Spoken Language Communication Group

Spoken Language Communications Research Laboratories

# Reducing Human Assessment of Machine Translation Quality to Binary Classifiers

September 8, 2007

Michael Paul, Andrew Finch, Eiichiro Sumita

NICT Spoken Language Communication Group,
ATR Spoken Language Communication Research Laboratories
Kyoto, Japan

# Assessment of Machine Translation Quality



| | Document | Sentence |
|---|---|---|
| | • set of sentence translations | • translation of a single input |
| Human | • average of sentence-level grades | • discrete evaluation grade<br>• median score of multiple human grades |
| | | ***metrics***: fluency, adequacy, … |
| Machine | • comparison to (multiple) reference translations<br>• assign single numerical score | • confidence estimation<br>• machine learning approach to predict human grades |
| | ***metrics***: BLEU, METEOR, … | ***classifiers***: SVM, DT, … |

# Assessment of Machine Translation Quality

| | Document | Sentence |
|---|---|---|
| **Usage** | • evaluation of MT system development progress<br>• MT system comparison (NIST, IWSLT, …) | • usability of given translation in a real-world application (post-editing, dialog translation, …) |
| **Problems** | • quality/coverage of reference translations<br>• "meaning" of (numerical) automatic evaluation scores | • **complexity of evaluation task (multi-class classification)**<br>• granularity of evaluation grades |

# Outline of Talk

## 1. Prediction of Sentence-Level Translation Quality:

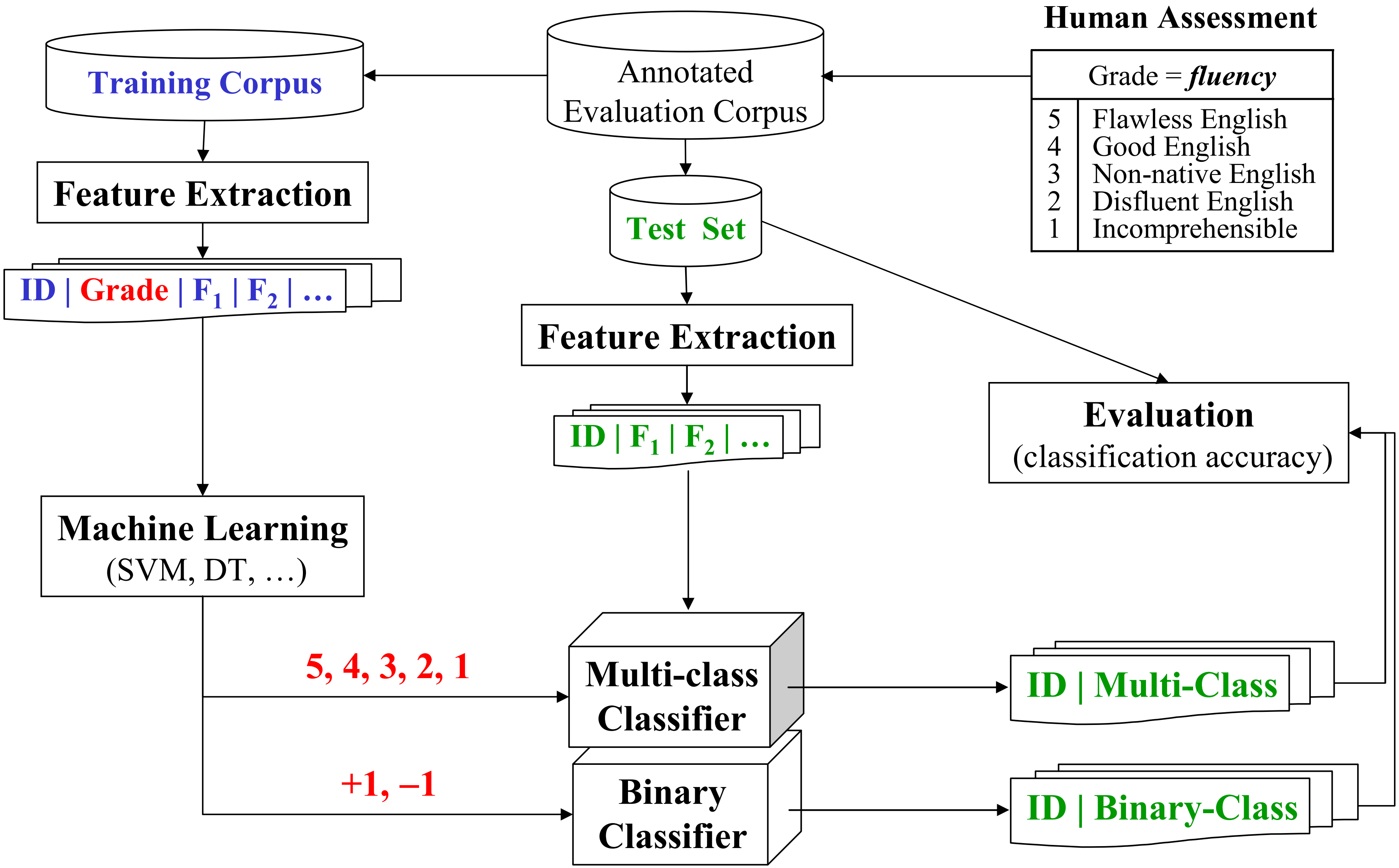
- decompose multi-class to binary classification
  - a *coding matrix*
- learn set of binary classifiers
  - feature selection, standard machine learning techniques
- predict multi-class label
  - compare binary classification results to *coding matrix*

## 2. Experimental Results:

- large-scale human-annotated evaluation corpus
- coding matrix optimization
- classification accuracy
- correlation to human assessments

# Prediction of Sentence-Level Translation Quality



# Classification Task



**Goal:** predict human evaluation grade (*fluency*, *adequacy*, …) for a given translation

→ **multi-class label**

**Multi-Class Classification:**

☺ direct prediction of multi-class label

☹ classification accuracy is low

**Binary-Class Classification:**

☺ classification accuracy is high

☹ multi-class label cannot be derived reliably

# Proposed Solution



## Reduction of Classification Complexity:

- decompose multi-class task into a set of binary classification problems
- apply standard learning algorithm to train binary classifiers
- combine results of binary classifiers using a "***coding matrix***" to predict multi-class label

