Phrase-level System Combination for Machine Translation Based on Target-to-Target Decoding

Columbia University

Wei-Yun Ma

Kathleen McKeown

Motivation of Combination

- Many successful MT approaches with very different techniques.
- Want to take advantage of the individual strengths and avoid the individual weakness of them

Motivation of Phrase-level Combination

- In Translation, Phrase-based MT is useful: several words should be translated as a whole.
- Similarly, in combination, several words should be substituted as a whole
- we develop a new phrase-level lattice decoding approach instead of phrase-based re-decoding approach (Rosti et al. 2007, Chen et al. 2008)
    - To utilize syntactic structure (or word order) of the best MT output

Outline

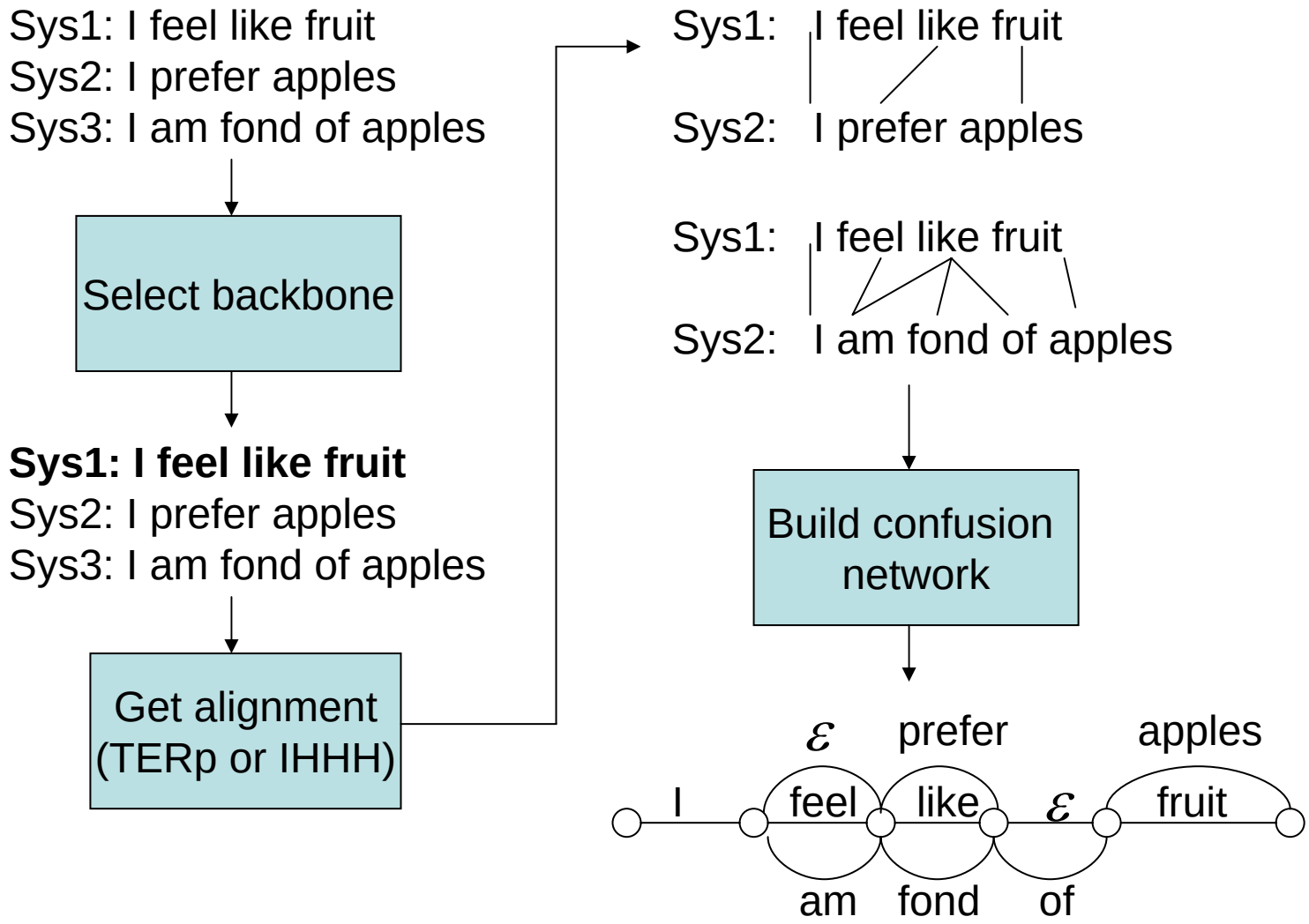
- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

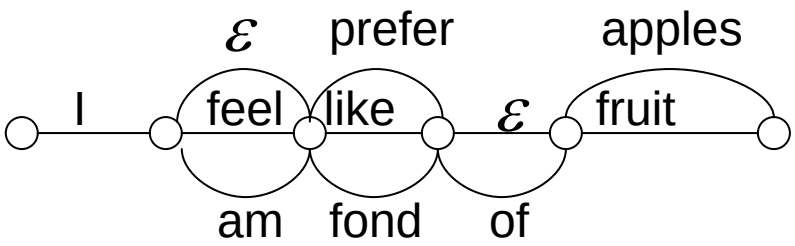
Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Related Work: Confusion Network

(Matusov et al., 2006; He et al. 2008; Rosti et al. 2007; Leusch and Ney, 2010)





