$$

where a = alignment, and $h_i(\mathrm{e}, \mathrm{f}, \mathrm{a})$ are suitable feature functions.

Special case: feature functions IBM-style:

$$h_1(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log p(\mathbf{e}) \quad \textbf{(target language model)}$$

$$h_2(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log p(\mathbf{a} \mid \mathbf{e}) \quad \textbf{(re-ordering model)}$$

$$h_3(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log p(\mathbf{f} \mid \mathbf{e}, \mathbf{a}) \quad \textbf{(lexicon model)}$$

Note: ML or minimum error training can be applied to estimate free parameters (λ's).

Phrase-based Machine Translation

- **A phrase is a sequence of one or more words without semantic/syntactic meaning**

- **Generative process:**
  1.
  2.
  3.
  4.

target words

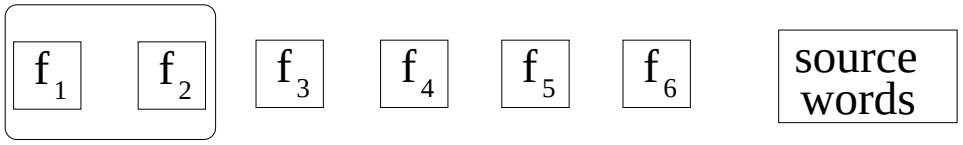
f_1 f_2 f_3 f_4 f_5 f_6

source words

Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

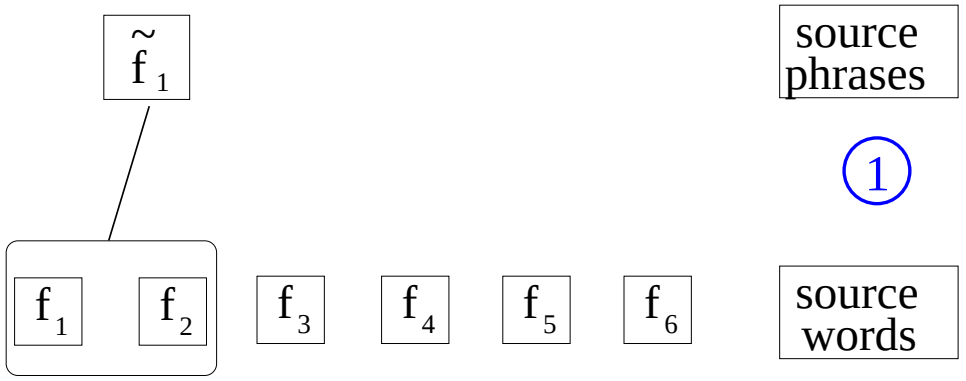
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2.
  3.
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

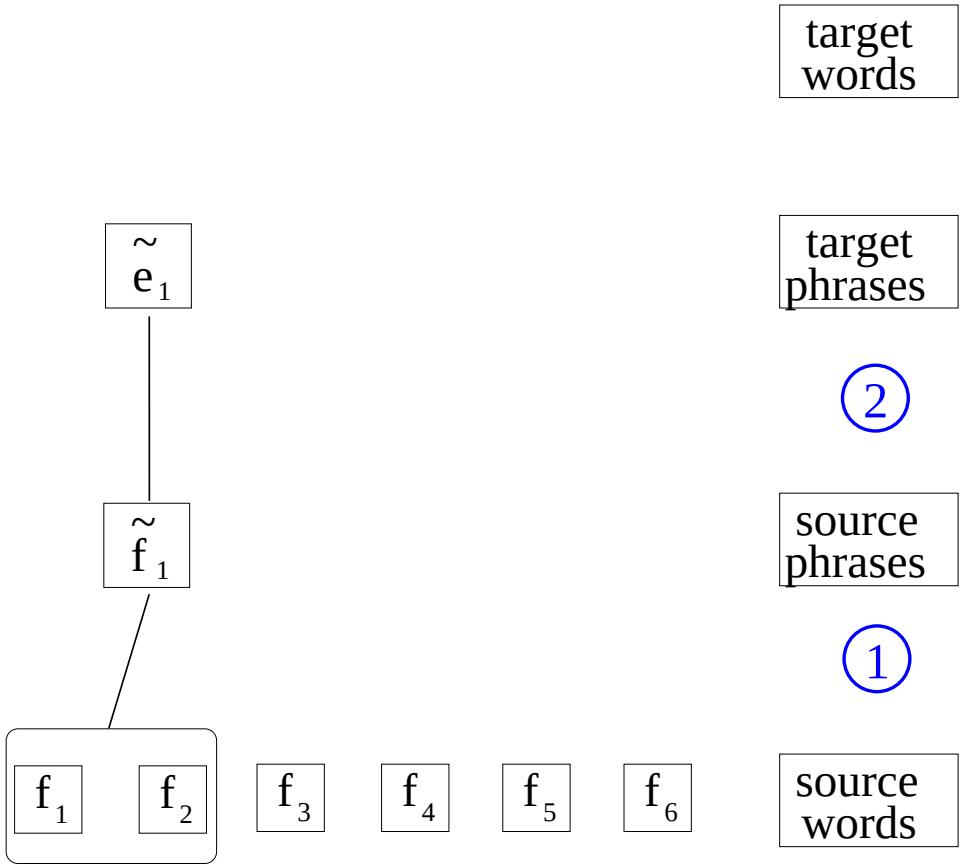
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2.
  3.
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

