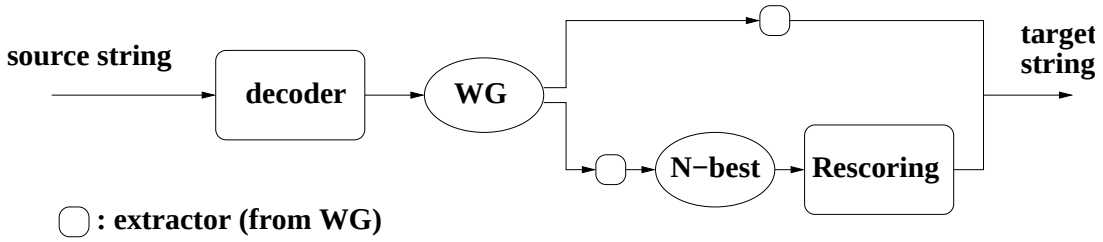
The ITC-irst SMT System for IWSLT-2006

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

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
38050 Povo (Trento), Italy

Two Pass Search Strategy



First Pass:

- Log-linear Phrase-based Model
- Dynamic programming algorithm
- Beam search decoder:
  - threshold and histogram pruning
- Non-monotone search constraints
  - max number of vacancies on the left (MVN)
  - max distance from left-most vacancy (MVD)

Second Pass:

- Extraction of 5,000-best
- Log-linear Model
- Re-ranking algorithm

Two Pass Search Strategy: First Pass

First Pass feature functions:

- Target 4-gram LM
- Fertility model target phrases
- Direct phrase-based lexicon
- Inverse phrase-based lexicon
- Direct word-based lexicon
- Inverse word-based lexicon
- Negative distortion
- Positive distortion
- $\tilde{e}_0$ fertility
- $\tilde{e}_0$ permutation
- Length Penalty

Training of Phrase-based model

Phrase-based model

- **Word-alignment: intersection alignment with expansion**
- **Phrase-extraction: max length 8, filtering**
- **Feature estimation: lexicon, fertility models (... by freq smoothing ...)**
- **Monotone and non-monotone search**

Training with additional alignments

Exploitation of additional word alignments [Chen & Federico, 2006]:

- IBM union word-alignments were added before phrase-extraction
- CLA word-alignments were added before phrase-extraction
  - IBM alignments are many-to-one
  - CLA alignments are one-to-one
  - CLA alignments have higher precision and lower recall

Experiments

- Task: Open Data Condition (only Supplied Resources)
- Lang: Chinese, Japanese, Arabic, Italian
- Dev set: devset4 text
- BLEU% and NIST: no-case without punctuation
- Weight optimization: Simplex algorithm
- Non-monotone search:
  - MVD=4 MVN=4 Arabic
  - MVD=6 MVN=6 Chinese
  - MVD=8 MVN=8 Japanese
- Monotone search:
  - MVD=0 MVN=0 Italian

Preprocessing and Postprocessing

- Preprocessing

Preprocessing	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	Chinese	English	Japanese	English	Arabic	English	Italian	English
tokenization	x	x	x	x	x	x	x	x
txt-to-digit	x	x	–	–	–	–	x	x
lower-casing	–	x	–	x	–	x	x	x

- Postrocessing
  - Punctuation restoration with hidden-ngram tool
  - Case restoration with disambig tool
  - Both are applied on the target side

Experimental Results: First Pass

Results of the optimization tecniques on the dev set
(BLEU% score and NIST score; without case nor punctuation)

System	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	16.42	5.800	15.63	5.894	23.07	5.825	41.57	8.865
+CLA alignment	16.85	5.977	16.40	6.000	23.50	5.978	41.97	8.873
+Union IBM	16.99	6.171	17.01	6.087	23.90	6.023	–	–
+non-monotonic	18.87	6.388	19.57	6.528	24.35	6.112	–	–

Two Pass Search Strategy

Second Pass feature functions:

- Direct IBM model 1 and 3 lexicon score
- CLA lexicon score [Chen, etc. IWSLT'2005]
- Question feature [Chen, etc. IWSLT'2005]
- Frequency of n-grams within n-best [Chen, etc. IWSLT'2005]
- Ratio of target source lengths
- 2,3,5-gram target LM
- N-gram post-probabilities [Zens, 2006]
- Sentence length posteriers [Zens, 2006]
- Phrase penalty
- Inverse IBM model 1 and 3 lexicon score
- Word/block reordering probabilities

Reordering Rules

Word/block reordering probabilities [Chen, etc. 2006]