→ **increase in classification accuracy**

→ **independent from learning algorithm**

# Proposed Solution



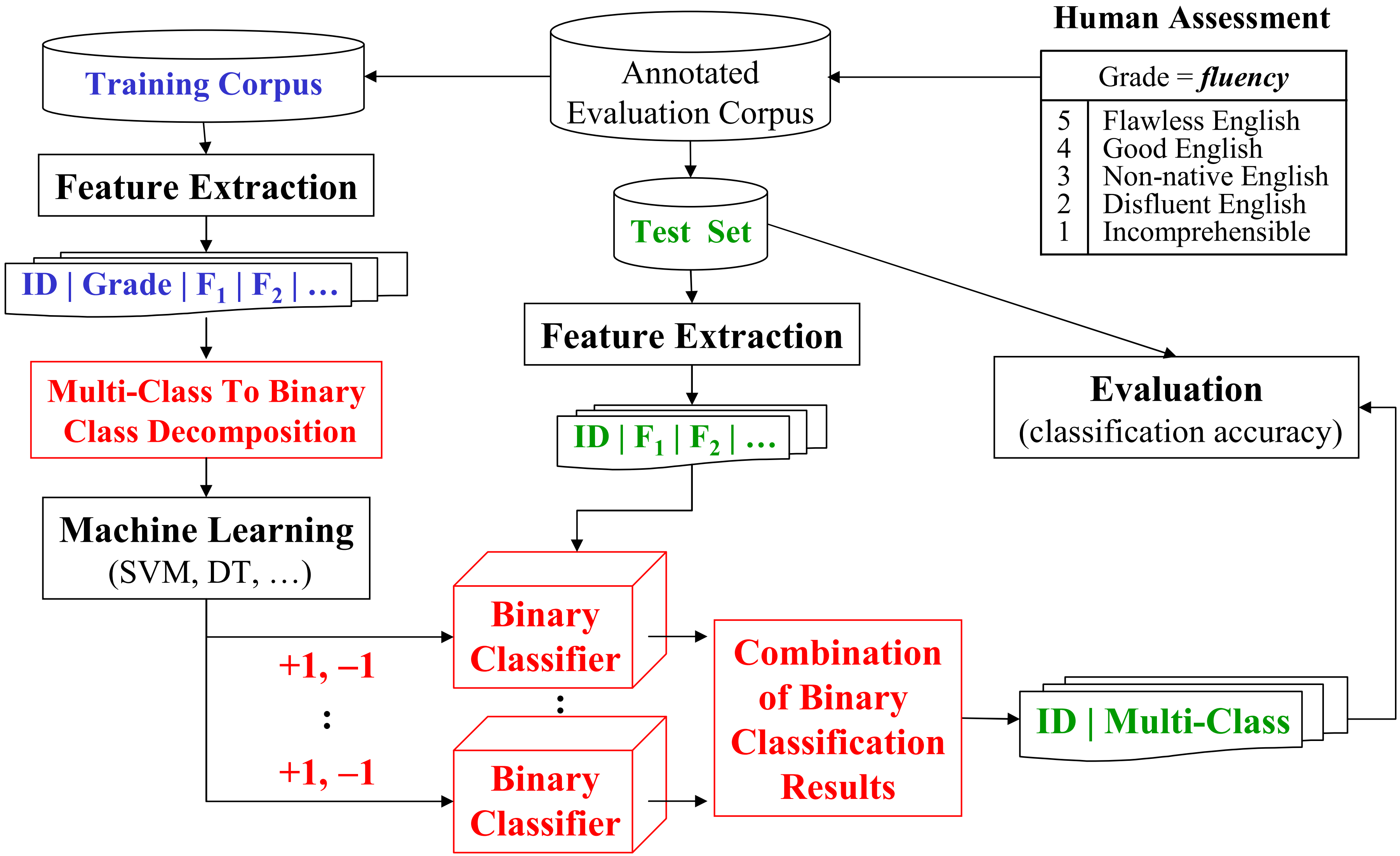
**Feature Selection for Translation Quality Prediction:**

- multiple automatic evaluation metric scores
  - BLEU
  - NIST
  - METEOR
  - WER
  - PER
  - TER
  - GTM
- metric-internal features
  - ngram-prec
  - length ratio
  - …

→ **takes into account different aspects of MT quality**

→ **independent from target language and MT system**

# Prediction of Sentence-Level Translation Quality



# Proposed Method



## 1. Decomposition Phase:

- **decompose multi-class** into set of binary classification tasks:

  ◦ ***one-against-all*** (**5**, **4**, **3**, **2**, **1**):

  *Example*:

  **5** : +1 → all training examples tagged with grade *5*

  –1 → all training examples tagged with grade *4* or *3* or *2* or *1*)

  ◦ ***boundary*** (54_321 , 543_21):

  *Example*:

  **54_321** : +1 → all training examples tagged with grade *5* or *4*

  –1 → all training examples tagged with grade 3 or 2 or 1

  ◦ ***all-pairs*** (**5_4** , **5_3** , **5_2** , **5_1** , **4_3** , **4_2** , **4_1** , **3_2** , **3_1** , **2_1**):

  *Example*:

  **5_4** : +1 → all training examples tagged with grade *5*

  –1 → all training examples tagged with grade *4*

# Proposed Method



## 2. Learning Phase:

- **learn binary classifier** for each decomposition task
  - ***feature set selection/extraction***

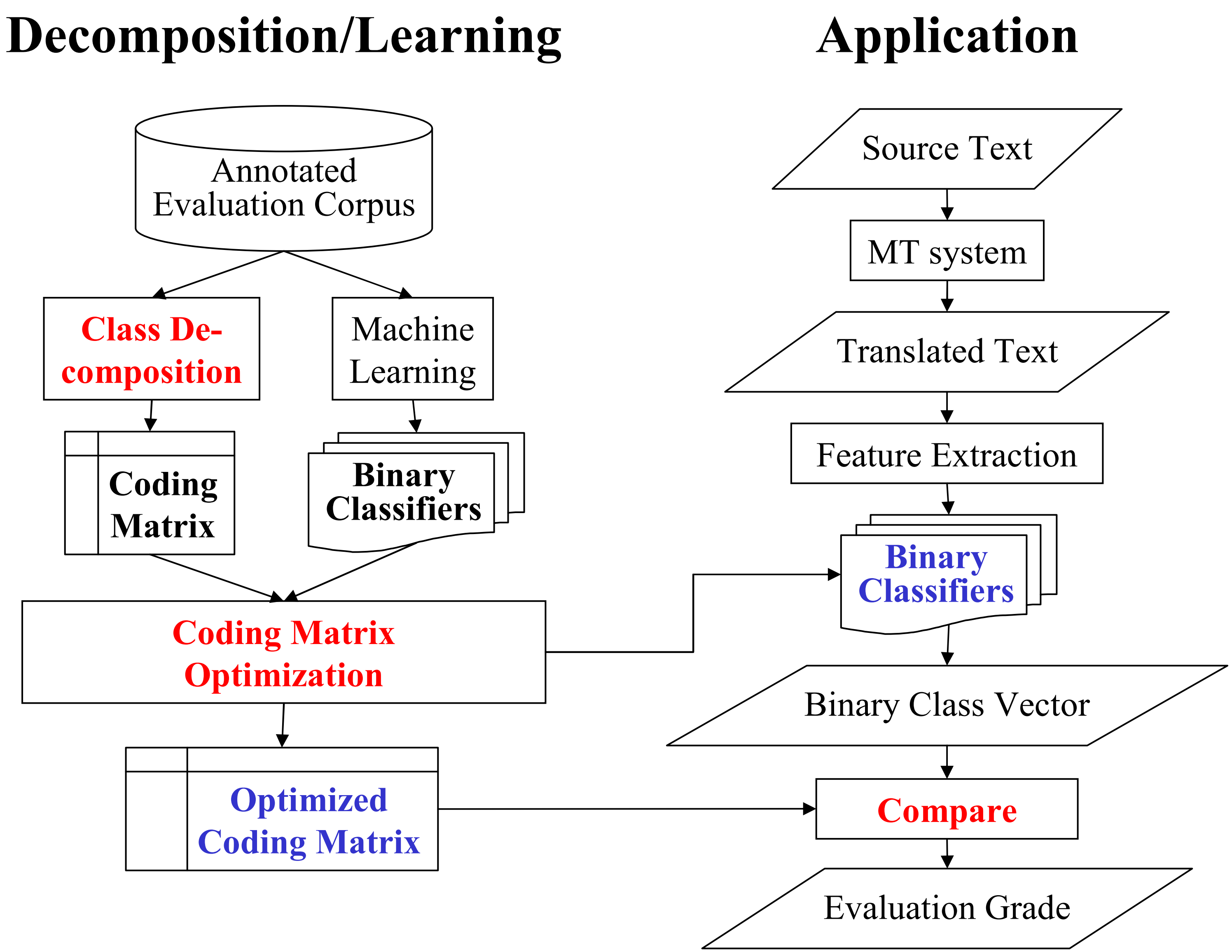
    *(exp)*: + 54 features (7 autom. eval. scores + metric-internal features)
  - ***classifier training***

    *(exp)*: + *fluency/adequacy*, DT classifier (+ SVM classifier)
- **identify optimal subset of binary classifiers**
- create ***coding matrix***
  - ***column***: class of pos./neg. training examples (for given binary classifier)
  - ***row***: correct binary classification result (for a given multi-class label)

## 3. Application Phase:

- apply all binary classifiers to given input → ***classification vector v***
- match *v* against *coding matrix* rows to identify ***multi-class label***

