TC-STAR: Open Lab 2006

Comparison of System Combination Methods for ASR

B. Hoffmeister, T. Klein, G. Heigold, R. Schlüter, H. Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
RWTH Aachen University
D-52056 Aachen, Germany

1 Overview

Overview:

- **Motivation**
- **Combination techniques**
  - **ROVER**
  - **Confusion Network Combination (CNC)**
  - **Frame Based System Combination**
  - **Discriminative Model Combination (DMC)**
- **Results**

2 System Overview

Corpus	Site	Description	WER[%]	
EPPS Eval05 Spanish	Limsi	Open Lab data	11.2	12.3
	RWTH	Open Lab data	12.6	13.2
	RWTH	3-gram LM w/o LDA	13.6	14.9
		3-gram LM	12.2	13.1
		3-gram LM with VTN	11.8	12.6
		4-gram LM w/o LDA	13.2	14.6
		4-gram LM	11.9	12.8
		4-gram LM with VTN	11.7	12.1
EPPS Eval06 English	Limsi	Evaluation system (ctm file + conf.)	10.2	8.3
	IBM	Evaluation system (ctm file)	10.7	8.7
	UKA	Evaluation system (ctm file)	12.8	9.9
	IRST	Evaluation system (ctm file)	13.1	11.0
	RWTH*	Baseline +CMLLR/MLLR	14.1	11.8
		+MMI	13.7	11.7
		+SAT	13.3	10.8
		+improved lexicon and LM/ Evaluation system	12.9	10.3

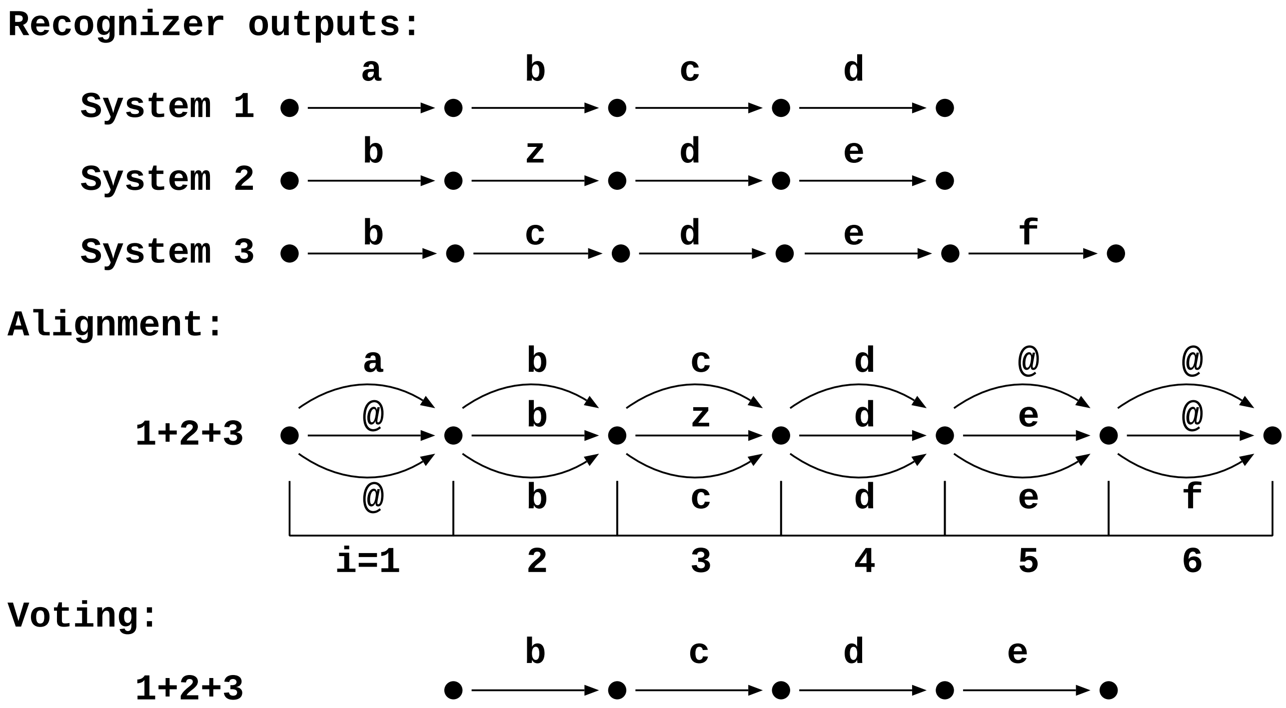
*Thanks to: Ch. Gollan, J. Lööf, Ch. Plahl, M. Bisani

