Review on WMT09 Evaluation and Recent Research on System Combination

Jinhua Du

NDP National Development Plan 2007 - 2013
sfi science foundation ireland fondúireacht eolaíochta éireann
DCU Dublin City University
UCD DUBLIN University College Dublin
University of Limerick
Trinity College Dublin

Outline

- MaTrEx System in WMT09
  - Tasks
  - Data
  - System Framework
  - Results
- Key Components
  - Data Cleaning
  - Rescore Model
  - TrueCaser
- Research on System Combination
  - Hypothesis Alignment Metrics: TER, HMM and IHMM
  - Source-driven Hypothesis Alignment
- Conclusion and Future Work

MaTrEx System in WMT09

- Tasks:
  - Language pairs: French-English & English-French
  - Tracks: Translation; System Combination
- Data:

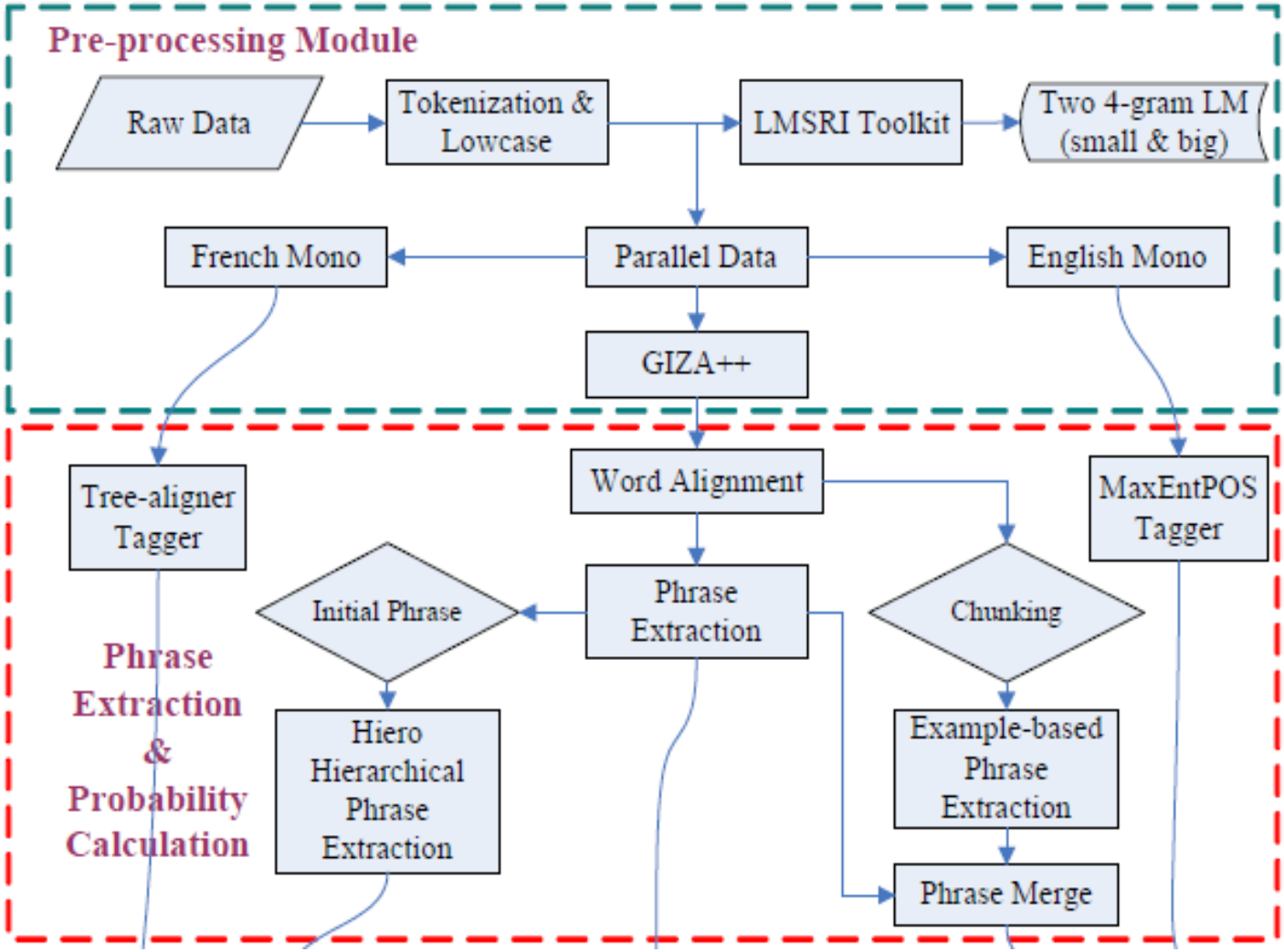
1) Parallel Data

Corpra	#Sen	#Token-En	#Token-Fr	#MaxSen_Length(Token)
Europarl	1.46m	39,240,672	42,252,067	80
Giga_news	2m	48,648,104	57,869,002	65