# Outline of Proposed Method



Spoken Language Communication Group

# Coding Matrix

Spoken Language Communications Research Laboratories

$$M = (m_{i,j})_{i=1,\dots,k\,;\,j=1,\dots,l}$$

$$m_{i,j} \in \{+1, -1, 0\}$$

( k=3 , l=3 )

***one-against-all***

| | $c_1 \bullet c_2\, c_3$ | $c_2 \bullet c_1\, c_3$ | $c_3 \bullet c_1\, c_2$ |
|---|---|---|---|
| $c_1$ | $+1$ | $-1$ | $-1$ |
| $c_2$ | $-1$ | $+1$ | $-1$ |
| $c_3$ | $-1$ | $-1$ | $+1$ |

***all-pairs***

| | $c_1 \bullet c_2$ | $c_1 \bullet c_3$ | $c_2 \bullet c_3$ |
|---|---|---|---|
| $c_1$ | $+1$ | $+1$ | $0$ |
| $c_2$ | $-1$ | $0$ | $+1$ |
| $c_3$ | $0$ | $-1$ | $-1$ |

# Combination of Binary Classifiers using a Coding Matrix

input

| | $c_1 \bullet c_2$ | $c_1 \bullet c_3$ | $c_2 \bullet c_3$ |
|---|---|---|---|
| ***v*** | +1 | +1 | –1 |

***all-pairs***

| | $c_1 \bullet c_2$ | $c_1 \bullet c_3$ | $c_2 \bullet c_3$ |
|---|---|---|---|
| $c_1$ | +1 | +1 | 0 |
| $c_2$ | –1 | 0 | +1 |
| $c_3$ | 0 | –1 | –1 |

**Hamming Distance**
(number of positions that differ)

**= 3**

# Combination of Binary Classifiers using a Coding Matrix

**input**

| | $c_1 \bullet c_2$ | $c_1 \bullet c_3$ | $c_2 \bullet c_3$ |
|---|---|---|---|
| ***v*** | +1 | +1 | −1 |

***all-pairs***

| | $c_1 \bullet c_2$ | $c_1 \bullet c_3$ | $c_2 \bullet c_3$ |
|---|---|---|---|
| $c_1$ | +1 | +1 | 0 |
| $c_2$ | −1 | 0 | +1 |
| $c_3$ | 0 | −1 | −1 |

| distance | select |
|---|---|
| 1 | |
| 3 | $c_1$ |
| 2 | |

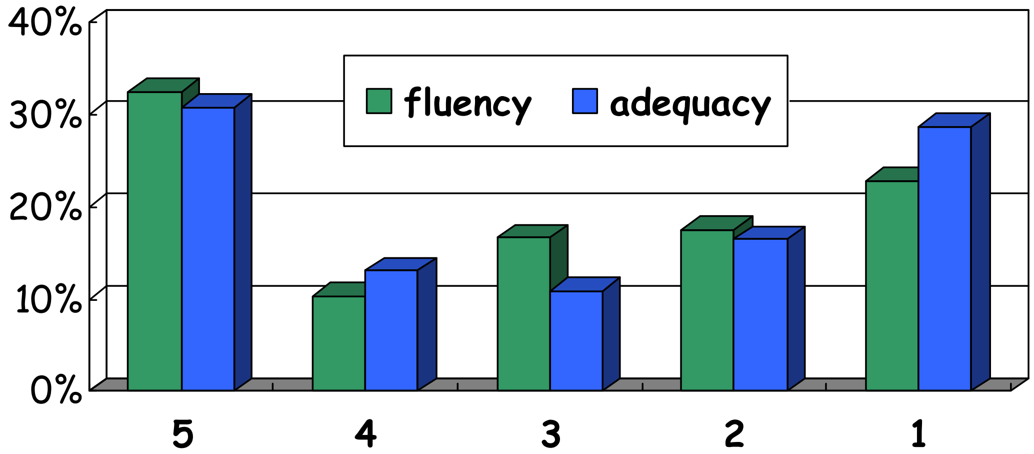
Spoken Language Communication Group

# Evaluation Corpus

Spoken Language Communications Research Laboratories

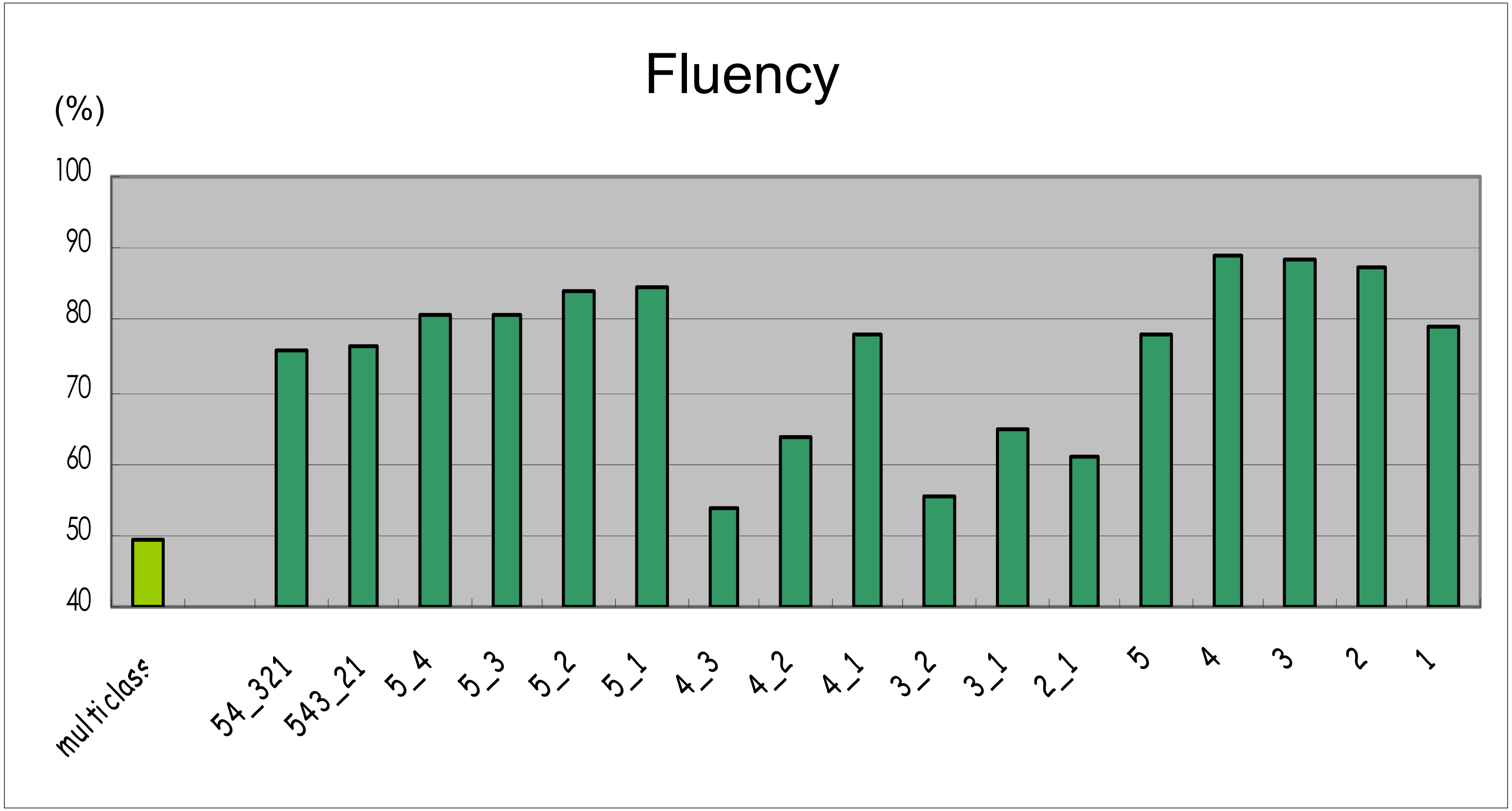
## Basic Travel Expression Corpus (BTEC):