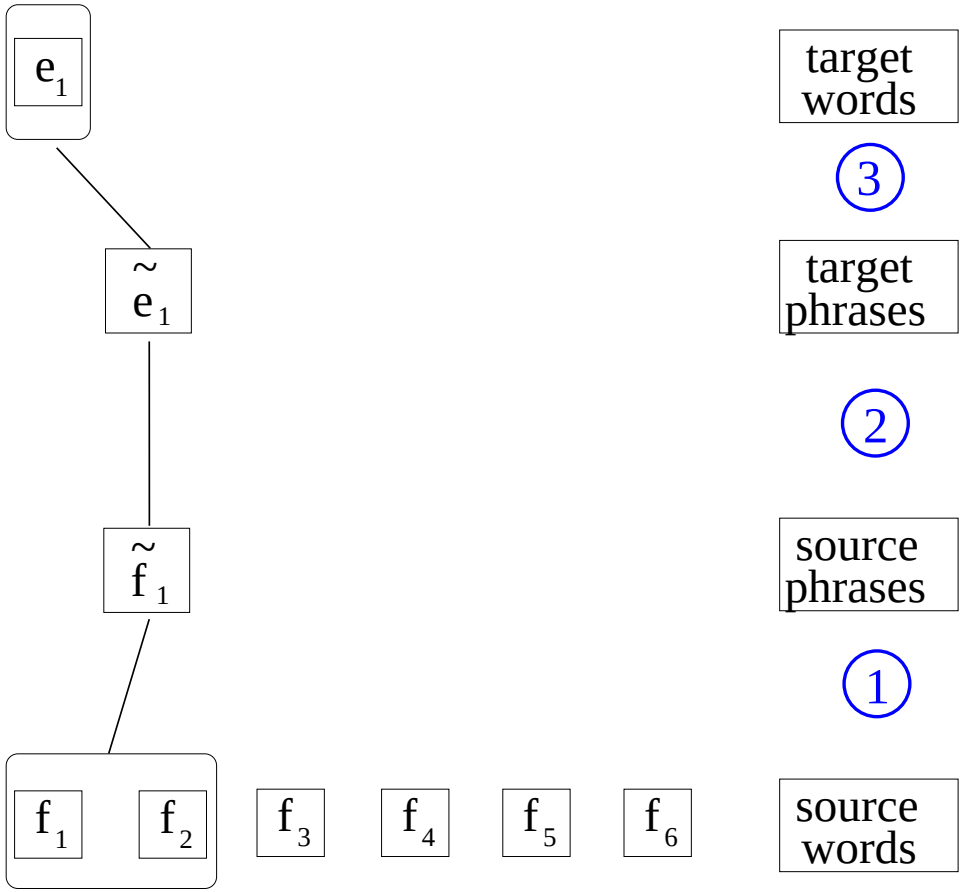
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3.
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

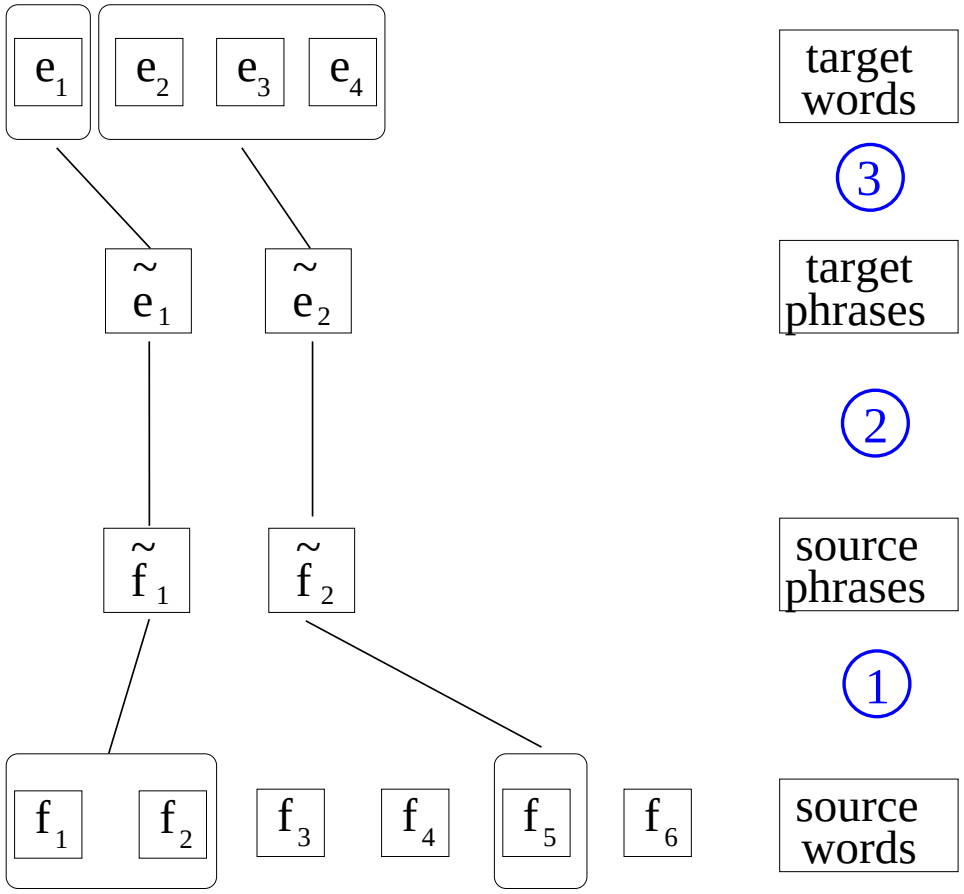
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

