기계번역 시스템 개발 실제 및 활용

김 성 동

한성대학교 컴퓨터공학부

Contents (1)

- MT approaches
- Problems in English-Korean Machine Translation

Contents (2)

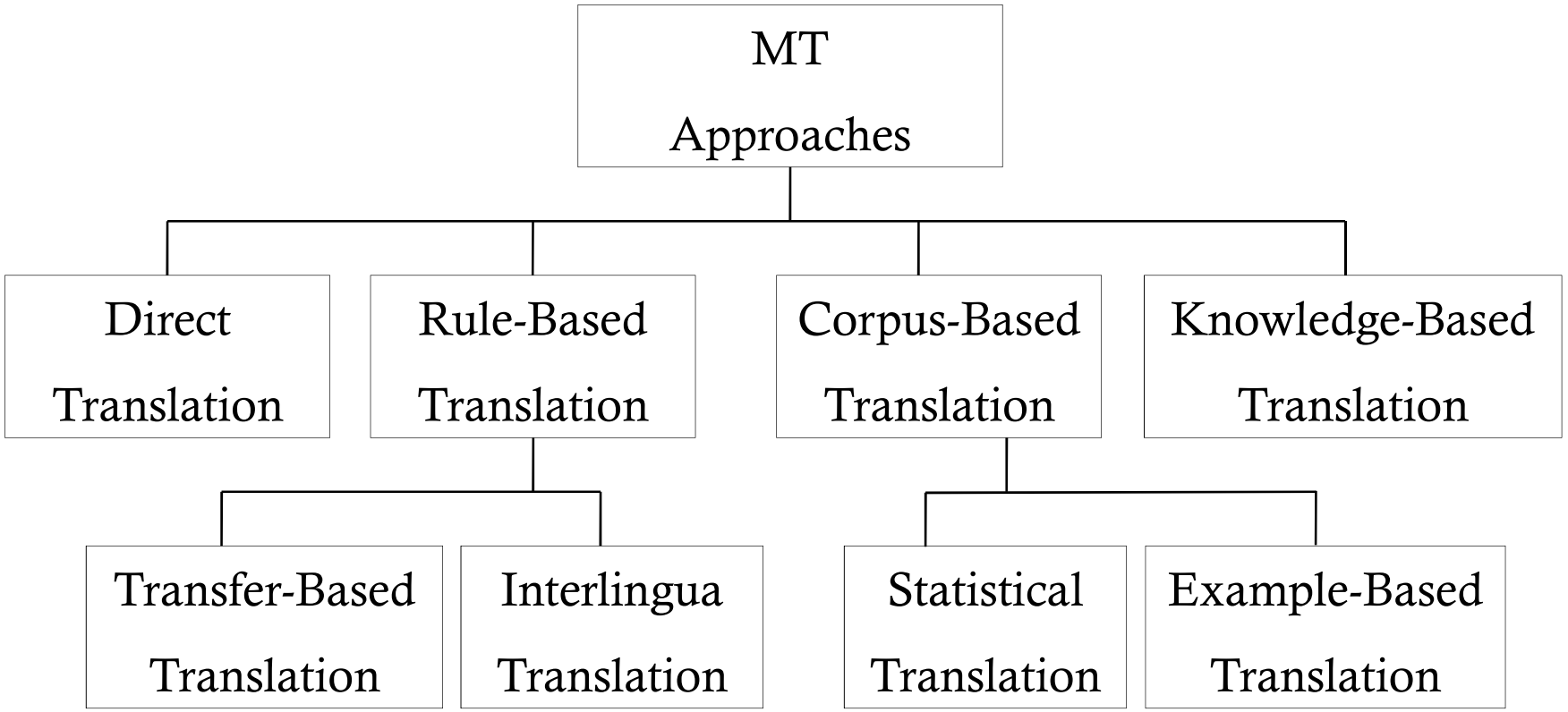
- English-Korean Machine Translation System
  - Applied approaches
  - EKMT system structure
  - EKMT system development
  - EKMT system problems & solutions
  - EKMT system structure / translation flow
  - Works for continuous improvement

Contents (3)

- Neural Machine Translation
- Google's Neural Machine Translation System
- MT applications

MT Approaches

MT Approaches (1)



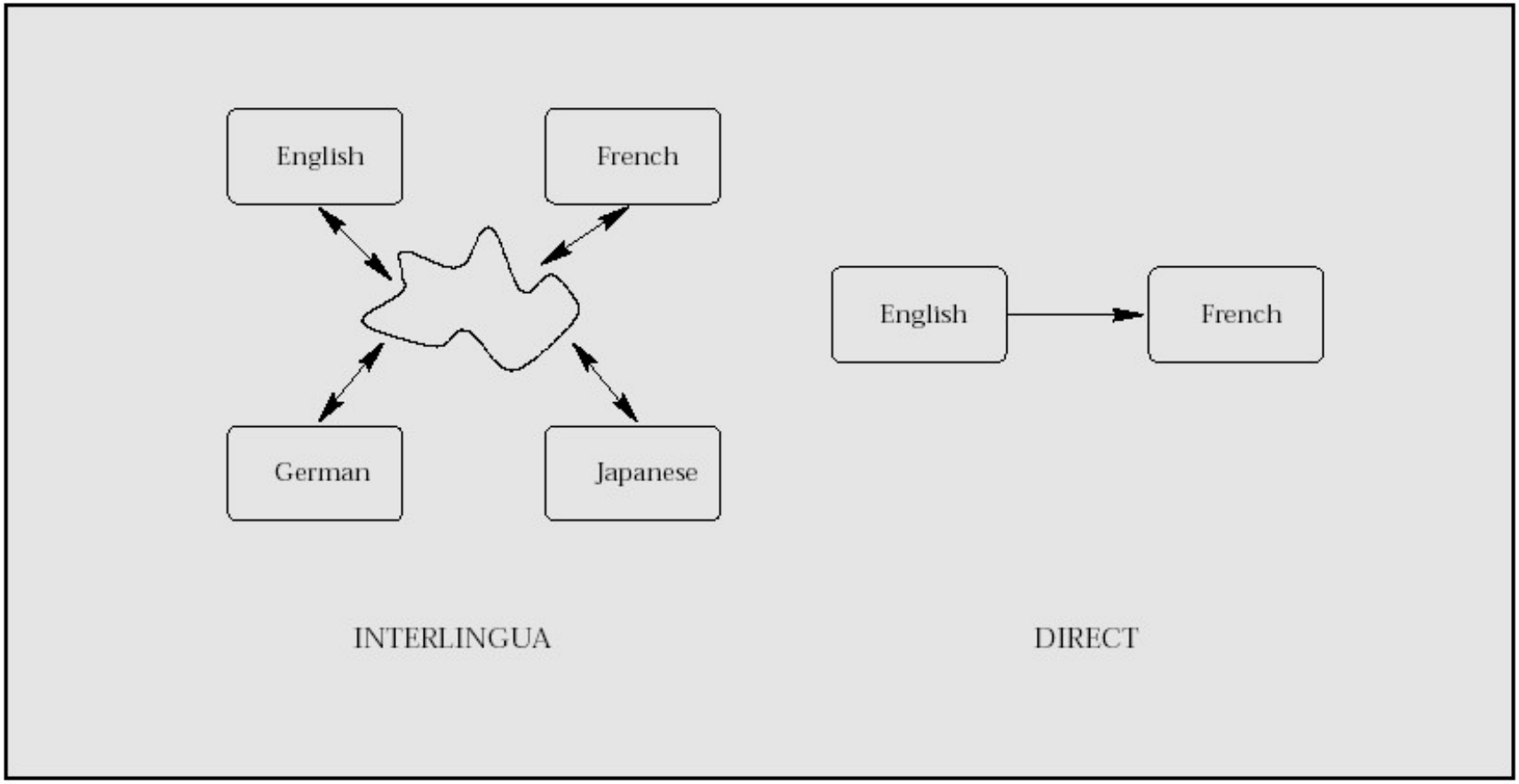
MT Approaches (2)

- 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)

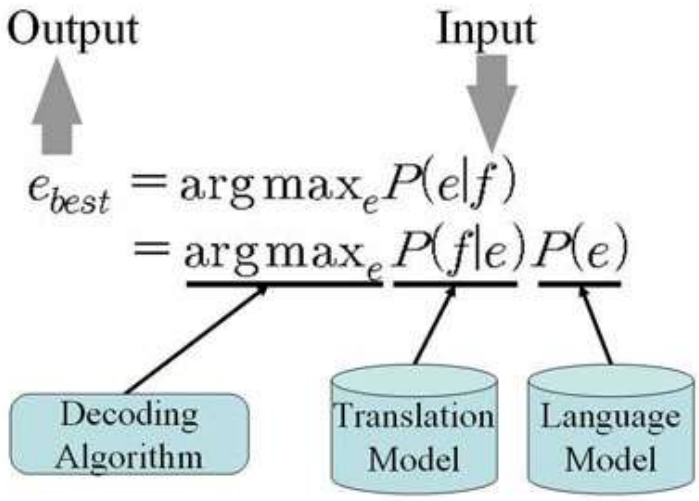


MT Approaches (5)

- Rule-based + transfer-based translation
  - Lexical rules
  - Syntactic rules
  - Transfer rules
  - Generation rules

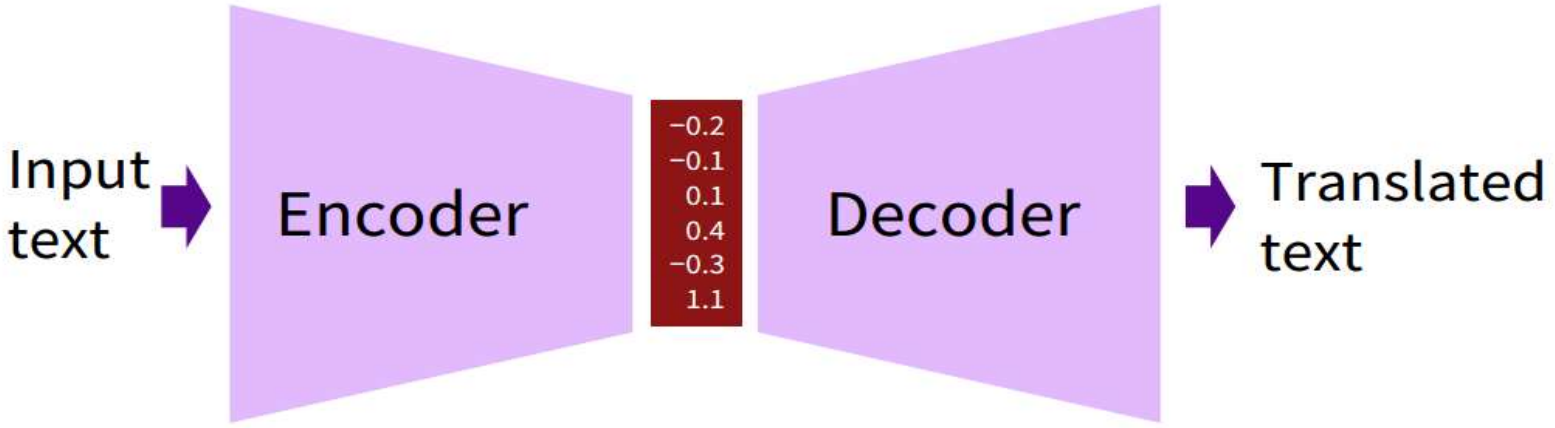
MT Approaches (6)

- Example-based translation
  - Translation example: pair of (source, target) sentences
- Statistical translation (French to English)
  - Use bilingual alignment corpus composed of (e, f)



MT Approaches (7)

