The RWTH Statistical Machine Translation System for the IWSLT 2007 Evaluation

Arne Mauser, David Vilar, Gregor Leusch, Yuqi Zhang, Hermann Ney

{lastname}@i6.informatik.rwth-aachen.de

International Workshop on Spoken Language Translation (IWSLT) 2007
October 16, 2007

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

RWTH System 2007

- **Main facts**
- **Models**
- **System combination**
- **Systems and Results**

Main Facts

Combination of different SMT system.

Each system: log-linear combination of a main translation model with several additional models.

- **Translation models**
  - **Phrase-based model**
  - **Bilingual n-gram model**
  - **Hierarchical phrase model**
- **Additional models**
  - **Target language model**
  - **Word lexicon**
  - **Phrase count features**
  - **Word penalty**
  - **Phrase penalty**
  - **Reordering model**

No additional training data.

From 2006 to 2007: Improvements

- System combination on site (last year: across TC-Star project sites)
- More robust system combination
- New hierarchical phrase model
- Add syntactical reordering model

In terms of automatic measures for Chinese-English on IWSLT 2005 (dev3):

System	BLEU [%]	NIST	WER [%]	PER [%]
2004	40.4	8.59	52.4	42.2
2005	46.3	8.73	47.4	39.7
2006	48.8	8.56	47.3	39.2
2006 (40k)	51.4	9.00	40.0	33.2
2007	62.4	9.64	30.7	26.0
2007 comb.	63.4	10.14	30.8	25.3

Models

Hierarchical model

Generalization of phrase-based-models

- **Allow for "gaps" in the phrases.**
- **Integration of reordering in the translation model.**

Similar to Chiang's approach

- **Rules of the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where**
  - **X is a non-terminal.**
  - **γ and α are strings of terminals and non-terminals.**
  - **$\sim$ is a one-to-one correspondence between the non-terminals of α and γ.**

Here: Rules are transformed into GNF-like form to allow for left-to-right generation

Extraction Process

- **Basic idea:**
  - **Extract standard phrases.**
  - **If the extracted phrases contain further sub-phrases, create “holes”.**

- **Main restrictions:**
  - **Maximum of two non-terminals.**
  - **Non-terminals must be non-adjacent in the source side.**
  - **Rules must have at least a terminal symbol.**

System Combination

System Combination

- Compute consensus translation by combining their outputs
- Idea: select words present in the majority of translations
  - Similar idea as ROVER for ASR
- Consider possible reordering of words/phrases
- High-quality alignment of different hypotheses required for voting
- Build confusion network from alignment
- Repeat Process N times for N systems

Improvements over the 2006 system combination

- Direct language model rescoring on the union of confusion networks
  - More efficient and better Bleu/TER score than N-best list rescoring
- use special language model in rescoring:
  - Train trigram LM on the outputs of all systems
  - Boost probability of n-grams present in the original phrases
- Optimize parameters automatically for Bleu
- Use 10-best lists from each system

Systems and Results

Italian-English Setup

Preprocessing

- lowercase and remove punctuation on the Italian training data
- bilingual n-gram model: keep punctuation in train and insert in test
- reduce case to the most frequent per word for english
- split punctuation and contractions: dell'albergo $\longrightarrow$ dell' albergo

Models and Training

- add dev1-3 to the training data
- same alignment for all models, refined heuristic
- use dev4 or dev5b for minimum error training of the systems
- used dev5a for tuning the system combination
- local or no reordering
- ASR: first-best

Italian-English Results

Model	opt-corpus	reorder	BLEU[%]	TER[%]
Phrase-based	dev4	no	41.6	44.5
		local	41.7	44.5
	dev5b	no	42.9	43.0
		local	42.8	43.0
Hierarchical	dev5b		42.5	43.7
n-gram	dev4	no	33.5	50.5
System Combination			45.3	41.4

n-gram model performed comparably on dev sets $\longrightarrow$ error in phrase extraction

2.4 Bleu-points improvement by system combination

Chinese-English Setup

Preprocessing and Training

- Chinese word segmentation using ICTClas or given segmentation
- Split punctuation marks and contractions: he'll → he will

