기계번역 시스템
개발 실제 및 활용

김성동
한성대학교 컴퓨터공학부
Contents (1)

- MT approaches
- Problems in English-Korean Machine Translation
Contents (2)

- English-Korean Machine Translation System
  - Applied approaches
  - EKMT system structure
  - EKMT system development
  - EKMT system problems & solutions
  - EKMT system structure / translation flow
  - Works for continuous improvement
Contents (3)

- Neural Machine Translation
- Google’s Neural Machine Translation System
- MT applications
MT Approaches
MT Approaches (1)

- Direct Translation
- Rule-Based Translation
- Corpus-Based Translation
- Knowledge-Based Translation
  - Transfer-Based Translation
  - Interlingua Translation
  - Statistical Translation
  - Example-Based Translation
MT Approaches (2)

- Direct translation
  - Simple word substitution
  - Some changes in ordering
  - Translation between similar languages
    - English ↔ French, Korean ↔ Japanese
MT Approaches (3)

• Interlingua translation
  • Translate source language into an underlying meaningful representation, interlingua
  • Generate target sentence from the internal representation
MT Approaches (4)

INTERLINGUA

DIRECT

English
French
German
Japanese

English
French
MT Approaches (5)

- Rule-based + transfer-based translation
  - Lexical rules
  - Syntactic rules
  - Transfer rules
  - Generation rules
MT Approaches (6)

- Example-based translation
  - Translation example: pair of (source, target) sentences

- Statistical translation (French to English)
  - Use bilingual alignment corpus composed of (e, f)

\[
e_{best} = \arg \max_e P(e|f) \\
= \arg \max_e P(f|e) P(e)
\]
MT Approaches (7)

• Neural translation
  • Model the entire MT process via one big artificial neural network

Problems in English-Korean MT
Problems in English-Korean MT (1)

- Part-of-Speech Ambiguity
- Structural Ambiguity
- Semantic Ambiguity
Problems in English-Korean MT (2)

• I saw bats  
  saw = NOUN ? VERB ?

• I saw bats in the park  
  I ? bat ? in the park

• I saw bats with the telescope  
  I ? bat ? with the telescope

• old men and women  
  men ? men and women ? old

• He reached the bank  
  둑? 은행? 어디에…
English-Korean MT system

- Applied approaches
English-Korean MT system
- Applied approaches (1)

- Direct translation
- Inter-lingual translation
- Transfer-based translation
- Rule-based translation
- Corpus-based translation
- Example-based translation
- Statistical translation
- Neural translation

Statistical methods using corpus

- Rule-based
- Transfer-based

+ 

- Idiom translation
- Sentence segmentation
English-Korean MT system
- Applied approaches (2)

• Idiom-based translation
  • English-Korean bilingual idiom
  • Idiom recognition before parsing
    • Reduce parsing complexity: idiom is treated one unit
    • Resolve translation ambiguity

<table>
<thead>
<tr>
<th>bread and butter</th>
<th>provide him with money (provide A with B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>빵과 버터</td>
<td>돈을 가지고 (가진) 그를 제공하다</td>
</tr>
<tr>
<td>→ 버터 바른 빵</td>
<td>→ 그에게 돈을 제공하다</td>
</tr>
</tbody>
</table>
English-Korean MT system
- Applied approaches (3)

• Sentence segmentation
  • Partial parsing: reduce parsing complexity → long sentence analysis
  • Maximum entropy probability model
    • Corpus tagged with segmentation positions
    • Learn the context of the segmentation positions

\[
p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{n} \lambda_i f_i(x, y))
\]

\(x\): context of a word, \(y: 0\) or 1
\(f\): feature, \(\lambda\): weight
\(Z(x)\): normalizing constant
English-Korean MT system

- Structure
English-Korean MT system
- Structure (1)

Translation Engine

Knowledge Base
- Dictionaries
- Rule Sets
English-Korean MT system
- Structure (2)

- Lexical Analysis
  - Lexical rules
  - Syntactic rules
  - Transfer rules
  - Generation rules

- Syntactic Analysis
  - Parsing

- Structure Transfer

- Korean Generation
  - English Lexical Dictionary
  - English-Korean Bilingual Dictionary
  - Korean Generation Dictionary
English-Korean MT system
- Development
English-Korean MT system
- Development – Knowledge Base (1)

English Lexical Dictionary

POS information
mist,1:I1;2:B11 8 1:A1:I1
mistakable,3:T3
mistake,1:N1;2:B8 9 21
2:N2:F40:A6:I96
mistaken,3:T3;2:F1:Lmistake:A6:I224

Frequency information
devalued VERB 8
devastate VERB 4
devastated VERB 23 ADJ 4
devastating VERB 22 ADJ 22
English-Korean MT system
- Development – Knowledge Base (2)

English-Korean Bilingual Dictionary

- General Dictionary
  - NOUN
  - VERB
  - ADJ
  - ADV
  - PRON
  - CONJ
  - PREP

- Domain Dictionary
  - Economy
  - Computer
  - Medical science
  - ...