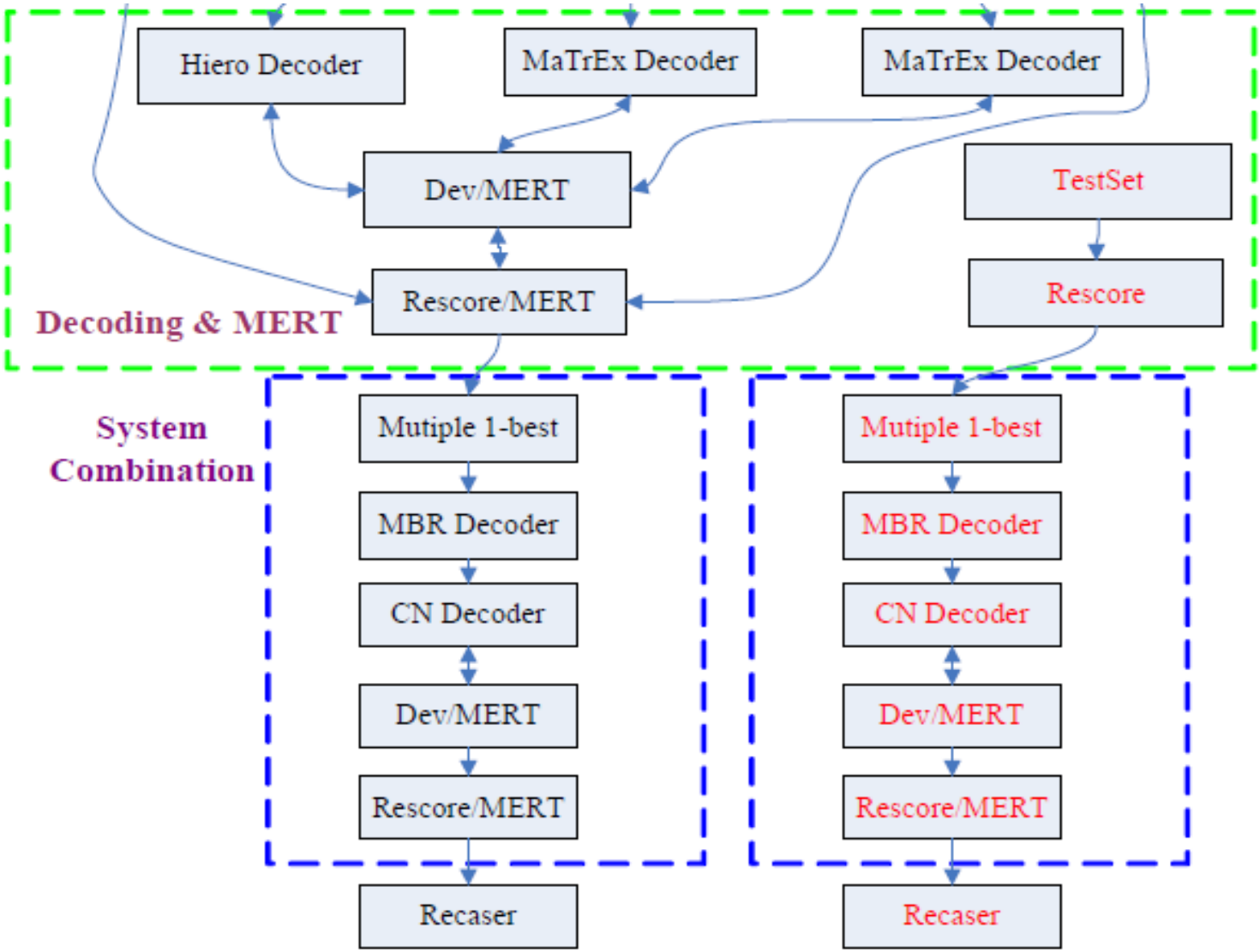
2) Monolingual Data

Language	#Sen	#Token	Source
English	9,966,838	240,849,221	Europarl/News/News_Commentary
French	9,966,838	260,520,313	Europarl/News/News_Commentary

System Framework – Preprocessing & Model Training



System Framework – Decoding & Combination



Hiero

- Parameters

Phrase Type	Length	Number	Time for Decoding			BLEU		
			$\alpha = 1.2$	$\alpha = 0.4$	$\alpha = 0.2$	$\alpha = 1.2$	$\alpha = 0.4$	$\alpha = 0.2$
IP	7	55.9m	NULL	NULL	NULL	NULL	NULL	NULL
HP	5	122m	10h	4h	10min	22.00	21.87	21.69

In table 4, 'IP' is the initial phrase table and 'HP' means the hierarchical phrase table.The decoding time and bleu score are evaluated by fr-en devset-a, which include 1025 sentences. From the statistics we can see, the pruning parameter α has a significant effect on speed rather than performance. Because of the time problem, we finally set $\alpha = 0.2$ for testset to get a faster decoding.

Results

- Experimental results on Dev and Test set

System	dev-a[BLEU-4]				testset[BLEU-4]			
	Fr-En		En-Fr		Fr-En		En-Fr	
	mert	test-b	mert	test-b	p-a	p-b	p-a	p-b
baseline(YH)	23.35/24.9	22.24	22.96	22.68	25.64/26.75	25.44	24.47/24.72	24.54/24.73
ebmt(SP)	22.58	22.04	22.78	22.12	25.67	25.41	24.43/24.94	24.17/24.35
hiero(JD)	21.92	21.69	22.58	21.12	25.20	25.11	24.19/24.61	24.14/24.48