3 Motivation

Motivation:

- **Different systems make different errors:**
  Example: EPPS Eval06en Evaluation Set
  Best system has 8.3 WER, oracle WER over all five submitted systems is 4.2.
  $\Rightarrow$ Combination of system outputs can improve recognition performance.
- **Additional information can improve the decision whether (a part of) a system's output is correct or not:**
  - **Confidence scores for best hypothesis.**
  - **Word graphs.**
- **Different methods for system combination are available:**
  - **ROVER**
  - **Confusion Network Combination (CNC)**
  - **Frame Based System Combination**
  - **Discriminative Model Combination (DMC)**

ROVER: Recognizer Output Voting Error Reduction

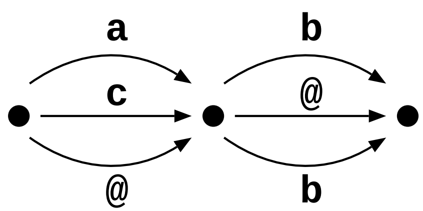


- J. G. Fiscus: A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), *IEEE ASRU Workshop*, 1997

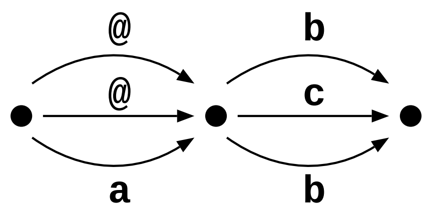
Alignment:

- **Alignment depends on system order:**

```
System 1: a b
System 2: c
System 3: b
Result  : a b
```




```
System 3: b
System 2: c
System 1: a b
Result  : b
```



- **Experience: Order of systems is important**
  - **Best system first**
  - **Use a parameter free voting to test all permutations.**
    Investigate only a reduced number of the best permutations.
- **Experience: Number of systems is important**
  - **Choose subset empirically.**
  - **Use a parameter free voting to test all subsets.**

6 ROVER

Voting:

- **Voting function: Score for word w at position i is**

$$\textbf{score}(w,i) \;=\; \frac{1}{S}\left(\alpha\sum_{s=1}^{S}\delta(w,w_{s,i}) + (1-\alpha)\sum_{s=1}^{S}\omega_s(\textbf{conf}_s(w,i))\right),$$

  S is the number of systems,
  $\alpha \in [0,1]$ interpolates majority vote and confidence scores, and
  $\omega_s(\cdot)$ is a system dependent confidence score weighting function.

- **Weighting functions**
  - **Majority vote:** $\omega_s(x) = \delta(w, w_{s,i})$ **or** $\alpha = 1$
  - **Unweighted confidence scores:** $\omega_s(x) = x$
  - **Linear weighted confidence scores:** $\omega_s(x) = \lambda_s x$
  - **Exp. weighted confidence scores:** $\omega_s(x) = x^{\lambda_s}$

7 ROVER

Parameter Optimization:

- **Free parameters:**
  - **None for majority vote.**
  - **Two for unweighted confidence scores: interpolation weight α and null confidence.**
  - **$S + 2$ for weighted confidence scores: additional S system weights.**
- **Multidimensional optimization problem:**
  - **Approaches: Grid search, Downhill simplex, Powell**
  - **Start Powell (Downhill Simplex) from 100 random start points.**
  - **Powell converges faster than Downhill simplex.**
  - **Difference between Grid search and Powell was $< 0.1\%$ WER (measured on a three-dimensional optimization problem).**

8 CNC

Confusion Network (CN) Decoding

- **Viterbi Decoding:**

$$\{w_1^N\}_{\text{viterbi}} = \underset{w_1^N}{\operatorname{argmax}}\, p(w_1^N | x_1^T),$$

$\{w_1^N\}_{\text{viterbi}}$ minimizes the Sentence Error Rate (SER).

