On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation

Authored by: **Liang Gu** and **Yuqing Gao**

Presented by: **Ruhi Sarikaya**

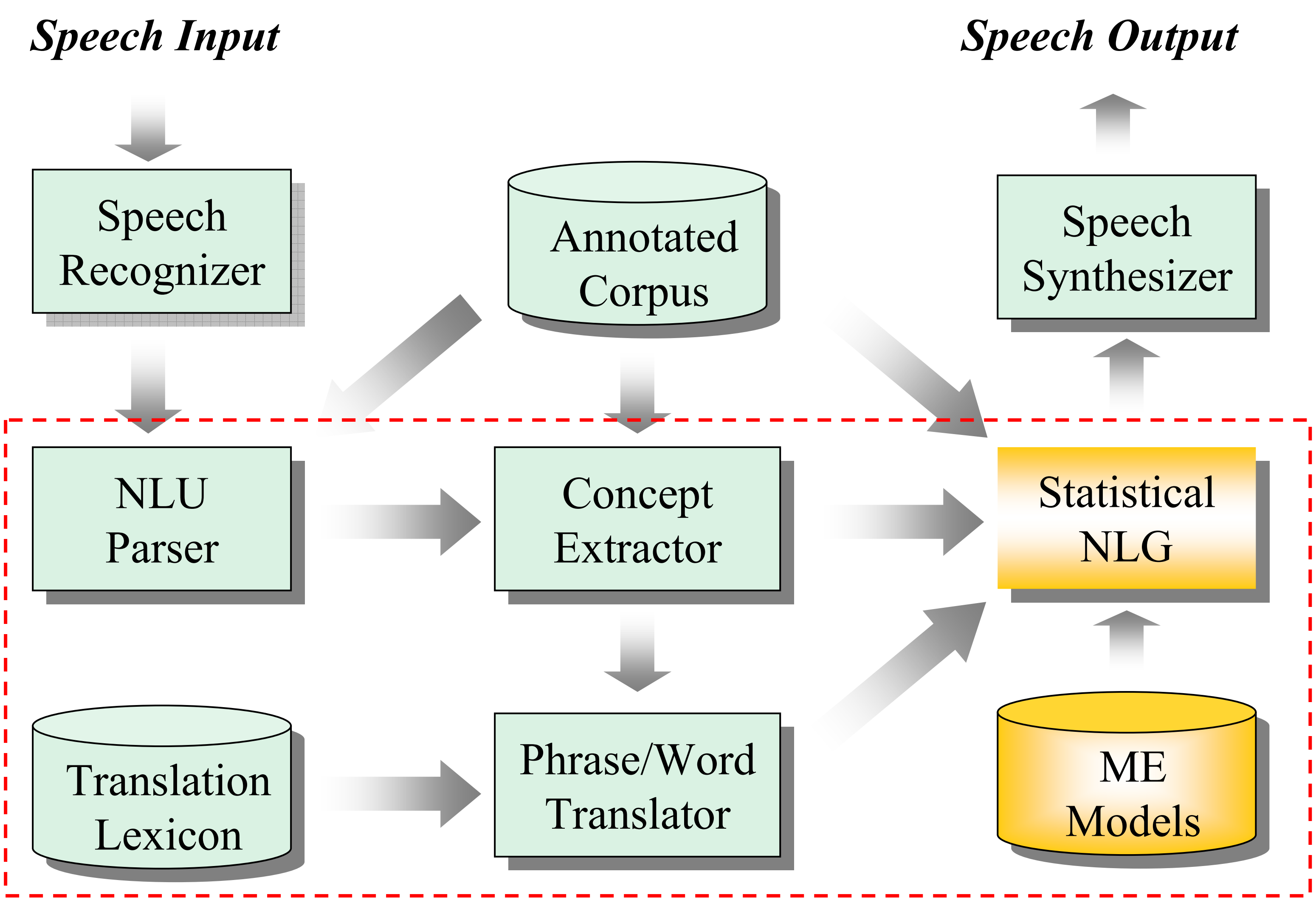
October 1st, 2004

IBM

Outline

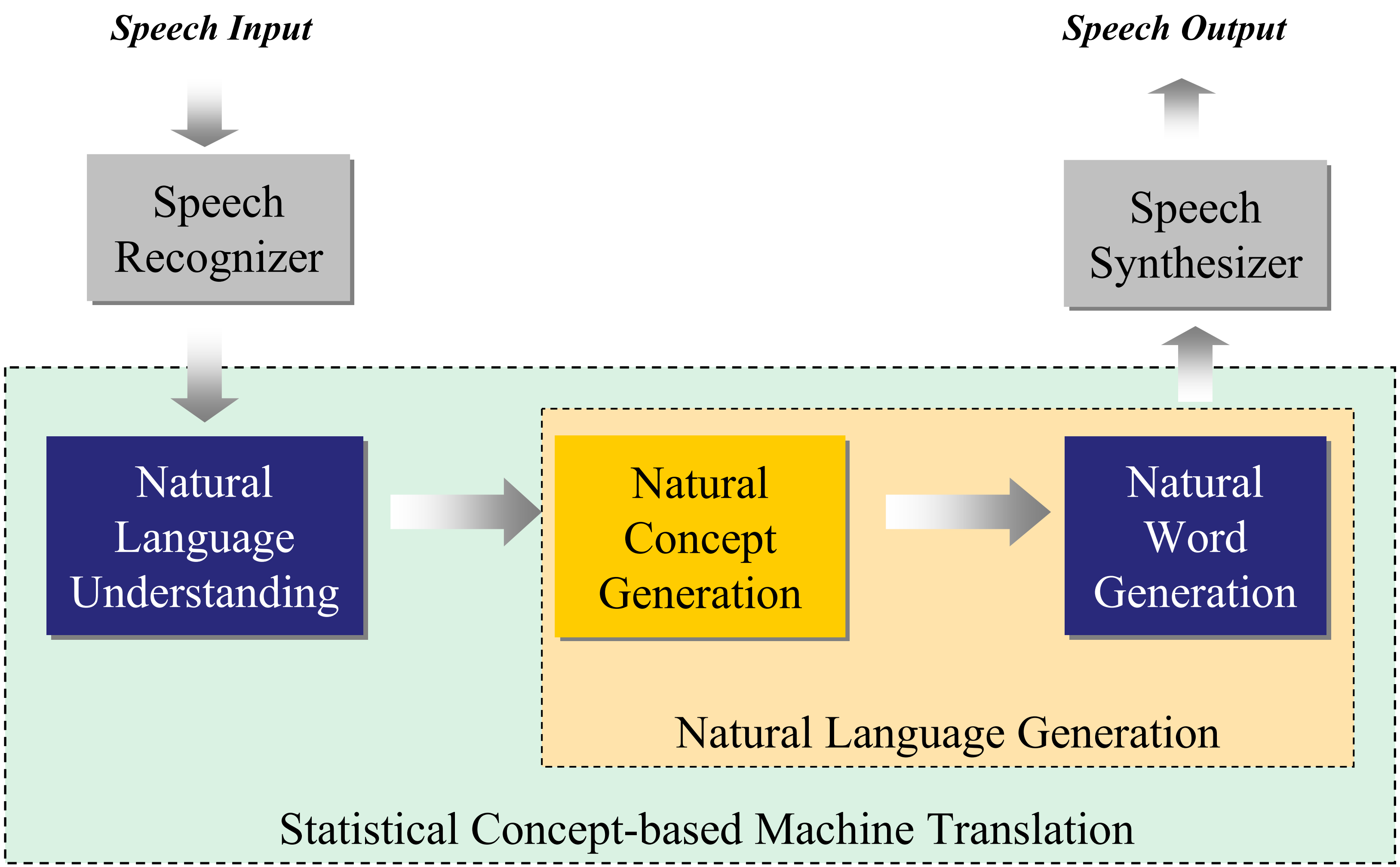
- Statistical Concept-based Speech Translation
- Natural Concept Generation (NCG)
- Feature Selection in Statistical NCG
  - Conciseness vs. Informativity of Concepts
  - Features using both Concept & Word Information
  - Multiple Feature Selection
- Experimental Results