mbr		22.33		22.68	26.97		25.08	
CN		22.45		22.76	27.03		25.11	
rescore		22.54		22.97	27.20		25.26	
Recasing					25.14		22.28	

Why Hiero didn't beat the Baseline

- No much time to adjust and optimize the parameters inside the decoder;
- Only used one 4-gram language model but phrase-based system used two language models;
- This is the first time we used Hiero in French-English pair, so there maybe some special linguistical phenomena we should consider.
- The initial phrase length is 7 words so that hiero can't generalize long phrase well.

Official Results – Human Evaluation

French–English

980 pairwise judgments per system

System	C?	≥others
GOOGLE •	no	.76
DCU ⋆	yes	.66
LIMSI •	no	.65
JHU ⋆	yes	.62
UEDIN ⋆	yes	.61
UKA	yes	.61
LIUM-SYSTRAN	no	.60
RBMT5	no	.59
CMU-STATXFER ⋆	yes	.58
RBMT1	no	.56
USAAR	no	.55
RBMT3	no	.54
RWTH ⋆	yes	.52
COLUMBIA	yes	.50
RBMT4	no	.47
GENEVA	no	.34

English-French

564 pairwise judgments per system

System	C?	≥others
LIUM-SYSTRAN •	no	.73
GOOGLE •	no	.68
UKA •⋆	yes	.66
SYSTRAN •	no	.65
RBMT3 •	no	.65
DCU •⋆	yes	.65
LIMSI •	no	.64
UEDIN ⋆	yes	.60
RBMT4	no	.59
RWTH	yes	.58
RBMT5	no	.57
RBMT1	no	.54
USAAR	no	.48
GENEVA	no	.38

C? indicates constraint condition, meaning only using the supplied training data and possibly standard monolingual linguistic tools (but no additional corpora).