- 36K English translations of 4K Japanese/Chinese inputs
- human assessments and automatic evaluation scores




| sentence count | fluency/ adequacy |
|---|---|
| *training* | 25,988 |
| *develop* | 2,024 ( 4 MT x 506) |
| *test* | 7,590 (15 MT x 506) |

# Coding Matrix Optimization
## (classification accuracy on DEV set)



# Coding Matrix Optimization (classification accuracy on DEV set)

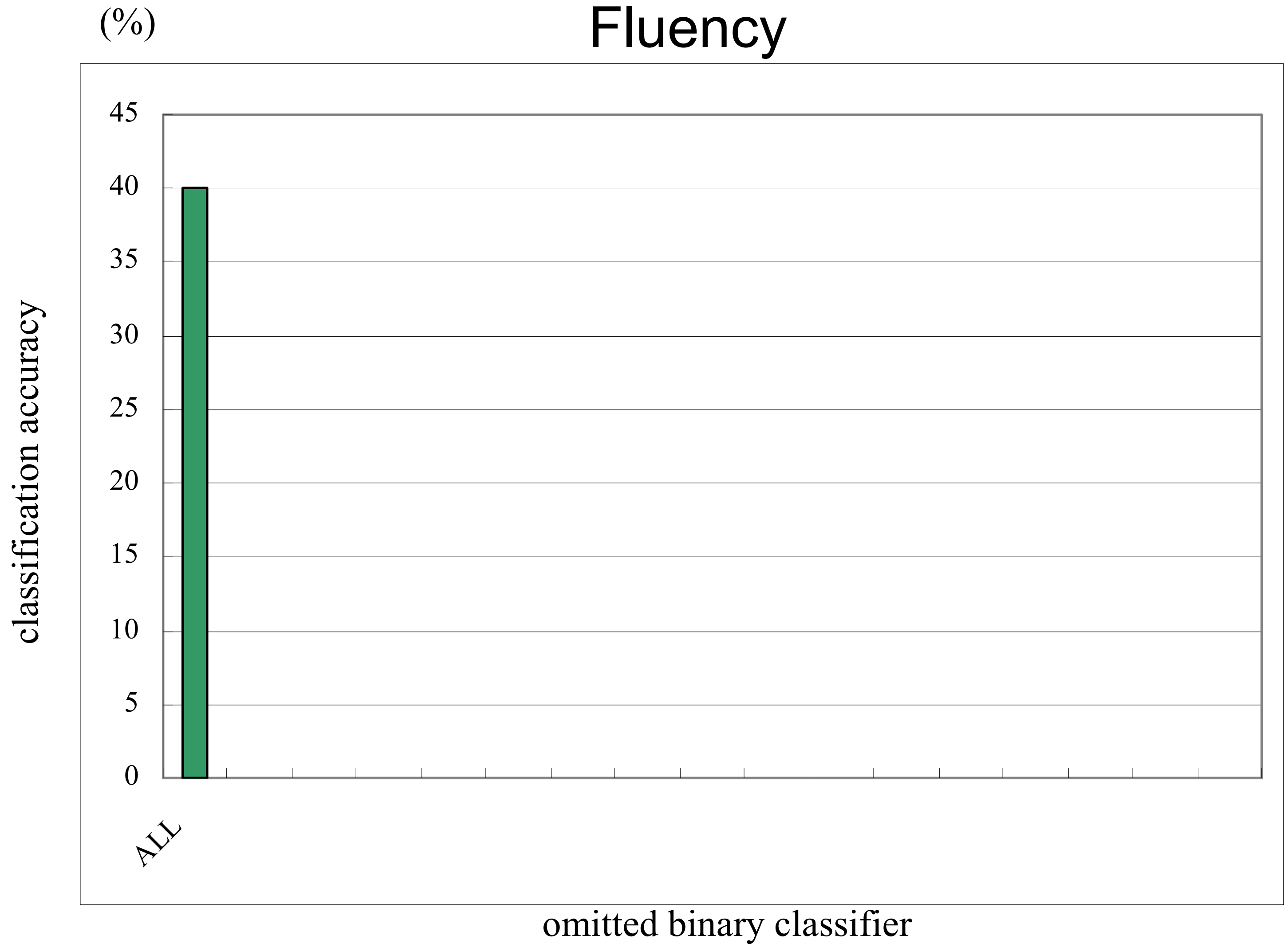
Coding Matrix

| | 54_ 321 | 543 _21 | 5_4 | 5_3 | 5_2 | 5_1 | 4_3 | 4_2 | 4_1 | 3_2 | 3_1 | 2_1 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5** | +1 | +1 | +1 | +1 | +1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | +1 | –1 | –1 | –1 | –1 |
| **4** | +1 | +1 | –1 | 0 | 0 | 0 | +1 | +1 | +1 | 0 | 0 | 0 | –1 | +1 | –1 | –1 | –1 |
| **3** | –1 | +1 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | +1 | +1 | 0 | –1 | –1 | +1 | –1 | –1 |
| **2** | –1 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | –1 | 0 | +1 | –1 | –1 | –1 | +1 | –1 |
| **1** | –1 | –1 | 0 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | –1 | –1 | –1 | –1 | –1 | –1 | +1 |