Models and Training

- Different alignment training
  - Number of wordclasses in GIZA++ training
  - Models used (HMM, HMM+IBM4)
- Phrase Extraction
  - different heuristic
  - Use minimum-weight edge cover algorithm on HMM and IBM4 probabilities
- All systems optimized on dev2 (IWSLT 2004)
- Optimization criterion: Bleu with minimum nearest/average reference length

Chinese-English - Results

Two combinations submitted (the wrong one as primary)

Model	Alignment	BLEU[%]	TER[%]
Phrase-based	IBM4, MWEC-Alignment	37.2	48.0
	HMM, MWEC-Alignment	36.7	48.4
	HMM, refined	34.7	52.8
+syntax-reordering	IBM4, MWEC-Alignment	33.6	54.2
Hierarchical	IBM4, refined	33.3	51.4
System Combination		38.5	47.2

Hierarchical system not fully optimized

- Recent results: Bleu: 35.0 TER: 50.5

Results

Evaluation submissions

Translation Direction	Input	Accuracy Measures		Error Rates		
		BLEU [%]	NIST	TER [%]	WER [%]	PER [%]
Italian-to-English	Clean	45.3	8.21	41.4	43.1	33.9
	ASR	41.3	7.74	44.9	46.5	36.9
Chinese-to-English (best RWTH)	Correct	37.1	6.75	50.4	51.4	45.0
	Correct	38.5	6.80	47.2	47.9	43.2

Summary

good results on both language pairs (Italian-English, Chinese-English)

robust, on-site system combination

combination of different translation systems / models

New approaches

- **Hierarchical phrase model**
- **Use syntax**

Thank you for your attention

Arne Mauser

mauser@i6.informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

Translation Models

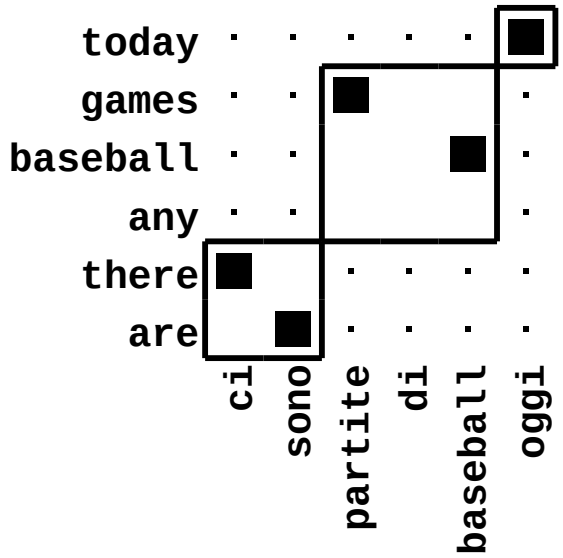
Phrase-Based Model

- standard phrase-based model
- training: bilingual phrase pairs extracted from word-aligned training data
- main features: log probabilities $p(\tilde{f}|\tilde{e})$ and $p(\tilde{e}|\tilde{f})$ estimated by relative frequencies

Bilingual n-gram model

Bilingual n-gram model

- model joint probability of source and target sentence $Pr(f_1^J, e_1^I)$
- segment the source and target sentences with the same restrictions given for the phrase-based model
- find the smallest units such that the resulting phrase segmentation is monotonic



extracted sentence:

`ci_sono|are_there partite_di_baseball|any_baseball_games`

`oggi|today` $\longrightarrow$ train smoothed n-gram language model [?]