- • indicates a **win** in the category, meaning that no other system is statistically significantly better at p-level≤0.1 in pairwise comparison.
- ⋆ indicates a **constraint win**, no other constraint system is statistically better.

Official Results – Automatic Evaluation

	RANK	LEU	LEU-CASED	LEU-TER	LEUSP	LEUSP4114	IST	IST-CASED	ER	ERP	VCD6P4ER	VPF	VPBLEU
English-French News Task													
DCU	0.65	0.24	0.22	–0.19	0.29	0.30	6.69	6.39	0.63	0.72	0.47	0.38	0.34
DCU-CMD	0.74	0.28	0.27	–0.15	0.33	0.34	7.29	7.12	0.58	0.67	0.44	0.42	0.38
GENEVA	0.38	0.15	0.14	–0.27	0.20	0.22	5.59	5.39	0.68	0.82	0.53	0.32	0.25
GOOGLE	0.68	0.25	0.24	–0.17	0.30	0.31	6.90	6.71	0.62	0.7	0.46	0.40	0.36
LIMSI	0.64	0.25	0.24	–0.17	0.3	0.31	6.94	6.77	0.60	0.71	0.46	0.4	0.35
LIUM-SYSTRAN	0.73	0.26	0.24	–0.17	0.31	0.32	7.02	6.83	0.61	0.71	0.45	0.40	0.36
RBMT1	0.54	0.18	0.17	–0.23	0.24	0.26	6.12	5.96	0.65	0.76	0.5	0.35	0.29
RBMT3	0.65	0.22	0.20	–0.20	0.27	0.28	6.48	6.29	0.63	0.72	0.48	0.38	0.33
RBMT4	0.59	0.18	0.17	–0.24	0.24	0.25	6.02	5.86	0.66	0.77	0.50	0.35	0.3
RBMT5	0.57	0.20	0.19	–0.21	0.26	0.27	6.31	6.15	0.63	0.74	0.49	0.36	0.31
RWTH	0.58	0.22	0.21	–0.19	0.27	0.28	6.67	6.51	0.62	0.75	0.48	0.38	0.32
SYSTRAN	0.65	0.23	0.22	–0.19	0.28	0.29	6.7	6.47	0.63	0.74	0.47	0.39	0.34
UEDIN	0.60	0.24	0.23	–0.18	0.29	0.30	6.75	6.57	0.62	0.71	0.47	0.39	0.35
UKA	0.66	0.24	0.23	–0.18	0.29	0.30	6.82	6.65	0.61	0.71	0.46	0.39	0.35
USAAR	0.48	0.19	0.18	–0.23	0.24	0.26	6.16	5.98	0.66	0.76	0.5	0.34	0.29
USAAR-CMD	0.77	0.27	0.25	–0.15	0.32	0.33	7.24	6.93	0.59	0.69	0.44	0.41	0.37
USAAR	0.48	0.19	0.18	–0.23	0.24	0.26	6.16	5.98	0.66	0.76	0.5	0.34	0.29
USAAR-CMD	0.77	0.27	0.25	–0.15	0.32	0.33	7.24	6.93	0.59	0.69	0.44	0.41	0.37