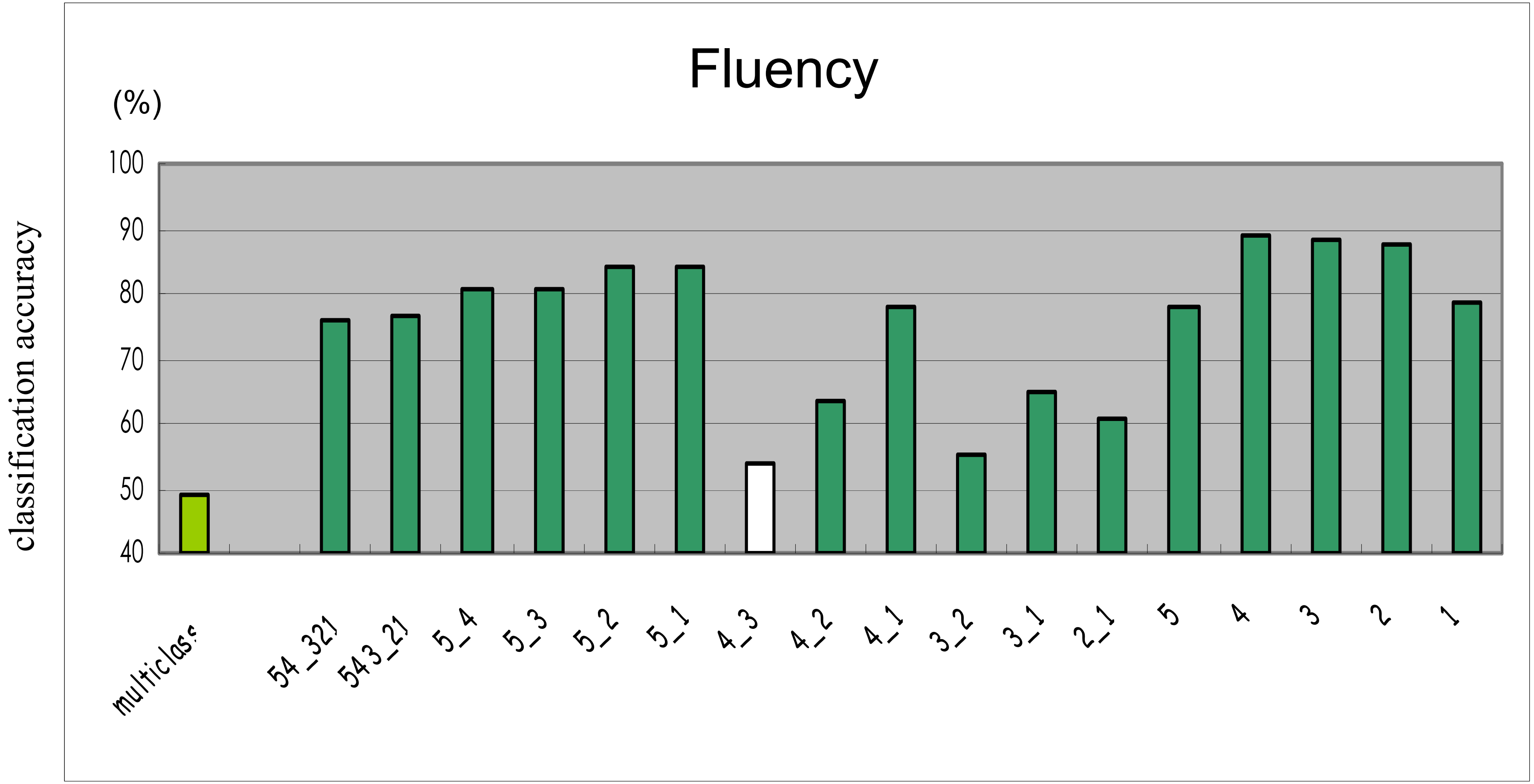
# Coding Matrix Optimization

## (omission of worst-performing classifier)





# Coding Matrix Optimization
## (classification accuracy on DEV set)



# Coding Matrix Optimization
## (classification accuracy on DEV set)

Coding Matrix

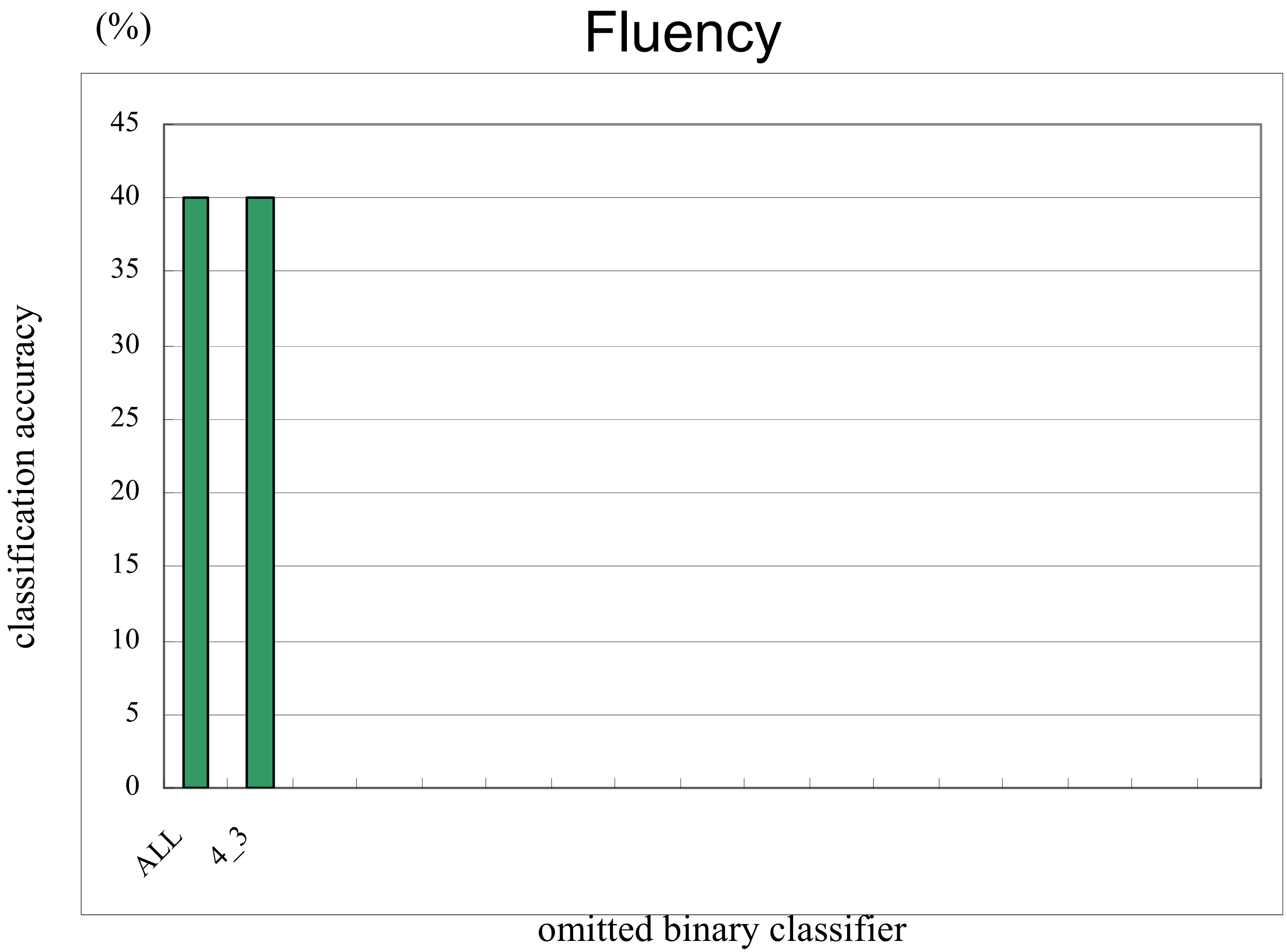
| | 54_321 | 543_21 | 5_4 | 5_3 | 5_2 | 5_1 | 4_3 | 4_2 | 4_1 | 3_2 | 3_1 | 2_1 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5** | +1 | +1 | +1 | +1 | +1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | +1 | –1 | –1 | –1 | –1 |
| **4** | +1 | +1 | –1 | –1 | 0 | 0 | +1 | +1 | +1 | 0 | 0 | 0 | –1 | +1 | –1 | –1 | –1 |
| **3** | –1 | +1 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | +1 | +1 | 0 | –1 | –1 | +1 | –1 | –1 |
| **2** | –1 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | –1 | –1 | +1 | –1 | –1 | –1 | +1 | –1 |
| **1** | –1 | –1 | 0 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | –1 | –1 | –1 | –1 | +1 |

Spoken Language Communication Group

Spoken Language Communications Research Laboratories

# Coding Matrix Optimization

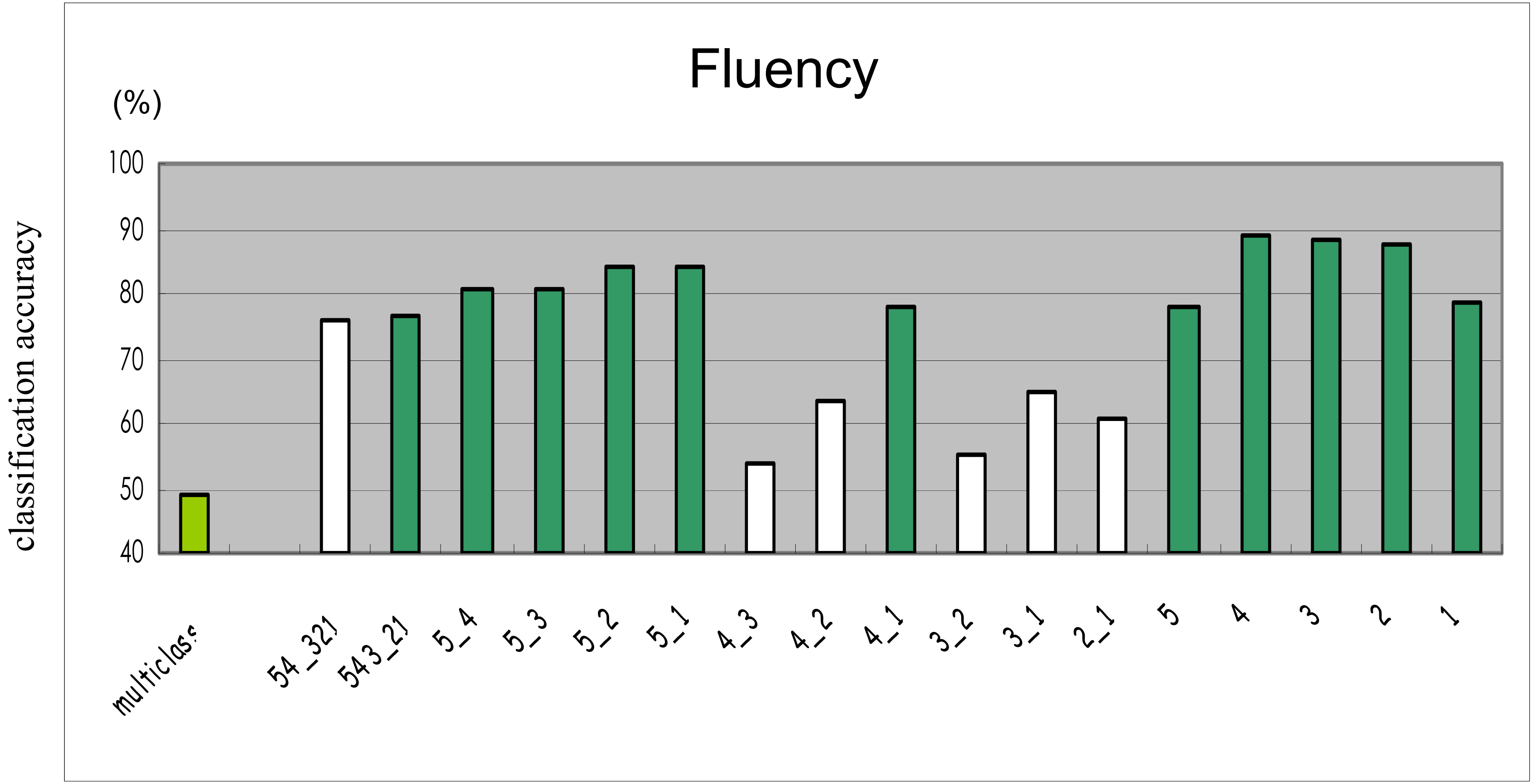
## (omission of worst-performing classifier)