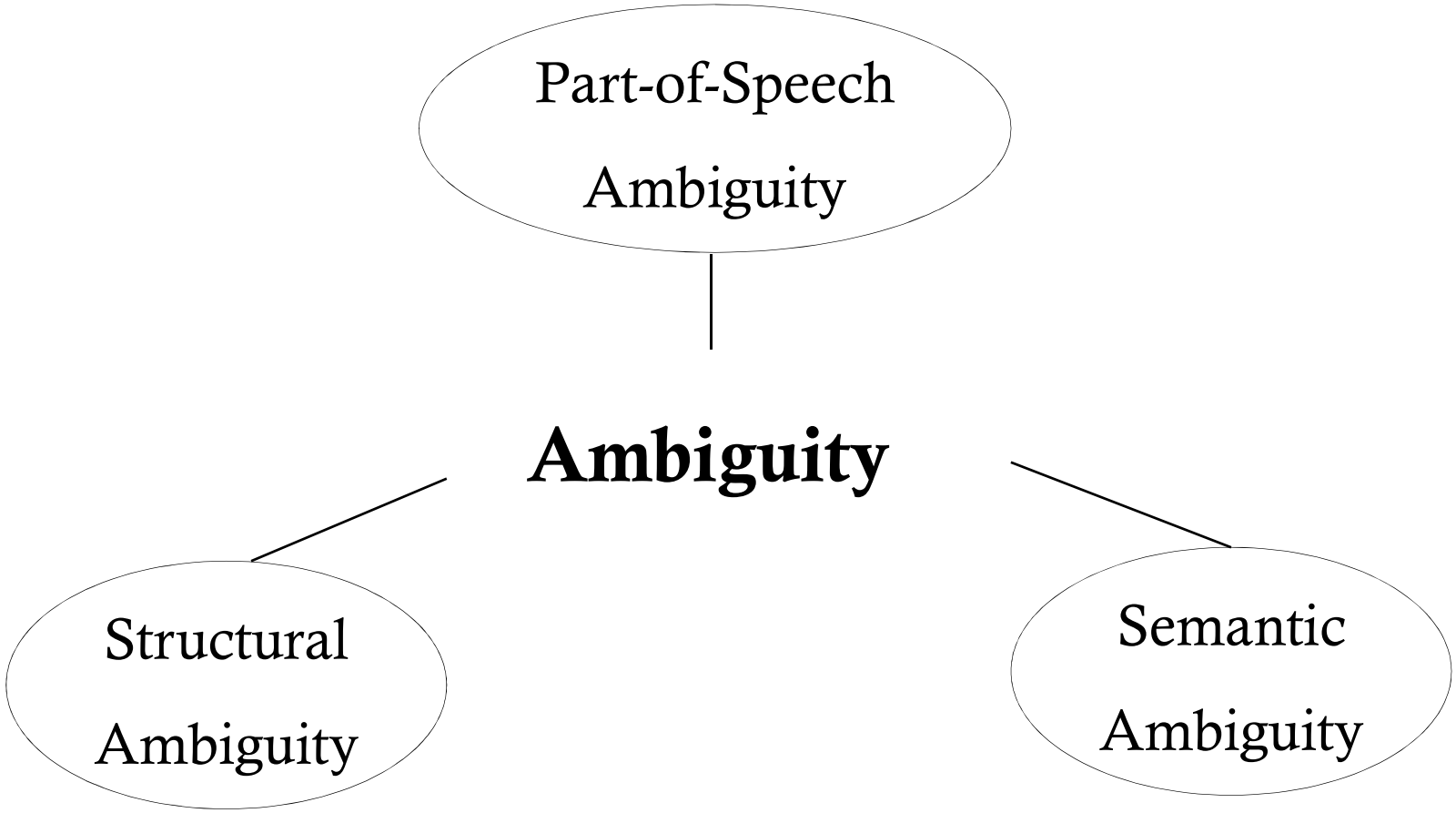
- Neural translation
  - Model the entire MT process via one big artificial neural network



Thang Luong, Kyunghyun Cho, and Cristopher Manning, Neural Machine Translation – Tutorial ACL 2016

Problems in English-Korean MT

Problems in English-Korean MT (1)



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

- Rule-based
- Transfer-based

+

- Idiom translation
- Sentence segmentation

Statistical methods using corpus

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

bread and butter

빵과 버터

→ 버터 바른 빵

provide him with money (provide A with B)

돈을 가지고 (가진) 그를 제공하다

→ 그에게 돈을 제공하다

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

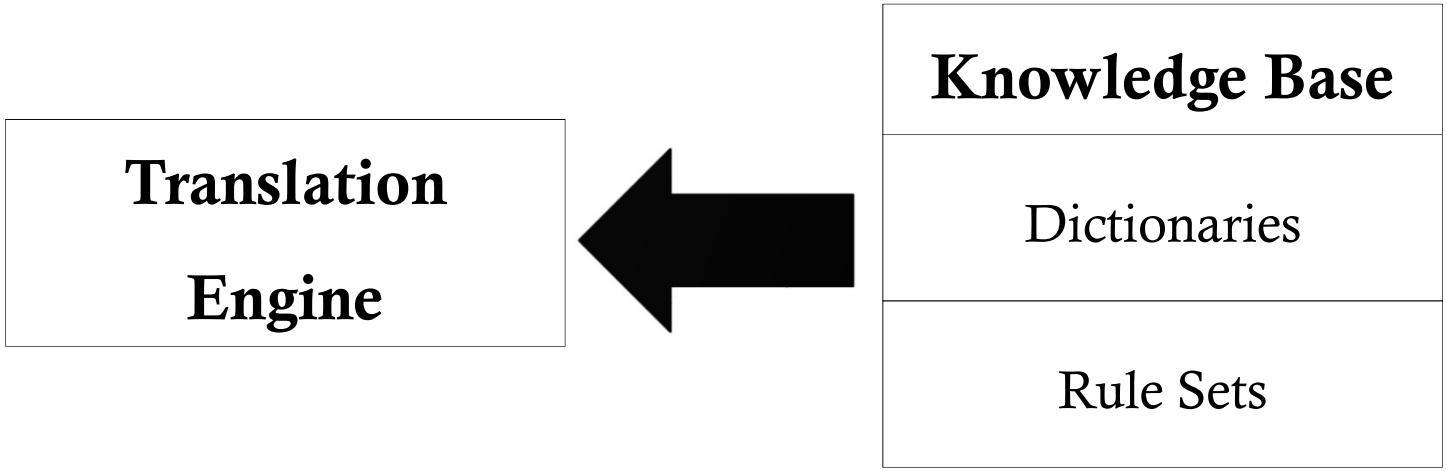
$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(x, y)\right)$$

x : context of a word, y : 0 or 1
f : feature, λ : weight
$Z(x)$: normalizing constant

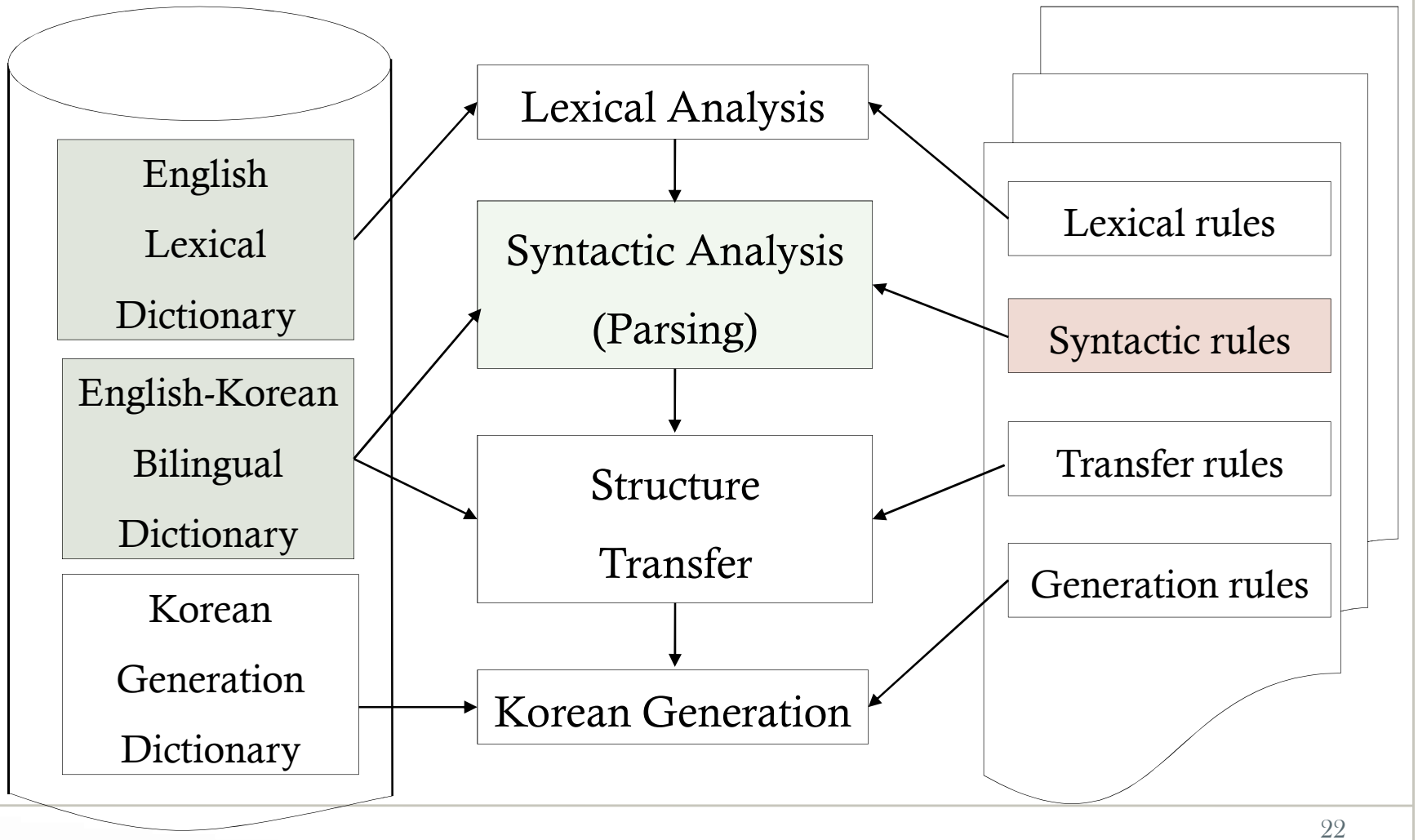
English-Korean MT system

- Structure

English-Korean MT system - Structure (1)



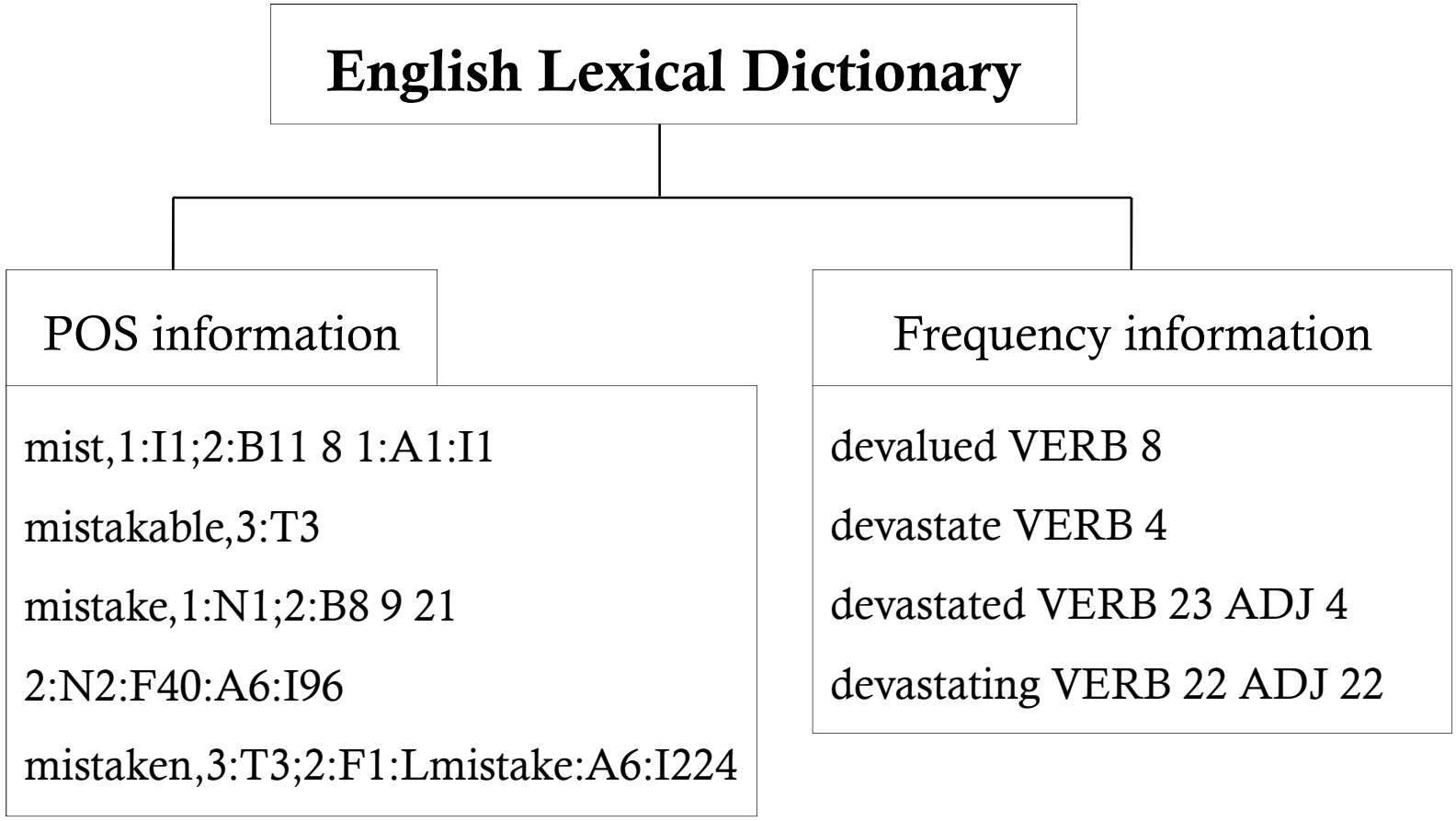
English-Korean MT system - Structure (2)