Official Results – Automatic Evaluation

	RANK	BLEU	BLEU-CASED	BLEU-TER	BLEUSP	BLEUSP4114	MAXSIM	METEOR-0.6	METEOR-0.7	METEOR-RANKING	NIST	NIST-CASED	RTE-ABSOLUTE	RTE-PAIRWISE	TER	TERP	ULC	WCD6P4ER	WPF	WPBLEU
French-English News Task																				
BBN-CMD	0.73	0.31	0.3	−0.11	0.36	0.38	0.54	0.59	0.64	0.45	7.88	7.58	0.14	0.12	0.2	0.20	0.36	0.40	0.41	0.37
CMU-CMD	0.66	0.3	0.29	−0.12	0.35	0.36	0.53	0.58	0.63	0.44	7.72	7.57	0.15	0.12	0.24	0.26	0.35	0.41	0.41	0.37
CMU-CMD-HYP	0.71	0.28	0.26	−0.14	0.33	0.35	0.53	0.57	0.61	0.43	7.40	7.15	0.1	0.08	0.31	0.33	0.34	0.42	0.4	0.35
CMU-STATXFER	0.58	0.24	0.23	−0.18	0.29	0.31	0.49	0.54	0.58	0.40	6.89	6.75	0.08	0.07	0.38	0.42	0.31	0.46	0.37	0.32
COLUMBIA	0.50	0.23	0.22	−0.18	0.29	0.30	0.49	0.54	0.58	0.40	6.85	6.68	0.07	0.07	0.36	0.39	0.31	0.46	0.36	0.31
DCU	0.66	0.27	0.25	−0.15	0.32	0.34	0.52	0.56	0.61	0.42	7.29	6.94	0.09	0.07	0.32	0.34	0.33	0.43	0.38	0.34
DCU-CMD	0.67	0.31	0.31	−0.11	0.36	0.37	0.54	0.59	0.64	0.44	7.84	7.69	0.14	0.12	0.21	0.22	0.35	0.41	0.42	0.38
GENEVA	0.34	0.14	0.14	−0.29	0.21	0.22	0.43	0.49	0.52	0.36	5.32	5.15	0.05	0.05	0.54	0.52	0.26	0.53	0.29	0.22
GOOGLE	0.76	0.31	0.30	−0.10	0.36	0.37	0.54	0.58	0.63	0.44	8	7.84	0.17	0.13	0.17	0.2	0.36	0.41	0.42	0.38
JHU	0.62	0.27	0.23	−0.15	0.32	0.33	0.51	0.56	0.6	0.41	7.23	6.68	0.08	0.05	0.33	0.36	0.32	0.43	0.37	0.32
LIMSI	0.65	0.26	0.25	−0.16	0.30	0.32	0.51	0.56	0.60	0.42	7.02	6.87	0.09	0.07	0.35	0.36	0.33	0.44	0.38	0.33
LIUM-SYSTRAN	0.60	0.27	0.26	−0.15	0.32	0.33	0.51	0.56	0.60	0.42	7.26	7.10	0.10	0.06	0.33	0.36	0.33	0.43	0.39	0.35
RBMT1	0.56	0.18	0.18	−0.25	0.24	0.25	0.48	0.53	0.57	0.4	5.89	5.73	0.07	0.06	0.51	0.45	0.3	0.50	0.34	0.26
RBMT3	0.54	0.2	0.19	−0.22	0.25	0.27	0.48	0.53	0.56	0.39	6.12	5.96	0.07	0.06	0.45	0.45	0.30	0.49	0.35	0.28
RBMT4	0.47	0.19	0.18	−0.24	0.24	0.26	0.48	0.52	0.56	0.39	5.97	5.83	0.07	0.06	0.46	0.45	0.3	0.49	0.34	0.27
RBMT5	0.59	0.19	0.19	−0.24	0.25	0.26	0.49	0.54	0.57	0.40	6.03	5.9	0.09	0.07	0.46	0.43	0.31	0.49	0.35	0.28
RWTH	0.52	0.25	0.24	−0.16	0.30	0.32	0.5	0.55	0.59	0.40	7.09	6.94	0.07	0.03	0.35	0.39	0.32	0.44	0.38	0.32
UEDIN	0.61	0.25	0.24	−0.16	0.31	0.32	0.50	0.55	0.59	0.41	7.04	6.85	0.08	0.04	0.35	0.38	0.32	0.44	0.38	0.33
UKA	0.61	0.26	0.25	−0.15	0.31	0.33	0.51	0.55	0.6	0.41	7.17	7.00	0.08	0.04	0.34	0.37	0.32	0.44	0.38	0.34
USAAR	0.55	0.19	0.18	−0.24	0.24	0.26	0.48	0.54	0.57	0.4	6.08	5.92	0.07	0.06	0.46	0.44	0.3	0.49	0.34	0.26
USAAR-CMD	0.57	0.26	0.25	−0.16	0.31	0.33	0.51	0.55	0.59	0.41	7.13	6.85	0.08	0.02	0.33	0.35	0.32	0.44	0.38	0.33