Overview of IBM MASTOR System



"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

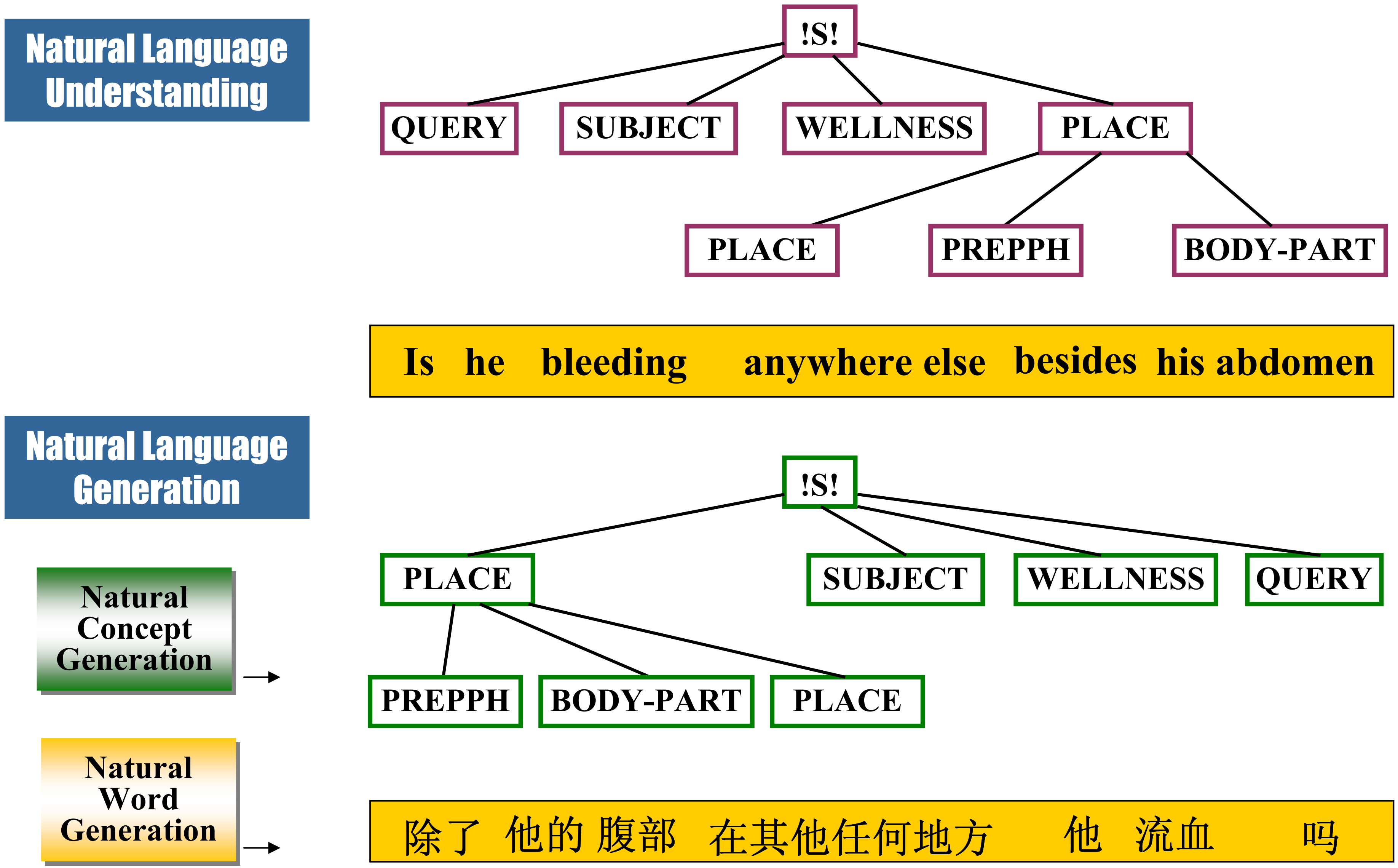
Spoken Language Translation via Concepts