- **Minimum Bayes Risk Decoding:**

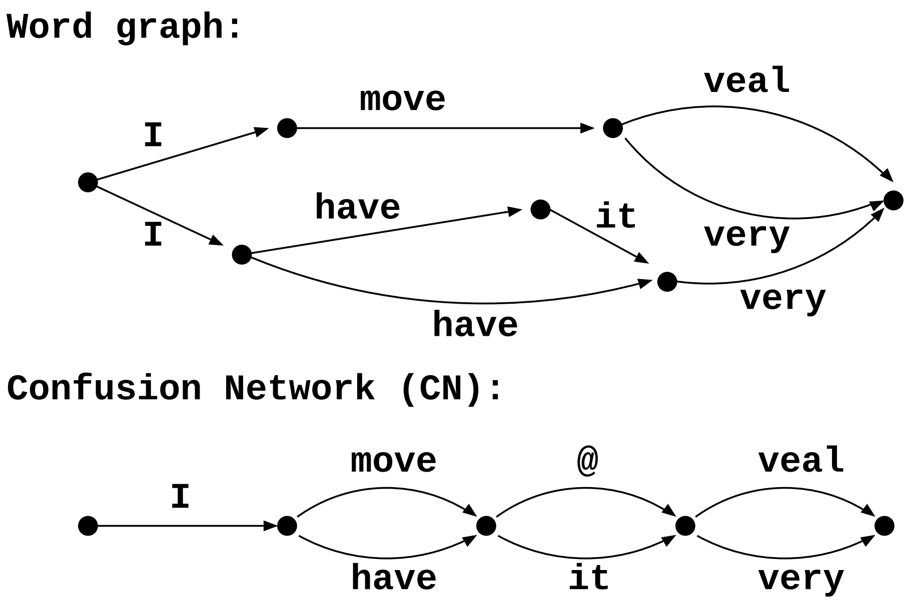
$$\{w_1^N\}_{\text{opt}} = \underset{w_1^N}{\operatorname{argmin}} \left\{ \sum_{v_1^M} C(w_1^N, v_1^M) p(v_1^M | x_1^T) \right\},$$

minimizes cost w.r.t. cost function C.
The Levenshtein distance as cost function minimizes the Word Error Rate (WER).

- **CN Decoding:**
  Idea:
  - **Approximate search space by CN, where the CNs are normally derived from word graphs.**
  - **For CN minimum WER decoding is simple.**

9 CNC

CN: Generation of CN from word graph

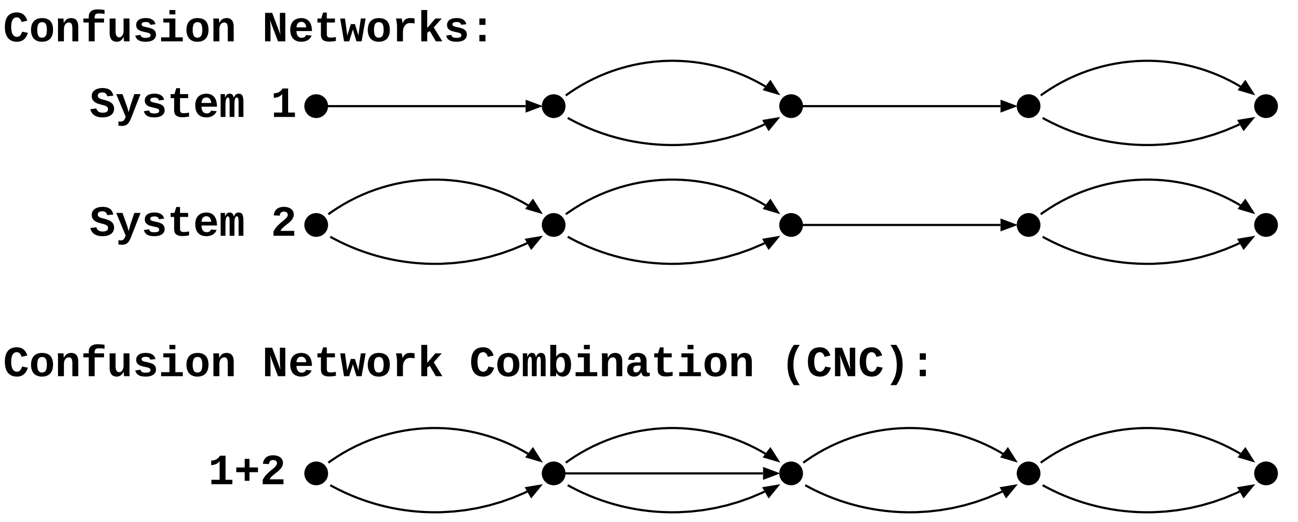


- L. Mangu, E. Brill and A. Stolcke (2000). Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks. In Computer, Speech and Language , 14(4):373-400, 2000.
- A. Stolcke (2002). SRILM - An Extensible Language Modeling Toolkit. In Proc. Intl. Conf. Spoken Language Processing, September 2002.