Decoding
using word consensus and LM

I feel like apples

Considering:
I feel like of apples
I feel like of fruit
I feel like apples
I feel like fruit
I prefer apples
I prefer fruit
I feel prefer apples
I am fond apples
I feel prefer apples.
I like apples
…

Confusion Network considers too many hypotheses.
→ Sometimes several words should be substituted as a whole with several other words

Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Basic Idea

- Phrase-level Lattice decoding
  - Robust Paraphrase Extraction Strategy
    - Independent with alignment approaches
  - Combination using Log-linear-model
    - Various phrase scoring functions
    - Use soft syntactic constraints
- Target-to-Target Decoding
  - "Translation" from the best MT output (backbone) to the combination result
  - Any MT decoder can serve this mission:
    - Ex, using Moses, Paraphrase Lattice can be modeled as Phrase Table (Target-to-Target pairs), and input is backbone
  - Capability of reordering

Outline

- Motivation
- Relevant Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Select the backbone

- Features
  - Sentence-level consensus by using TER
  - A general LM
  - Length smoothing

$$\log p(E_i) = \sum_{s=1}^{Ns} \left(\lambda_s * \log(1 - TER(E_i, E_s))\right) + \lambda^l * \log(LM(E_i)) + \lambda^w * Length(E_i)$$

Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Monolingual Word Alignment: TERp tool

E_b : w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10} w_{11}

E_h : $\overline{w}_1$ $\overline{w}_2$ $\overline{w}_3$ $\overline{w}_4$ $\overline{w}_5$ $\overline{w}_6$ $\overline{w}_7$ $\overline{w}_8$ $\overline{w}_9$ $\overline{w}_{10}$

TERp tool reorders E_h and generate word alignment :

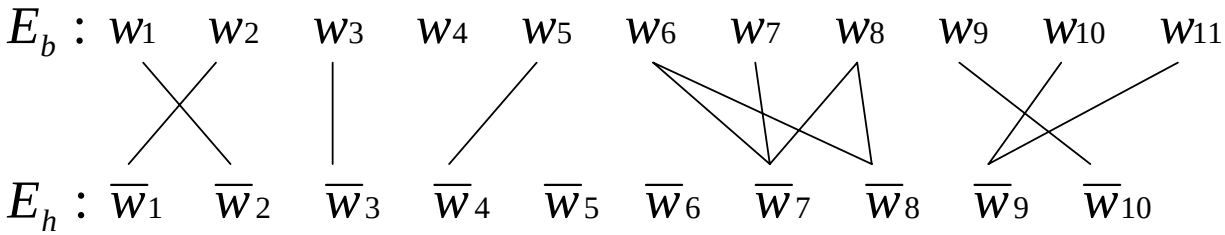
E_b : w_1 w_2 w_3 w_4 w_5 ε ε $[w_6$ w_7 $w_8]$ w_9 $[w_{10}$ $w_{11}]$

S P T D Y I I P M P

E_h : $\overline{w}_2$ $\overline{w}_1$ $\overline{w}_3$ ε $\overline{w}_4$ $\overline{w}_5$ $\overline{w}_6$ $[\overline{w}_8$ $\overline{w}_7]$ $\overline{w}_{10}$ $[\overline{w}_9]$

M (Exact Match),
I (Insertion),
D (Deletion),
S (Substitution),
T (Stem Match),
Y (Synonym Match)
P (Paraphrase)

After we reorder E_h back to its original order :

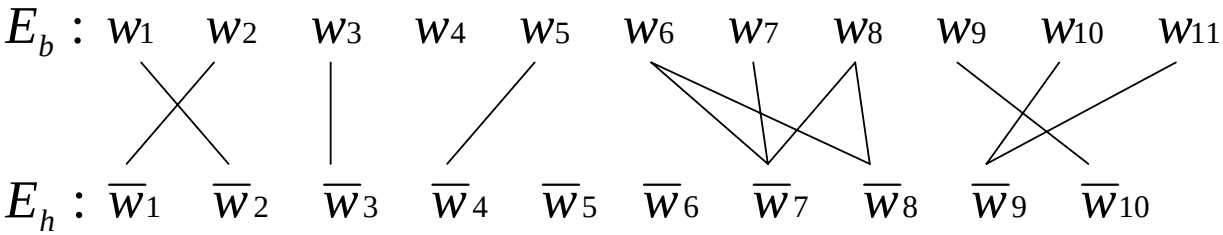


Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

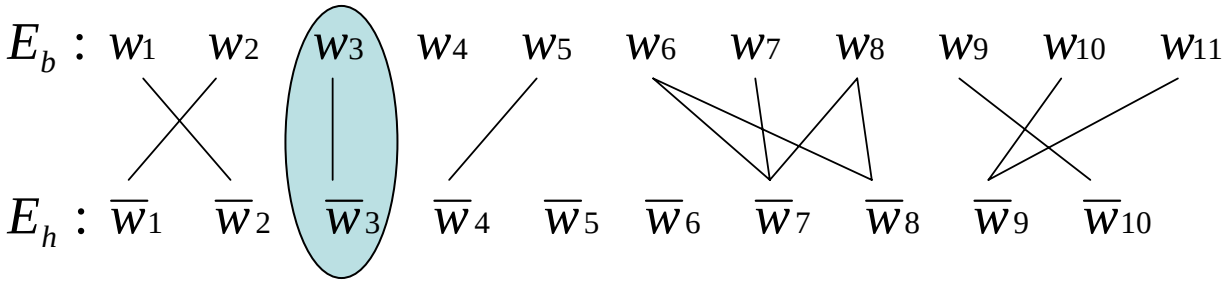
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment