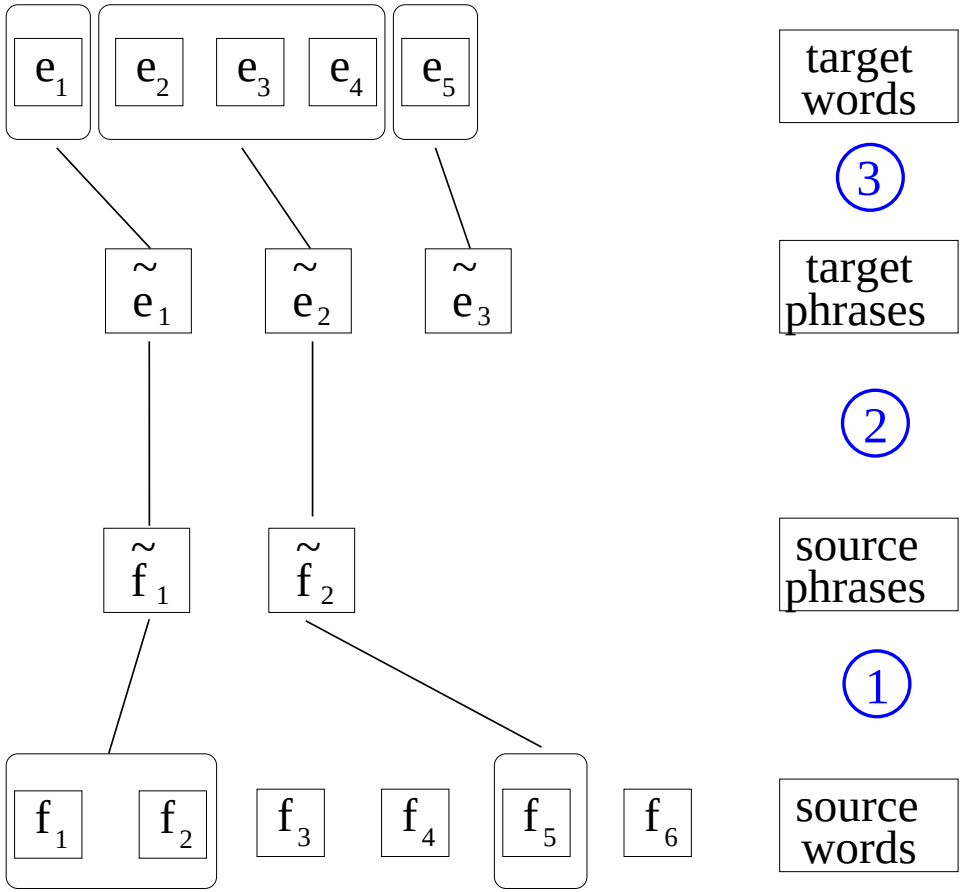
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

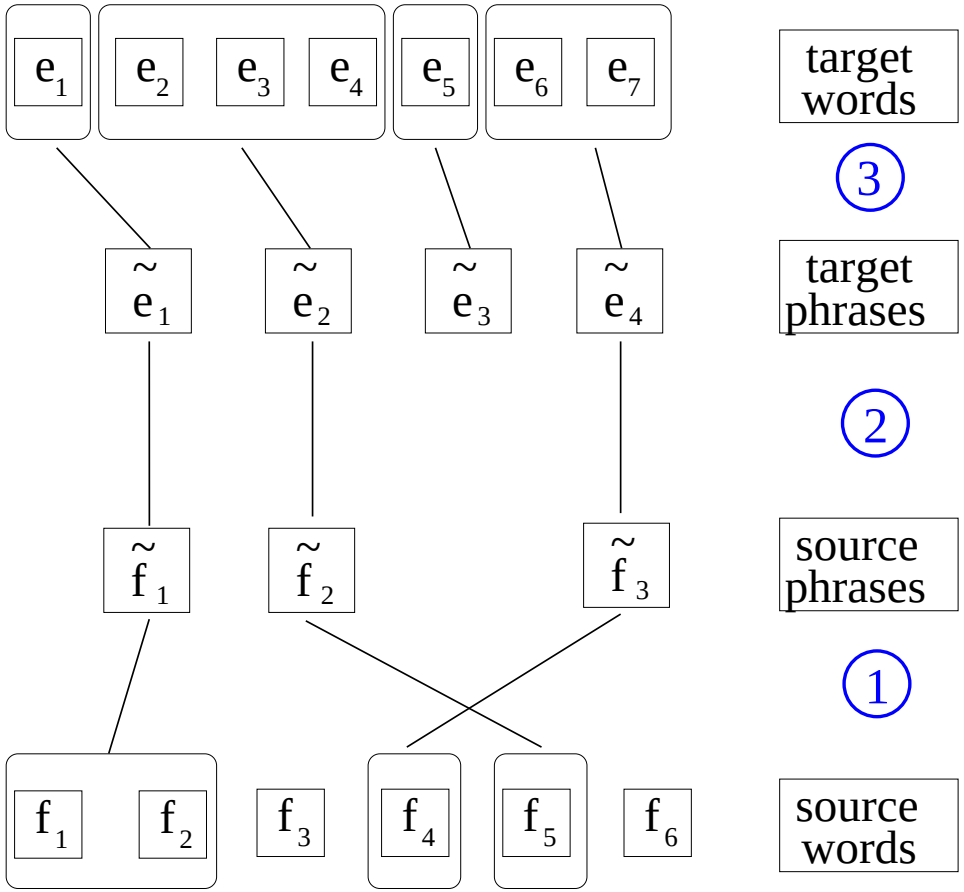
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