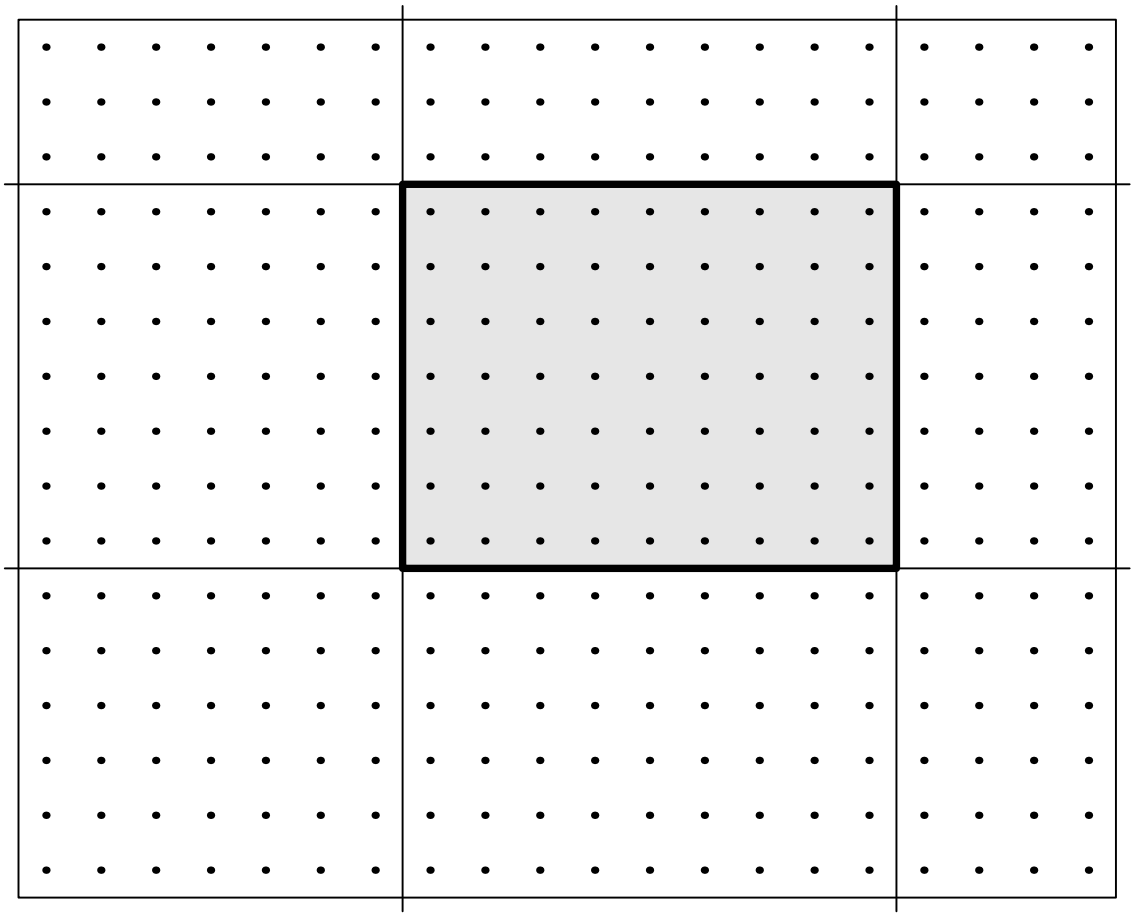
Chiang's approach

- Formalization as a synchronous CFG.
- Rules of the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where
  - X is a non-terminal.
  - γ and α are strings of terminals and non-terminals.
  - $\sim$ is a one-to-one correspondence between the non-terminals of α and γ.
- Example:

