TOSHIBA

Leading Innovation >>>

# The TCH Machine Translation System for IWSLT 2008

**Haifeng Wang**

**Toshiba (China) R&D Center**

**Oct. 20, 2008**

# Outline

- Overview
- Modules
  - Dictionary, segmentation, alignment, NE
  - LM, Punctuation and case restoration
- Tasks
  - CE: CT, BTEC
  - EC: CT
  - CS: BTEC
  - CES: PIVOT
- Evaluation Results
- Summary

# Introduction

- Tasks
  - BTEC tasks: BTEC_CE, BTEC_CS
  - Challenge tasks: CT_CE, CT_EC
  - Pivot task: PIVOT_CES
- Methods
  - SMT
  - RBMT
  - Pivot SMT
  - Combination
  - Module improvement
- Resources
  - Supplied resources provided for each data track
  - Other publicly available resources

# MT Methods

- SMT
  - Phrase-based SMT: Moses
- Pivot SMT

  (Wu and Wang, ACL 2007)
  - Phrase translation probability
  - Lexical weight
- RBMT
  - Publicly available software:  Dr. eye
- Combination of RBMT and SMT

  (Hu, Wang and Wu, EMNLP 2007)
  - Using SMT system as the main MT system
  - Using RBMT system to produce synthetic bilingual corpus
  - The SMT system is trained using both real and synthetic corpus
- Translation selection
  - 5-gram LM method (Chen et al. IWSLT 2006)
  - Target sentence average length method

# Modules

- Dictionary
- Chinese Word Segmentation
- Word Alignment
- Named Entity Translation
- Language Model
- Punctuation Restoration
- Case Restoration

# Bilingual Dictionary

- Existed bilingual dictionary
  - General dictionary
    - LDC Chinese-English translation lexicon
  - NE dictionary
    - LDC Chinese <-> English Name Entity Lists
    - Person names and location names
- Dictionary extracted from corpus
  - Automatically extracted from in-domain corpus
    - Bidirectional word alignment
    - Filtering
      - Translation probability
      - Co-occurring frequency
  - Check

# Chinese Monolingual Dictionary

- Dictionaries
  - General dictionary
    - Extracted from LDC Chinese-English lexicon
  - NE dictionary
    - Extracted from LDC Chinese-English NE list
  - In-domain dictionary
    - Extracted from in-domain corpus
- Word granularity
  - Tune the word unit referring to its translation in target language
    - Word
    - Multi-word expression

# Chinese Word Segmentation

- Initial experiments
  - Segmentation ambiguity in domain-specific spoken language is not serious
- Segmentation method
  - Forward maximum-matching
    - Basic segmentation method
  - Back one character method
    - To indentify ambiguous fragment
  - Ambiguous fragments database
    - For disambiguation
- Word normalization
  - To deal with data sparseness
  - Extract a synonym list from translation dictionary and corpus
  - Only used when Chinese is source language

# Word Alignment

- **Alignment algorithm**
  - Bidirectional word alignment using IBM models
  - Keep links in the intersection set
  - Keep links occurring in bilingual dictionaries
  - Delete links conflicting with the links in the final alignment set
  - Keep remained links
  - Different alignment heuristics
    - Grow-diag (CE, CS), grow (EC), grow-diag-final (ES)
- **Resources**
  - Bilingual corpus
  - General dictionary
  - Domain-specific dictionary

# Named Entity Translation

- NE recognition and translation

| | Digit | Date | Time | Person name | Location name |
|---|---|---|---|---|---|
| Method | Rule | Rule | Rule | Dictionary | Dictionary |

- NE processing in SMT
  - Training
    - Replace NEs in the training data with NE tags
    - Train model on the data with NE tags
  - Translating
    - Replace NEs in the input sentence with NE tags
    - Translate
    - Restore the NE tags with their translations

# Language Model

- In-domain corpus
  - Target language part of the provided corpus for a given track
- Out-of-domain corpus
  - Publicly available corpus
  - Selection
    - Perplexity
    - Sentence
    - Using the in-domain LM
- Interpolation
  - Linear interpolation using SRILM
  - Weight tuned on development sets