# Coding Matrix Optimization

## (classification accuracy on DEV set)



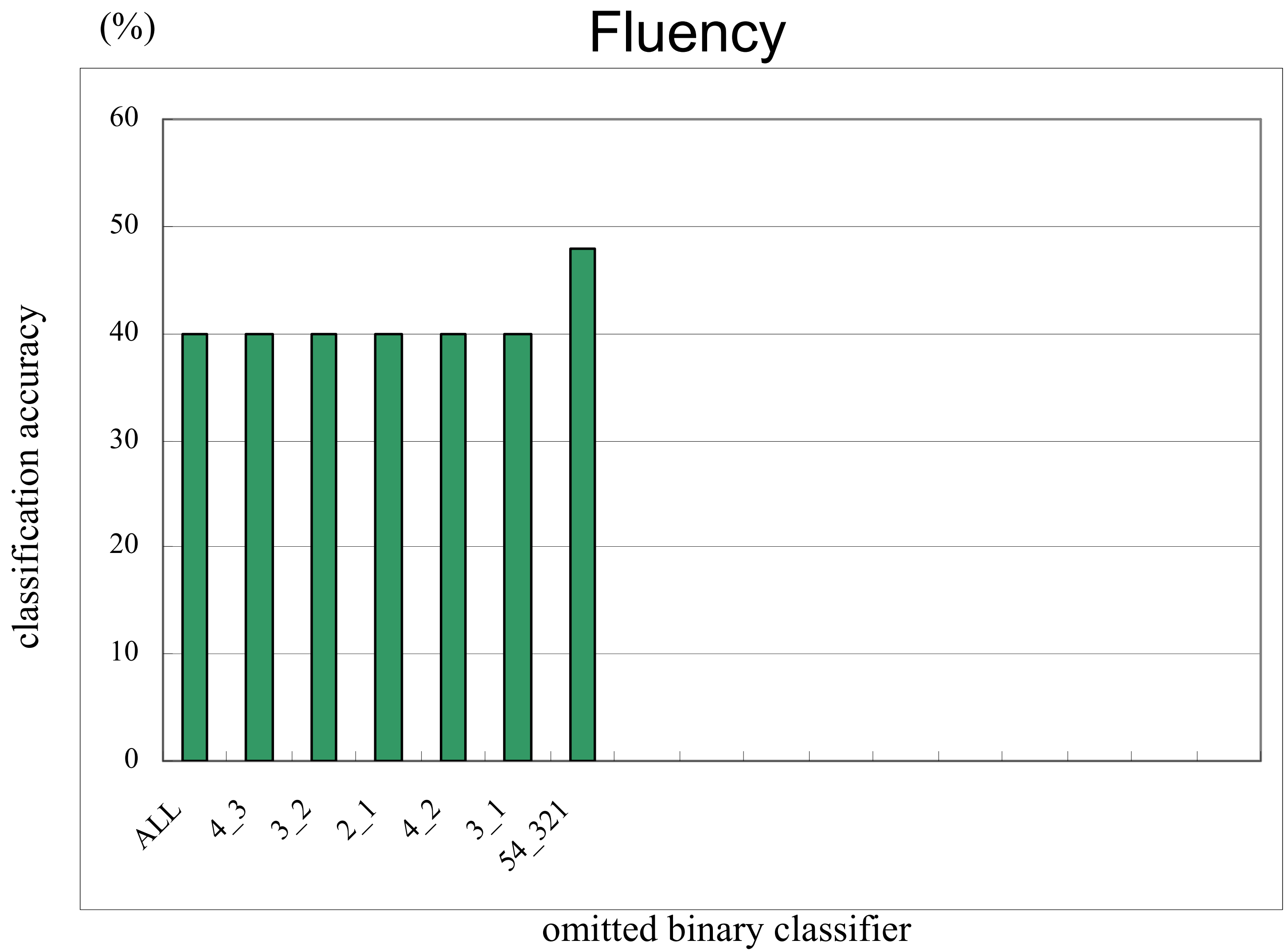


# Coding Matrix Optimization
## (classification accuracy on DEV set)

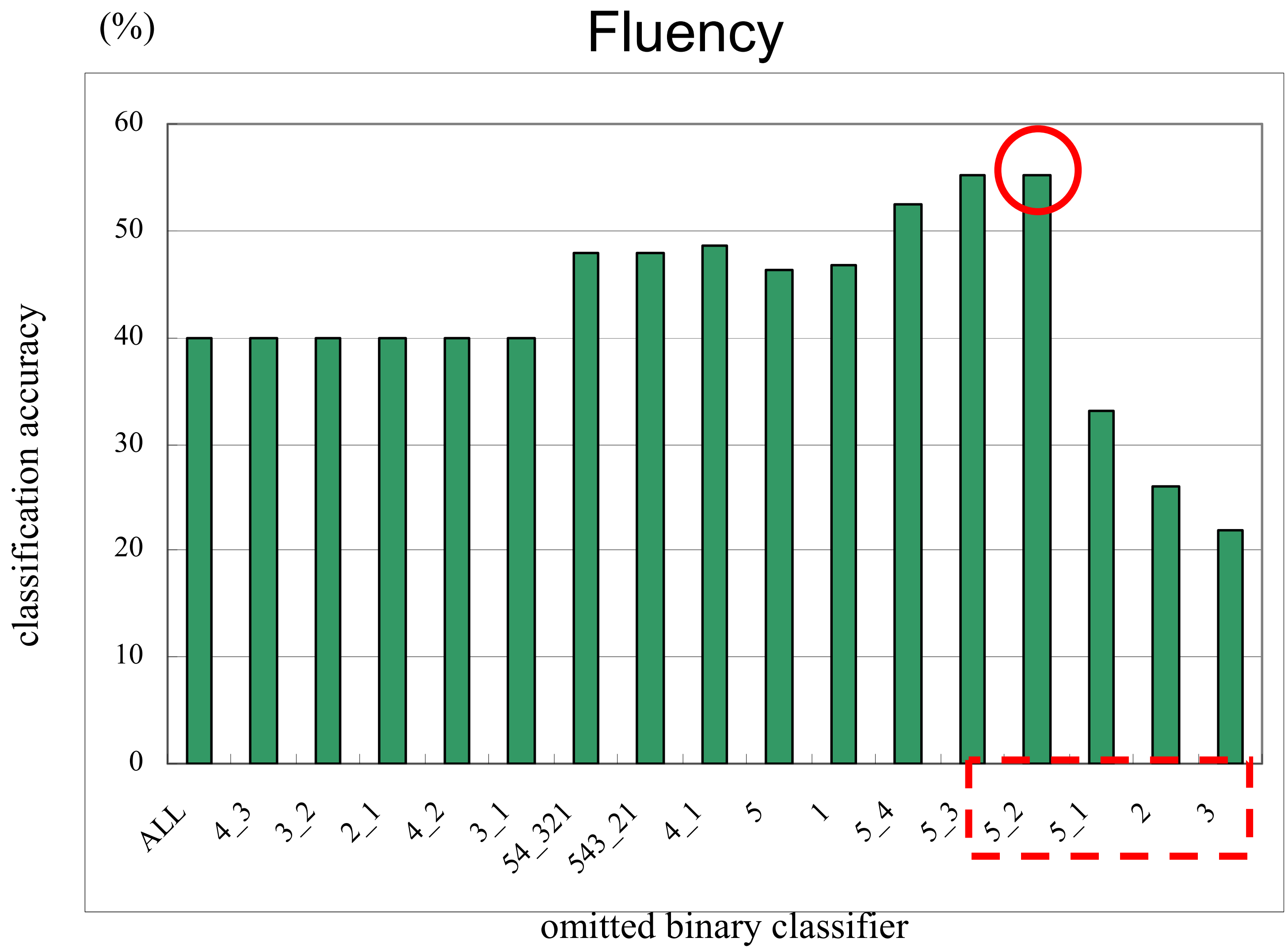
Coding Matrix

| | 54_321 | 543_21 | 5_4 | 5_3 | 5_2 | 5_1 | 4_3 | 4_2 | 4_1 | 3_2 | 3_1 | 2_1 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5** | +1 | +1 | +1 | +1 | +1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | +1 | –1 | –1 | –1 | –1 |
| **4** | +1 | +1 | –1 | –1 | 0 | 0 | +1 | +1 | +1 | 0 | 0 | 0 | –1 | +1 | –1 | –1 | –1 |
| **3** | –1 | +1 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | +1 | +1 | 0 | –1 | –1 | +1 | –1 | –1 |
| **2** | –1 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | –1 | –1 | +1 | –1 | –1 | –1 | +1 | –1 |
| **1** | –1 | –1 | 0 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | –1 | –1 | –1 | –1 | +1 |