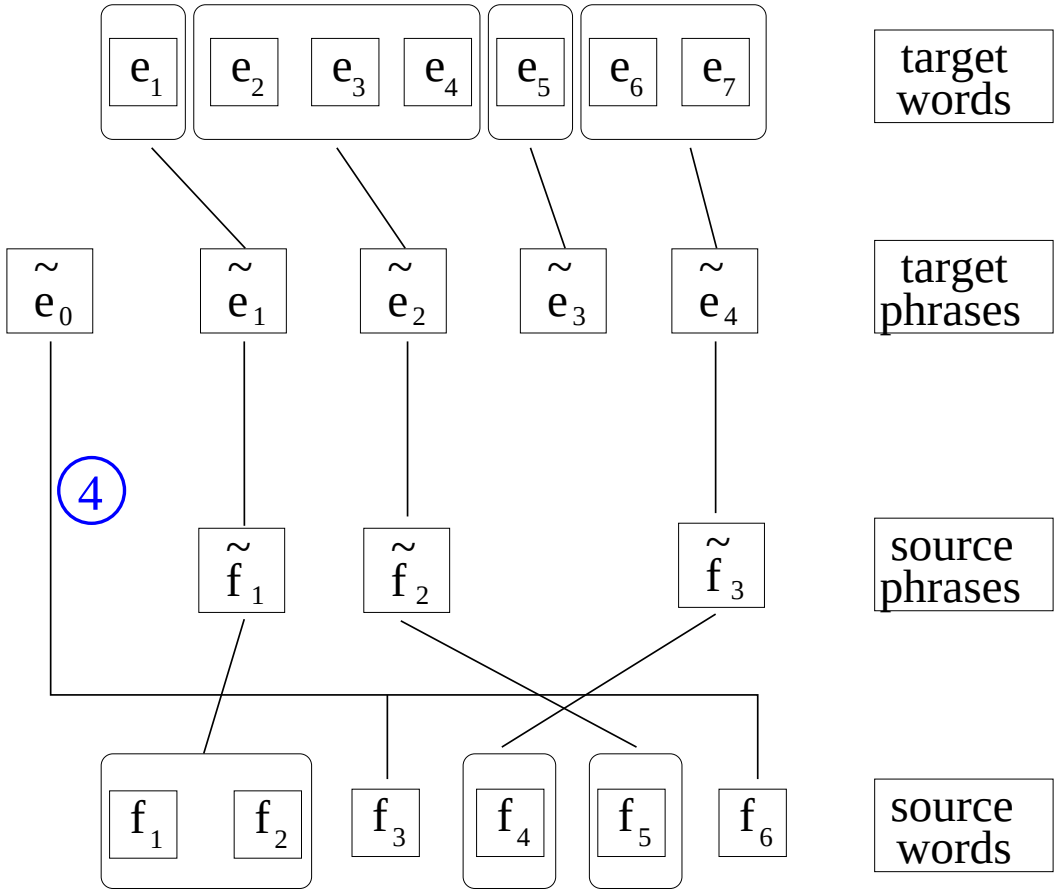
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4.



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

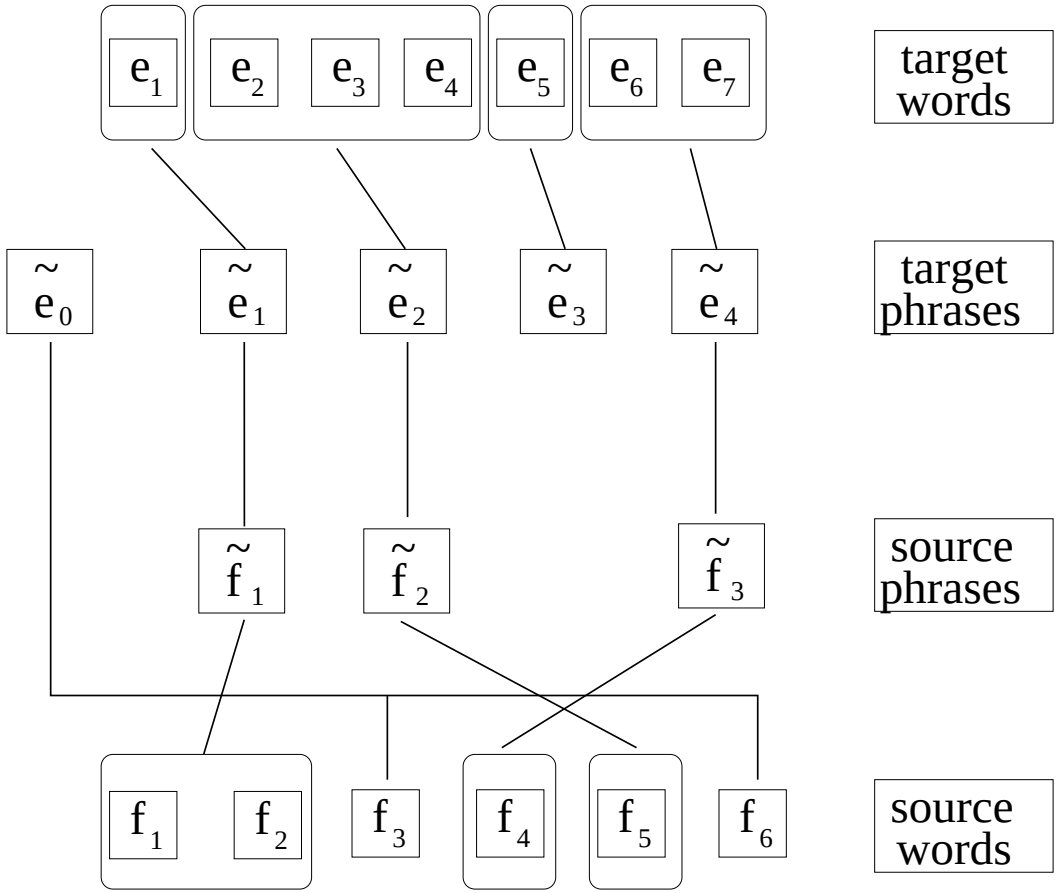
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4. untranslated words (lexicon model)



Phrase-based Machine Translation

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

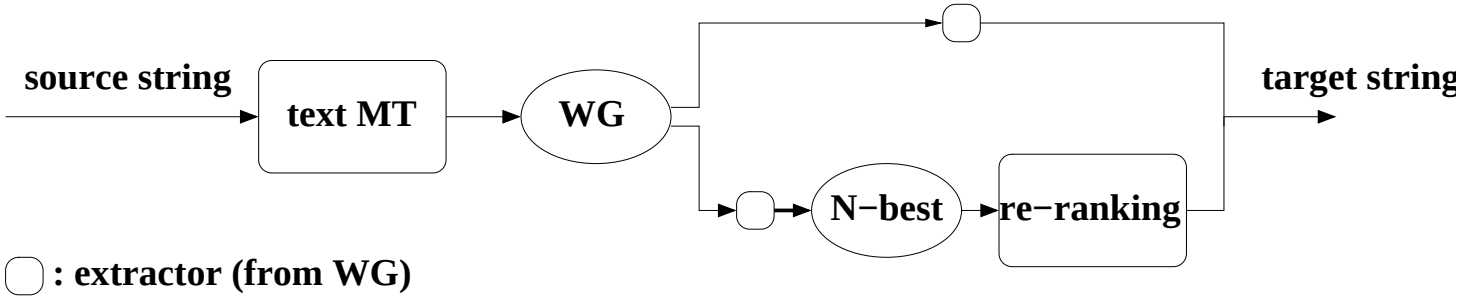
- Generative process:
  1. cover new source positions (re-ord. mdl)
  2. link to target phrase (lexicon model)
  3. add target phrase (language model)
  4. untranslated words (lexicon model)