- User Dictionary
prove (VERB)

Default meaning

vt (증명하다)
vi (증명하다)

Collocation

검사하다 (ore 광성)
보이다 (strength 좋은 성능)
입증하다 (guilt 유죄) (innocence 결백함)

Idiom

prove A TO_INF : A가 TO_INF 함을 증명하다
prove to be A : A라고 판명되다
Domain Dictionary

"finance"
- finance: 금융
- finance bill: 금융 어음
- finance company: 금융회사
- finance corporation: 공고
- finance for dead stock: 기초재고금융
- finance lease: 파이낸스 리스
"간선하다" VERB 여불
"간섭되다" VERB 정
"간섭하다" VERB 여불 PASSIVE: "에 의해" "간섭되다"
"간소하다" VERB 여불
"간소화되다" VERB 정
"간소화하다" VERB 여불 PASSIVE: "에 의해" "간소화되다"
English-Korean MT system
- Development – Knowledge Base (6)

**Syntactic Rules**

NP020: NP(%NP, DETS=DET.DETS) →

  DET

  NP(DEF=0, INDEF=0, PSMOD=0, pos!='VERB)

NP032: NP(%NP, AMODS->AJP) →

  AJP(QUESFLAG=1)

  NP(pos!='PRON, DETS=0, RESTRIC=0, MODS->0)
English-Korean MT system
- Development – Translation Engine (1)

• Lexical analyzer
  • Input sentence → word stream
  • Search word information: POS, base form, …
  • Calculate POS probability
I like the books

<table>
<thead>
<tr>
<th>Word</th>
<th>Part of Speech</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>PRON</td>
<td>1.0</td>
</tr>
<tr>
<td>like</td>
<td>VERB</td>
<td>0.59</td>
</tr>
<tr>
<td>PREP</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>CONJ</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
<td>1.0</td>
</tr>
<tr>
<td>books</td>
<td>VERB</td>
<td>0.04</td>
</tr>
<tr>
<td>NOUN</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>
English-Korean MT system
- Development – Translation Engine (3)

• Parser = Syntactic Analyzer
  • Chart parser
  • Context free English grammar
    • $O(n^3)$
Grammar

1. \( S \rightarrow NP \ VP \)
2. \( NP \rightarrow ART \ ADJ \ N \)
3. \( NP \rightarrow ART \ N \)
4. \( NP \rightarrow ADJ \ N \)
5. \( VP \rightarrow AUX \ VP \)
6. \( VP \rightarrow V \ NP \)

Input: The large can can hold the water

- the: ART
- large: ADJ
- can: N, AUX, V
- hold: N, V
- water: N, V
<table>
<thead>
<tr>
<th>ART1</th>
<th>ADJ1</th>
<th>AUX1</th>
<th>AUX2</th>
<th>N3</th>
<th>ART2</th>
<th>N4</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>large</td>
<td>can</td>
<td>can</td>
<td>hold</td>
<td>the</td>
<td>water</td>
</tr>
</tbody>
</table>

S → NP ⋊ VP

S → NP ⋊ VP

VP → AUX ⋊ VP

VP → AUX ⋊ VP
<table>
<thead>
<tr>
<th>S1 (rule 1 with NP1 and VP2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2 (rule 1 with NP2 and VP2)</td>
</tr>
<tr>
<td>VP3 (rule 5 with AUX1 and VP2)</td>
</tr>
<tr>
<td>VP2 (rule 5)</td>
</tr>
<tr>
<td>VP1 (rule 6)</td>
</tr>
<tr>
<td>NP2 (rule 4)</td>
</tr>
<tr>
<td>NP1 (rule 2)</td>
</tr>
<tr>
<td>N1</td>
</tr>
<tr>
<td>V1</td>
</tr>
<tr>
<td>ART1</td>
</tr>
<tr>
<td>1 the 2 large 3 can 4 can 5 hold 6 the 7 water</td>
</tr>
</tbody>
</table>
English-Korean MT system
- Development – Translation Engine (4)

I – PRON
like – VERB
PREP
ADJ
CONJ
the – DET
books – VERB
NOUN

like
+— [SUBJ] → i
+—[OBJ] → books
English-Korean MT system
- Development – Translation Engine (5)

• Structure transfer
  • Resolve differences between English and Korean
    • It (real subject) – TO_INF (pseudo subject)
    • It (real object) – TO_INF (pseudo object)
    • It (real subject) – THAT_CLAUSE (pseudo subject)
  • 비인칭 주어 “it”
  • 부가 의문문
  • ...