- **Word/block reordering rules consist of two sides:**
  - **lhs: word-based pattern**
  - **rhs: possible reordering of the pattern**
- **Rules are extracted from the word aligned training data**
- **Rules are weighted according to observed statistics**
- **Compute the score:**

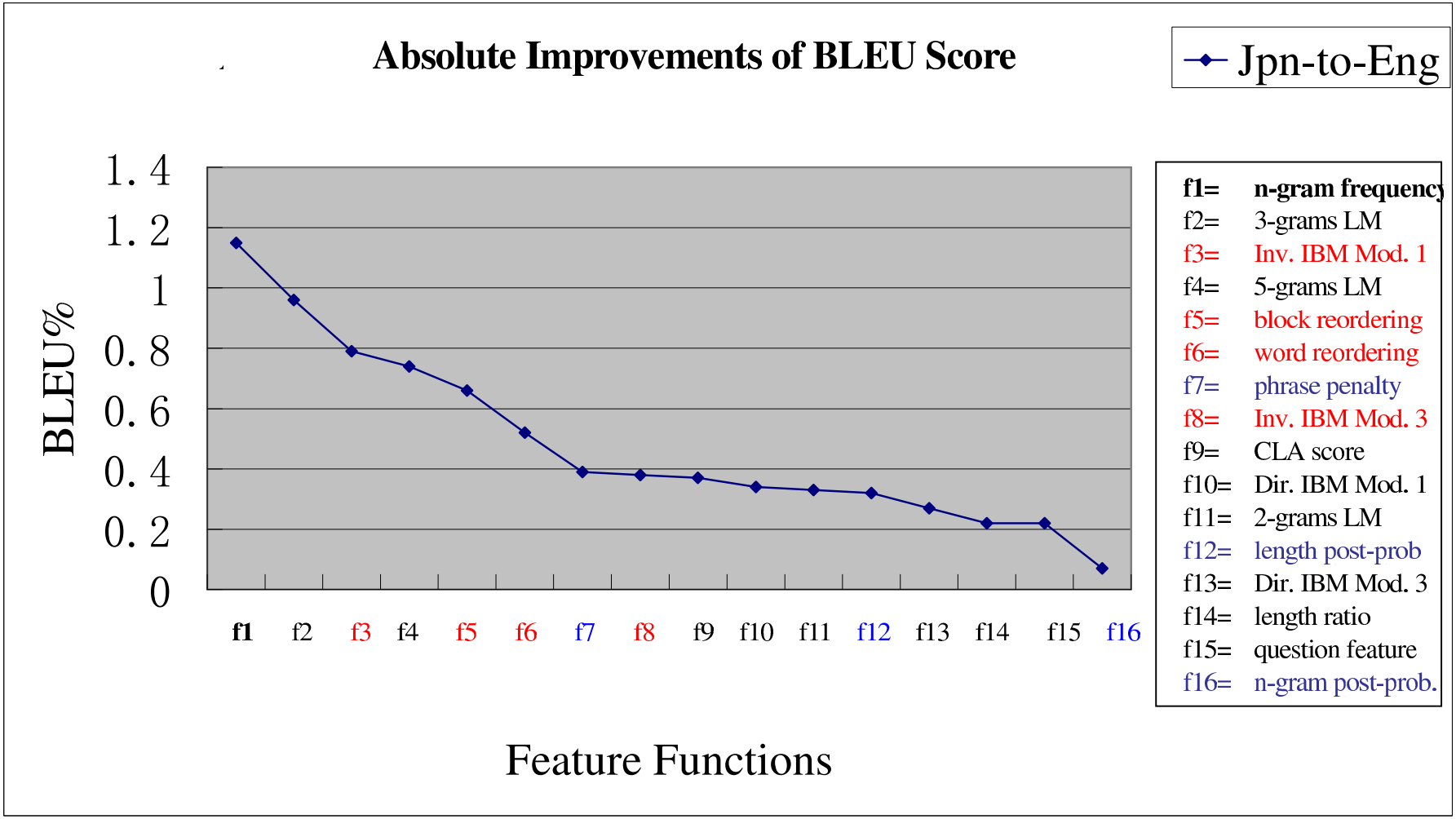
$$h_{\texttt{rules}}(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a}) = \frac{1}{K} \sum_{i=1}^{K} \log \Pr(r_i) \qquad \textbf{(1)}$$

where r_i is a matching rule, $\Pr(r_i)$ its probability and K the number of the reordering patterns matching the given source/target pair.