# Coding Matrix Optimization
## (omission of worst-performing classifier)



# Coding Matrix Optimization
## (omission of worst-performing classifier)

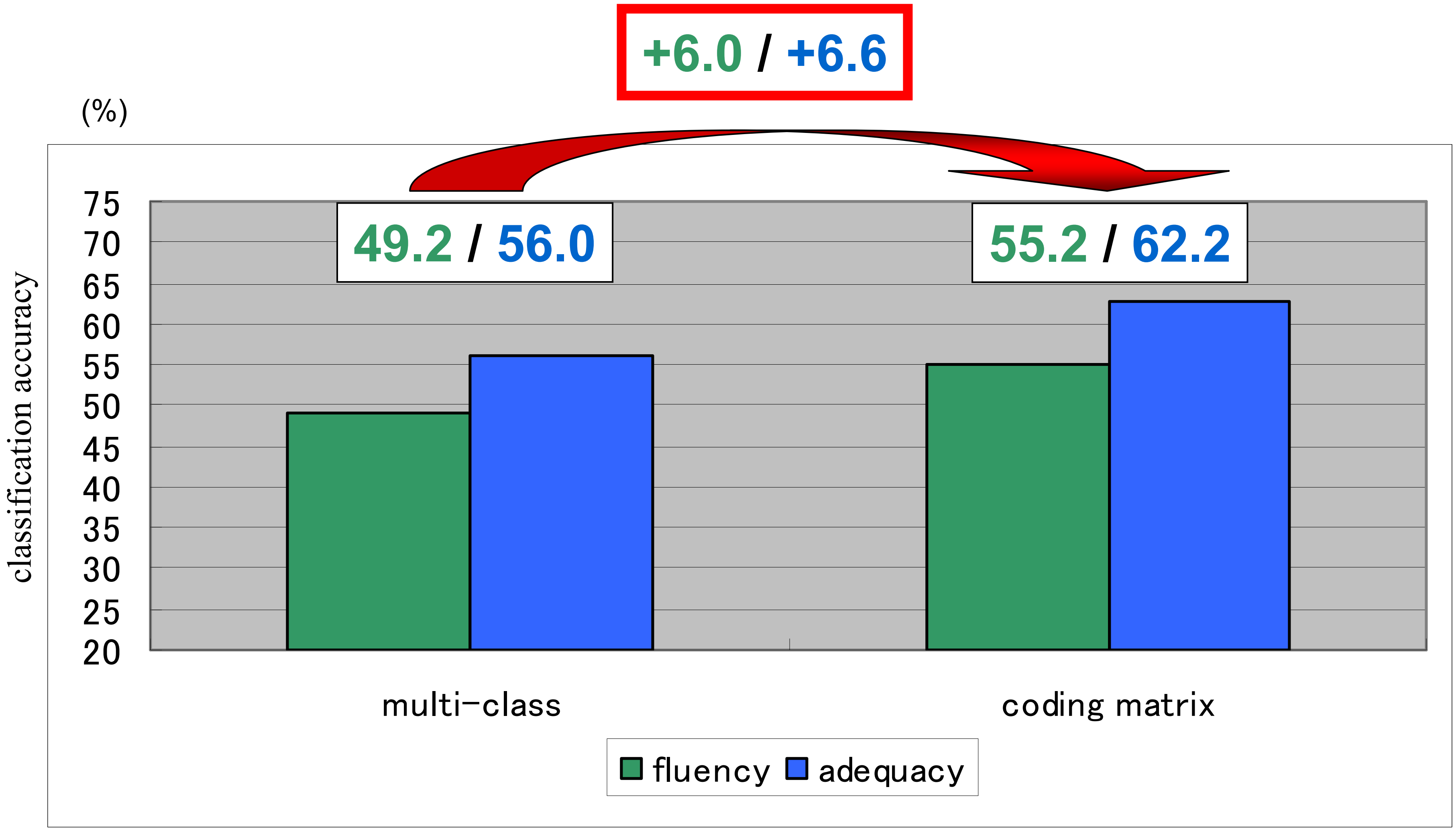


# Coding Matrix Optimization
## (classification accuracy on DEV set)

Optimized Coding Matrix

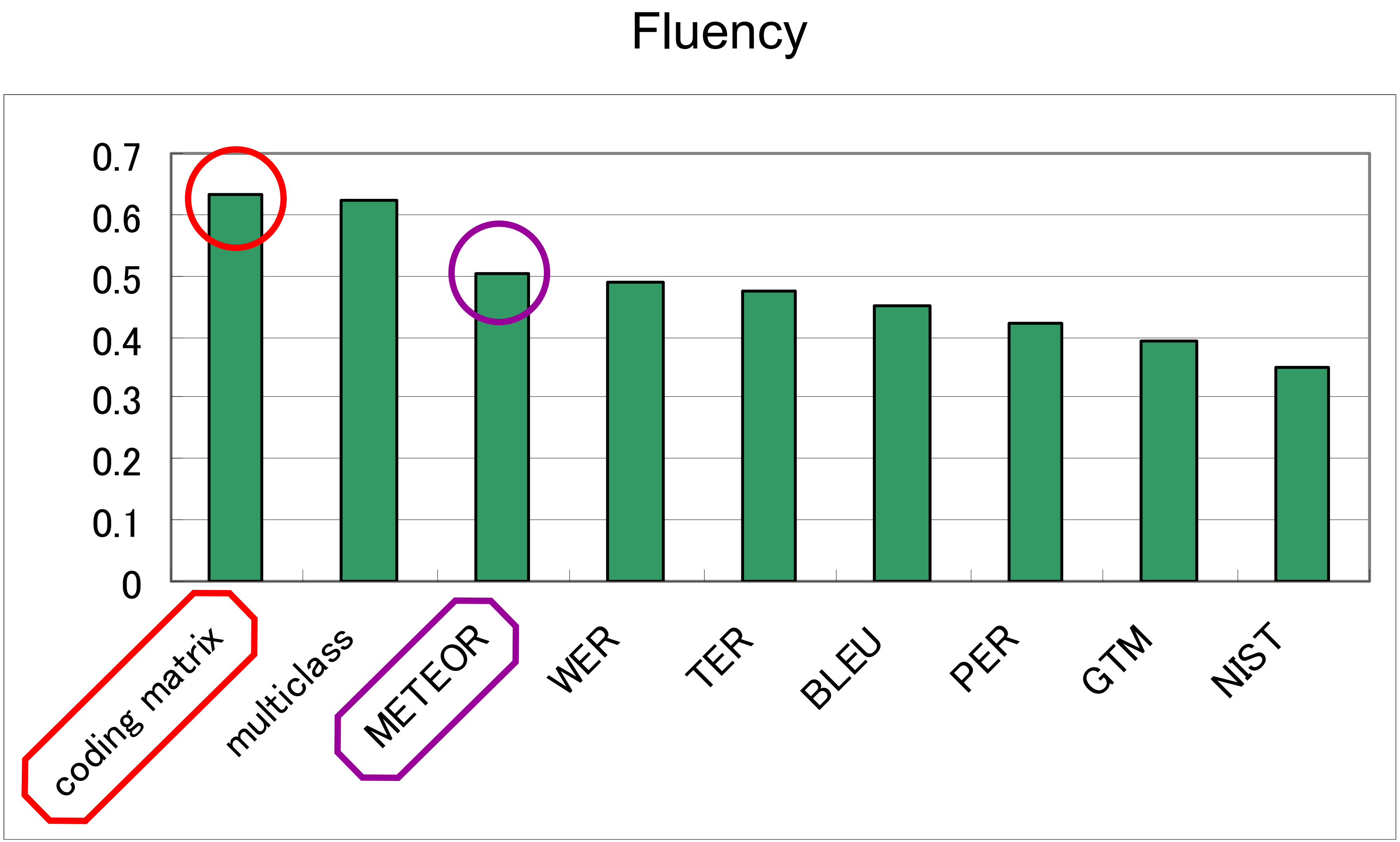
| | 54_321 | 543_21 | 5_4 | 5_3 | 5_2 | 5_1 | 4_3 | 4_2 | 4_1 | 3_2 | 3_1 | 2_1 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5** | +1 | +1 | +1 | +1 | +1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | +1 | –1 | –1 | –1 | –1 |
| **4** | +1 | +1 | –1 | –1 | 0 | 0 | +1 | +1 | +1 | 0 | 0 | 0 | –1 | +1 | –1 | –1 | –1 |
| **3** | –1 | +1 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | +1 | +1 | 0 | –1 | –1 | +1 | –1 | –1 |
| **2** | –1 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | –1 | –1 | +1 | –1 | –1 | –1 | +1 | –1 |