“On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation,” IWSLT, Kyoto, Japan, 2004

IBM

Spoken Language Translation via Concepts



"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

IBM

Concept-based Speech Translation

Concepts

- **Language-independent** representation of intended meanings
- Parsed from source language
- Organized in a **language-dependent** tree-structure
- Comparable to interlingua

Merits

- More flexible meaning preservation
- Wider sentence coverage
- Easier portability between different domains

Challenges

- Design and selection of concepts
- **Generation of concepts**

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

Natural Concept Generation (NCG)

Purpose

- Correct set of concepts in target language
- Appropriate order of concepts in target language

Approaches

- Statistical model-based generation
  - Trained on Maximum-Entropy models

Challenges

- Design of generation procedure
  - Generation of concept sequences
  - Transformation of semantic parse tree
- **Selection of features**

Statistical NCG on Sequence Level

English	QUERY	PRON	POSSESS	WEAPON
Chinese	PRON	POSSESS	WEAPON	QUERY

Statistical NCG on Sequence Level

Source ***C***	C_1	C_2	…	C_M
Target ***S***	S_1	S_2	…	S_n	S_{n+1}

$$p\left(s \middle| c_m, s_n, s_{n-1}\right) = \frac{\prod_k \alpha_k^{\,g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{\,g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}}$$

α_k : probability weight corresponding to feature $\vec{f}_k$

$\vec{f}_k$: feature of concepts

g : binary test function

$$g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right) = \begin{cases} 1 & if\ \vec{f}_k = \left(s, c_m, s_n, s_{n-1}\right) \\ 0 & otherwise \end{cases}$$

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

Statistical NCG on Sequence Level (Cont.)

Source *C*	C_1	C_2	…		C_M
Target *S*	S_1	S_2	…	S_{n+1} …	S_N

Select the concept candidate with highest probability:

Generation

$$s_{n+1} = \arg\max_{s \in V} \left\{ \prod_{m=1}^{M} p\left(s \middle| c_m, s_n, s_{n-1}\right) \right\}$$