(detailed description will be given in the afternoon's presentation)

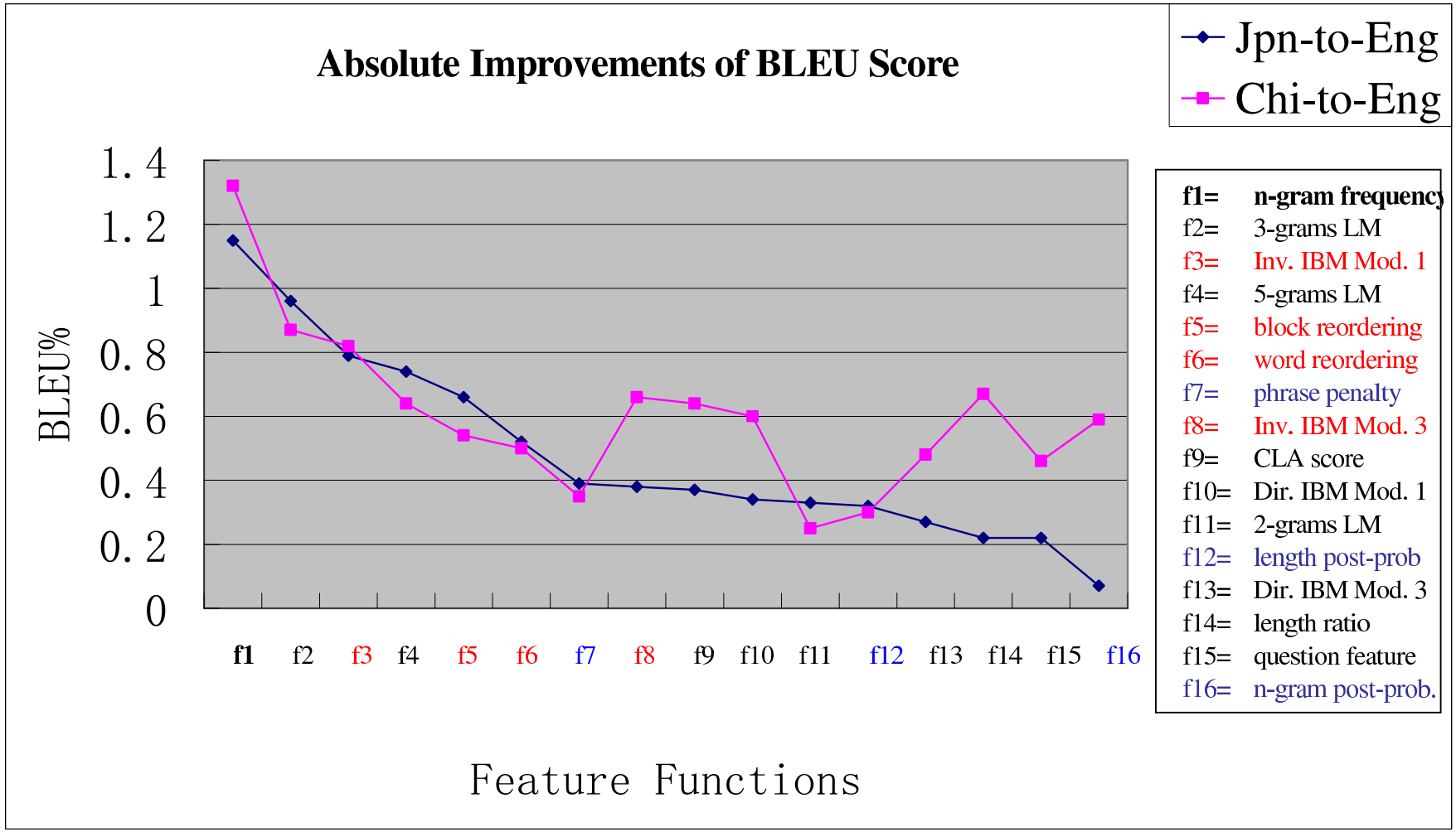
Experimental Results: Re-scoring Stage (1)

Contribution of each feature function on Japanese Dev set (text input)