Search is over strings of phrases:

$$\tilde{\mathbf{e}}^* = \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a}} \sum_i \lambda_i h_i(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})\}$$

Two Pass Search Strategy



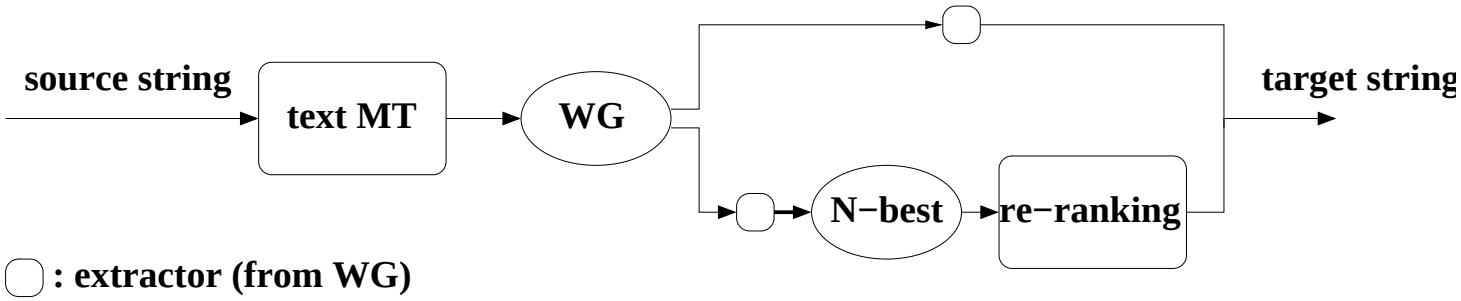
First Pass:

- Log-linear Model
- Dynamic programming algorithm
- Beam search decoder:
  - threshold and histogram pruning
- Non-monotone search constraints
  - max number of vacancies on the left
  - max distance from left-most vacancy
- Output: word-graph (WG)

Second Pass:

- Extraction of 1,000-best
- Log-linear Model
- Re-ranking algorithm

Two Pass Search Strategy



First Pass feature functions:

- Target 3-gram LM
- Fertility model target phrases
- Direct phrase-based lexicon
- Inverse phrase-based lexicon
- Negative distortion
- Positive distortion
- $\tilde{e}_0$ fertility
- $\tilde{e}_0$ permutation

Second Pass (additional) feature functions:

- IBM model 1 lexicon score
- IBM model 3 lexicon score
- CLA lexicon score
- Question feature
- Frequency of n-grams within n-best
- ratio of target source lengths
- 2-,4-,5-gram target LMs

Spoken Language Translation

Translating speech implies:

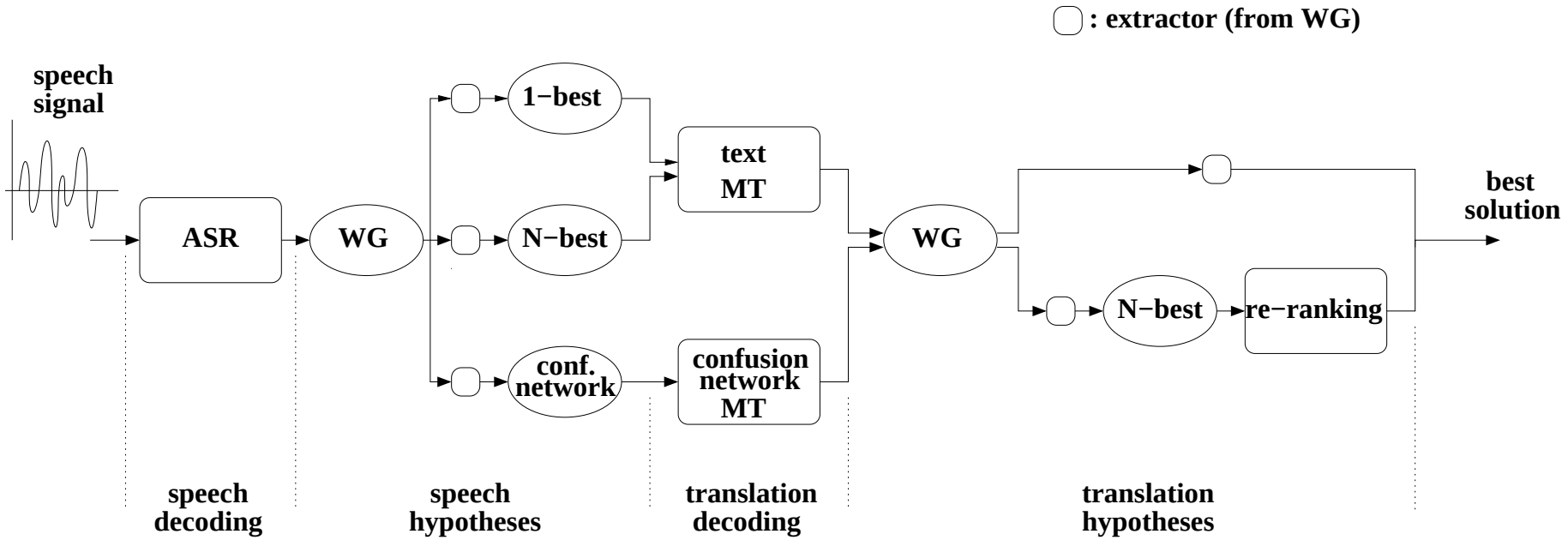
- **speech recognition**
- **spontaneous language**
- **fast response time**

Technical problems:

- **error propagation**
- **complexity**
- **spontaneous speech, poor syntax**
- **training mismatch**