$$s_0 = s_{-1} = \text{START}$$

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

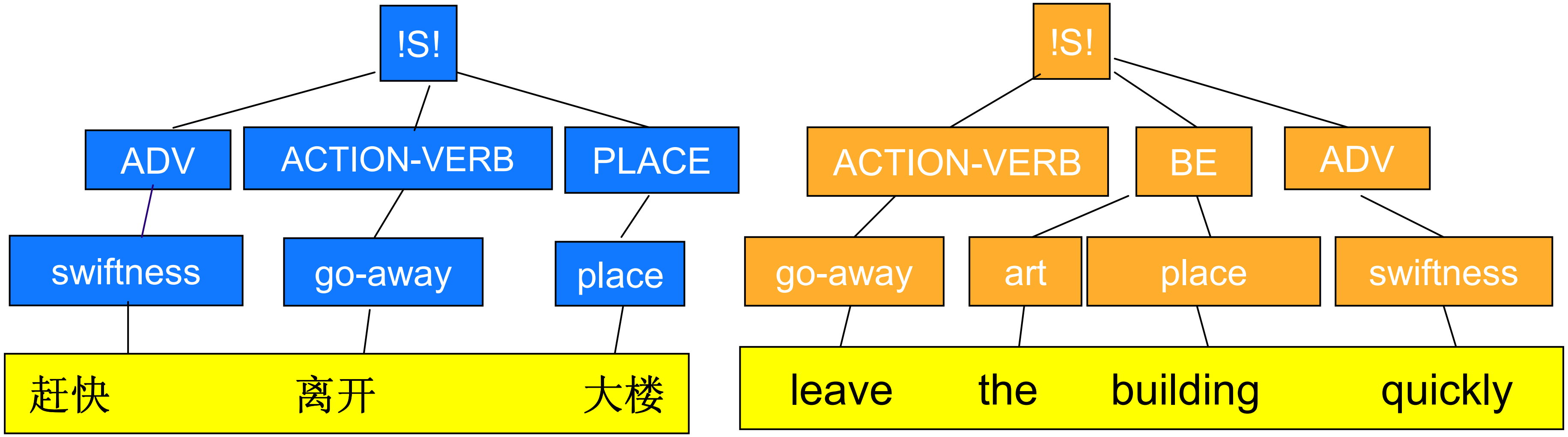
Model Training by Maximizing Entropy

$$p\left(s \middle| c_m, s_n, s_{n-1}\right) = \frac{\prod_k \alpha_k^{g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}}$$

$$\alpha_k = \arg\max_{\alpha} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{m} \log\left[p\left(s \middle| c_m, s_n, s_{n-1}\right)\right]$$

$Q = \{q_l, 1 \le l \le L\}$: total set of concept sequences

Structural Concept Sequence Generation



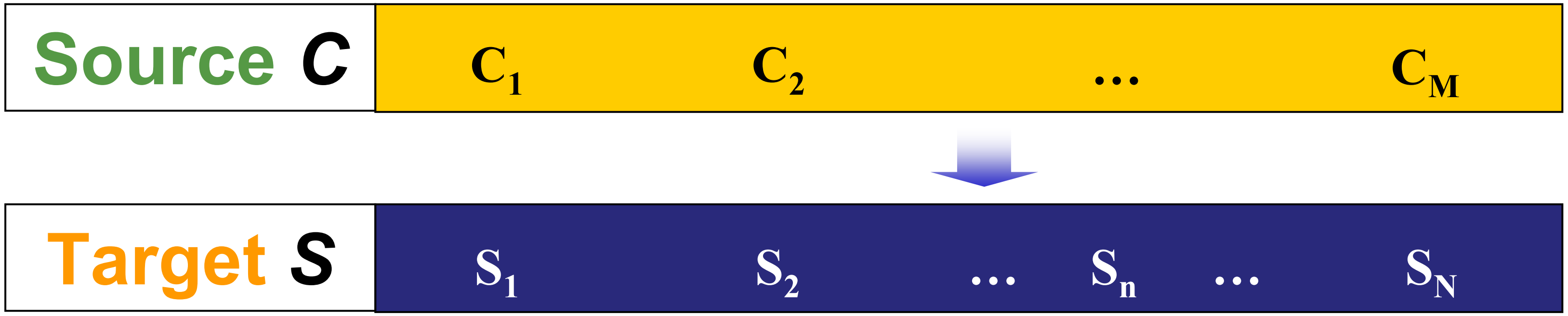
Traverse the semantic parse tree in a bottom-up left-to-right breath-first mode

For each un-processed concept unit on a parse tree, generate an optimal concept unit in target language via the procedure

Repeat until all units in the parse tree in the source language are processed

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

Feature Selection in Maximum-Entropy-based Statistical NCG



Baseline Features

$$\vec{f}_k^{(4)} = \left(s_{+1}^k, c^k, s_0^k, s_{-1}^k\right)$$

Augmented Features on Parallel Corpora

$$\vec{f}_k^{(5)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, s_0^k, s_{-1}^k\right)$$

- new features derived from both source language and target language
- Trained on **parallel tree-bank**
- Strengthen the link between source language and target language

“On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation,” IWSLT, Kyoto, Japan, 2004