Key Components

- Clean data

1) We splitted the Giga corpus into even segments, each segment consisted of 20 lines.
2) We trained an SVM classifier with positive examples from the monolingual training data (in domain news data) and negative examples from noisy sentence (menus, numbers, meaningless word combinations, etc.) from the Giga corpus. The features were: adverb frequency, long word (length > 6) frequency, 'therefore' frequency, 'me' frequency, present tense frequency, 'I' frequency, average length of words, 'it' frequency, 'which' frequency, average length of sentences, frequency of capitalized characters, frequency of digit characters and frequency of 'if'. We filtered the segments from Giga corpus with this classifier. The last 25 percent of segments that were assigned the lowest scores were removed from the corpus.
3) We selected 1600 words that has the highest mutual information with monolingual training data against the Giga corpus.
4) We selected 100000 segments that these words occurred most frequently. However the sentence was dropped if the length ratio between en and fr is larger than 1.5 or less than 0.67. The selected sentences were processed by tools provided by WMT09.

Rescore Model – Global Features

- Direct and inverse IBM model;
- 3, 4-gram target language model;
- 3, 4, 5-gram POS language model;
- Sentence length posterior probability;
- N-gram posterior probabilities within the N-Best list;
- Minimum Bayes Risk probability;
- length ratio between source and target sentence;

N-Gram Posterior Probabilities & Sentence Length Posterior Probability (Zens, 2006)

- Length Posterior Probability

$$p(I|f_1^J) = \sum_{e_1^I} p(e_1^I|f_1^J)$$

- N-gram Posterior Probabilities

$$h_n(f_1^J, e_1^I) = \frac{1}{I} \log \left(\prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}, f_1^J) \right)$$

$$p(e_i|e_{i-n+1}^{i-1}, f_1^J) = \frac{C(e_{i-n+1}^{i}, f_1^J)}{C(e_{i-n+1}^{i-1}, f_1^J)}$$

TrueCaser

- The Truecase training model in moses is weak

Format	Fr-En	En-Fr
Before Cased	27.20	25.26
Cased	25.14	22.28
Down	2.06	2.98

- We should have a powerful TrueCaser

Sub-conclusion

- We won out in WMT09 evaluation
- For future evaluation and application, what needs to be done or refined:
  - Preprocessing: remove noises, get more parallel data from comparable corpus, tokenisation, filter corpus;
  - Name Entity Recognition and Translation: significant impact on translation quality, especially for News domain
  - Parallel Training and decoding: ICHEC?
  - TrueCaser
  - Should have more different types of MT systems, especially the decoders
  - Try to use more existent components into the system