# Punctuation Restoration

- Restore punctuation in source language
- English
  - Hidden-ngram (SRILM toolkit)
  - Rules
    - By hand
    - Based on some keywords, e.g. a sentence begin with “could”
- Chinese
  - Maximum entropy model
  - 2 steps
    - Position determination
    - Punctuation determination
  - Features
    - Words around a boundary
    - Words at the beginning or end of a sub-sentence

# Case Restoration

- Restore case in target language
  - English
  - Spanish
- Method
  - recaser
    - In the training scripts of Moses
    - As a MT problem
    - Trained on the corpus with case information
  - Lexicon based post-processing
    - To process English words that should be capitalized
      Such as proper nouns
    - The lexicon is extracted from some available resources
      Such as training text in respective tasks, HIT corpus, Tanaka corpus

# Tasks

- **Five tasks**
  - Chinese-English
    - Challenge task (CT_CE)
    - BTEC task (BTEC_CE)
  - English-Chinese
    - Challenge task (CT_EC)
  - Chinese-Spanish
    - BTEC task (BTEC_CS)
  - Chinese-English-Spanish
    - Pivot task (PIVOT_CES)
- **Input**
  - Spontaneous speech (SS)
  - Read speech (RS)
  - Correct recognition result (CRR)

# Chinese-English Tasks – Data

- Dictionary

| Type | General | Domain | NE |
|---|---|---|---|
| Source | LDC2002L27 | Extracted from In-domain corpus | LDC2005T34 |
| Number | 54,170 | 38,620 | 47692 |

- Training Corpus

| Corpus | BTEC | HIT | CLDC | Tanaka |
|---|---|---|---|---|
| # sentence pairs | 19,972 | 80,868 | 200,732 | 149,207 |
| # source words | 177,168 | 802,454 | 2,113,534 | - |
| # target words | 182,627 | 822,508 | 2,096,731 | 1,351,645 |

  – Selection and preprocessing

- Development set
  – devset1, devset2, devset4
- Test set
  – devset3 (2005), devset5 (2006), devset6 (2007)

# Chinese-English Tasks – Experimental Results

- Results (Case sensitive BLEU score, CRR input)

| | devset3 | devset5 | devset6 |
|---|---|---|---|
| RBMT | 0.4253 | 0.2020 | 0.2086 |
| Baseline | 0.5186 | 0.2013 | 0.2807 |
| Our segmenter | 0.5425 | 0.2047 | 0.3029 |
| +HIT | 0.5697 | 0.2323 | 0.3416 |
| +Dic | 0.5819 | 0.2375 | 0.3456 |
| +NE | 0.5838 | 0.2396 | 0.3537 |
| +CLDC | 0.5891 | 0.2445 | 0.3554 |
| +RBMT | 0.6091 | 0.2536 | 0.3570 |
| +LM Inter. | 0.6223 | 0.2516 | 0.3823 |

- Translation selection
  - Mert
    - Default: default in Moses
    - Mert1: best on devset5
    - Mert2: Stable
  - Selection metric: voting, length

| | devset3 | devset5 | devset6 |
|---|---|---|---|
| Default | 0.5927 | 0.2547 | 0.3453 |
| Mert1 | 0.6061 | 0.2679 | 0.3837 |
| Mert2 | 0.6274 | 0.2551 | 0.3863 |
| Select | 0.6260 | 0.2627 | 0.3882 |

# English-Chinese Tasks – Data

- Dictionary
  - General dictionary, domain dictionary, NE dictionary

  (Same as CE tasks)
- Training Corpus

| Corpus | BTEC | HIT |
|---|---|---|
| # sentence pairs | 19,972 | 89,318 |
| # source words | 189,041 | 945,010 |
| # target words | 178,339 | 914,121 |

  - Selection
  - Preprocessing
    - English abbreviation restoration
    - Without Chinese word normalization
- Development and test set
  - devset, devset3
  - No MERT