-Development – Translation Engine (5)
### English-Korean MT system
**- Development – Translation Engine (6)**

<table>
<thead>
<tr>
<th>English</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>is</td>
<td>is</td>
</tr>
<tr>
<td>+— [SUBJ] (\rightarrow) it</td>
<td>+— [SCOMP] (\rightarrow) good</td>
</tr>
<tr>
<td>+— [SCOMP] (\rightarrow) good</td>
<td>+— [TOINF] (\rightarrow) know</td>
</tr>
<tr>
<td></td>
<td>+— [OBJ] (\rightarrow) truth</td>
</tr>
<tr>
<td>SCOMP is</td>
<td>SCOMP is</td>
</tr>
<tr>
<td>+— [SCOMP] (\rightarrow) good</td>
<td>+— [SUBJ] (\rightarrow) know</td>
</tr>
<tr>
<td></td>
<td>+— [OBJ] (\rightarrow) truth</td>
</tr>
</tbody>
</table>
English-Korean MT system - Development – Translation Engine (7)

- Translation and attachment determination for **preposition**
- Attachment determination for **to-infinitive** clause
- Homonym (동음이의어) translation
English-Korean MT system
- Development – Translation Engine (8)

- Korean Generator

Sentence structure → Transform the structure → Word ordering → Morpheme insertion → Morpheme synthesis → Korean sentence

Morpheme generation → Structure generation
English-Korean MT system

- Problems & Solutions
English-Korean MT system
- Problems & Solutions (1)

<table>
<thead>
<tr>
<th>Speed</th>
<th>Memory, Structural ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsing complexity $= O(n^3)$</td>
<td>Too many structures during parsing</td>
</tr>
</tbody>
</table>

- POS determination
- Intra-sentence segmentation
- 3-phase parsing
English-Korean MT system
- Problems & Solutions (2)

• POS determination
  • Reduce POS ambiguity $\rightarrow$ efficient parsing

```
I/PRON like/VERB/PREP/ADJ/CONJ the/DET
books/VERB/NOUN

I/PRON like/VERB the/DET books/NOUN
```
English-Korean MT system - Problems & Solutions (3)

- Intra-sentence segmentation
  - Long sentence $\rightarrow$ several short segments $\rightarrow$ parsing unit
  - Try to reduce $n$ in $O(n^3)$ $\rightarrow$ make parsing faster

The chef cooks the soup, and I enjoy it.

The chef cooks the soup and I enjoy it.
We also analyze the effect of various choices while inducing word embeddings on “downstream” POS induction results.
3-phase parsing

and I enjoy it
English-Korean MT system
- Problems & Solutions (5)

- 3-phase parsing
  - Syntactic rule classification
  - Rule acquisition for 3rd-phase parsing
Various language differences

- Multi sentences separated by ‘;’, ‘:’, ...
- Enclosed parts by “”, ( ), < >,…
- Composition words: composite NOUN, VERB, …
- Special patterns managed by rules or idioms: [not only ~ but also], [ ~ so that ~ ], …
- …
English-Korean MT system
- Problems & Solutions (7)

• Preprocessing – post-processing
  • Pre-lexical analysis preprocessing
  • Post-lexical analysis preprocessing
  • Post-segmentation preprocessing
  • Post processing
English-Korean MT system

- Structure / Translation Flow
English-Korean MT system - Structure / Translation flow (1)

English Sentence

Before Lexical Preprocessing

Lexical Analysis

POS Determination

After Lexical Preprocessing
English-Korean MT system
- Structure / Translation flow (2)

Segmentation by Commas

segment₁ → ... → segmentⁿ

1st-phase parsing

segment₁ 1st structure → ... → segmentⁿ 1st structure
English-Korean MT system
- Structure / Translation flow (3)

2nd-phase parsing

\[ \text{segment}_1 \text{ structure} \quad \ldots \quad \text{segment}_n \text{ structure} \]

3rd-phase parsing

Sentence structure
English-Korean MT system
- Structure / Translation flow (4)

Structural Transfer

Korean Generation

Korean Sentence
English-Korean MT system

- Works for continuous improvement
English-Korean MT system
- Works for continuous improvement (1)

Performance of EKMT system

- Speed
- Accuracy

- Hardware
- Translation Knowledge
English-Korean MT system
- Works for continuous improvement (2)

Building
Translation Knowledge

General Purpose EKMT ?