Research on System Combination

- What is system combination

E_1	he have good car
E_2	he has nice sedan
E_3	it a nice car
E_4	a sedan he has

(a) hypothesis set

$$E_B = \arg\min_{E' \in \mathbf{E}} \sum_{E \in \mathbf{E}} TER(E', E)$$

e.g., $E_B = E_1$

(b) backbone selection

E_B he have ε good car

E_4 a ε sedan he has

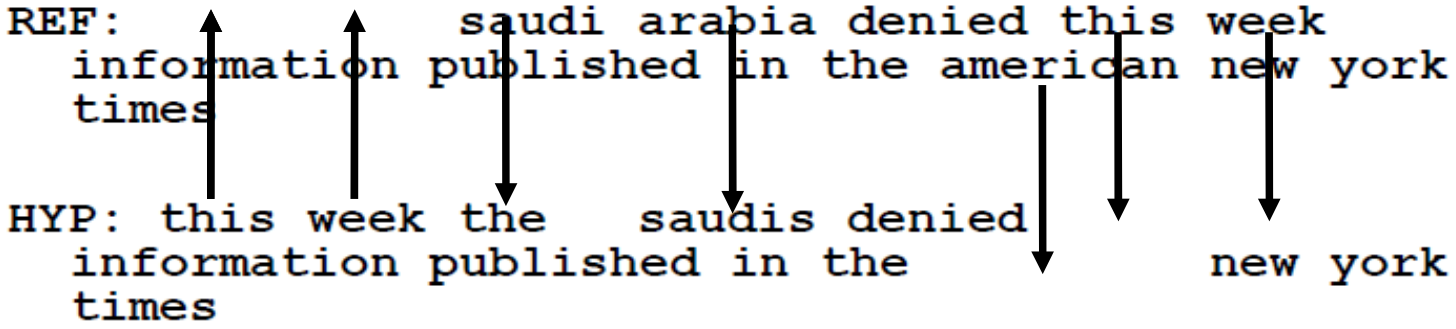
(c) hypothesis alignment

he	have	ε	good	car
he	has	ε	nice	sedan
it	ε	a	nice	car
he	has	a	ε	sedan

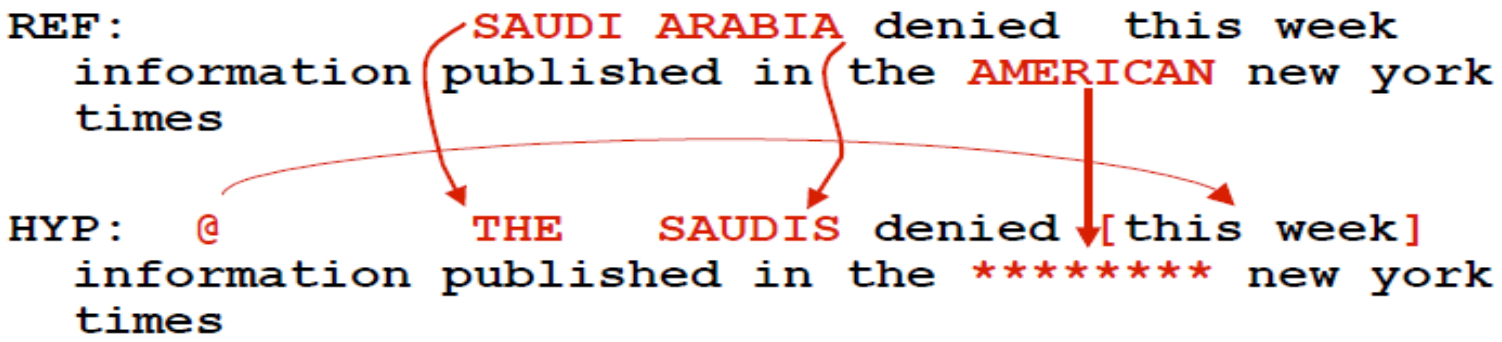
(d) confusion network

Hypothesis Alignment Metrics

- Standard – TER (translation error/edit rate) (Snover,2006)



WER = (7/13)=53.85%



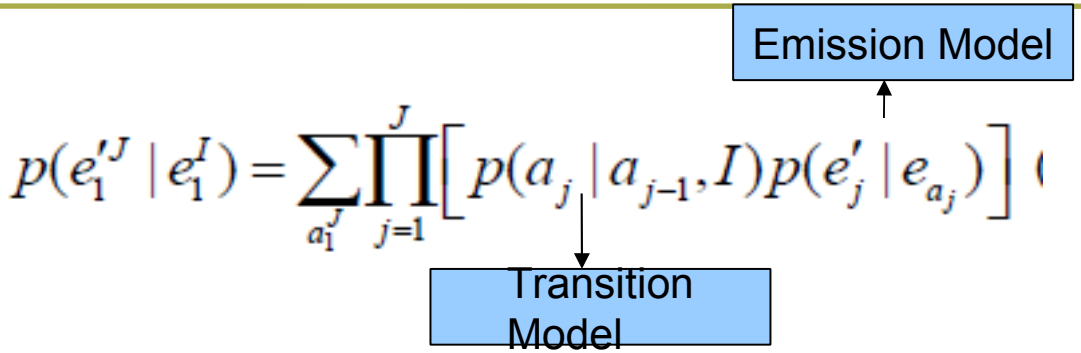
Edits:

- **Shift** "`this week`" **to after** "`denied`"
- **Substitute** "`Saudi Arabia`" **for** "`the Saudis`"
- **Insert** "`American`"