Conciseness vs. Informativity of Concepts

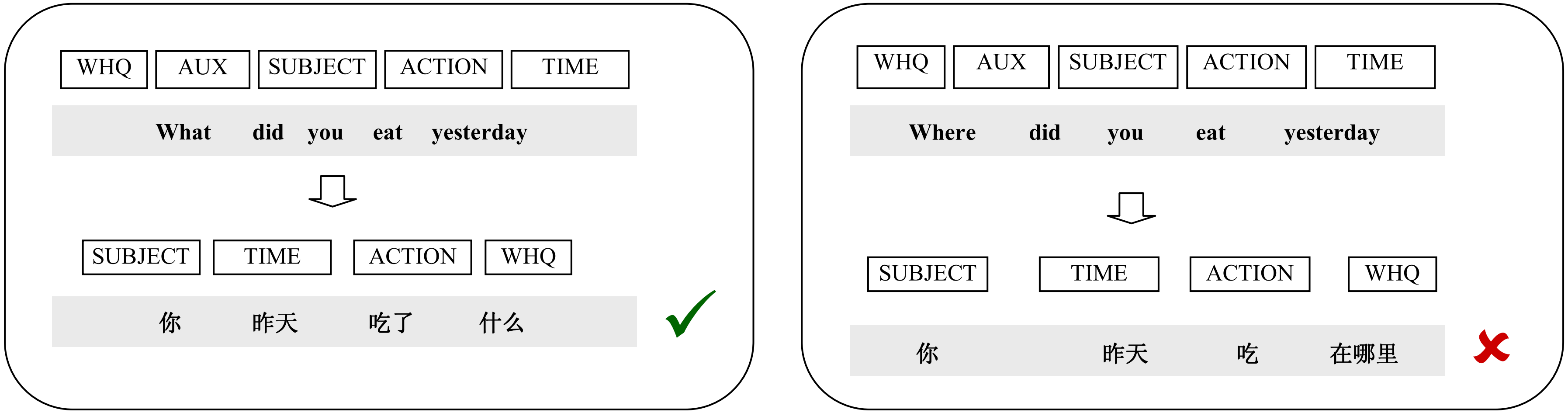
Conciseness

- Define minimum number of distinct concepts
- Reduce labor-extensive, time-consuming annotation process
- **Improve NLU parsing**

Informativity

- Define concepts as informative as possible
- Concept generation largely relies on the sufficient information provided by each concept
- **Improve NCG**

Examples of NCG with too Concise Concepts



- Two input English sentences with **SAME** set and order of concepts generate two **DIFFERENT** concept sequences
- The concept WHQ is **too concise** that it is **not informative enough** to discriminate the different generation behavior between (WHQ, what) and (WHQ, where)
- Features of $\vec{f}_k^{(4)} = \left(s_{+1}^k, c^k, s_0^k, s_{-1}^k\right)$ and $\vec{f}_k^{(5)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, s_0^k, s_{-1}^k\right)$ not helpful

Using both Concept & Word Information

Concept information only

WHQ	AUX	SUBJECT	ACTION	TIME
What	did	you	eat	yesterday

SUBJECT	TIME	ACTION	WHQ
你	昨天	吃了	什么

✓

WHQ	AUX	SUBJECT	ACTION	TIME
Where	did	you	eat	yesterday

SUBJECT	TIME	ACTION	WHQ
你	昨天	吃	在哪里

✗

Concept & Word information

WHQ-what	AUX	SUBJECT	ACTION	TIME
What	did	you	eat	yesterday

SUBJECT	TIME	ACTION	WHQ-what
你	昨天	吃了	什么

✓

WHQ-where	AUX	SUBJECT	ACTION	TIME
Where	did	you	eat	yesterday

SUBJECT	TIME	WHQ-where	ACTION
你	昨天	在哪里	吃的

✓

“On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation,” IWSLT, Kyoto, Japan, 2004

Features using both Concept & Word Information

Concept Sequence: $C = \{c_1, c_1, \cdots, c_M\}$

Word Sequence: $W = \{\overline{w}_1, \overline{w}_2, \cdots, \overline{w}_M\}$

$$p\left(s \middle| c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right) = \frac{\prod_k \alpha_k^{g\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}}$$

$$\alpha_k = \arg\max_{\alpha} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{m=1}^{M-1} \log\left[p\left(s \middle| c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)\right]$$

Optimized on parallel treebank: $QQ = \{u_l, v_l \mid 1 \le l \le L\}$

Multiple Feature Selection

Problem

- Sparser data because of higher dimensional features

Strategy

- Additional sets of features in ME-based concept generation
- Multiple sets of features represent context information in both the source and the target language at different levels