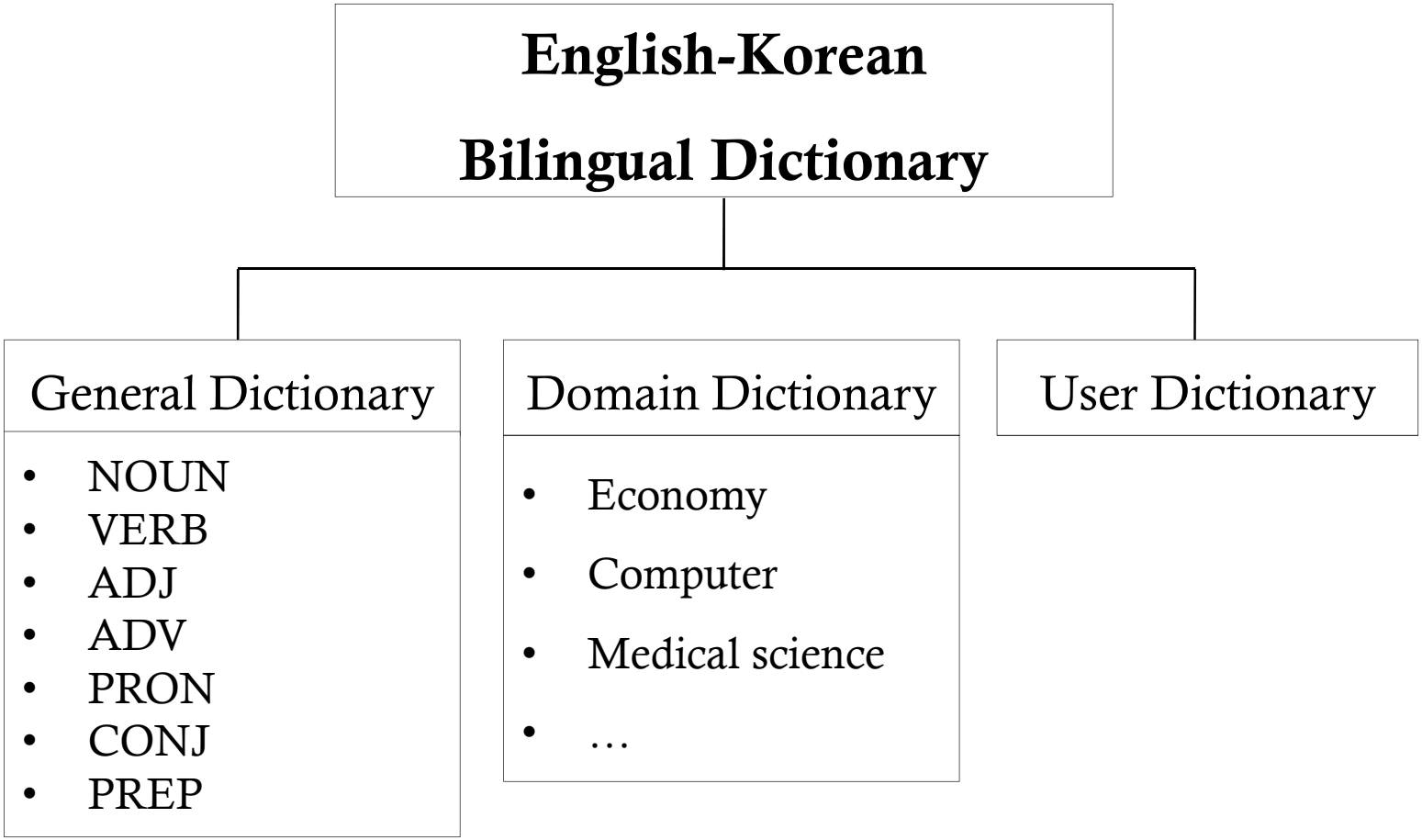
English-Korean MT system

- Development

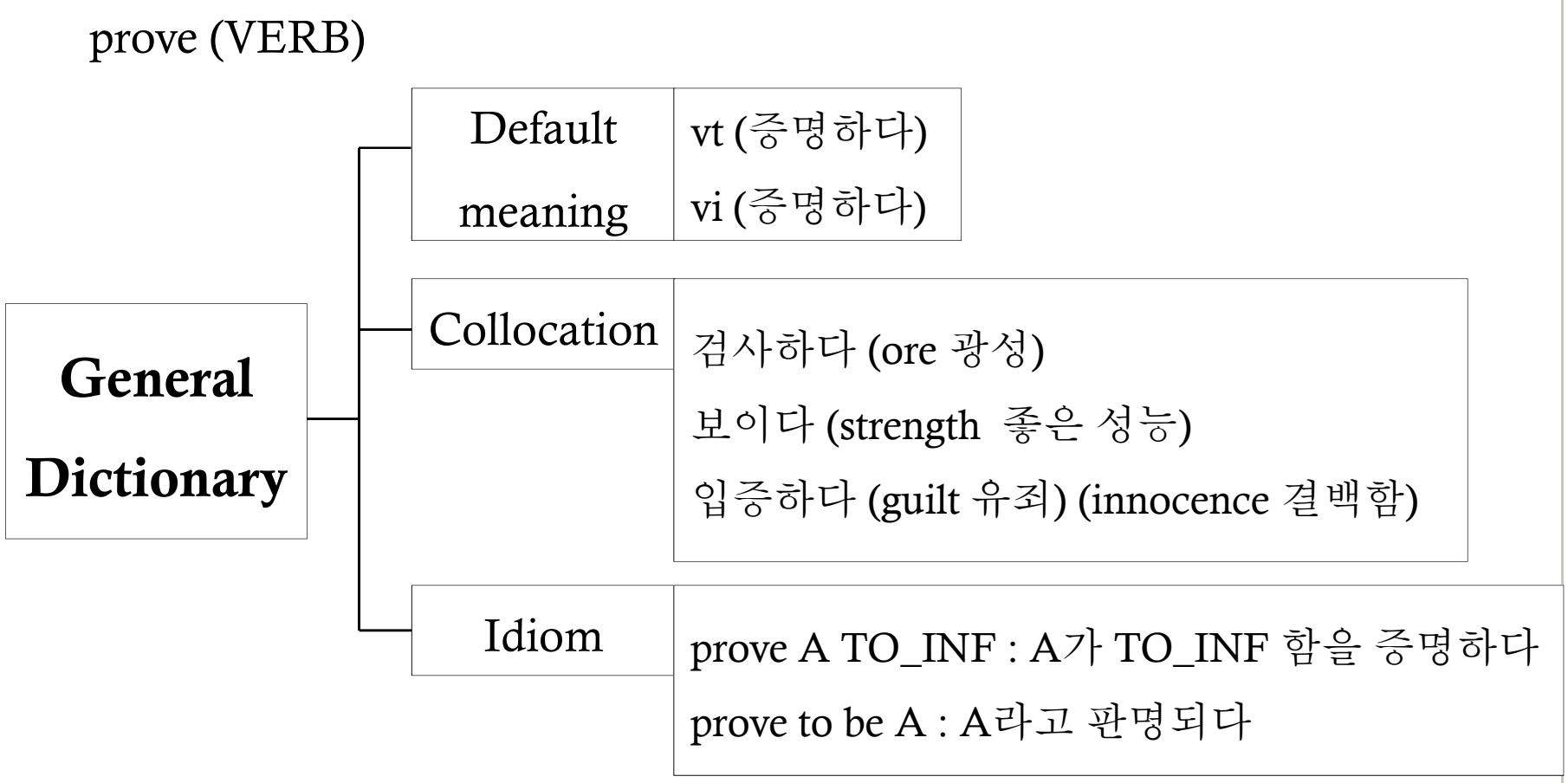
English-Korean MT system - Development – Knowledge Base (1)



English-Korean MT system - Development – Knowledge Base (2)



English-Korean MT system
- Development – Knowledge Base (3)



English-Korean MT system
- Development – Knowledge Base (4)

Domain Dictionary

"finance"

- finance : 금융
- finance bill : 금융 어음
- finance company : 금융회사
- finance corporation : 공고
- finance for dead stock : 기초재고금융
- finance lease : 파이낸스 리스

English-Korean MT system
- Development – Knowledge Base (5)

Korean Generation Dictionary

"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

English-Korean MT system
- Development – Translation Engine (2)

I like the books

I – PRON (1.0)
like – VERB (0.59) PREP (0.34) ADJ (0.059) CONJ (0.05)
the – DET (1.0)
books – VERB (0.04) NOUN (0.96)

English-Korean MT system - Development – Translation Engine (3)

- Parser = Syntactic Analyzer
  - Chart parser
    - T. Winograd, “Language as a Cognitive Process: Syntax”, volume 1, Addison-Wesley, 1983
    - Context free English grammar
      - $O(n^3)$

James Allen, “Natural Language Understanding”