Integration of ASR and MT: Interfaces and Search Methods

- 1-best hypothesis $\longrightarrow$ MT decoder
- N-best hypotheses $\longrightarrow$ MT decoder + re-ranking
- Word-Graph/Confusion Network $\longrightarrow$ Finite State Transducer, CN decoder (ITC-irst)



TC-STAR

Technology and Corpora for Speech to Speech Translation: integrated project funded by the European Union within the 6th framework

- **Partners: ITC-irst, RWTH, LIMSI, UPC, UKA, IBM, Nokia, Siemens, Sony, ELDA, RU-SPEX**
- **Translation directions (first evaluation, March 2005):**
  - **Spanish $\Leftrightarrow$ English (European Parliament)**
  - **Chinese $\Rightarrow$ English (agency news)**
- **EPPS Data:**
  - **Spanish speeches of the European Parliament Plenary Sessions**
  - **train: $\approx$ 30Mw**
  - **test: $\approx$ 20Kw**

EPPS Results

output	DEV					TEST				
	input size	input WER	BLEU	BLEU +opt	decoding time	input size	input WER	BLEU	BLEU +opt	decoding time
verbatim	1	0	49.65	51.26	0.6	1	00.00	44.69	45.82	1.7
1-best	1	11.77	43.92	45.50	0.6	1	14.60	39.74	40.78	2.1
100-best	38	5.07	43.77	45.73	33.2	56	09.09	39.75	40.99	135.3
cn-p00	1	11.67	43.39	45.70	4.0	1	14.46	39.23	40.75	28.4
cn-p60	194	8.41	43.59	45.84	6.7	1284	10.71	39.43	40.87	37.9

IWSLT 2005

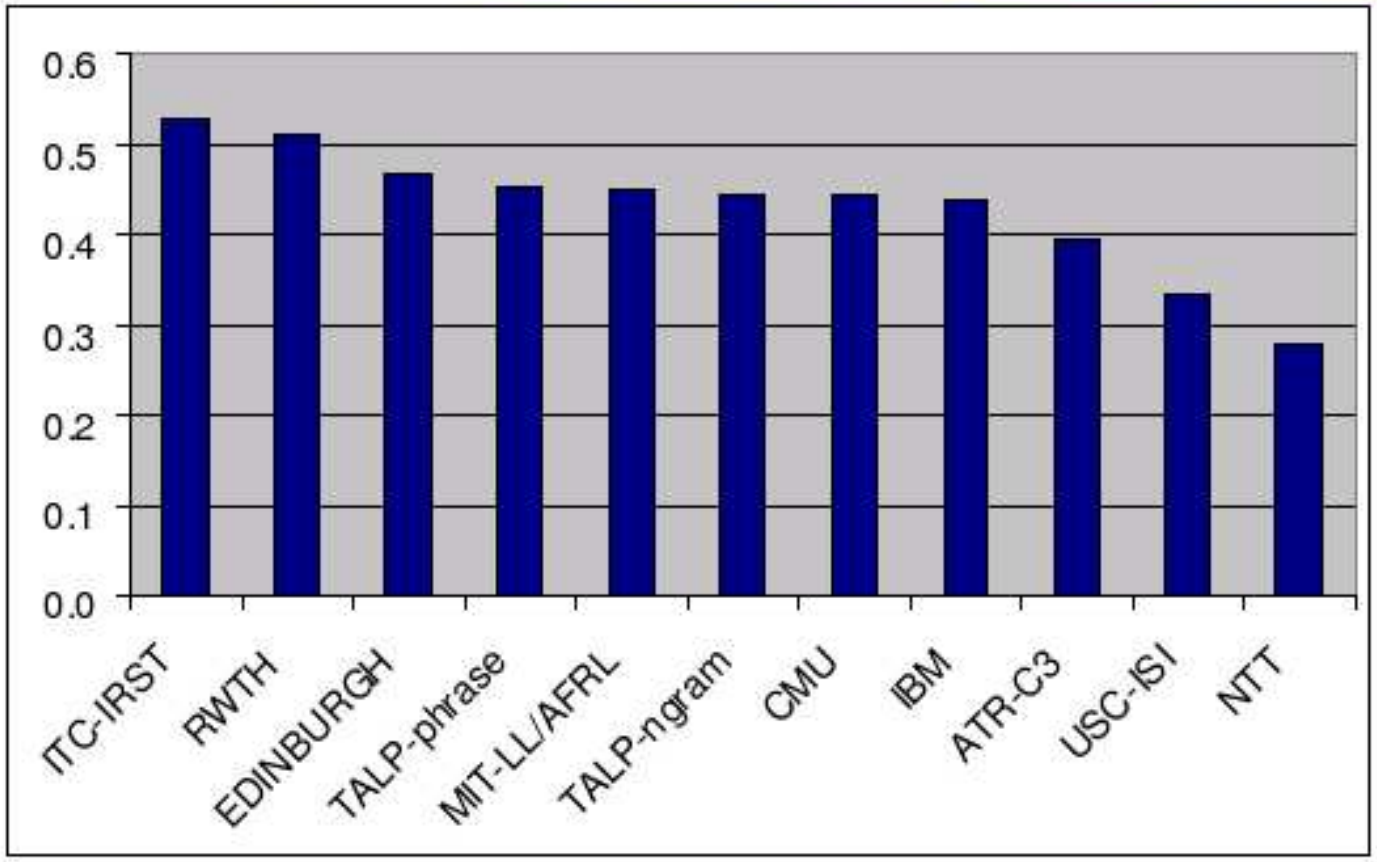
The International Workshop on Spoken Language Translation 2005 (Pittsburgh, Oct 2005) organized an evaluation campaign, which ITC-irst took part in:

- **Translation directions:**
  - **Chinese $\Leftrightarrow$ English**
  - **Japanese $\Rightarrow$ English (manual transcription + ASR)**
  - **Korean $\Rightarrow$ English**
  - **Arabic $\Rightarrow$ English**