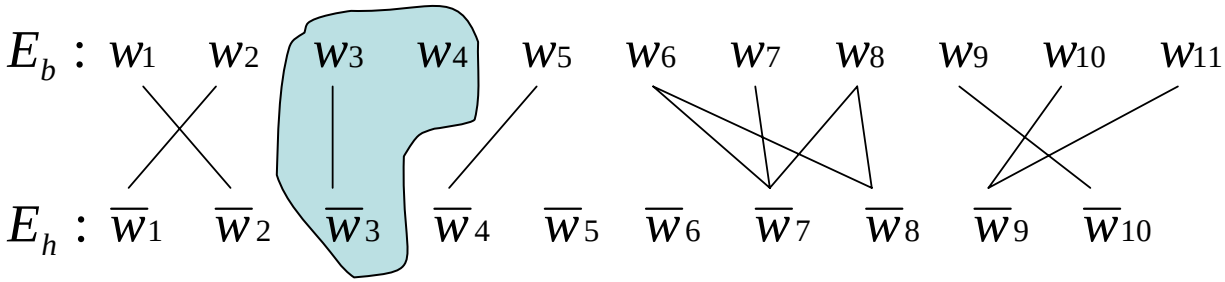
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment



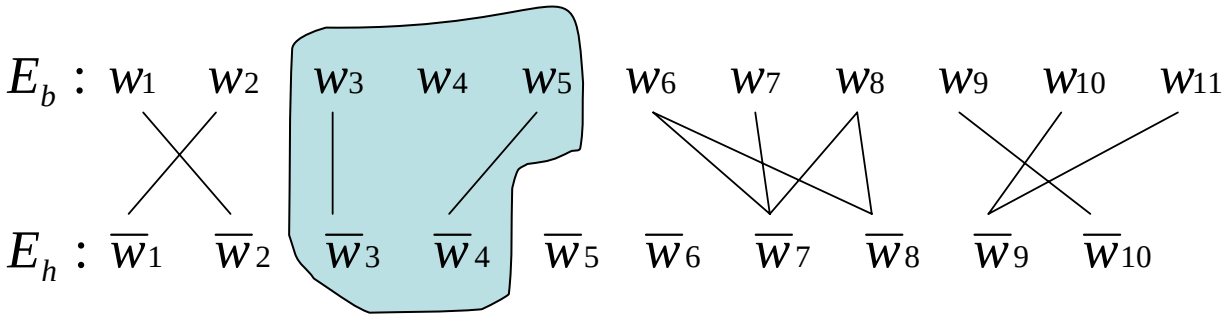
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment



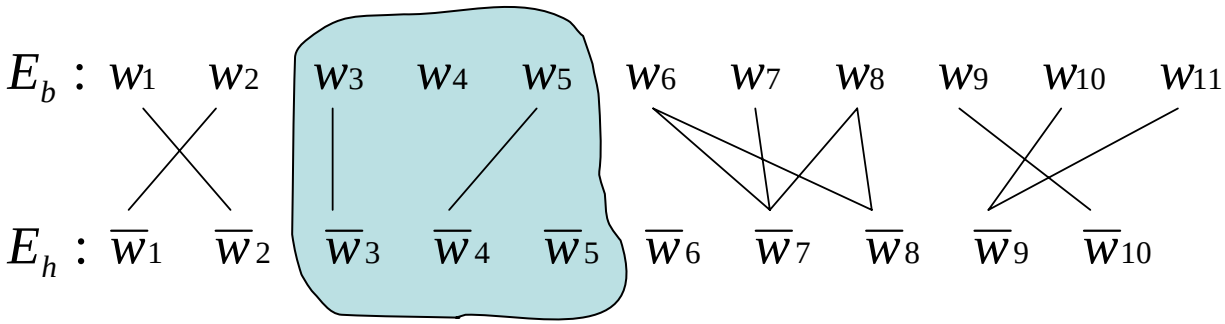
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment



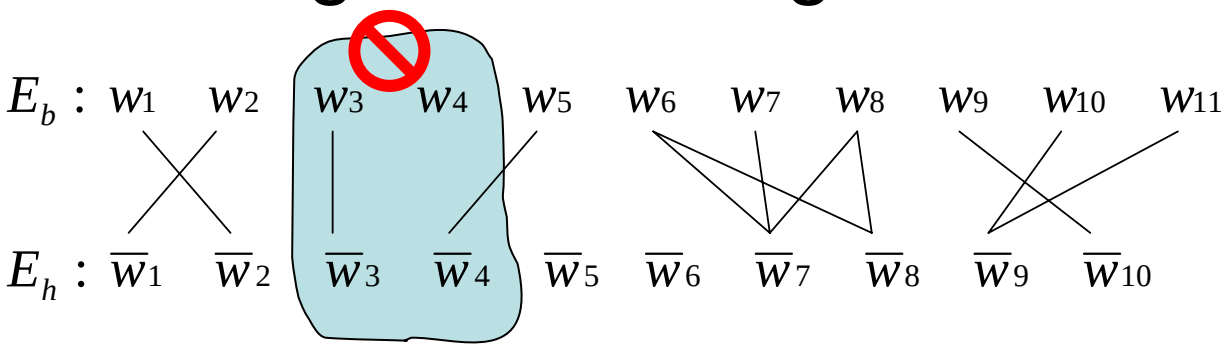
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment



Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment

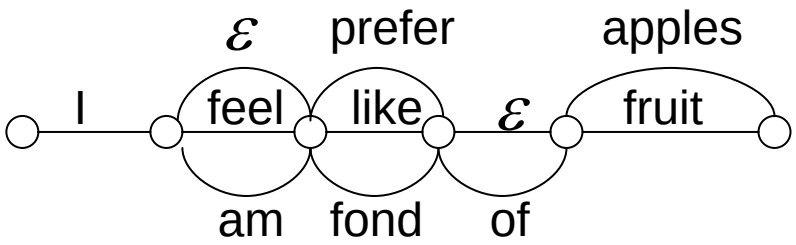


Sys1: I feel like fruit

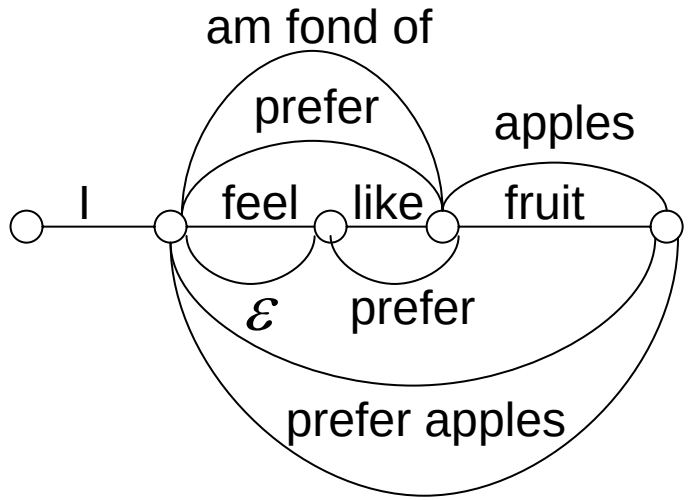
Sys2: I prefer apples

Sys1: I feel like fruit

Sys2: I am fond of apples



CN



TTD

Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Combination Model

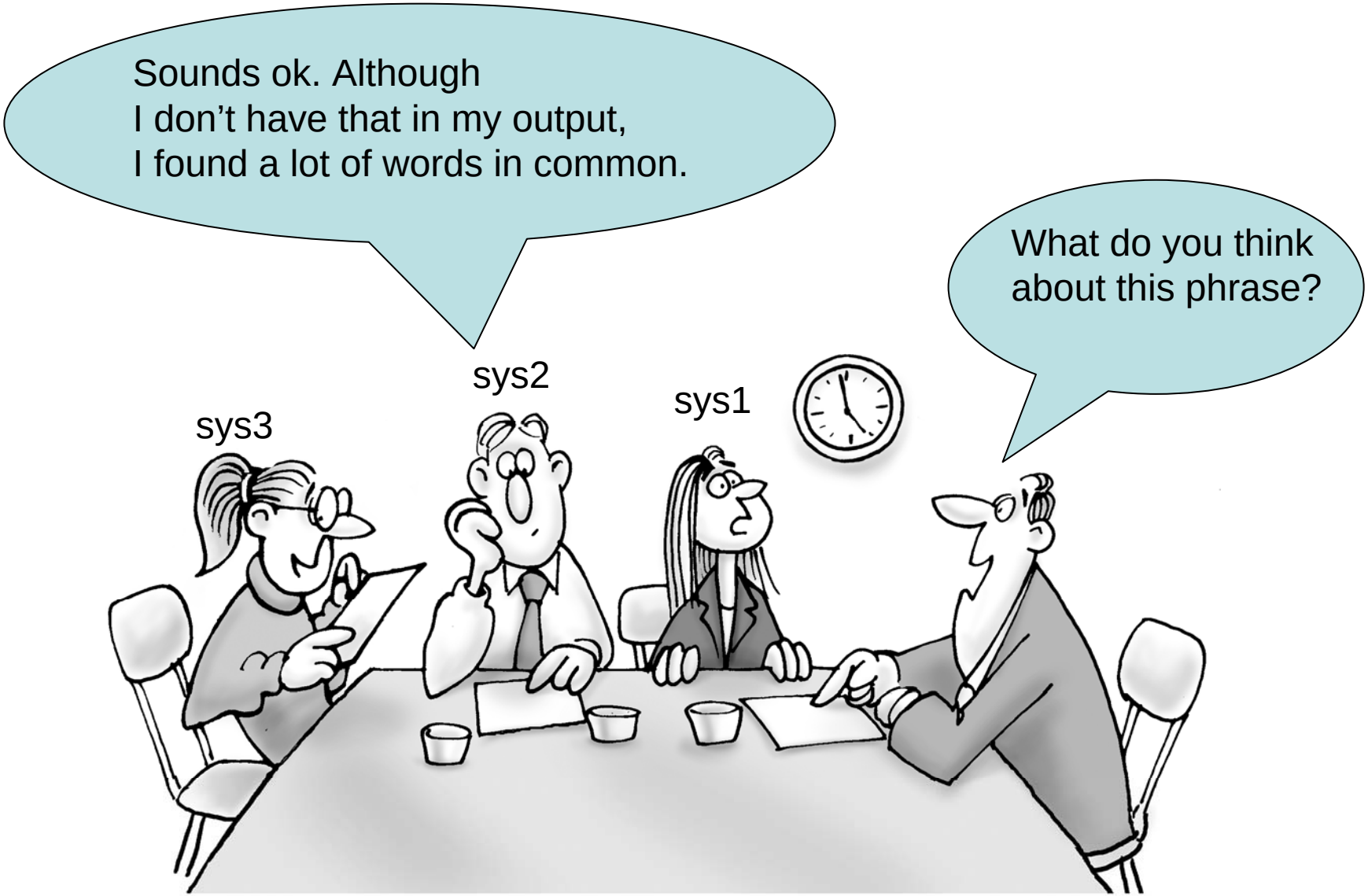
- Phrase Scoring Functions
  - Paraphrase confidence scores (cs)
  - Lexical weighting (lex)
  - Syntactic indicators of whether paraphrases are syntactic constituents (syn)
  - Word and phrase penalty
  - Reordering model (r)
  - General language model
  - System-specific LMs for employing N-gram consensus information. (sl)