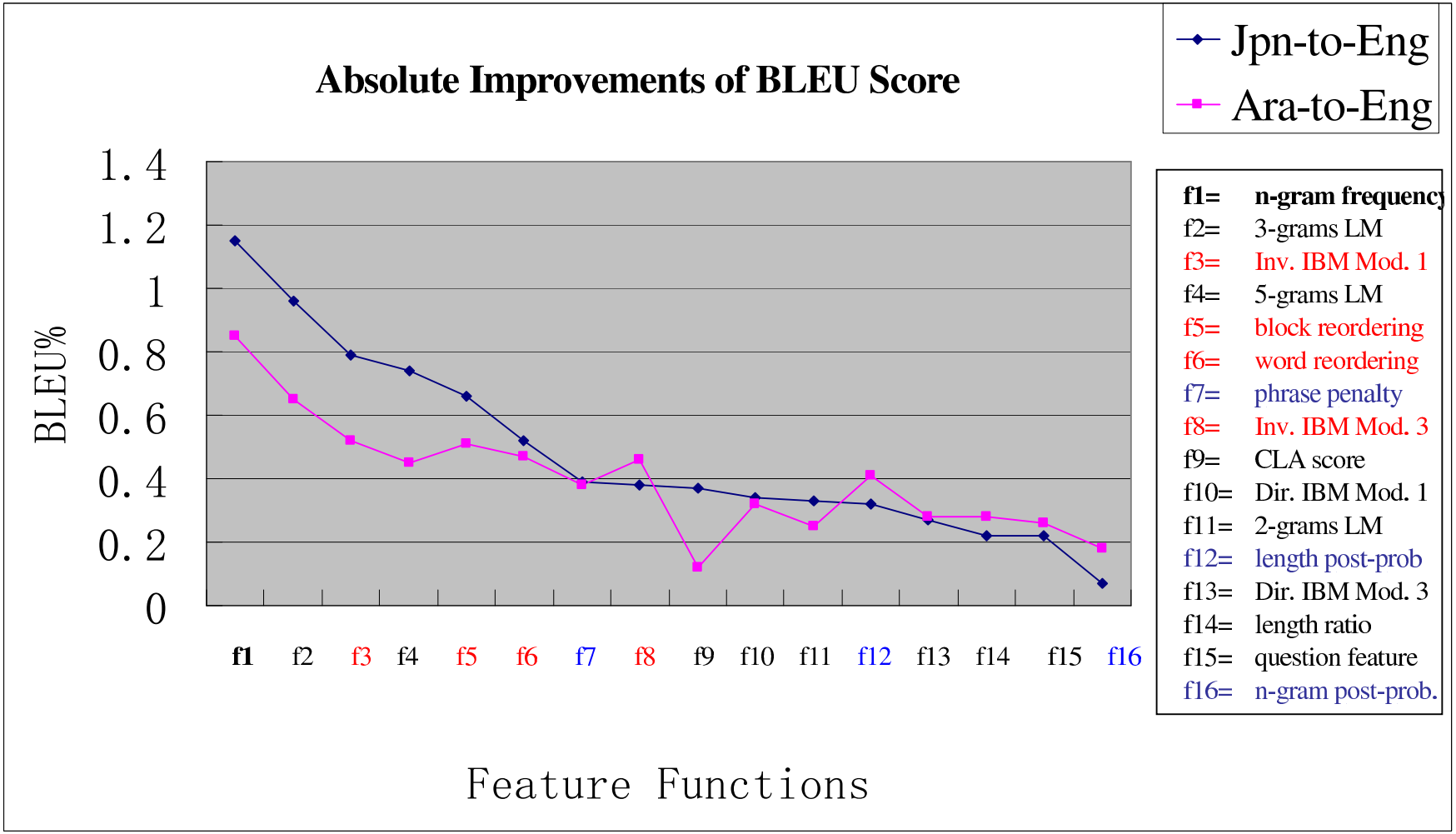
Experimental Results: Re-scoring Stage (2)

Contribution of each feature function on Chinese Dev set (text input)