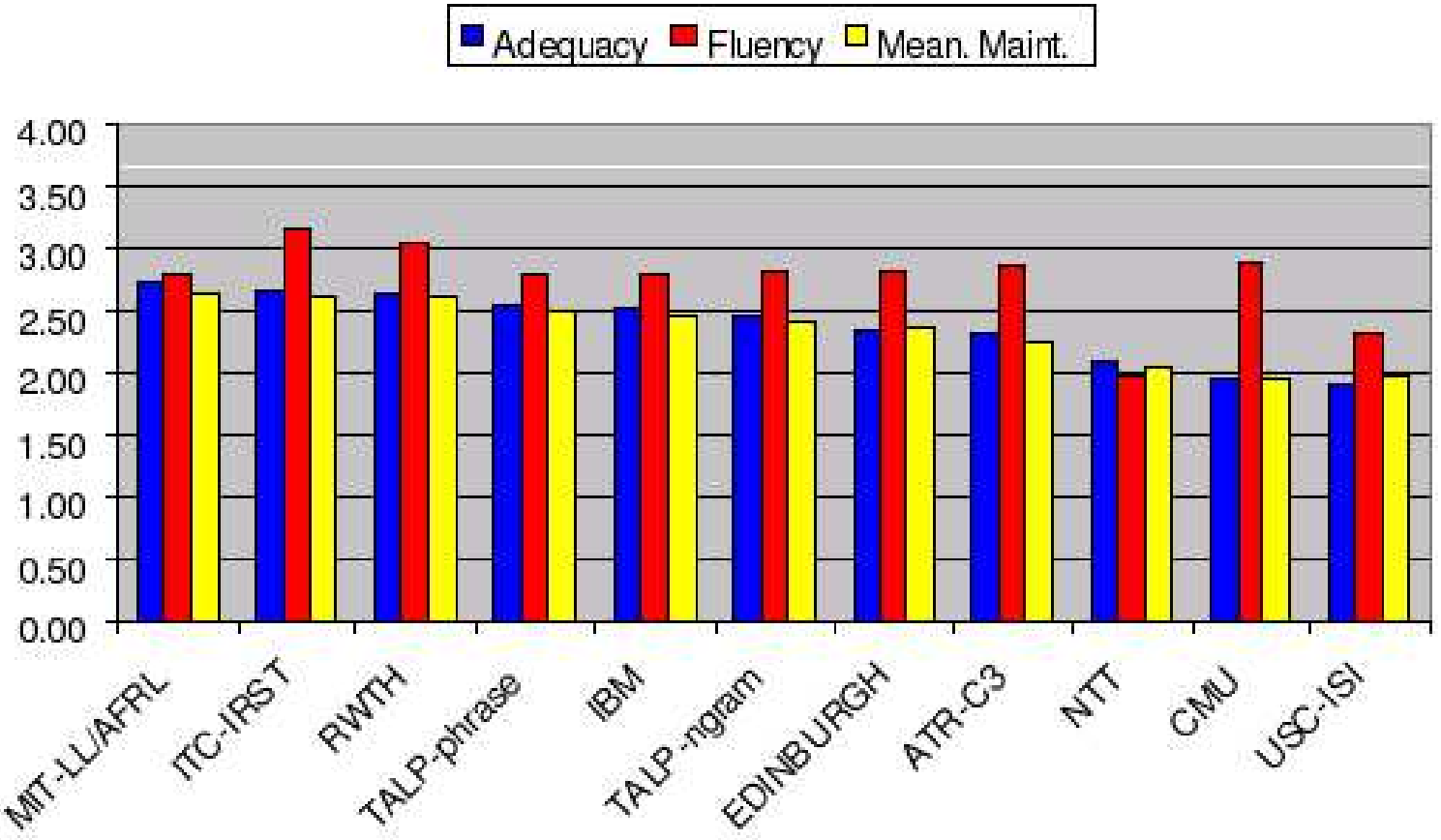
- **Data:**
  - **BTEC, Basic Traveling Expressions Corpus: a collection of traveling expressions**
  - **train: $\approx$ 500Kw**
  - **test: $\approx$ 25Kw**

