# The ICT Machine Translation System for IWSLT 2010

**Hao Xiong, Jun Xie, Hui Yu, Kai Liu, Wei Luo**
**Haitao Mi, Yang Liu, Yajuan Lv, Qun Liu**

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology
**Chinese Academy of Sciences**
No.6 Kexueyuan South Road, Haidian District
P.O. Box 2704, Beijing, China, 100080
http://nlp.ict.ac.cn/english/

## Main Contributions

☺ Improved Single Systems

- 👍 Silenus=>SuperSilenus
- ✌ Chiero=>John
- ☝ Bruin=>TemBruin

☺ Refined Word Segmentation

☺ Multi Word Alignment

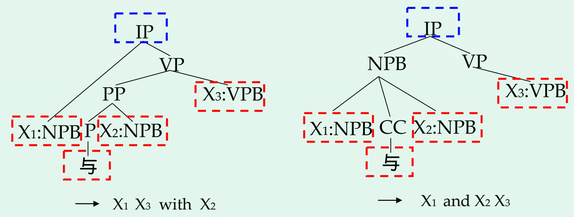
☺ Fuzzy Matching

☺ New Method for ASR Translation

## SuperSilenus

➲ Forest based tree-to-string model (Mi et., 2009)

🎧 Fuzzy rule matching




🍉 Matching of root node and leaf-nodes

🌟 Using tree-kernel (Collins and Duffy 2001) to compute similarity between two rules

## John

➲ Hierarchical phrase-based model (Chiang 2005)

🎧 Joint Tokenization and translation (Xiao et., 2010)

## TemBruin

➲ Maximum entropy based phrase reordering model (Xiong et., 2006)

🎧 Manually written rules

- ☌ 这不是息票是登记收据
- ☍ 这不是#1是#2 => this is only #2, not #1

## Chinese Word Segmentation

☺ **ORI++**: refine the original segmentation by the ICTCLAS segmentation

😐 **ICTCLAS**: segmentation by the open toolkit ICTCLAS

☹ **COMB**: combine ORI++ and ICTCLAS in the training set

**Trick**: Combine the temporal and numerical expressions using heuristic rules

## Alignment

😐 **Giza++:** grow-diag-final

☹ **Berkeley:** http://nlp.cs.berkeley.edu/Main.html#WordAlignment

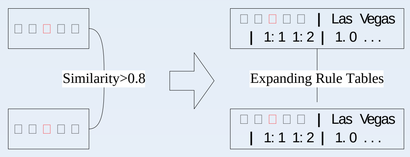
☺ **COMB**: combine Giza++ and Berkeley in the training set

All English letters are in lowercase

## Fuzzy Matching

👎 Training Set: 拉斯韦加斯

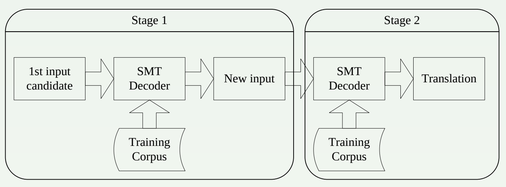
👍 Test Set:拉斯维加斯

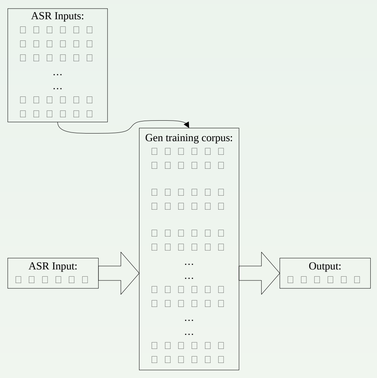



## ASR Translation

☹ Most given input candidates are incorrect

☺ Generate better input through MT




## Experiments

### Different Segmentation Strategies

| | ORI++ | ICTCLAS | COMB |
|---|---|---|---|
| **C2E** | **54.33** | 53.80 | 53.59 |
| **E2C** | **47.62** | 46.92 | 46.28 |

### Different Word Alignment Strategies

| | Giza++ | Berkeley | COMB |
|---|---|---|---|
| **C2E** | 54.03 | 52.78 | **54.33** |
| **E2C** | 45.09 | 43.97 | **47.62** |

### Test Set

| Task | Input | System | BLEU |
|---|---|---|---|
| C2E | CRR | Rescoing(John) | **24.58** |
| | | John | 23.77 |
| | | TemBruin | 23.70 |
| C2E | ASR | Rescoing(John) | 22.20 |
| | | John | **22.27** |
| | | TemBruin | 19.35 |
| E2C | CRR | Rescoing(SuperSilenus) | **37.67** |
| | | SuperSilenus | 35.16 |
| | | Moses | 33.44 |
| E2C | ASR | Rescoing(SuperSilenus) | **30.80** |
| | | SuperSilenus | 28.96 |
| | | Moses | 28.17 |

### Progress Test

| Task | Input | System | BLEU |
|---|---|---|---|
| C2E | CRR | Last year | 31.85 |
| | | This year | **36.70** |
| C2E | ASR | Last year | 28.53 |
| | | This year | **33.34** |
| E2C | CRR | Last year | 39.98 |
| | | This year | **49.61** |
| E2C | ASR | Last year | 29.85 |
| | | This year | **38.57** |

## Conclusions

☺ Refine the data preprocessing

☺ Adopt a fuzzy matching technique to reduce the number of OOV

❄ Apply a novel method for the ASR task

❄ Improve the performance of single decoder

**Tricks**: 你=> 您 | 先生=>Mr. 女士=>Ms. 小姐=>Ms.