Combination Model

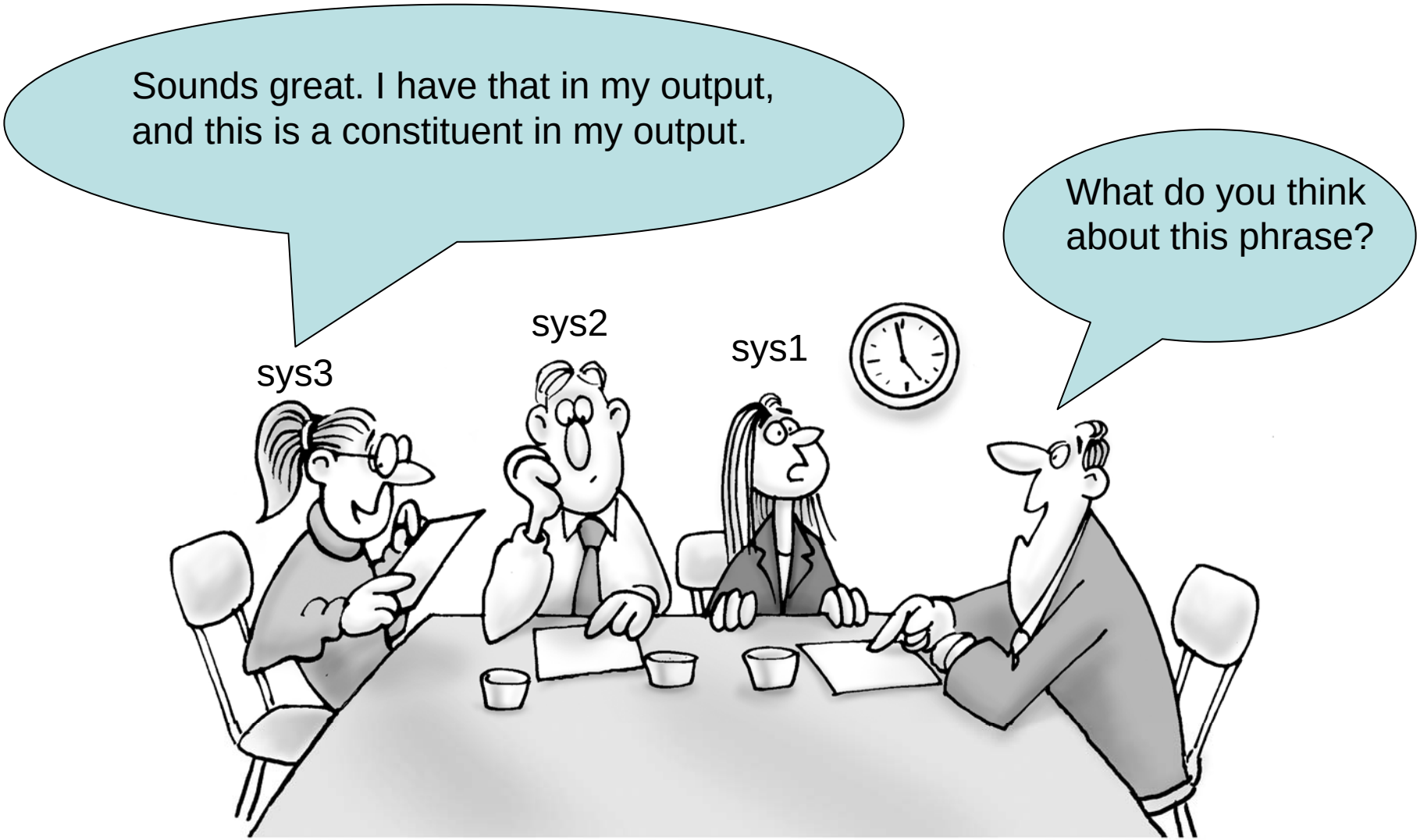
- Phrase Scoring Functions
  - Paraphrase confidence scores (cs)
  - Lexical weighting (lex)
  - Syntactic indicators of whether paraphrases are syntactic constituents (syn)
  - Word and phrase penalty
  - Reordering model (r)
  - General language model
  - System-specific LMs for employing N-gram consensus information. (sl)