10 CNC

CNC: Confusion Network Combination Decoding

- **Idea: Combine CNs from several systems to super CN.**



- **Give the CNs system dependent weights.**

- **G. Evermann, P.C. Woodland (2000). Posterior Probability Decoding, Confidence Estimation and System Combination, In Proc. Speech Transcription Workshop, College Park, MD. 20**

11 Frame Based System Combination

Time Frame Error (fWER)

- **Minimum Bayes Risk Decoding:**

$$\{w_1^N\}_{\text{opt}} = \operatorname*{argmin}_{w_1^N} \left\{ \sum_{v_1^M} C(w_1^N, v_1^M) p(v_1^M | x_1^T) \right\}$$

- **Time Frame Error:**

$$C([w;t]_1^N, [v;\tau]_1^M) = \sum_{n=1}^{N} \left\{ \left(\sum_{\substack{\hat{t}=t_{n-1}+1; \\ v_{\hat{t}} \leftarrow [v;\tau]_m: \\ \tau_{m-1} < \hat{t} \leq \tau_m}}^{t_n} 1 - \delta(w_n, v_{\hat{t}}) \right) \Big/ \Big(1 + \alpha(t_n - t_{n-1} - 1) \Big) \right\}$$

  - **Experimental results show strong relation between WER and fWER.**

- **F. Wessel, R. Schlüter, and H. Ney (2001). "Explicit Word Error Minimization using Word Hypothesis Posterior Probabilities". In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 33-36, Salt Lake City, Utah, May 2001.**

12 Frame Based System Combination

Minimum Time Frame Error Decoding

- **Minimum fWER decoding:**

$$\{[w;t]_1^N\}_{\text{opt}} = \underset{[w;t]_1^N}{\operatorname{argmin}} \sum_{n=1}^{N} \left\{ \left(\sum_{\hat{t}=t_{n-1}+1}^{t_n} \left[1 - p(w_n|\hat{t}, x_1^T)\right] \right) \Big/ \Big(1 + \alpha(t_n - t_{n-1} - 1)\Big) \right\}$$

- **Minimum fWER decoding on word graphs:**
  - **A word graph is a directed and acyclic graph with E edges.**
  - **Complexity:**
    - $p(\cdot|t, x_1^T)$ **can be efficiently calculated with a modified Fwd./Bwd.-Algorithm** $\rightarrow O(E)$**.**
    - **Decode over all hypotheses in word graph** $\rightarrow O(E)$**.**

  $\Rightarrow$ **fWER decoding is efficient; at no stage an alignment is required.**
  - **Robustness:**
    - **Robust w.r.t word graph density. From 20 to several hundred no significant deviations.**
    - **Not robust w.r.t to word graph production, e.g. distorted probability distributions by "noise/silence clouds".**

13 Frame Based System Combination

Minimum Time Frame Error Decoding

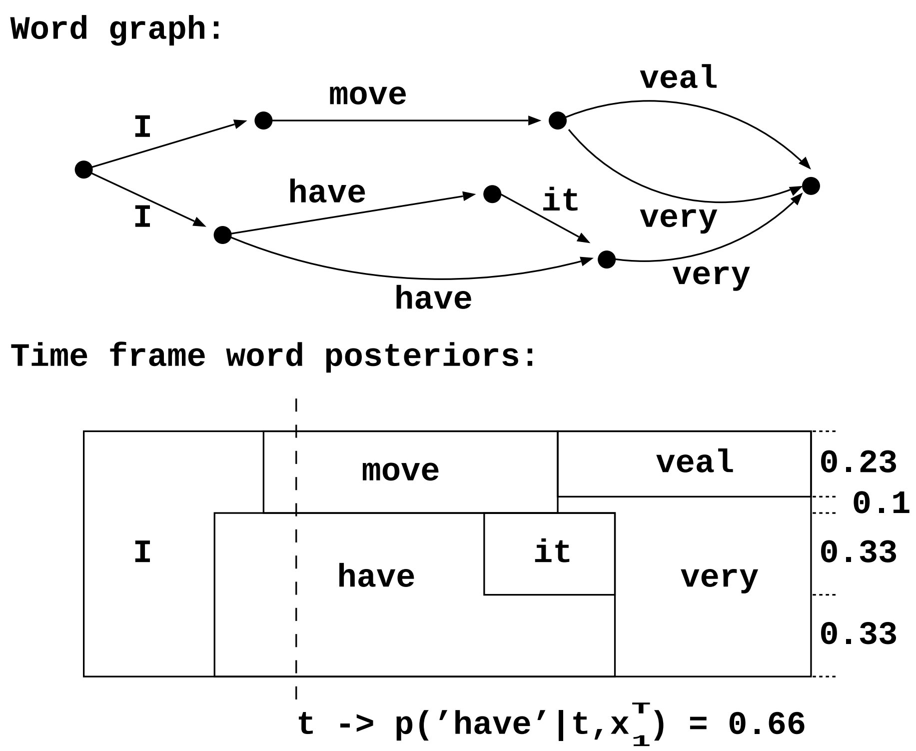


Illustration of the calculation of $p(\cdot|t, x_1^T)$ from a word graph.