Grammar

(1) $S \rightarrow NP\ VP$

(2) $NP \rightarrow ART\ ADJ\ N$

(3) $NP \rightarrow ART\ N$

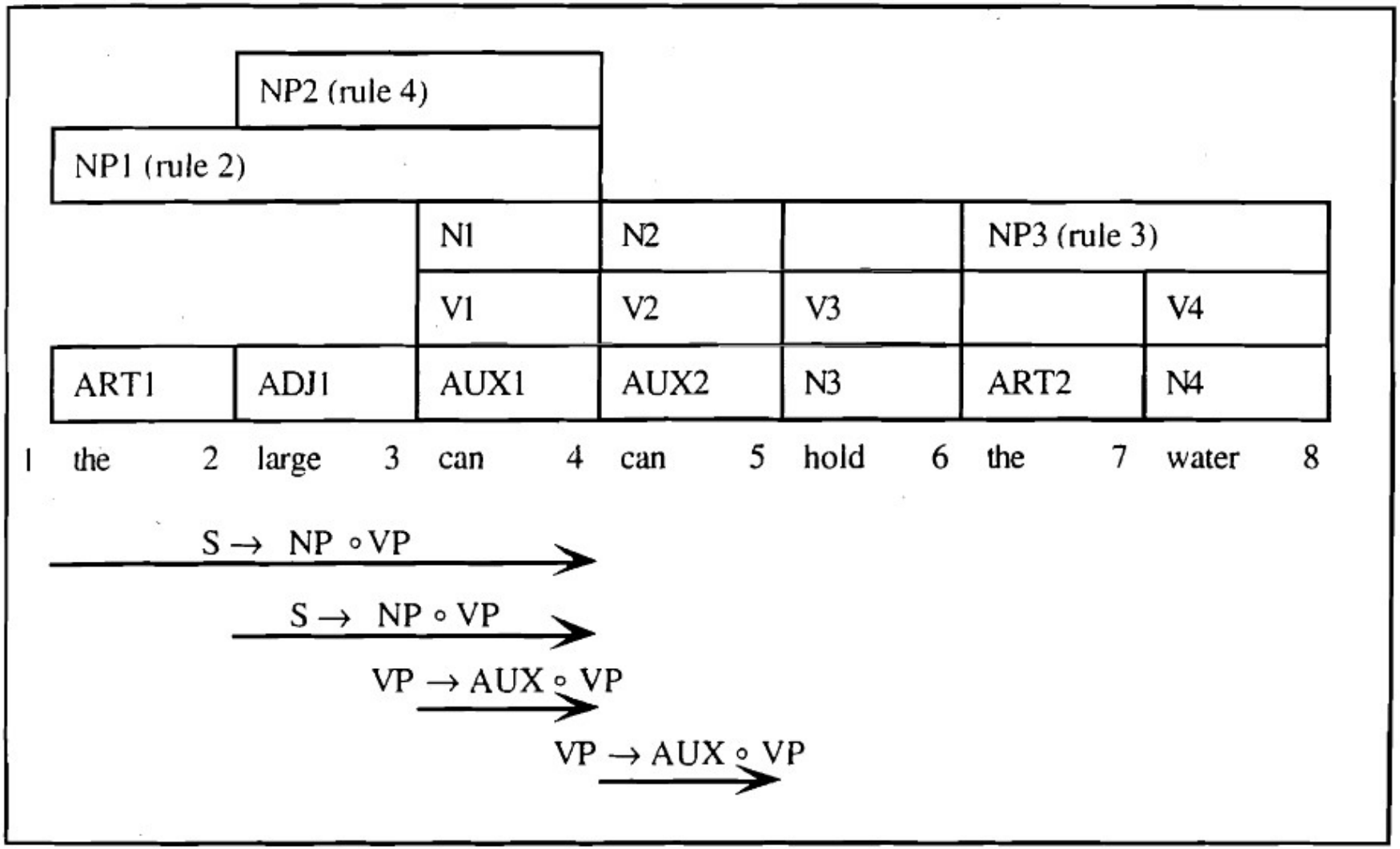
(4) $NP \rightarrow ADJ\ N$

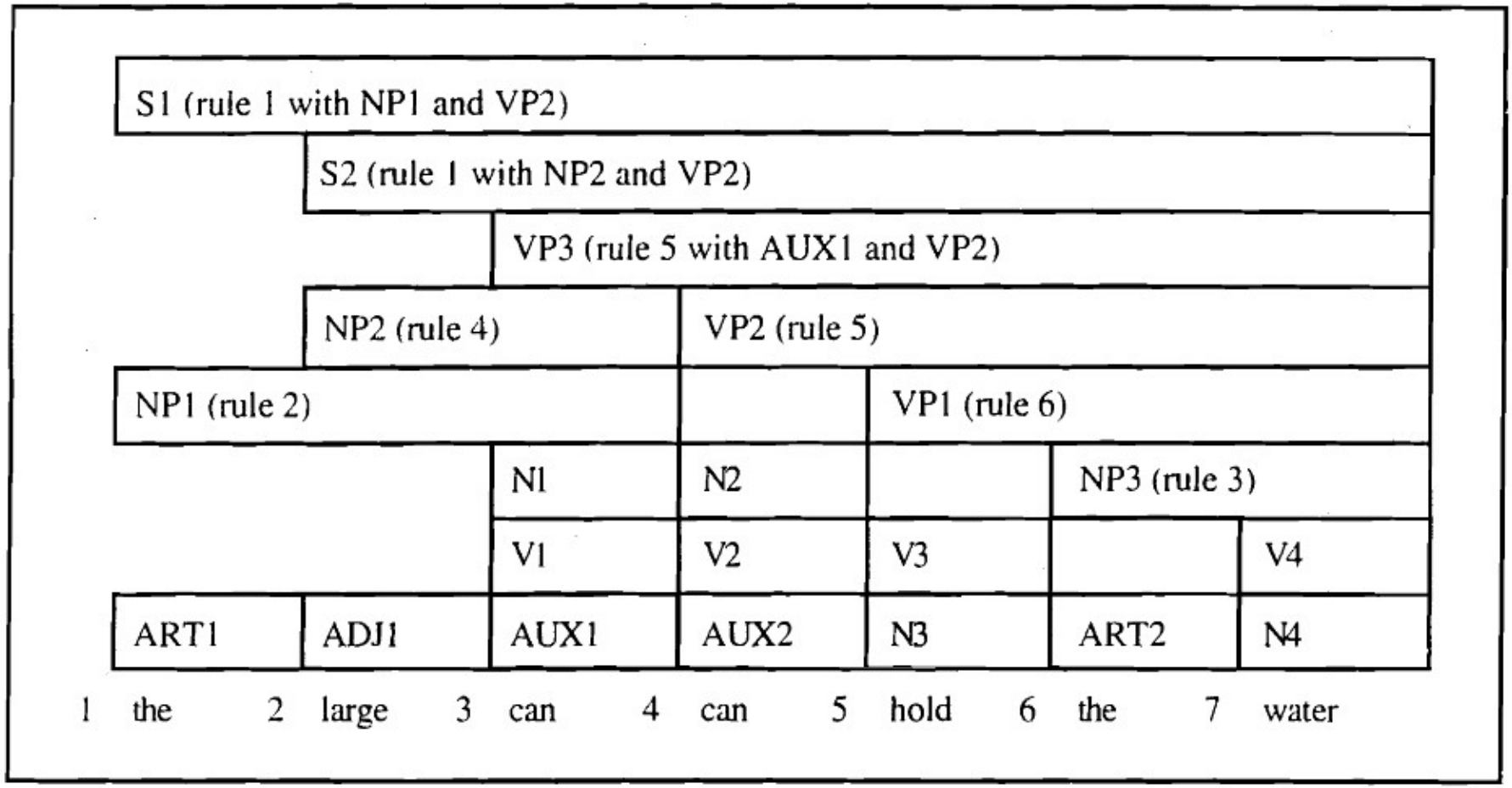
(5) $VP \rightarrow AUX\ VP$

(6) $VP \rightarrow V\ NP$

Input: The large can can hold the water

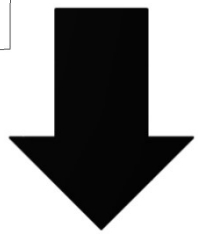
- the: *ART*
- large: *ADJ*
- can: *N, AUX, V*
- hold: *N, V*
- water: *N, V*


NP2 (rule 4)
NP1 (rule 2)
N1
N2
NP3 (rule 3)
V1
V2
V3
V4
ART1
ADJ1
AUX1
AUX2
N3
ART2
N4
1 the 2 large 3 can 4 can 5 hold 6 the 7 water 8
S → NP ∘ VP
S → NP ∘ VP
VP → AUX ∘ VP
VP → AUX ∘ VP


S1 (rule 1 with NP1 and VP2)
S2 (rule 1 with NP2 and VP2)
VP3 (rule 5 with AUX1 and VP2)
NP2 (rule 4)
VP2 (rule 5)
NP1 (rule 2)
VP1 (rule 6)
N1
N2
NP3 (rule 3)
V1
V2
V3
V4
ART1
ADJ1
AUX1
AUX2
N3
ART2
N4
1 the 2 large 3 can 4 can 5 hold 6 the 7 water

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
    - …

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

is

+— [SUBJ]→ it

+— [SCOMP] → good

+— [TOINF] → know

+— [OBJ] → truth

SCOMP is

+—[SCOMP] → good

+—[SUBJ] → know

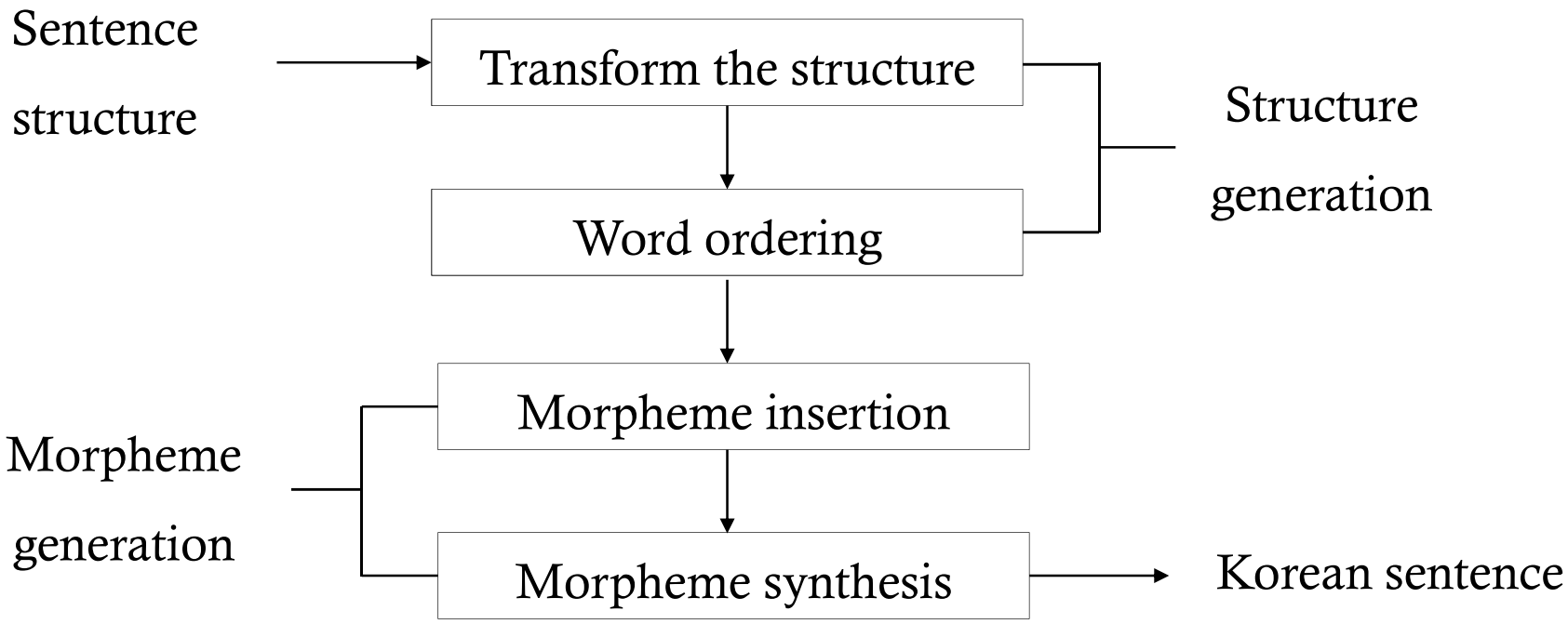
+— [OBJ] → truth

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



English-Korean MT system

- Problems & Solutions

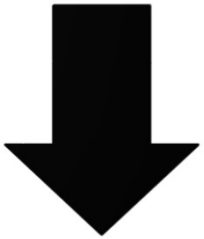
English-Korean MT system - Problems & Solutions (1)

Speed

Parsing complexity = $O(n^3)$

Memory, Structural ambiguity

Too many structures during parsing



- **POS determination**
- **Intra-sentence segmentation**
- **3-phase parsing**

English-Korean MT system - Problems & Solutions (2)

- POS determination
  - Reduce POS ambiguity → 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 → several short segments → parsing unit
  - Try to reduce *n* in $O(n^3)$ → 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.