# English-Chinese Tasks – Experimental Results

- Results

| | devset3 | devset |
|---|---|---|
| RBMT | 0.4362 | 0.4425 |
| Baseline | 0.4455 | 0.4511 |
| Our segmenter | 0.4528 | 0.4564 |
| +Dic | 0.4551 | 0.4684 |
| +NE | 0.4558 | 0.4773 |
| +HIT | 0.4830 | 0.5325 |
| +RBMT | 0.5131 | 0.5426 |
| +Select | 0.5133 | 0.5551 |

- Translation selection
  - 2 Candidates
    - Without RBMT
    - With RBMT
  - Selection metric: LM

# Chinese-Spanish Tasks

- Training Corpus
  - BTEC data provided for this task
  - Preprocessing similar as CE task
- Dictionary
  - Extracted from the training corpus (9990 entries)
- Test set
  - Devset3
- Post-processing
  - Rule-based, such as question mark "?" and "¿"
- Experimental Results

| | Baseline | Our segmenter | +dic |
|---|---|---|---|
| BLEU | 0.3596 | 0.3726 | 0.3839 |

# Chinese-English-Spanish – Data

- Dictionary
  - LDC CE dictioanry
  - CE dictionary extracted from BTEC and HIT CE corpus (39010)
  - ES dictionary extracted from BTEC and Europarl ES corpus (10426)
- Training Corpus

| Corpus | BTEC CE | HIT CE | BTEC ES | Europarl ES | Tanaka |
|---|---|---|---|---|---|
| # sentence pairs | 20,000 | 80,868 | 19,972 | 400,000 | 149,207 |
| # source words | 164,957 | 802,454 | 182,627 | 8,485,253 | - |
| # target words | 182,793 | 822,508 | 185,527 | 8,219,380 | 1,351,645 |

  - Selection and preprocessing
- Test set
  - devset3

# Chinese-English-Spanish – Experimental Results

- Results

| | Baseline | +dic+HIT+Europarl | +RBMT |
|---|---|---|---|
| Pivot model | 0.2791 | 0.3616 | 0.4136 |
| Transfer model | 0.3243 | 0.4139 | 0.4423 |
| Trans. selection | - | - | 0.4510 |

- RBMT
  - Translate the English part of ES corpus into Chinese -> synthetic CE corpus
  - Synthetic CE corpus is used in pivot and transfer model
- Transfer model is better than pivot model
  - CE translation is quite good (0.6024)
  - English and Spanish are more similar than Chinese and Spanish
  - pivot model contains much more noise than the transfer model
- Translation selection
  - Selection metric: length

# IWSLT 2008 Evaluation Results

| | | (Bleu + Meteor)/2 | Bleu | Meteor | Human Eval. |
|---|---|---|---|---|---|
| CT_EC | SS | 0.5647 | 0.4818 | 0.6476 | 0.3906 |
| | CRR | 0.6566 | 0.5912 | 0.7219 | - |
| CT_CE | SS | 0.5257 | 0.4166 | 0.6347 | 0.4516 |
| | CRR | 0.5909 | 0.4980 | 0.6837 | - |
| BTEC_CE | RS | 0.5358 | 0.4474 | 0.6241 | 0.4730 |
| | CRR | 0.5887 | 0.5085 | 0.6688 | - |
| BTEC_CS | RS | 0.3273 | 0.3218 | 0.3328 | 0.4316 |
| | CRR | 0.3597 | 0.3582 | 0.3611 | - |
| PIVOT_CES | RS | 0.3620 | 0.3657 | 0.3583 | 0.4624 |
| | CRR | 0.4044 | 0.4157 | 0.3931 | - |

# Summary

- Tasks
  - BTEC_CE, BTEC_CS, CT_CE, CT_EC, PIVOT_CES
- Resources
  - Supplied resources provided for each data track
  - Other Publicly available resources
- Methods
  - Adaptation of Chinese word segmentation
  - Word alignment refinement using dictionary and various heuristics
  - Named entities translation
  - Additional corpus (In-domain, Out-of-domain)
  - Combination of SMT and RBMT
  - Translation selection

Thanks!