- **1 Shift, 2 Substitutions, 1 Insertion**
  - **4 Edits (TER = 4/13 = 31%)**

$$\mathrm{TER}(E, E_r) = \frac{\mathrm{Ins} + \mathrm{Del} + \mathrm{Sub} + \mathrm{Shft}}{N_r} \times 100\%$$

HMM Alignment (Matusov, 2006)



- Steps:
  - Select the 'backbone' and build the parallel corpus
  - Use GIZA++ to obtain the HMM alignment
  - Carry out the word re-ordering for all the rest of hypotheses
- Problems
  - Small training data
  - Complicated than TER

IHMM Alignment (He, 2008)

- Emission Model

$$p(e'_j \mid e_i) = \alpha \cdot p_{sem}(e'_j \mid e_i) + (1-\alpha) \cdot p_{sur}(e'_j \mid e_i)$$

where $p_{sem}(e'_j \mid e_i)$ and $p_{sur}(e'_j \mid e_i)$ reflect the semantic and surface similarity between e'_j and e_i, respectively, and α is the interpolation factor.

$$p_{sem}(e'_j \mid e_i) = \sum_{k=0}^{K} p(f_k \mid e_i) p(e'_j \mid f_k, e_i) \approx \sum_{k=0}^{K} p(f_k \mid e_i) p(e'_j \mid f_k)$$

$$p_{sur}(e'_j \mid e_i) = \exp\left\{\rho \cdot \left[s(e'_j, e_i) - 1\right]\right\}$$

$$s(e'_j, e_i) = \frac{M(e'_j, e_i)}{\max(|e'_j|, |e_i|)}$$

- Transition Model

$p(a_j = i \mid a_{j-1} = i', I)$ depend only on the jump distance $(i - i')$ (Vogel et al., 1996):

$$p(i \mid i', I) = \frac{c(i - i')}{\sum_{l=1}^{I} c(l - i')} \qquad (5)$$

$$c(d) = \left(1 + |d - 1|\right)^{-\kappa},\ d = -4, \ldots, 6$$

$$\tilde{p}(i \mid i', I) = \begin{cases} p_0 & \text{if } i = \textit{null} \text{ state} \\ (1 - p_0) \cdot p(i \mid i', I) & \text{otherwise} \end{cases}$$

Comparison Experiments

Set	Sentence	Pair	Source	Total System
Dev	502	En-Fr	WMT09	16
Test	2525	En-Fr	WMT09	16

System	DevSet			TestSet		
	BLEU	NIST	Imp.	BLEU	NIST	Imp.
Best Single	0.2723	6.4150	---	0.2543	6.9917	---
Oracle	0.3566	7.2577	8.43	0.3385	8.0356	8.42
MBR-Uniform	0.2823	6.5441	1.23	0.2654	7.1696	1.11
MBR-Mert	0.2950	6.6230	2.27	0.2728	7.2367	1.85
TER	0.2921	6.6495	1.98	0.2756	7.3312	2.13
HMM	0.2948	6.6893	2.25	0.2764	7.3806	2.21
IHMM	0.2927	6.6934	2.04			
WER	0.2949	6.6396	2.26	0.2751	7.3023	2.08
Com-MBR+	0.2996	6.7002	2.73	0.2778	7.3626	2.35

Sourceside Context Driven Alignment

- When decoding, keep the source phrase span
- Trace the alignment points of source-target phrase pair
- Locally align and reorder the word order of hypothesis

Source: l' inflation , en europe , a dérapé sur l' alimentation

Target: inflation |0-1| in |2-3| europe , has |4-6| dérapé |7-7| on |8-8| food |9-10|

Sentence ID:Seg-1
REF: inflation , in europe , has dérapé on food
HYP: inflation ***** in europe , has dérapé on food

Experimental Results

- Data: Europarl 1.5M
- Divided into 5 parts, each is 300K
- Training 5 models and get 5 systems
- Dev set: wmt-dev2009a (1025) and wmt-dev2009b (1026)
- Test set: wmt2009 test set (3027)

system	BLEU	NIST
Best Single	0.2303	6.8781
TER	0.2362	7.0185
HMM	0.2376	7.0623
Source-info	0.2368	7.0572

Conclusion and Future work

- Refine alignment model
- Improve the combination model of different alignment metrics