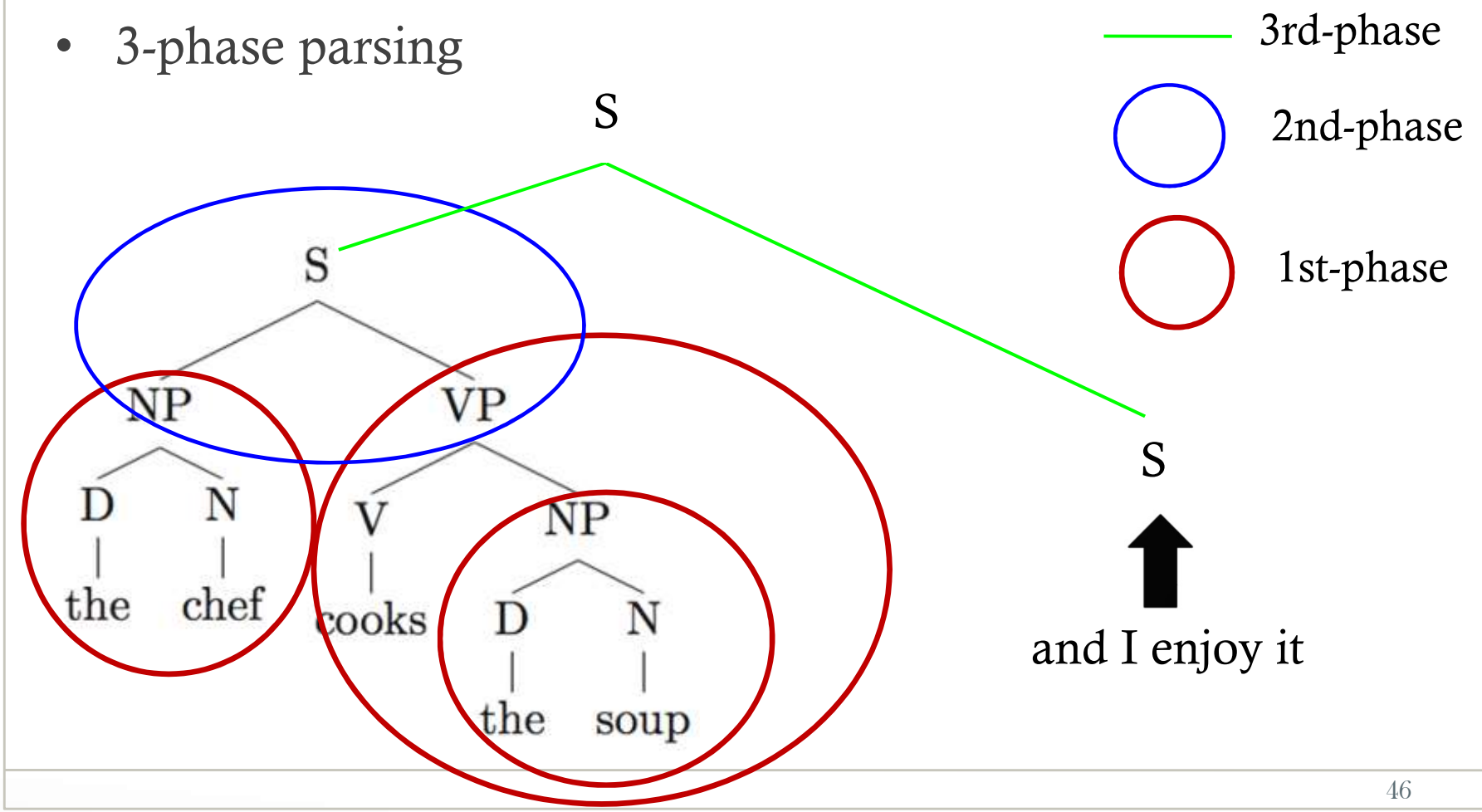
We also analyze the effect of various choices

+

while inducing word embeddings on “downstream” POS induction results.

English-Korean MT system - Problems & Solutions (4)

- 3-phase parsing



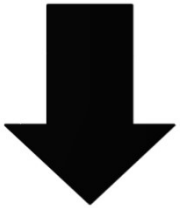
English-Korean MT system - Problems & Solutions (5)

- 3-phase parsing
  - Syntactic rule classification
  - Rule acquisition for 3rd-phase paring

English-Korean MT system - Problems & Solutions (6)

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 ~], …
- …



Preprocessing

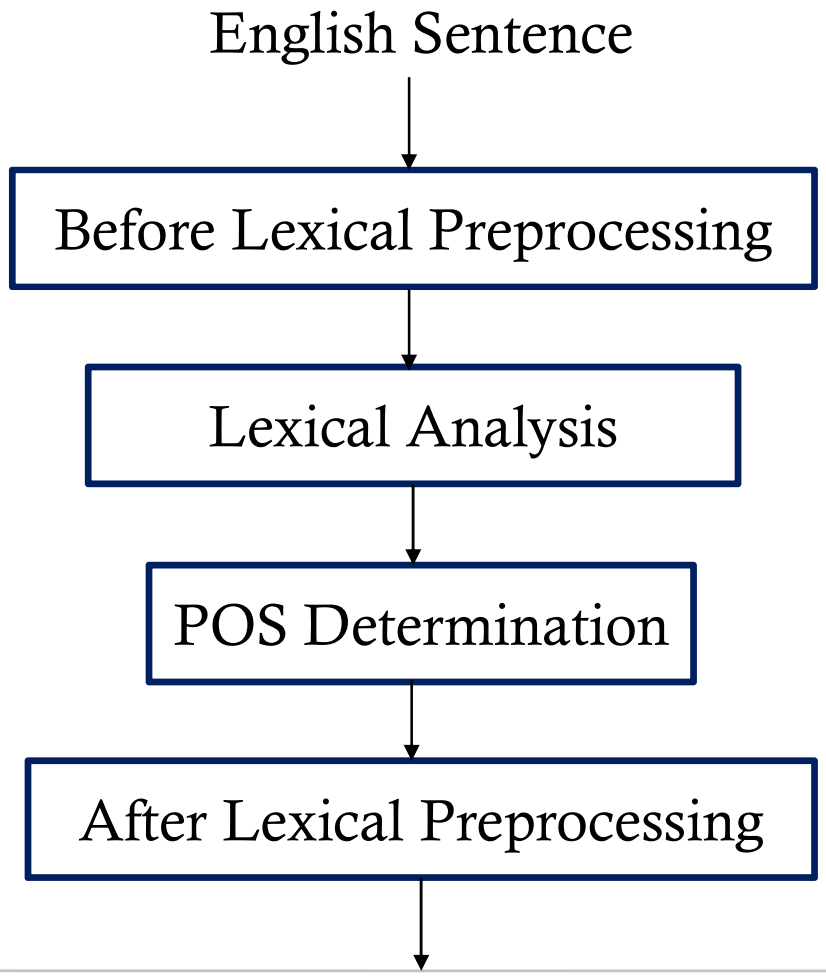
English-Korean MT system - Problems & Solutions (7)

- Preprocessing – post-processing
  - Pre-lexical analysis preprocessing
  - Post-lexical analysis preprocessing
  - Post-segmentation preprocessing
  - Post processing

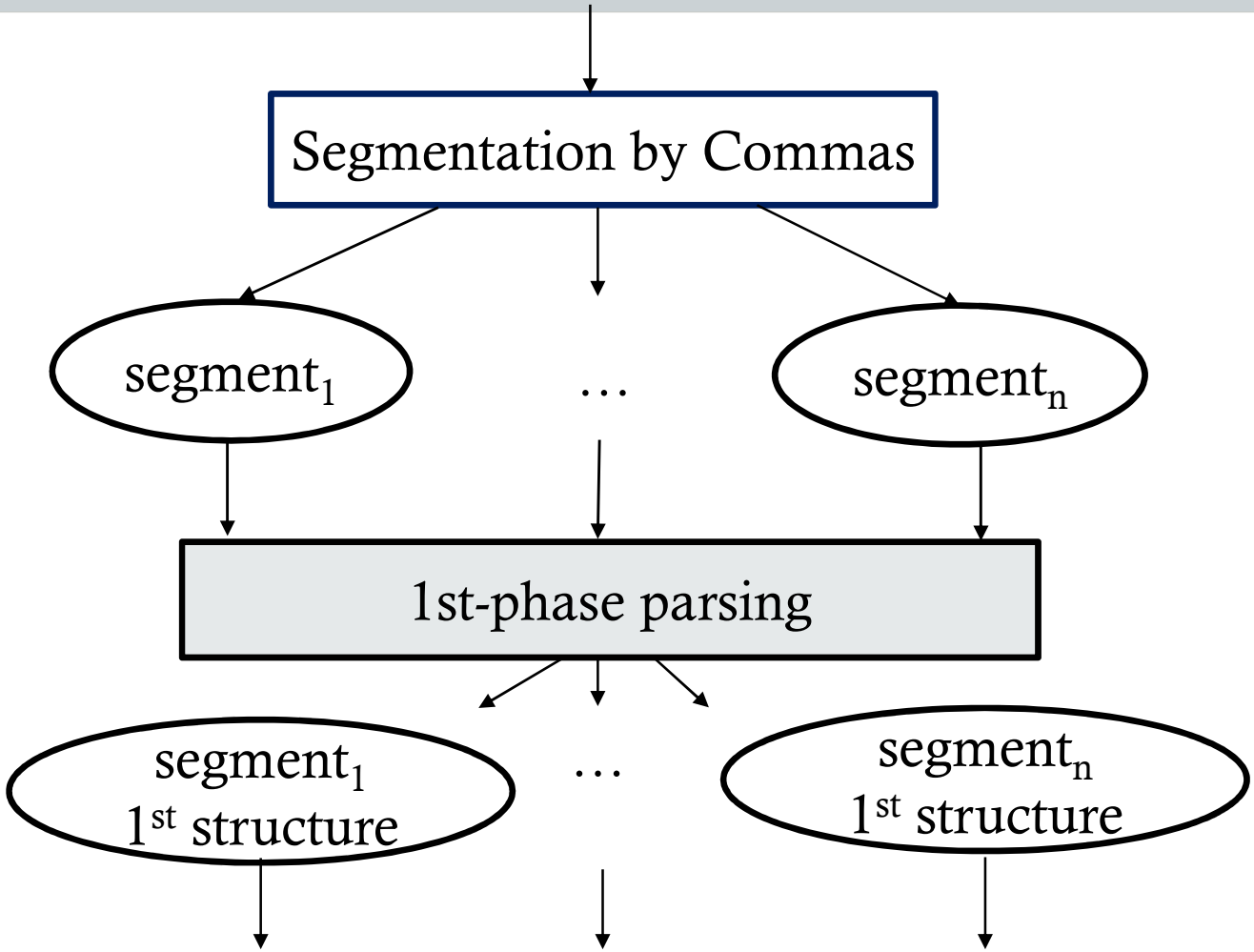
English-Korean MT system

- Structure / Translation Flow

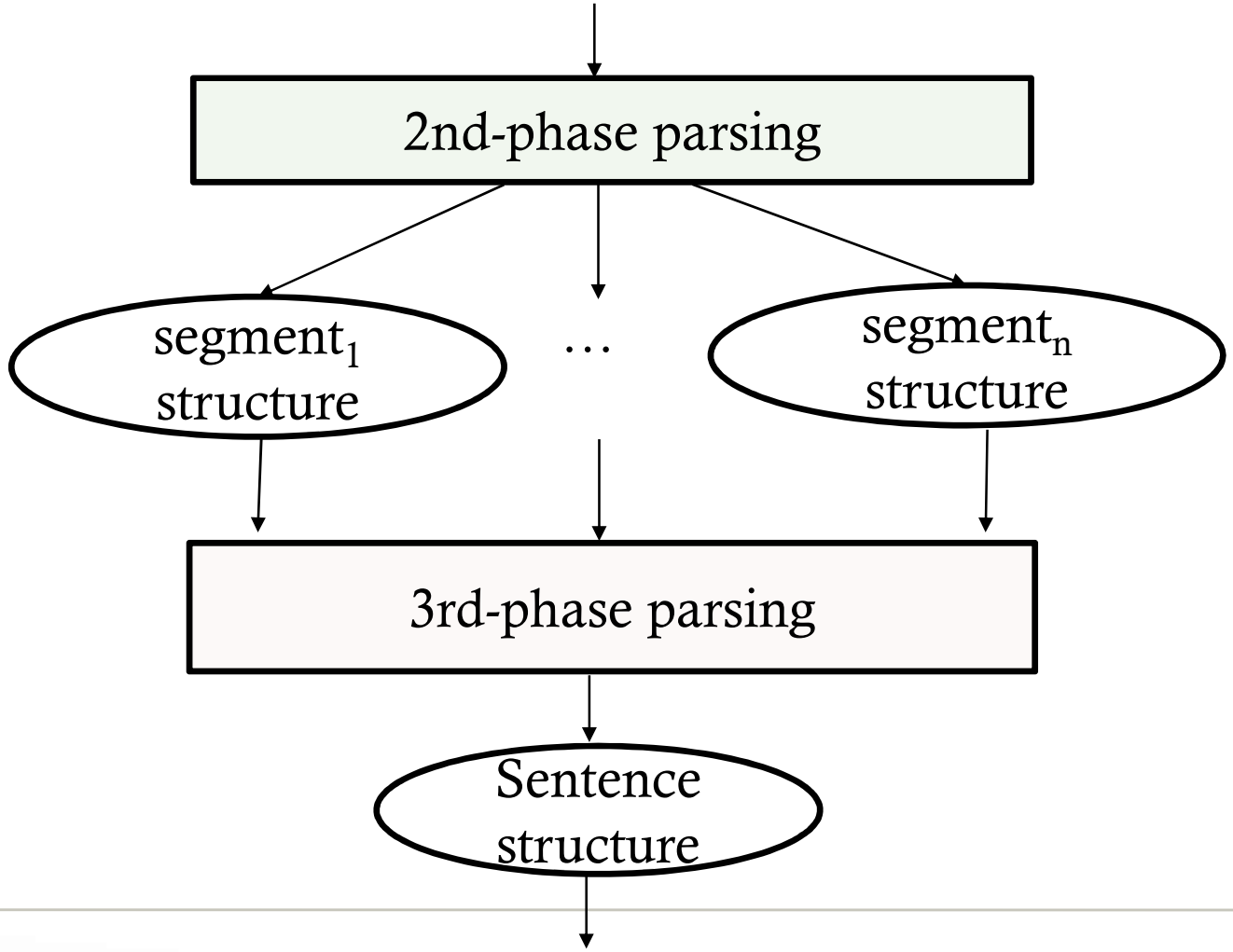
English-Korean MT system - Structure / Translation flow (1)