| **1** | –1 | –1 | 0 | 0 | 0 | –1 | 0 | 0 | –1 | 0 | 0 | –1 | –1 | –1 | –1 | –1 | +1 |

# Evaluation
## (classification accuracy on TEST set)

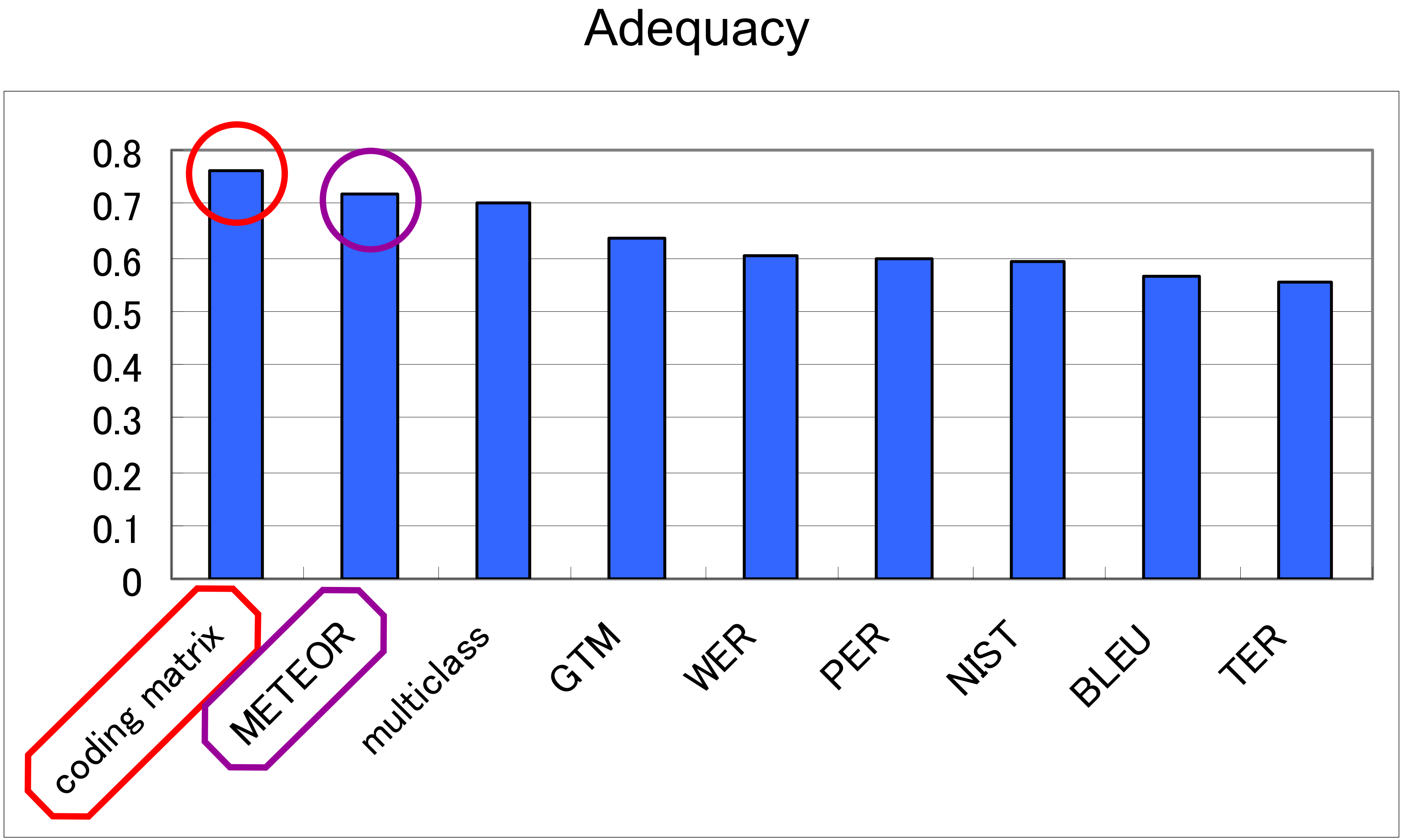


# Correlation to Human Assessment on Sentence-Level



# Correlation to Human Assessment on Sentence-Level



# Summary



## Multiclass reduction to binary:

- **robust and reliable** method to predict human assessments on sentence-level
- **high correlation to human judges** outperforming commonly used automatic evaluation metrics
- **outperforms standard classification methods**
  → gains: **+6.0** (*fluency*) and **+6.6** (*adequacy*) in classification accuracy

## Extension of proposed method:

- apply learning method to **select features used to build the coding matrix**
- investigate in the use of **additional features** that increase binary classification accuracy and thus boost overall multi-class prediction accuracy