What do you think about this phrase?
sys3
sys2
sys1

Paraphrase confidence scores



Lexical weighting



Syntactic indicator

Paraphrase confidence scores (cs)

$$cs_s(\bar{e} \mid e) = \begin{cases} \dfrac{\text{MTYP\# of } (e, \bar{e})}{\text{MTYP\# of } (e, \bar{e}) + \text{IDS\# of } (e, \bar{e})} & \text{if } (e, \bar{e}) \text{ can be extracted in s} \\ 0 & \text{otherwises} \end{cases}$$

Lexical weighting (lex)

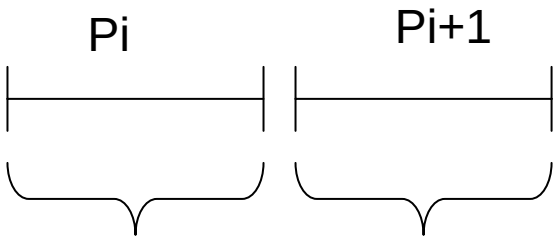
$$lex_s(\bar{e} \mid e) = \frac{\text{Common } \textit{Word\# of } \bar{e} \text{ and } \mathrm{A}_s(e)}{|\bar{e}| + |\mathrm{A}_s(e)|}$$

M (Exact Match),
I (Insertion),
D (Deletion),
S (Substitution),
T (Stem Match),
Y (Synonym Match)
P (Paraphrase)

Syntactic indicator (syn)

$$syn_s(\bar{e}_i \mid e_i) = \begin{cases} 1 & \text{if } (e, \bar{e}) \text{ can be extracted in system s and} \\ & e \text{ and } \bar{e} \text{ are both syntactic constituents} \\ 0 & \text{otherwise} \end{cases}$$

Consensus Model:

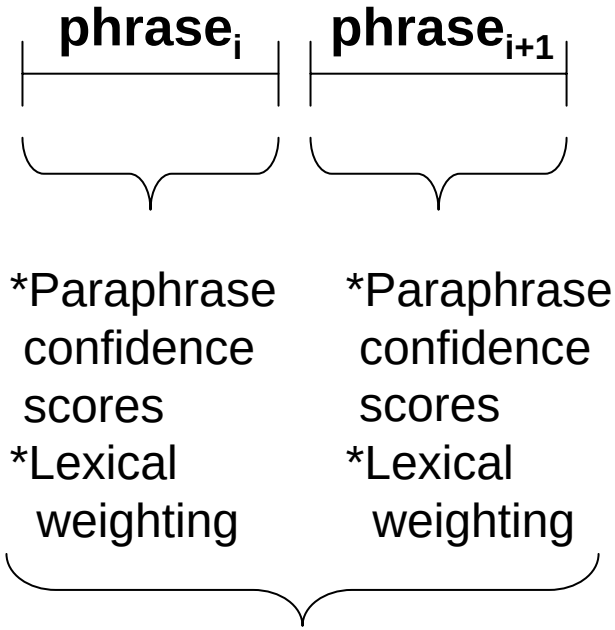


*Paraphrase confidence scores
*Lexical weighting

*Paraphrase confidence scores
*Lexical weighting

System-specific LM

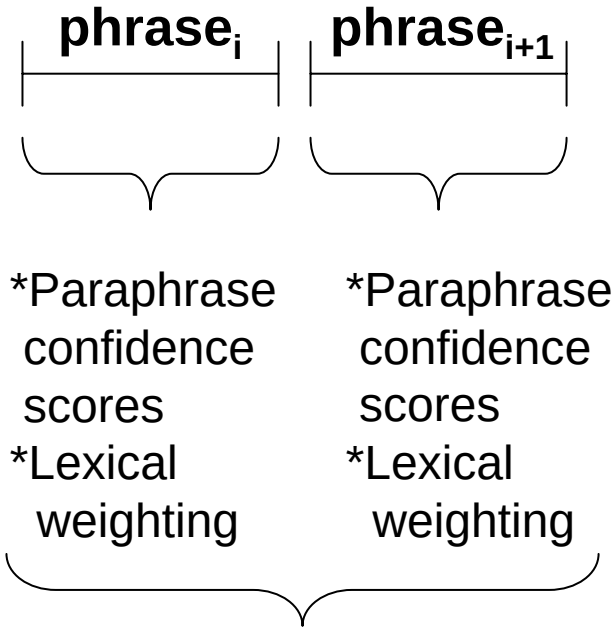
Consensus Model:



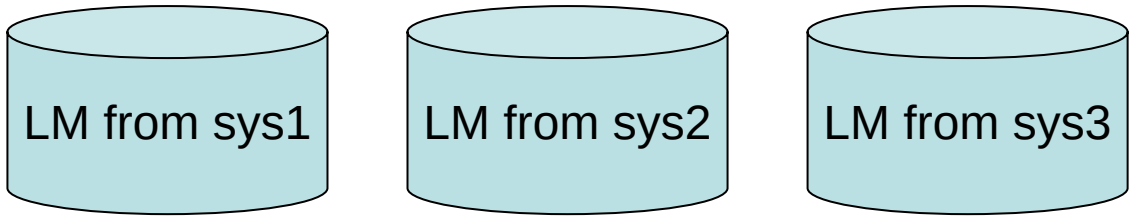
We also want to model consensus between phrases