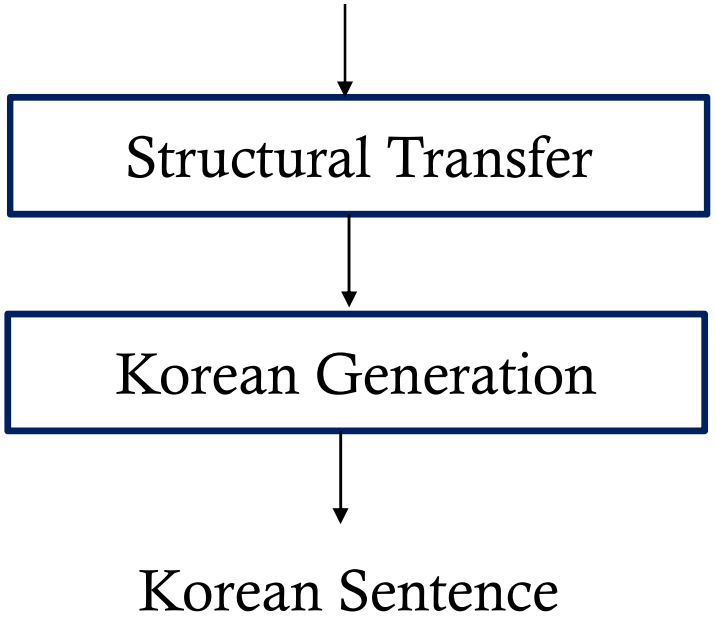
English-Korean MT system - Structure / Translation flow (2)



English-Korean MT system - Structure / Translation flow (3)



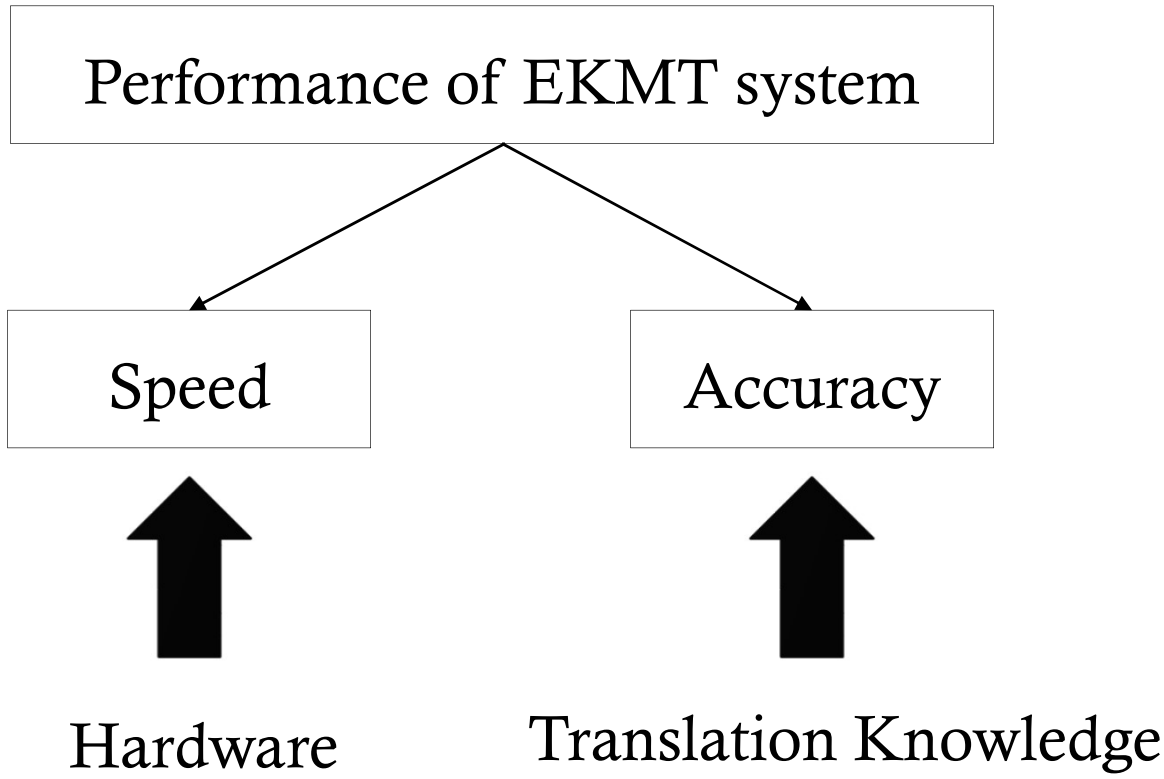
English-Korean MT system - Structure / Translation flow (4)



English-Korean MT system

- Works for continuous improvement

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



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

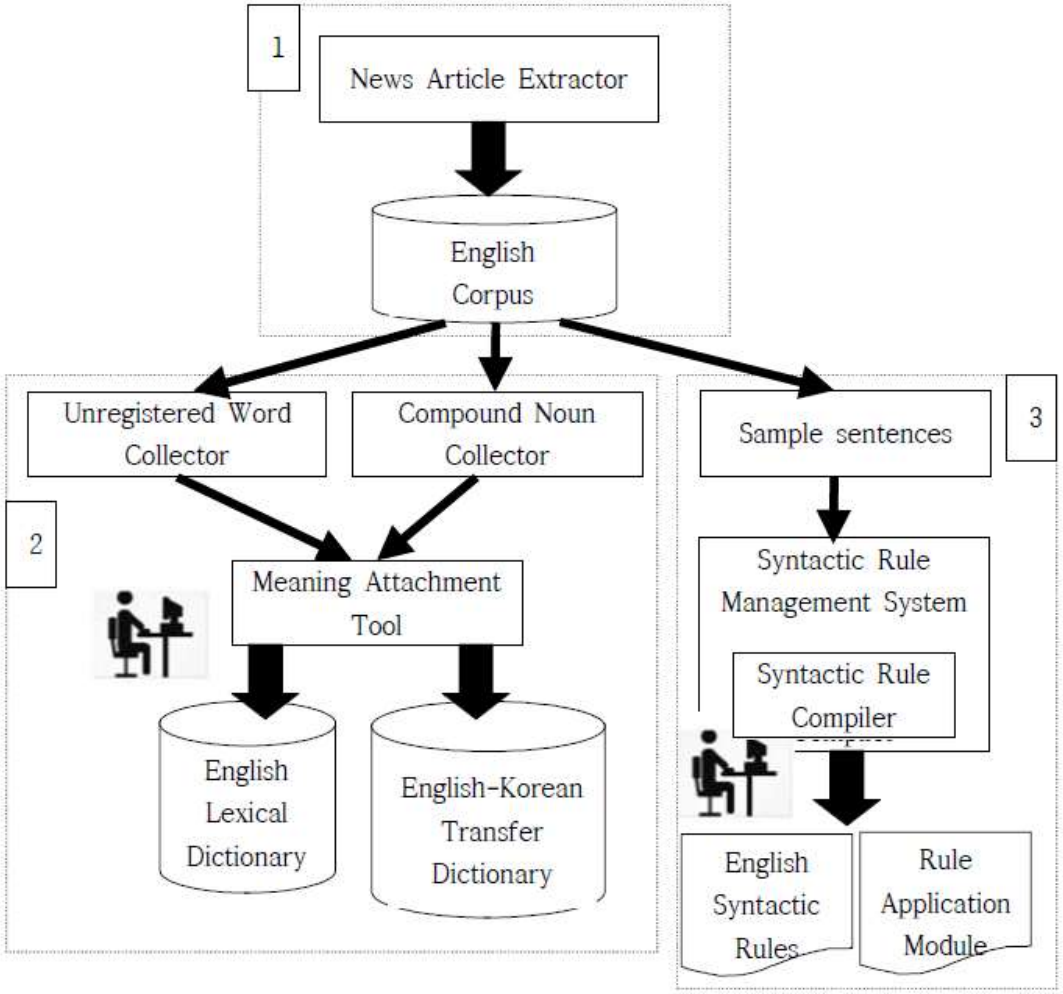
Building

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

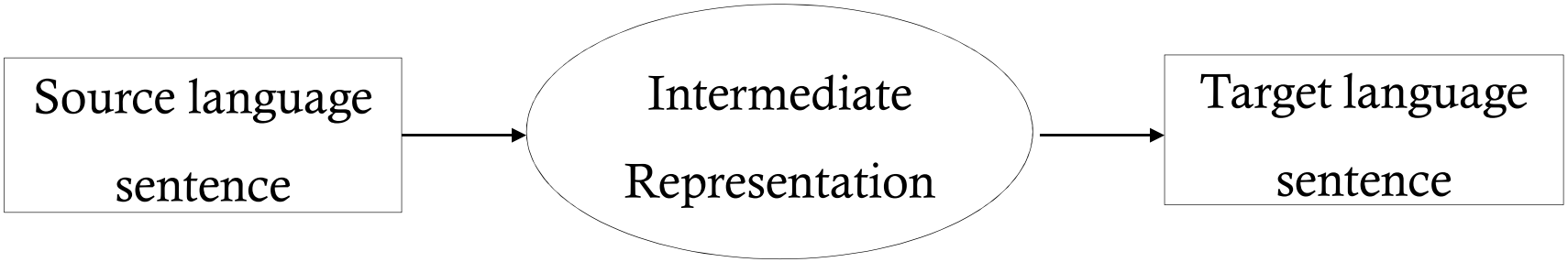
Neural Machine Translation

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
  - Cho et al., “Learning phrase representation using RNN encoder-decoder for statistical machine translation,” EMNLP 2014
  - Cho et al., “On the properties of Neural Machine Translation: Encode-Decoder Approaches,” 2014

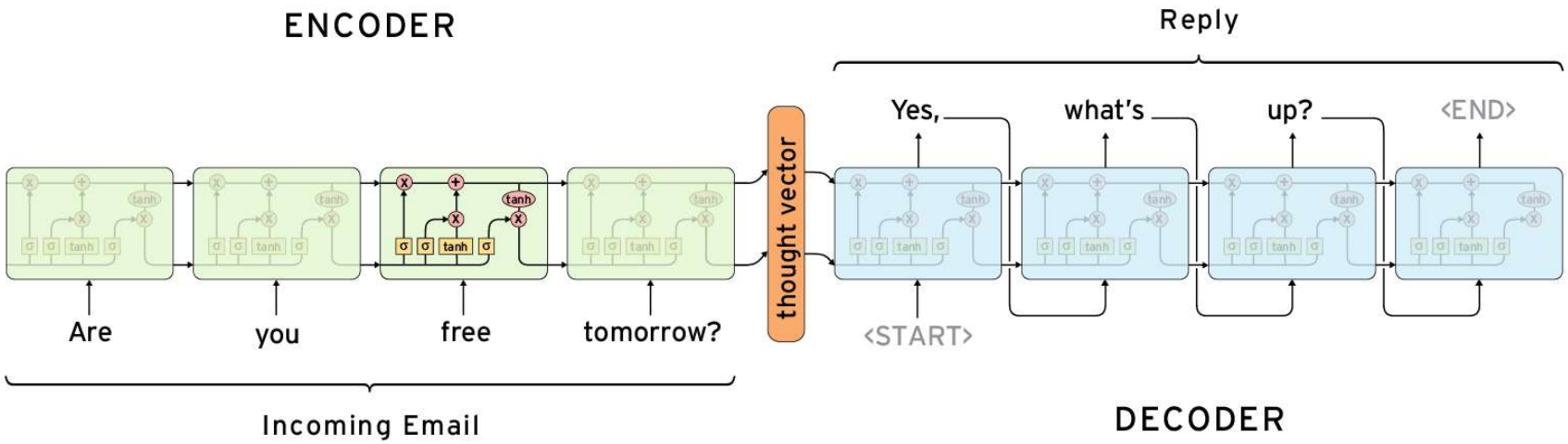
Neural Machine Translation (2)