Example

Feature A: $\vec{f}_k^{(5)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, s_0^k, s_{-1}^k\right)$

Feature B: $\vec{f}_k^{(7)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, \overline{w}_0^k, \overline{w}_{+1}^k, s_0^k, s_{-1}^k\right)$

$$\alpha_k = \arg\max_{\alpha} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{m=1}^{M-1} \left\{ \log \frac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(5)}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k\left(\vec{f}_k^{(5)}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}} + \log \frac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}} \right\}$$

Experimental Setup

MT Method	statistical interlingua-based speech translation
Language Pair	English – Chinese (Mandarin)
Domain	medical and force protection
Corpora	10,000 annotated parallel sentences
Vocabulary size	3000
Size of Concept Set	68

Experiments on ME-based Statistical NCG

Description

- Evaluate on primary concept sequences that represents the top-layer concepts in a semantic parser tree
- Concept sequences containing only one concept are removed as they are easy to generate
- Specific set of parallel concept sequences that contain the same set of concepts in both languages
- 5600 concept sequences are selected, 80% for training and 20% for testing
- Random partitioning of training and test set for 100 times
- Average error rates were recorded
- Worst-case test: sequences appear in the training corpus are not allowed to appear in the test corpus
- Normal-case test: sequences appear in the training corpus may appear in the test corpus

ME-NCG Experiments with Forward Models

NCG Methods	Training set (SER / CER)	Test set (SER / CER)
Baseline NCG with basic feature $\vec{f}_k^{(4)}$	14.0% / 8.8%	28.0% / 18.9%
+ feature on parallel corpora $\vec{f}_k^{(5)}$	6.2% / 3.5%	21.7% / 14.1%
+ concept-word features $\vec{f}_k^{(7)}$	0.7% / 0.4%	20.2% / 13.1%
+ multiple feature selection $\left(\vec{f}_k^{(5)} + \vec{f}_k^{(7)}\right)\vec{f}_k^{(4)}$	**0.7% / 0.4%**	**17.4% / 11.4%**

- A concept sequence is considered to have an error during measurement of sequence error rate if one or more errors occur in this sequence
- Concept error rate, on the other hand, evaluates concept errors in concept sequences such as substitution, deletion and insertion

ME-NCG Experiments with Forward-Backward Models

NCG Methods	Training set (SER / CER)	Test set (SER / CER)
Baseline NCG with basic feature $\vec{f}_k^{(4)}$	9.1% / 5.5%	24.4% / 16.4%
+ feature on parallel corpora $\vec{f}_k^{(5)}$	5.7% / 3.2%	17.8% / 11.6%
+ concept-word features $\vec{f}_k^{(7)}$	0.5% / 0.3%	17.7% / 11.5%
+ multiple feature selection $\left(\vec{f}_k^{(5)} + \vec{f}_k^{(7)}\right)\vec{f}_k^{(4)}$	**0.5% / 0.3%**	**15.8% / 10.4%**

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

Experiment on Statistical Interlingua-based S2S

Translation Methods	$\vec{f}_k^{(4)}$	$\vec{f}_k^{(5)}$	$\left(\vec{f}_k^{(5)} + \vec{f}_k^{(7)}\right)$
Text-to-Text	0.536	0.578	**0.605**
Speech-to-Text	0.437	0.469	**0.489**

Bleu metric (proposed by Kishore et. al.) measures MT performance by evaluating n-gram accuracy with a brevity penalty

$$Bleu = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

p_n : n-gram precision rate w_n : n-gram weight BP : Brevity penalty

"On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation," IWSLT, Kyoto, Japan, 2004

Summary

- Attack the problems of feature selection during maximum-entropy-based model training and concept generation
- New concept-word features proposed that exploit both information at both concept level and word level
- Multiple feature selection algorithm combines different features in maximum-entropy models to alleviate data-sparseness-caused over-training problem
- Significant improvements are achieved in both concept sequence generation test and speech translation experiments