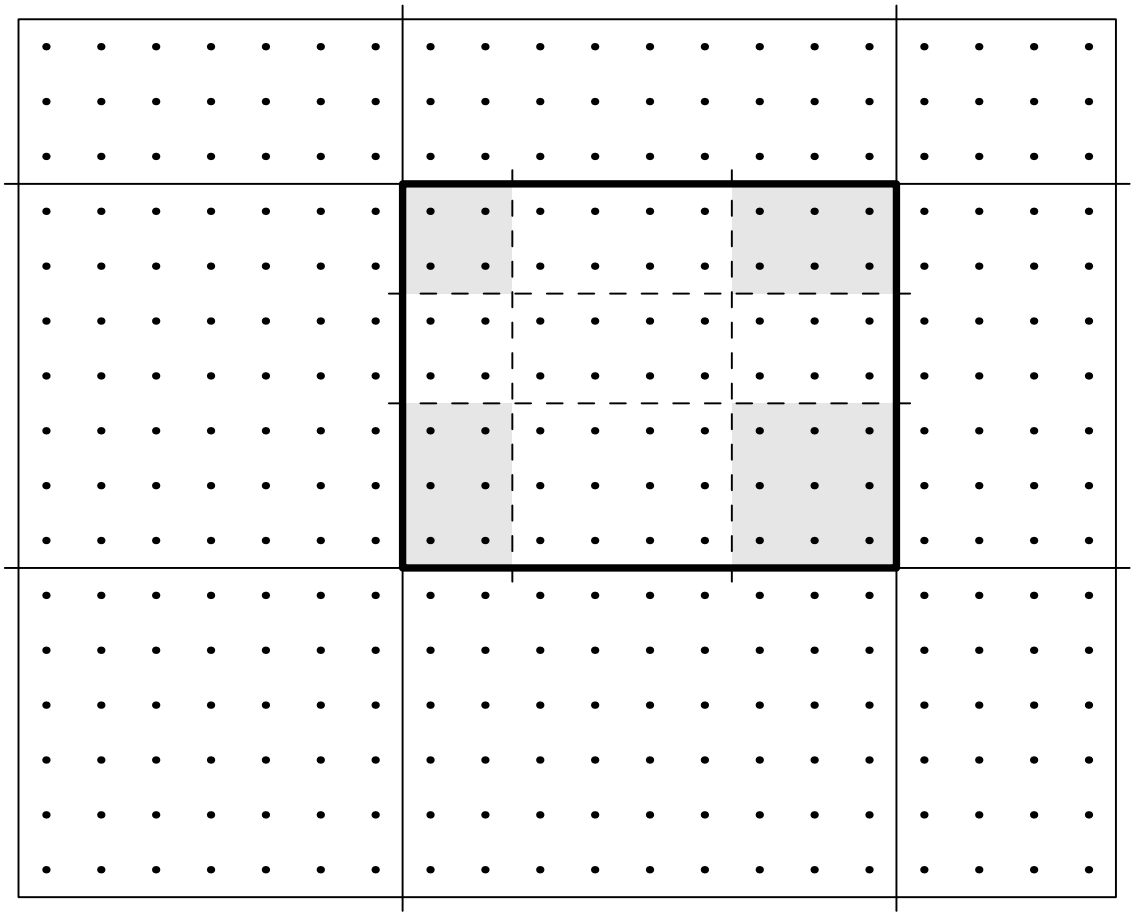
$$X \rightarrow \langle \text{yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}}, \text{ have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle$$
$$X \rightarrow \langle X_{\boxed{1}} \text{ de } X_{\boxed{2}}, \text{ the } X_{\boxed{2}} \text{ that } X_{\boxed{1}} \rangle$$

- Additionally: Glue rules

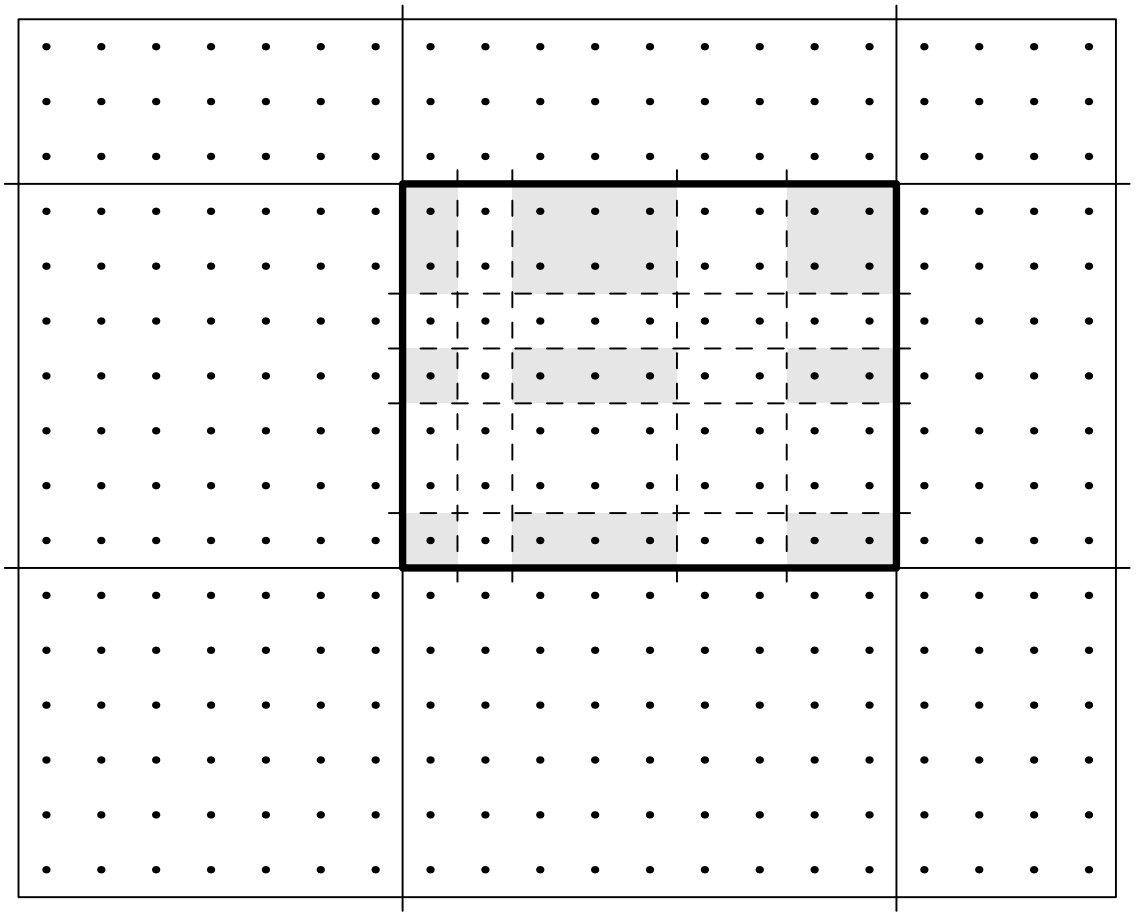
$$X \rightarrow \langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \rangle$$
$$X \rightarrow \langle X_{\boxed{1}}, X_{\boxed{1}} \rangle$$