System-specific LM

Consensus Model:



We also want to model consensus between phrases



For each single system, we build its system-specific LM based on the whole tuning/test corpus of all translation

$$\log p(\overline{E} \mid E) =$$

Paraphrase confidence scores

$$\sum_{i=1}^{I}\left(\sum_{s=1}^{Ns}\left(\lambda_s^{pc} * cs_s(\overline{e}_i \mid e_i) + \lambda_s^{lex} * lex_s(\overline{e}_i \mid e_i) + \lambda_s^{syn} * syn_s(\overline{e}_i \mid e_i)\right)\right)$$

Lexical weighting

Syntactic indicators of whether paraphrases are syntactic constituents

$$+ \sum_{s=1}^{Ns}\left(\lambda_s^{sl} * \log(LM_s(\overline{E}))\right)$$

System-specific LM

$$+ \sum_{i=1}^{I}\left(\lambda^{d} * d(start_i, end_{i-1})\right)$$

$$+ \lambda^{l} * \log(LM(\overline{E}))$$

Reordering model

$$+ \lambda^{w} * length(\overline{E})$$

$$+ \lambda^{p} * I$$

Target-to-Target Decoding Outline

- Motivation
- Related Work
- Basic Idea
- Methodology
  - Select the backbone
  - Monolingual Word Alignment
  - Paraphrase Extraction
  - Combination Model
- Experiments

Two Environments

- Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)
- German-English combination shared task held by the WMT in 2011

Two Environments

- Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)
- German-English combination shared task held by the WMT in 2011

Chinese-English of NIST 2008

- Four human reference translations and corresponding machine translations for the NIST Open MT08 test sets
- Manually select Top5 systems out of 23 systems
- Tuning: 524 sentences
- Testing: 788 sentences

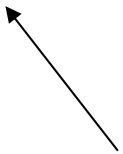
Result of backbone selection

	BLEU	TERp	METEOR
System 03	**30.16**	**63.04**	**51.94**
System 15	30.06	62.82	51.80
System 20	28.15	65.39	49.72
System 22	29.94	63.19	51.51
System 31	29.52	61.70	51.89
backbone	**30.89**	**61.28**	**52.65**

Impact of Feature Combination

	BLEU	TERp	METEOR
System 03	**30.16**	**63.04**	**51.94**
backbone	**30.89**	**61.28**	**52.65**
P+cs	31.74	**60.11**	53.59
P+cs+sl	32.63	60.49	53.53
P+cs+lex	31.81	60.32	53.53
P+cs+syn	31.74	60.22	53.55
P+cs+sl+lex+syn	**32.85**	60.32	**53.76**

P: phrase
W: word



Best feature setting

Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)

Impact of Using Phrase and Word under no re-ordering

	BLEU	TERp	METEOR
System 03	30.16	63.04	51.94
backbone	30.89	61.28	52.65
W+cs	**30.98**	**60.98**	**52.90**
W+cs+sl	**31.29**	**61.36**	**52.70**
P+cs	**31.74**	**60.11**	**53.59**
P+cs+sl	**32.63**	**60.49**	**53.53**
P+cs+lex	31.81	60.32	53.53
P+cs+syn	31.74	60.22	53.55
P+cs+sl+lex+syn	32.85	60.32	53.76

Under the same settings,
phrase is always better than word

Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)

Impact of Word Reordering

	BLEU	TERp	METEOR
System 03	**30.16**	**63.04**	**51.94**
backbone	**30.89**	**61.28**	**52.65**
P+r+cs	31.80	60.21	53.71
P+r+cs+sl	**32.80**	60.13	**53.86**
P+r+cs+lex	31.76	**60.12**	53.54
P+r+cs+syn	31.72	60.37	53.38
P+r+cs+sl+lex+syn	32.75	60.48	53.63

Best feature setting

	BLEU	TERp	METEOR
P+cs+sl+lex+syn	32.85	60.32	53.76

Re-ordering (r)
Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)

Impact of Using Phrase and Word under re-ordering

	BLEU	TERp	METEOR
System 03	30.16	63.04	51.94
backbone	30.89	61.28	52.65
W+r+cs	**31.13**	**60.99**	**53.01**
W+r+cs+sl	**31.33**	**61.72**	**52.55**
P+r+cs	**31.80**	**60.21**	**53.71**
P+r+cs+sl	**32.80**	**60.13**	**53.86**
P+r+cs+lex	31.76	60.12	53.54
P+r+cs+syn	31.72	60.37	53.38
P+r+cs+sl+lex+syn	32.75	60.48	53.63

Under the same settings,
phrase is always better than word

Re-ordering (r)
Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)

Two Environments

- Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)
- German-English combination shared task held by the WMT in 2011

German-English combination shared task (WMT 2011)

- One human reference translations and corresponding machine translations for the WMT 2011 test sets
- 10 combination system results are provided
- Manually select Top 6 systems out of 26 systems
- Tuning: 524 sentences
- Testing: 788 sentences

Result of backbone selection (WMT 2011)