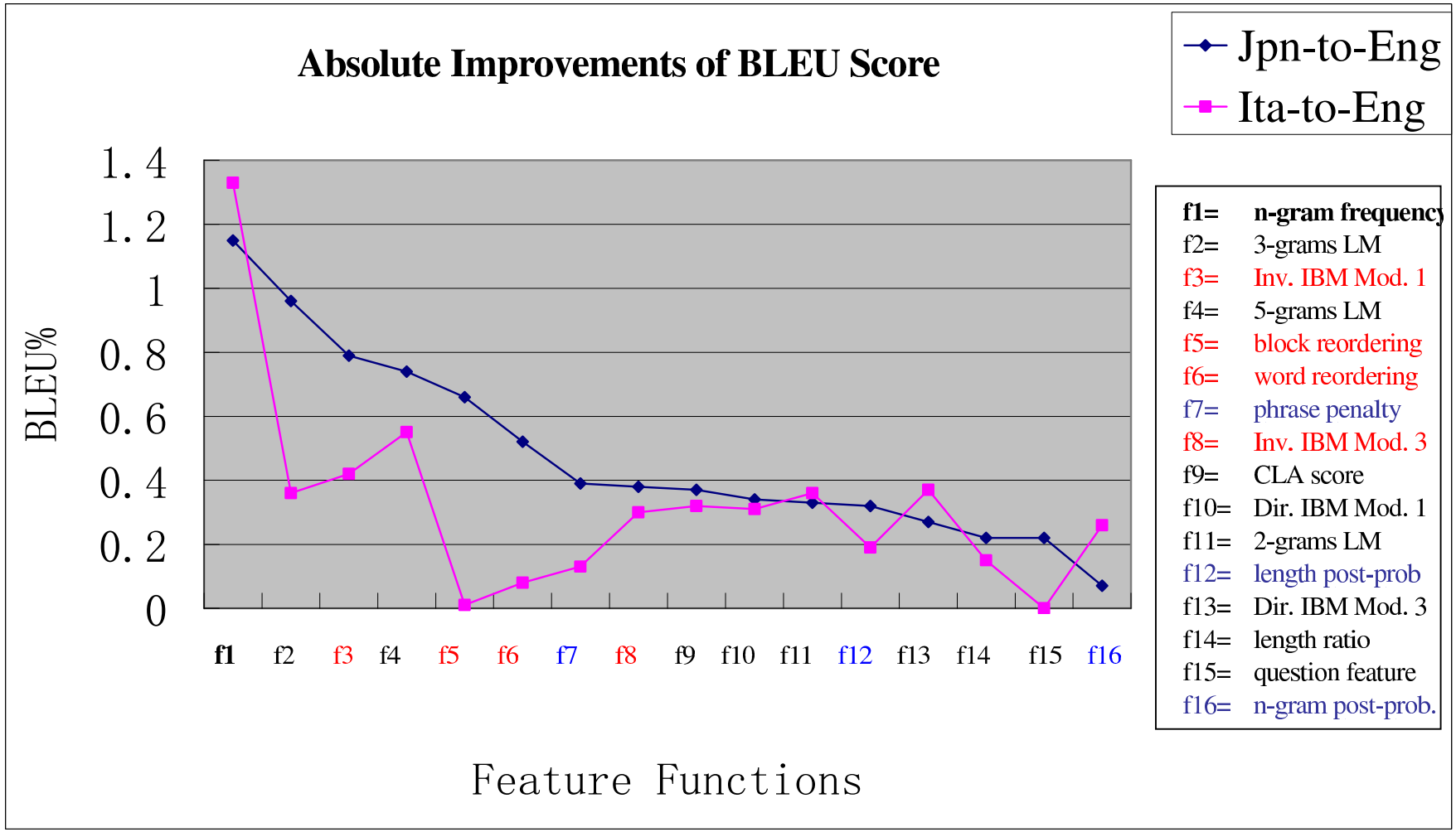
Experimental Results: Re-scoring Stage (3)

Contribution of each feature function on Arabic Dev set (text input)



Experimental Results: Re-scoring Stage (4)

Contribution of each feature function on Italian Dev set (text input)



Experimental Results: Re-scoring Stage

– Official scores of the 1-pass and 2-pass on the test sets (BLEU% and NIST scores; without case nor punctuation).

Language	Test Set	BLEU%		NIST	
		1-pass	2-pass	1-pass	2-pass
Chi-Eng	Text	16.94	19.92	6.043	6.426
	Read Speech	14.61	16.98	5.272	5.744
	Spont. Speech	13.44	15.77	4.961	5.480
Jpn-Eng	Text	16.61	18.82	6.010	6.320
	Read Speech	14.34	16.17	5.508	5.833
Ara-Eng	Text	19.75	20.48	5.524	5.604
	Read Speech	17.05	17.80	5.011	5.190
Ita-Eng	Text	37.10	37.97	8.489	8.619
	Read Speech	28.78	29.69	7.176	7.260

Comparison of Systems

Comparison of the ITC-irst systems of the years 2005 and 2006 for the supplied data track (2005) and open data track (2006) on IWSLT04 and IWSLT05 test sets.

	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng	
BLEU%	2005	2006	2005	2006	2005	2006
IWSLT'04	46.37	53.23	50.11	54.01	56.37	58.14
IWSLT'05	52.75	59.91	43.13	54.83	56.22	57.57

Conclusions

- Use of alternative word-alignment models improve the translation performance
  - IBM Intersection alignment
  - CLA word alignment
  - IBM Union alignment
- New feature functions used for n-best re-scoring:
  - Inverse IBM model 1 and 3
  - Word/block reordering rules

The End ... Thank You!