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



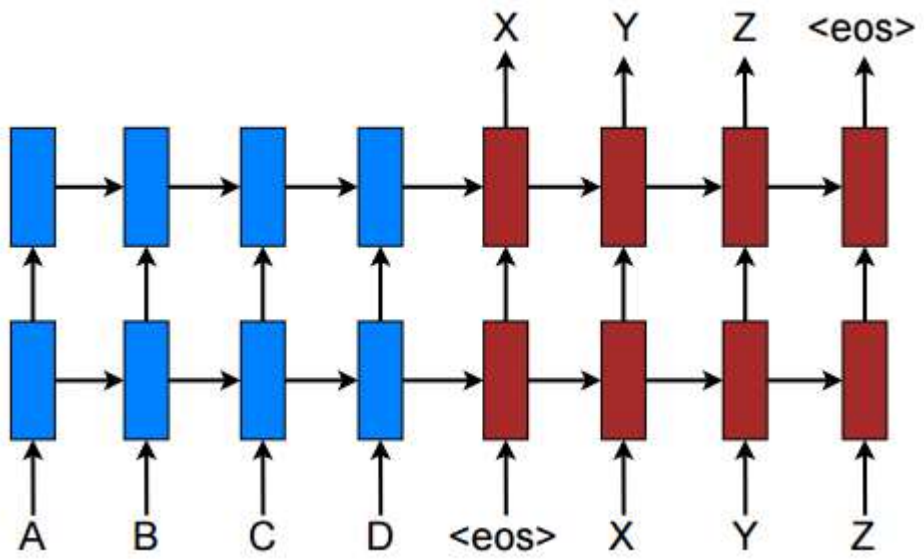
Neural Machine Translation (3)

- Basic framework: Encoder/Decoder



https://research.googleblog.com/2015/11/computer-respond-to-this-email.html

Neural Machine Translation (4)



Minh-Thang Luong, et al., "Effective Approaches to Attention-based Neural Machine Translation," EMNLP, 2015

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 → number
  - Relationship among words

$$vector(King) - vector(Man) + vector(Woman) \approx vector(Queen)$$

$$vector(Paris) - vector(France) + vector(Italy) \approx vector(Rome)$$

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, 0271]

Roelof Pieters, "Deep Learning for NLP: An Introduction to Neural Word Embeddings," 2014

Neural Machine Translation (7)

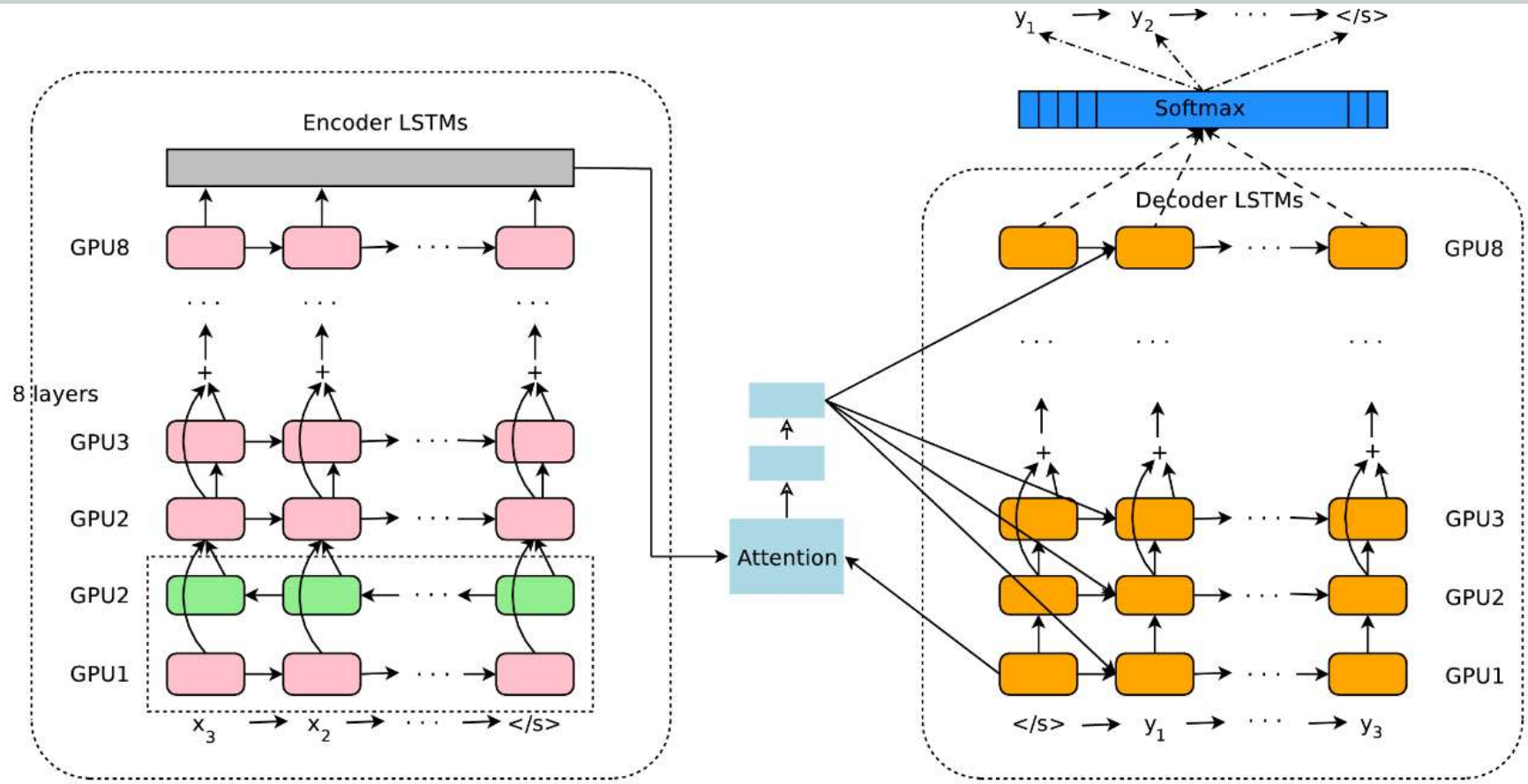
- Word2vec
  - Google
  - CBOW (continuous bag of words) model architecture
  - Skip-gram model architecture
  - Mikolov, T., Chen, K., Corrado, G., & Dean, J., “Efficient Estimation of Word Representations in Vector Space,” 2013

Google's NMT System

Google's NMT System (1)

- November, 2016
- "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," Technical Report, 2016

Google's NMT System (2)



Yonghui Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", 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)

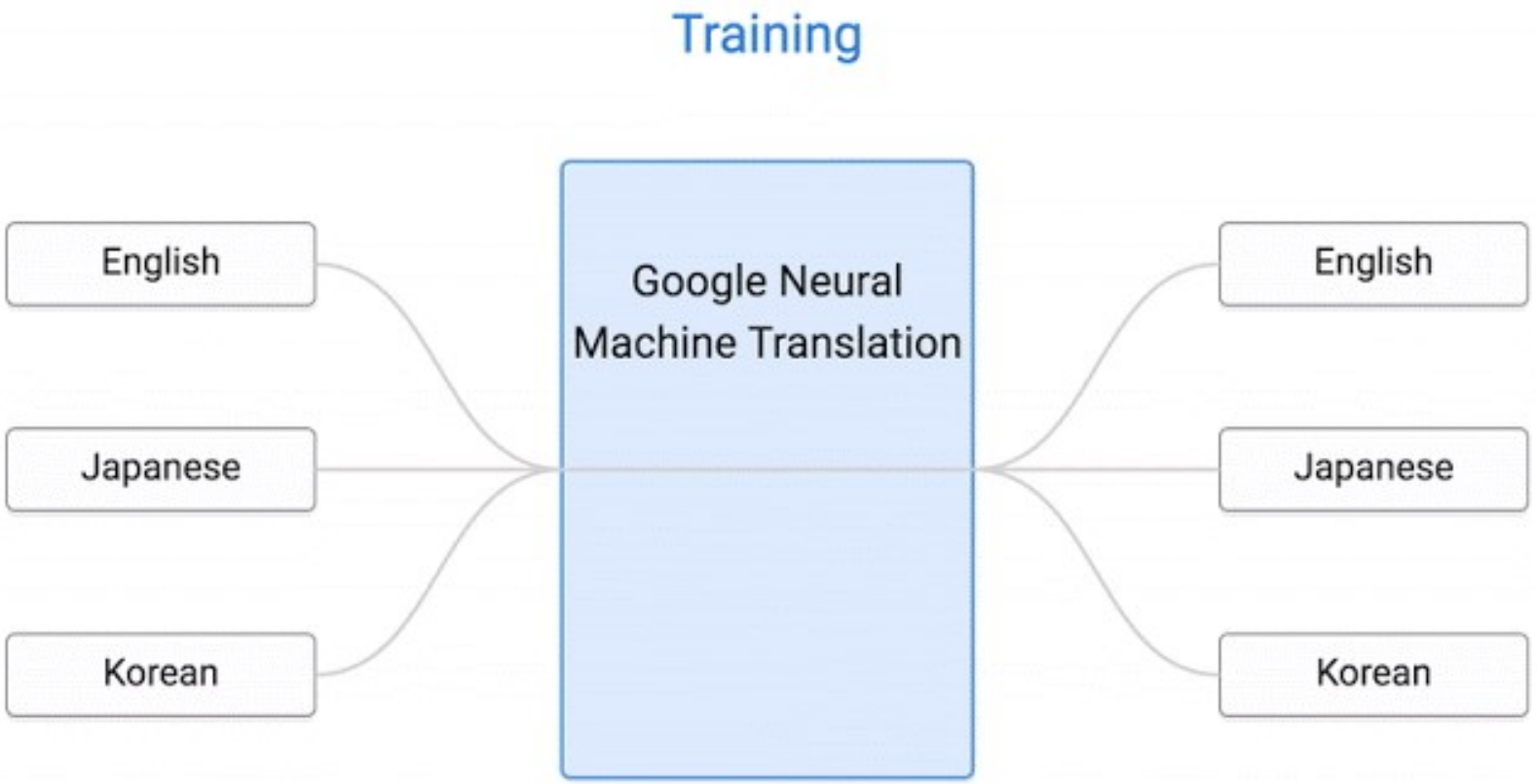
- Performance
  - Human-rated side-by-side comparison

GNMT	≈	Average bilingual human translator

  - Roughly 60% reduction in translation errors compared to the PBMT

Google's NMT System (6)

- Features
  - Zero-shot translation
    - Melvin Johnson et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," 2016
    - Training with examples English-Korean and English-Japanese, GNMT automatically does Japanese → Korean reasonably well
    - Transfer the "translation knowledge" → **transfer learning**



https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Google's NMT System (7)

- TensorFlow – ML toolkit
- Tensor Processing Unit

MT Applications

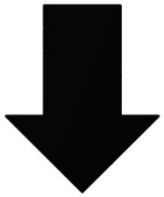
MT applications (1)

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

File format handling

MT applications (2)

- Video caption translation



Caption generation + video file handling

- Image translation

Optical character recognition

MT applications (3)