IWSLT 2005

BLEU: Chinese ⇨ English – MT – Supplied Data

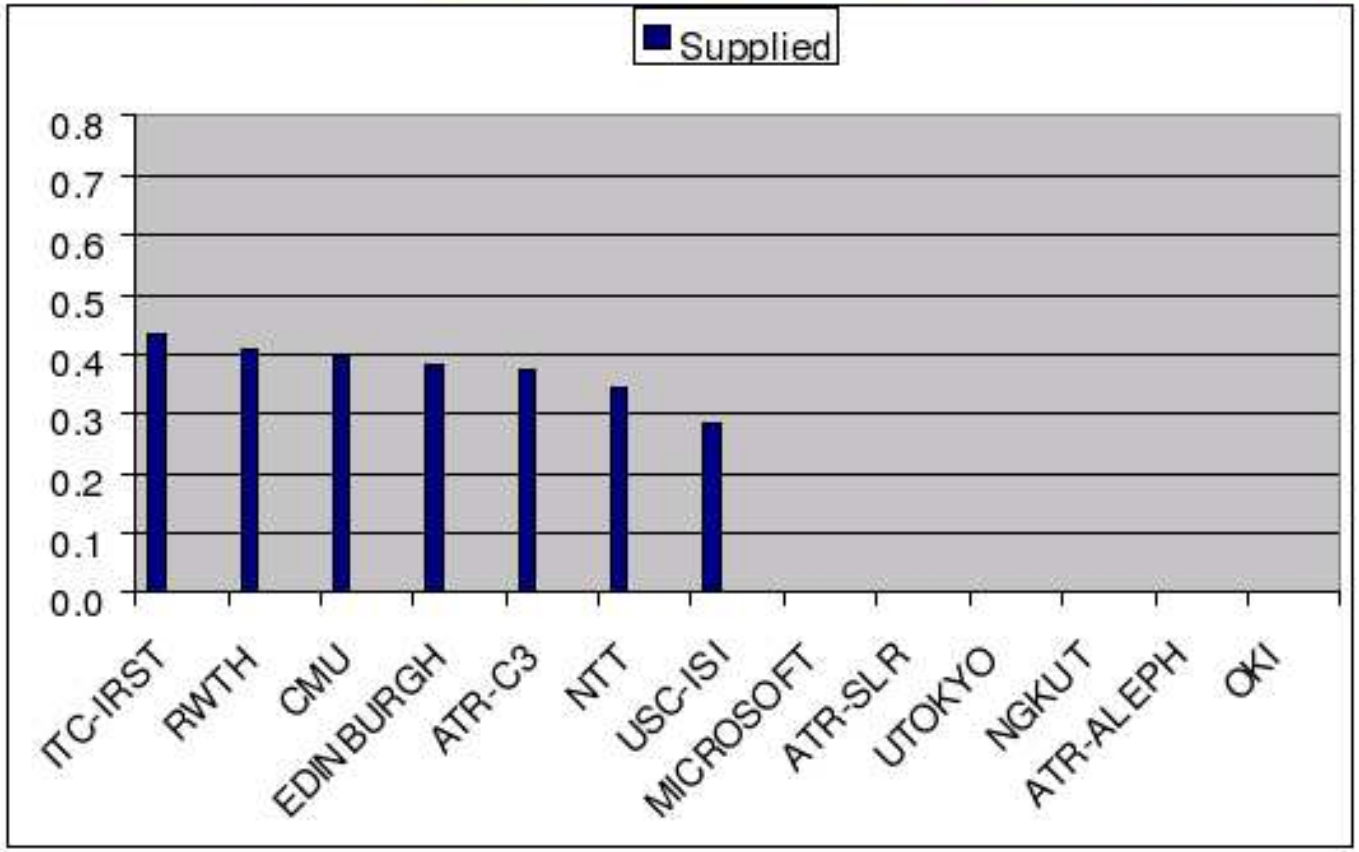


IWSLT 2005: Subjective Evaluations (CHI⇒ENG)



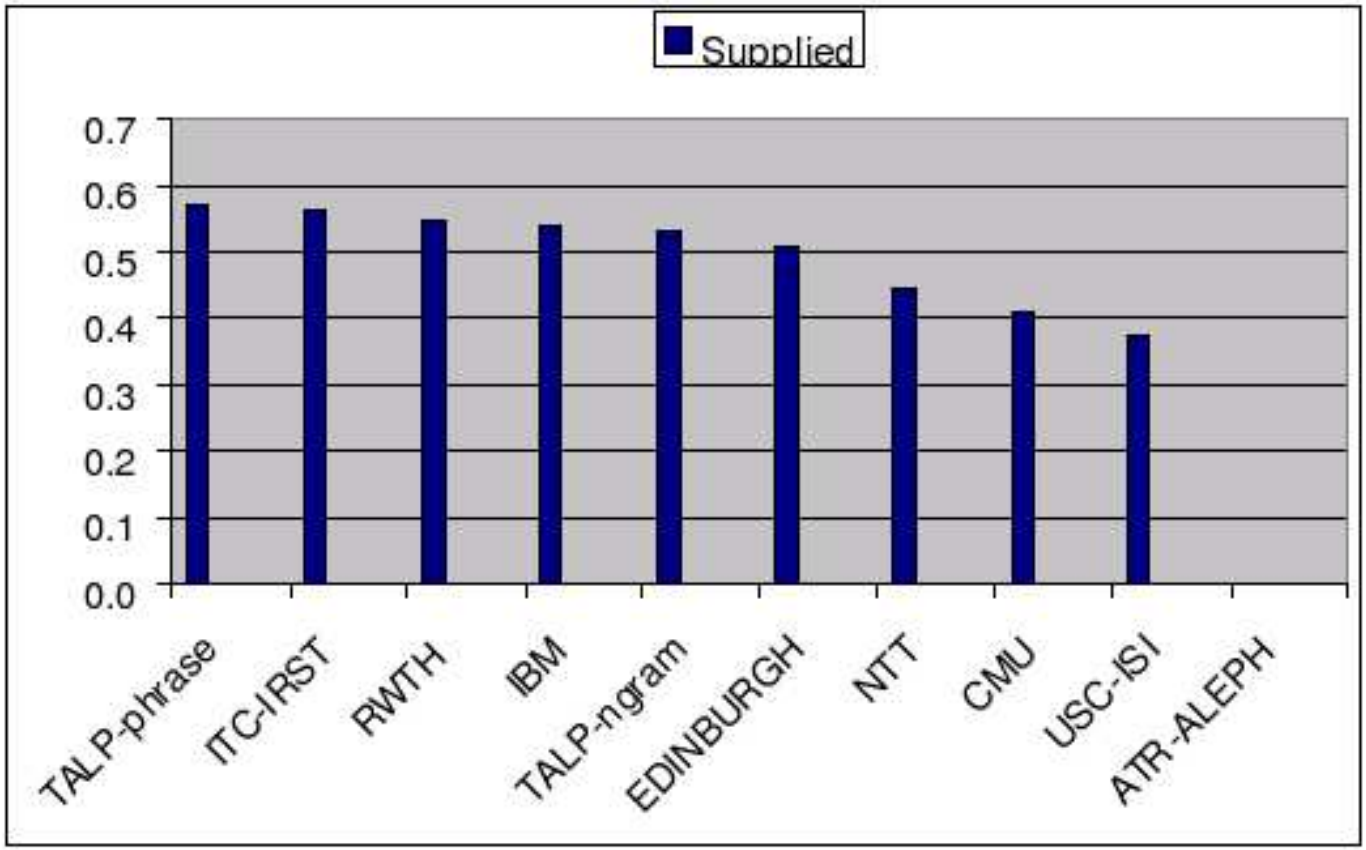
IWSLT 2005

Japanese ⇨ English – BLEU Scores



IWSLT 2005

Arabic ⇨ English – BLEU Scores



The End ... Thank You!