Special Purpose EKMT
English-Korean MT system
- Works for continuous improvement (3)
English-Korean MT system
- Works for continuous improvement (4)

- News Article Extractor (NAE)
  - Extracts English news articles from web
  - Construct corpus for target translation domain
English-Korean MT system
- Works for continuous improvement (5)

- Dictionary Enhancement
  - Unregistered Word Collector (UWC)
  - Compound Noun Collector (CNC)
  - Meaning Attachment Tool (MAT)
    - Help human input the meaning of new (compound) word
    - Integrate new entry with the existing dictionary
English-Korean MT system
- Works for continuous improvement (6)

- Syntactic Rule Improvement
  - Syntactic rule management system
    - Help human improve English syntactic rules during the translation test
    - Assist in searching, comparing, and managing syntactic rules
Neural Machine Translation
Neural Machine Translation (1)

• Early researches
  • Nal Kalchbrenier and Phil Blunsom, “Recurrent continuous translation models,” EMNLP 2013
  • Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, “Sequence to sequence learning with neural networks,” NIPS 2014
Neural Machine Translation (2)

- Features
  - Require minimal domain knowledge
  - Conceptually simple model
  - Sequence to sequence translation

- Source language sentence
- Intermediate Representation
- Target language sentence
Neural Machine Translation (3)

- Basic framework: Encoder/Decoder

https://research.googleblog.com/2015/11/computer-respond-to-this-email.html
Neural Machine Translation (5)

- What to study (know)
  - Neural network
  - Deep learning
  - Word embedding
  - Recurrent neural network
    - GRU (Gated Recurrent Unit)
    - LSTM (Long Short Term memory)
  - ...

Neural Machine Translation (5)

- Word embedding
  - Word: symbol $\rightarrow$ number
  - Relationship among words

\[
\text{vector}(\text{King}) - \text{vector}(\text{Man}) + \text{vector}(\text{Woman}) \approx \text{vector}(\text{Queen})
\]
\[
\text{vector}(\text{Paris}) - \text{vector}(\text{France}) + \text{vector}(\text{Italy}) \approx \text{vector}(\text{Rome})
\]
Neural Machine Translation (6)

- Word representation
  - One hot representation (encoding)
    - candy = [0, 0, 0, 1, 0, 0, ...]
- Distributional representation
  - Neural word embedding
    - Candy = [0.286, 0.792, -0.177, -0.107, -0.109, -0.542, 0.349, 0.271]

Neural Machine Translation (7)

- Word2vec
  - Google
  - CBOW (continuous bag of words) model architecture
  - Skip-gram model architecture
Google’s NMT System
Google’s NMT System (1)

- November, 2016

Google’s NMT System (3)

- **Structure**
  - Encoder: 8 LSTM layer (1 bidirectional, 7 unidirectional)
  - Decoder: 8 LSTM layer (unidirectional)
  - Attention layer
  - Softmax layer
Google’s NMT System (4)

- Training
  - WMT [ En → Fr ] data set → 36M sentence pairs
  - 96 NVIDIA K80 GPUs
  - Around 6 days

- RL based refinement
  - Around 3 days
Google’s NMT System (5)

- Performance
  - Human-rated side-by-side comparison

\[ \text{GNMT} \approx \text{Average bilingual human translator} \]

- Roughly 60% reduction in translation errors compared to the PBMT
Google’s NMT System (6)

• Features
  • Zero-shot translation
    • Training with examples English-Korean and English-Japanese, GNMT automatically does Japanese $\rightarrow$ Korean reasonably well
    • Transfer the “translation knowledge” $\rightarrow$ transfer learning
Training

Google Neural Machine Translation

English
Japanese
Korean

<table>
<thead>
<tr>
<th>Google’s NMT System (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• TensorFlow – ML toolkit</td>
</tr>
<tr>
<td>• Tensor Processing Unit</td>
</tr>
</tbody>
</table>
MT Applications
MT applications (1)

- Document translation
  - PDF, Power point, word, …
  - Technical documents

File format handling
MT applications (2)

- Video caption translation
- Image translation

Caption generation + video file handling

Optical character recognition
MT applications (3)

- Web translation

HTML Document Handing
MT applications (4)

- Speech translation

Hello

English Speech -> Speech Recognition -> English Text

안녕

Korean Speech -> Speech Synthesizer -> Korean Text

EMKT
MT applications (5)

- Chat translation
- E-mail translation
- ...

MT applications (6)

- Application areas
  - Study
    - Document translation
    - Video caption translation
    - Coursera, MOOC, SlideShare, …
  - Travel
    - Speech translation
    - Image translation
MT applications (7)

- Information retrieval
  - Web translation
- Communication
  - E-mail translation
  - Chatting translation
- Localization
  - Document translation
MT applications (8)

- ...
Why old-fashioned MT approach?

I ate bread an butter

나는 빵과 버터를 먹었다.
나는 버터 바른 빵을 먹었다.
Why old-fashioned MT approach?

A good book gives the peace of mind.

좋은 책은 마음의 평화를 준다

양서는 마음의 평화를 준다
감사합니다!
sdkim@hansung.ac.kr