Standard phrase



Phrase with one gap



Phrase with two gaps

Experimental Results

IWSLT 2007, Chinese-to-English task

System	BLEU	TER	WER	PER
PBT monoton	29.6	56.0	58.3	48.9
best PBT	37.2	48.0	48.7	44.3
Hierarchical	35.0	50.5	51.3	46.4

Example translations:

PBT	The sightseeing. Where is it?
Hierarchical	Where is the tourist information office?
Reference	Where is the tourist information office?

PBT	Please tell me about this form, I do not know what to fill out?
Hierarchical	Please tell me how to fill out this form?
Reference	Can you tell me how to fill out this form?

Phrase count features

Motivation:

- rare phrases are overestimated
- estimated probabilities not reliable

Idea:

- adjust probabilities of rare phrases
- "mark" phrases with a occurrence count below a given threshold
- include these marker as a binary feature in the log-linear translation model

$$h_{\mathbf{C},\tau}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} [N(\tilde{f}_k, \tilde{e}_k) \leq \tau]$$

τ: threshold, $N(\tilde{f}_k, \tilde{e}_k)$: bilingual phrase count,
i_1^K: segmentation of the source sentence

Algorithm: Idea

- Align different MT system outputs for each source sentence:
  - Allow word reordering
  - Take context of whole test document into account
  - Get more reliable alignment by using an iterative alignment procedure
- Construct confusion network from the reordered hypotheses
- Use system prior probabilities and other statistical models to select consensus translation from network