	BLEU	TERp	METEOR
cmu-dyer	22.72	60.89	55.09
dfki-xu	22.44	62.31	53.89
kit	22.75	60.82	54.81
online-A	23.16	58.96	56.34
online-B	**24.27**	**57.89**	**56.93**
rwth-fre-c	21.86	62.82	53.46
backbone	**25.38**	**57.05**	**57.72**

Result of Combination (WMT 2011)

	BLEU	TERp	METEOR
Online B	**24.27**	**57.89**	**56.93**
backbone	**25.38**	**57.05**	**57.72**
koc-combo	23.41	61.83	54.08
quaero-combo	23.37	60.86	55.03
rwth-leusch-combo	25.62	57.44	57.20
jhu-combo	25.08	57.81	56.87
jhu-combo-contrastive	24.46	57.20	57.26
bbn-combo	**26.73**	**56.13**	**58.30**
cmu-heafield-combo	25.31	57.27	57.71
cmu-heafield-combo-contrastive	25.24	57.37	57.68
upv-prhlt-combo	24.65	59.25	56.24
uzh-combo	24.55	58.47	56.76
P+r+cs+sl	**25.81**	**56.89**	**57.88**
P+cs+sl+lex+syn	**25.96**	**57.18**	**57.64**

We try our best two settings, TTD is Top 2 out of 11 combination systems

Discussion

- Under same feature setting, feature is better than word
- Effect of phrase confidence score and System-specific LM is significant
- Effect of Lexical weighting, syntactic indicator and reordering is not very significant

Conclusion

- A new phrase-level combination technique
  - A novel perspective: “translation” from one target (backbone) to another target (combination output)
  - Several phrase confidence estimations are presented, such as phrase confidence score, lexical weighting and syntactic indicator
  - Introduce the capability of word re-ordering
  - System-specific LM is proposed

Future Work

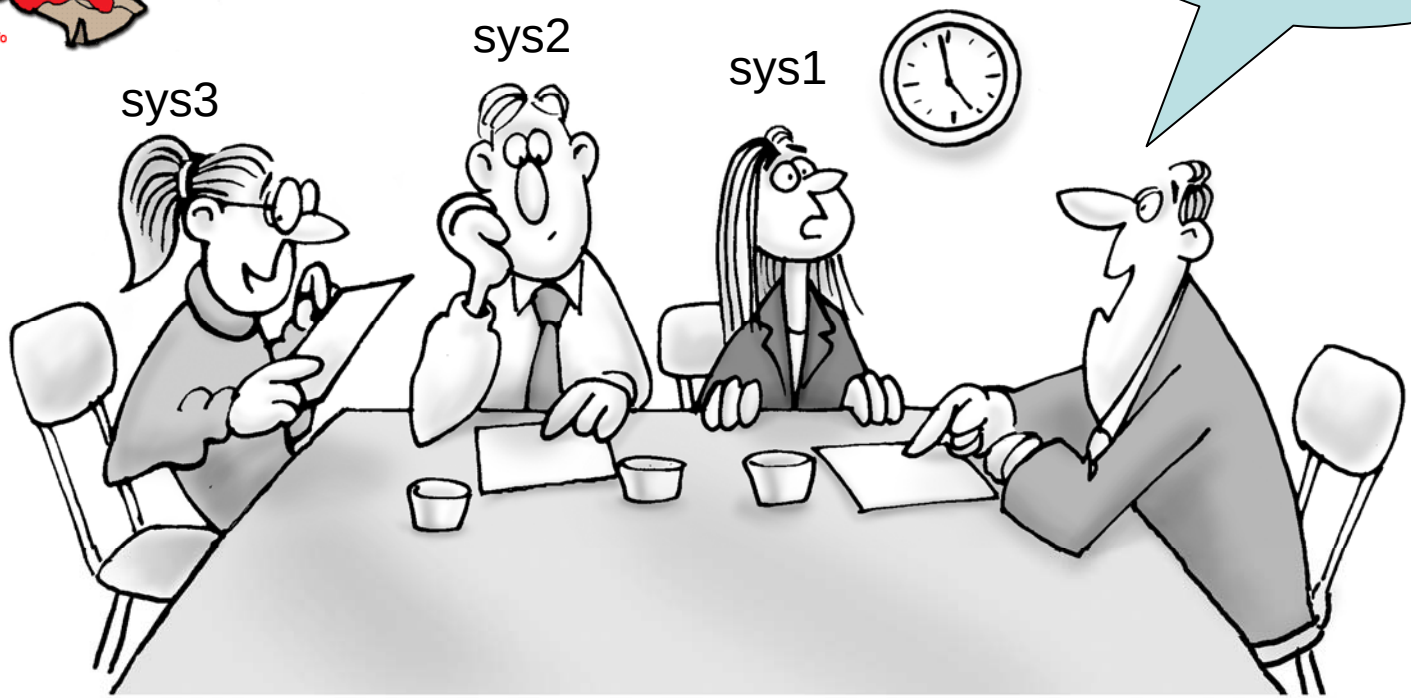
- Exploit information from the source
  - What do you think about this phrase?

    Ask the source.

Mrs. Source

It seems ok. The phrase in either sys1 or sys3 preserves the semantics (relation) of the source

What do you think about this phrase?



Future Work

- “Translation” from backbone to the combination result motivates that we can try other more comprehensive “translation” model than Moses
  - Ex: Hierarchical phrase-based model

    Syntax-oriented phrase-based model