- Web translation

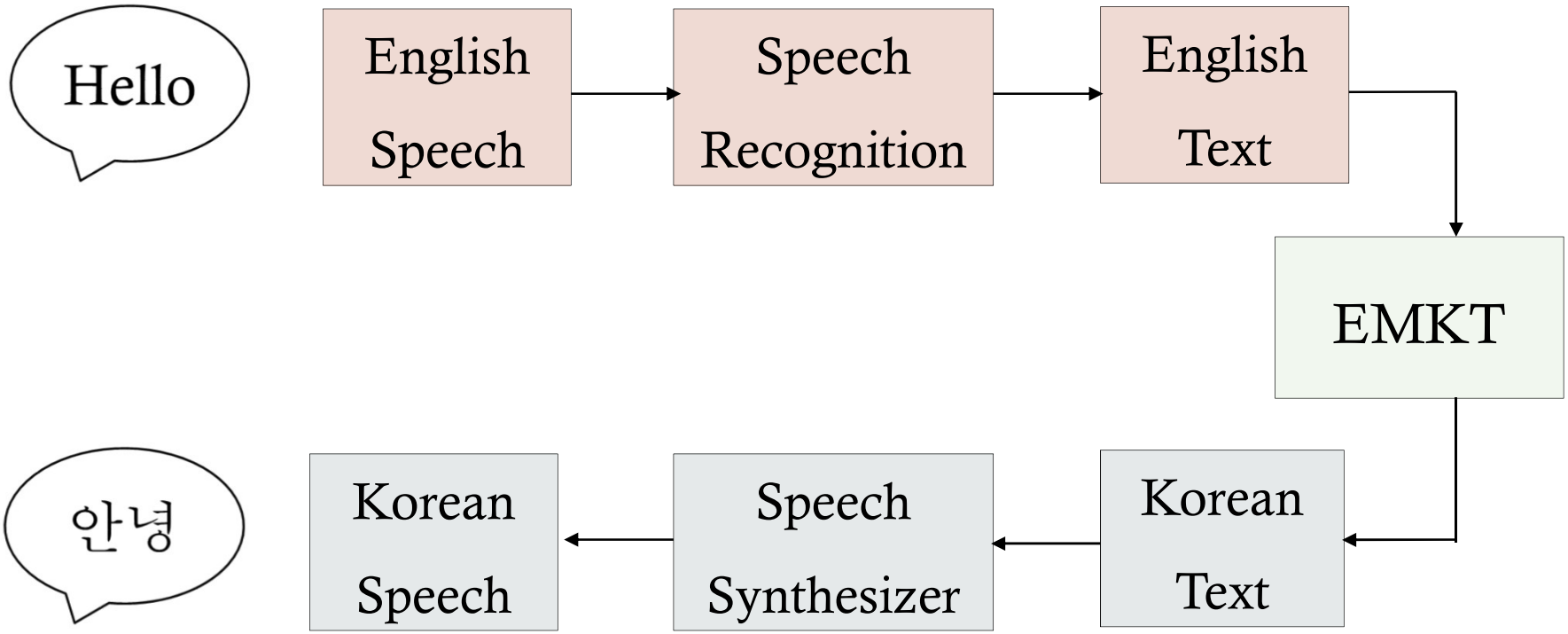
HTML

Document

Handing

MT applications (4)

- Speech translation



MT applications (5)

- Chat translation
- E-mail translation
- …

MT applications (6)

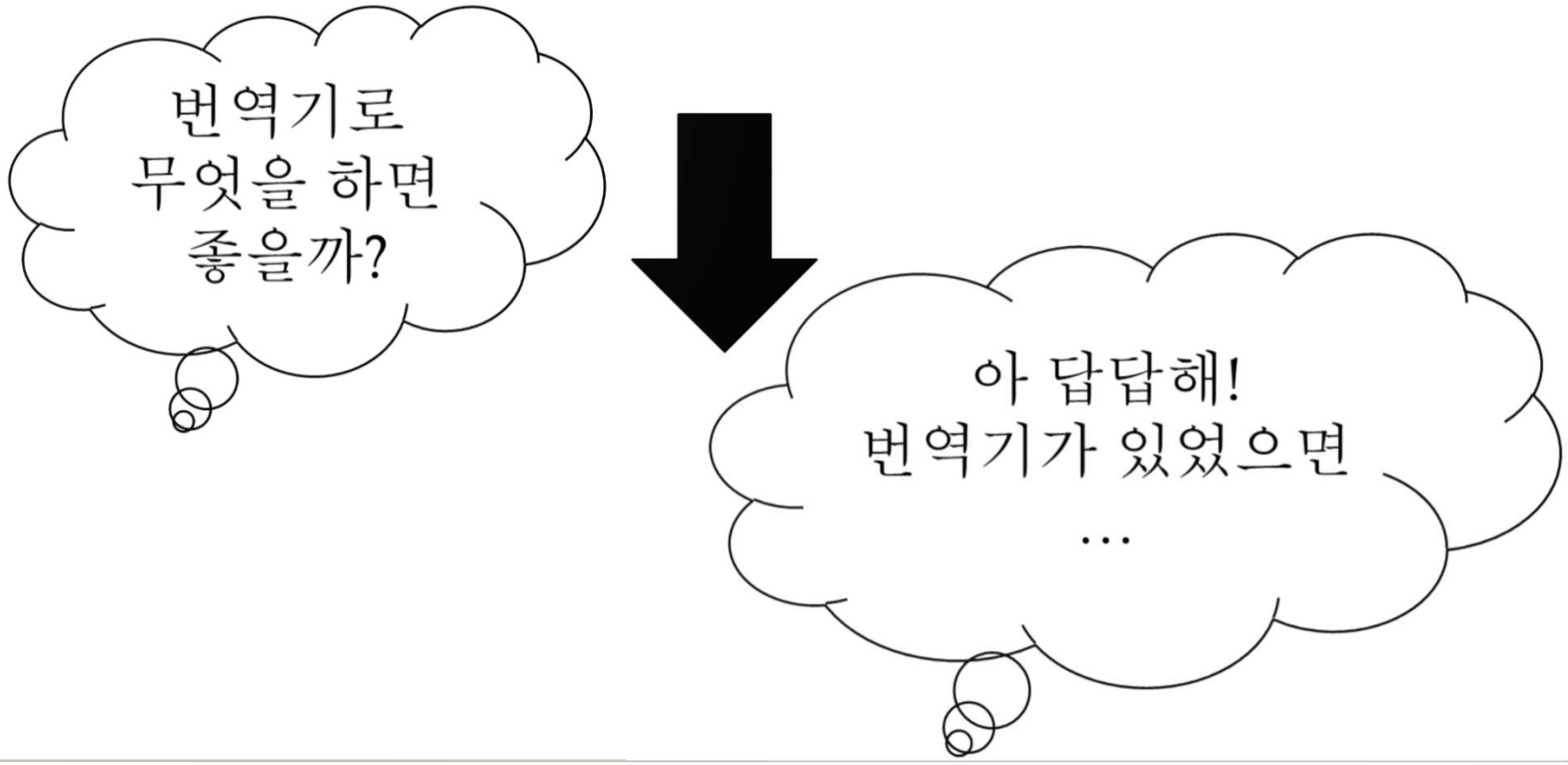
- Application areas
  - Study
    - Document translation
    - Video caption translation
    - Coursera, MOOC, SlideShare, …
  - Travel
    - Speech translation
    - Image translation

MT applications (7)

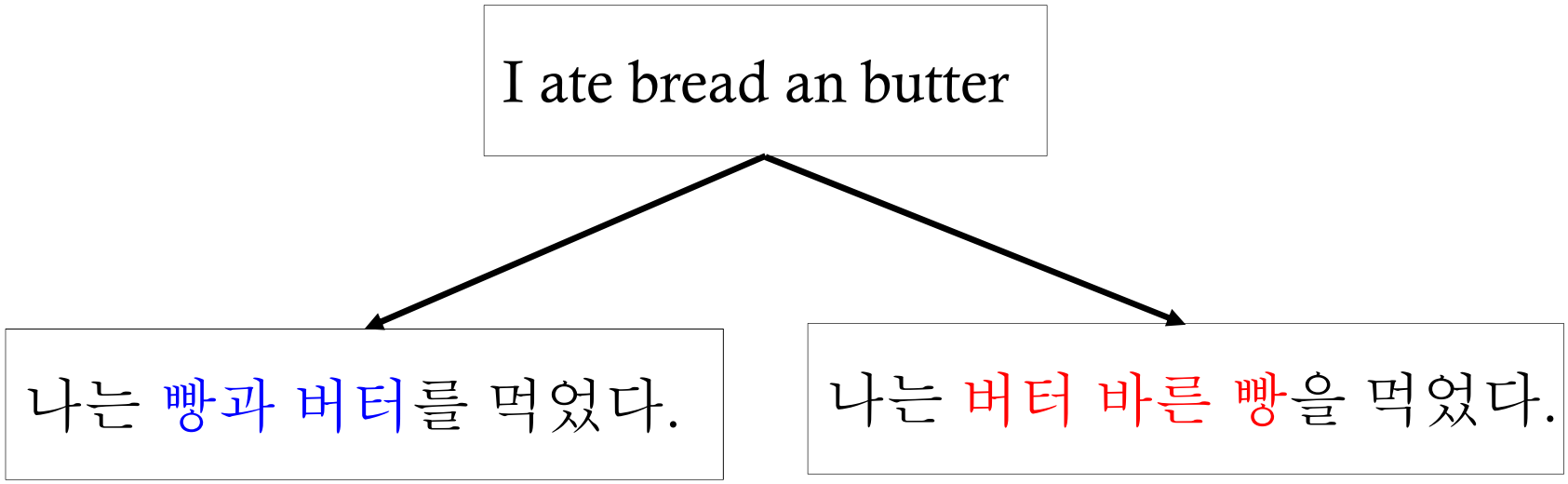
- Information retrieval
  - Web translation
- Communication
  - E-mail translation
  - Chatting translation
- Localization
  - Document translation

MT applications (8)

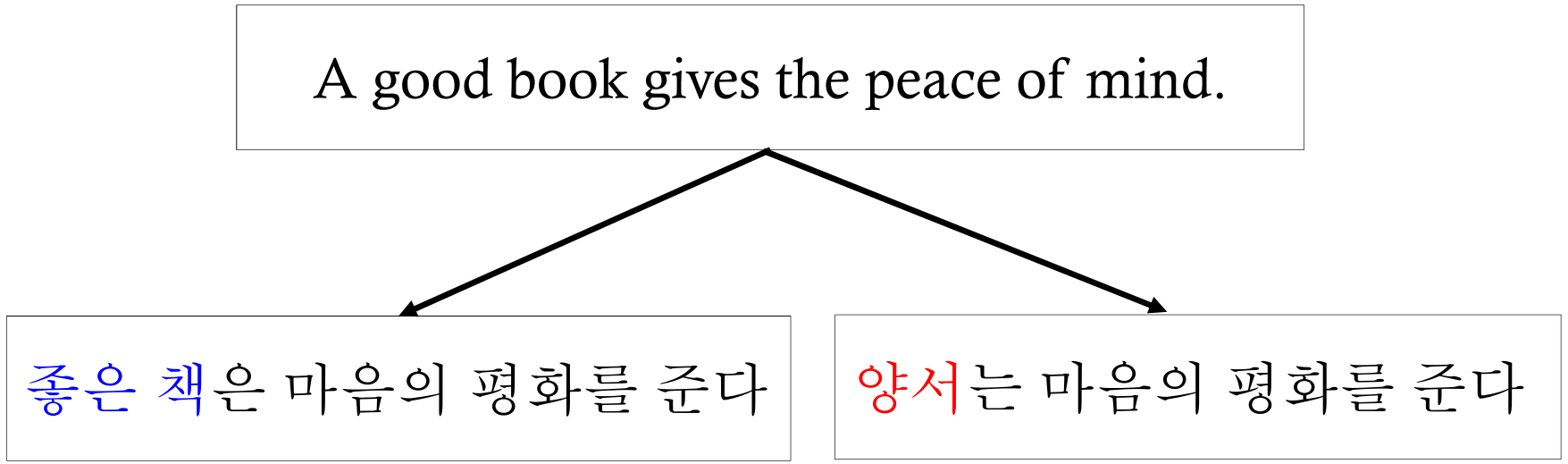
- …



Why old-fashioned MT approach ?



Why old-fashioned MT approach ?



감사합니다 !

sdkim@hansung.ac.kr