Alignment and Reordering

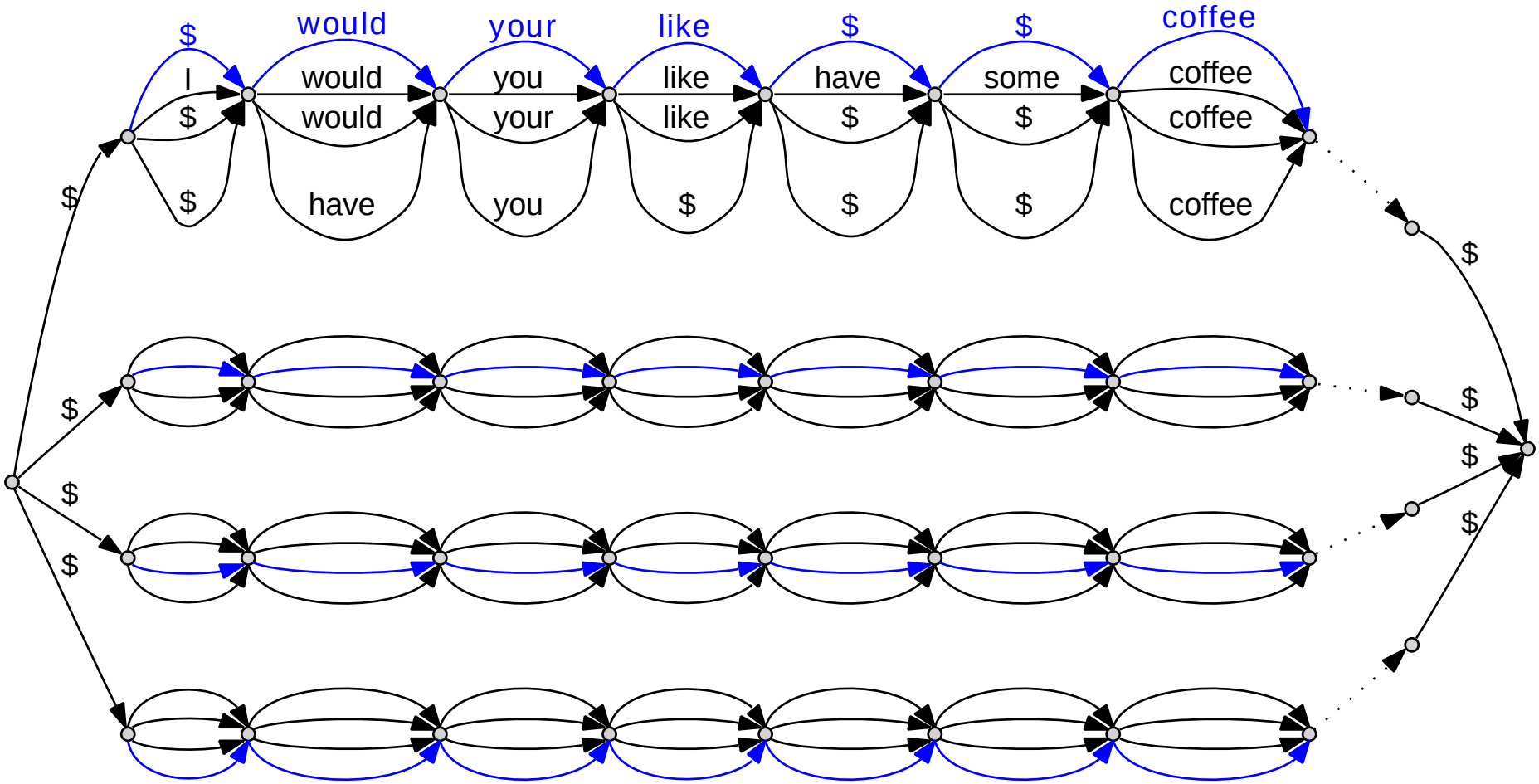
Alignment

- Pairwise alignment of the output of M systems for N test sentences
- IBM Model 1 and HMM alignment models
- Note: Alignment can be improved by adding more translated data from involved MT systems

Reordering

- Select a primary hypothesis E_m
- Reorder each other MT output E_n based on alignment with E_m
- Construction of confusion network from alignments

Example: Confusion Network



Example

system hypotheses	0.25	would your like coffee or tea							
	0.35	have you tea or coffee							
	0.10	would like your coffee or							
	0.30	I have some coffee tea would you like							
alignment and reordering	have|would you|your $|like coffee|coffee or|or tea|tea								
	would|would your|your like|like coffee|coffee or|or $|tea								
	I|$ would|would you|your like|like have|$ some|$ coffee|coffee $|or tea|tea								
confusion network	$	would	your	like	$	$	coffee	or	tea
	$	have	you	$	$	$	coffee	or	tea
	$	would	your	like	$	$	coffee	or	$
	I	would	you	like	have	some	coffee	$	tea
voting	$/0.7	would/0.65	you/0.65	$/0.35	$/0.7	$/0.7	coffee/1.0	or/0.7	tea/0.9
	I/0.3	have/0.35	your/0.35	like/0.65	have/0.3	some/0.3		$/0.3	$/0.1
consensus translation	would you like coffee or tea								