14 Frame Based System Combination

Frame Based System Combination:

- **Time frame-wise word posterior distributions**

$$p(w|t,x_1^T) \;=\; \sum_{s=1}^{S} p(s|t,x_1^T)p(w|s,t,x_1^T),$$

  - **Approximate $p(s|t,x_1^T)$ by a system dependent constant λ_s.**
  - **Optimize λ_s per Grid Search or Powell.**

- **Decoding**
  - **Decode over the union of the system dependent word graphs.**
  - **Build a single time-conditioned word graph from the union.**
    $\Rightarrow$ Slight improvements over all corpora.

15 Log-Linear Model Combination

DMC: Discriminative Model Combination

- **Log-linear combination of models:**

$$p_\Lambda(w_1^N|x_1^T) = \frac{\exp(\sum_s \lambda_s f_s(x_1^T, w_1^N))}{\sum_{v_1^M} \exp(\sum_s \lambda_s f_s(x_1^T, v_1^M))}$$

- **Negative logarithm of respective model probabilities as feature functions: emission models, time distortion penalties, language models, ...**
- **Parameter estimation: minimization of expected word error on n-best lists**

- **P. Beyerlein (1998). Discriminative Model Combination, In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Seattle, WA.**

16 Log-Linear Model Combination

DMC: Discriminatively trained log-linear model combination

- **Promising results in feature combination experiments on Epps05en:**

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	MFCC with voicedness features	14.3	14.8
	VTN with voicedness features	13.8	14.0
DMC	(MFCC+voicedness)+(VTN+voicedness)	13.6	13.5

- **So far, fails to generalize for Epps06en:**

combination method	systems	WER[%] opt. set	WER[%] test set
single systems	MCE	16.0	15.8
	SAT+MCE+CMLLR+MLLR	13.1	12.9
DMC		12.9	13.1

17 Results: Overview

- **EPPS Eval05es**
  - **Internal System Combination**
  - **Internal System Combination + Limsi**
  - **Official Workshop Data**
- **EPPS Eval06en**
  - **Internal System Combination**
  - **External System Combination**

18 Eval05es: RWTH Internal Systems + Limsi

Baseline results: Epps 2005 Spanish

	Viterbi WER[%]		CN WER[%]		fWER [%]		graph WER[%]		avg. density	
	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.
3-gram LM										
w/o LDA	13.6	14.9	13.6	14.8	13.4	14.9	3.2	4.0	152	163
	12.2	13.1	12.2	13.0	12.1	13.0	4.8	5.5	46	54
with VTN	11.8	12.6	11.9	12.5	11.7	12.5	5.0	5.8	33	42
4-gram LM										
w/o LDA	13.2	14.6	13.2	14.5	13.2	14.6	3.4	4.2	119	129
	11.9	12.8	11.9	12.8	11.9	12.9	5.2	6.0	36	42
with VTN	11.7	12.1	11.7	12.1	11.5	12.2	4.9	5.8	32	40
Limsi	11.2	12.3	11.2	12.2	-	-	4.0	5.0	15	15

19 Eval05es: RWTH Internal Systems + Limsi

Internal System Combination: Epps 2005 Spanish

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	lm4 w/o LDA	13.2	14.6
	lm3	12.2	13.1
	lm4 with VTN	11.7	12.1
Oracle*		8.1	8.7
ROVER		11.3	12.2
	+ conf. scores	11.2	12.0
	+ linear weighted conf. scores	11.2	11.9
	+ exp. weighted conf. scores	11.2	12.1
CNC		11.3	12.2
	+ weights	11.3	12.1
Frame Based		11.2	12.2
	+ weights	11.1	12.1

* ROVER's oracle WER on single systems best hypotheses.

20 Eval05es: RWTH Internal Systems + Limsi

Internal System Combination + Limsi: Epps 2005 Spanish

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	RWTH (lm3)	12.2	13.1
	RWTH (lm4 with VTN)	11.7	12.1
	Limsi	11.2	12.3
Oracle*		6.6	7.3
ROVER		10.4	11.4
	+ conf. scores	10.3	11.2
	+ linear weighted conf. scores	10.0	10.8
	+ exp. weighted conf. scores	10.3	11.1
CNC		10.6	11.3
	+ weights	10.3	11.2
	best RWTH internal comb.	11.2	11.9

* ROVER's oracle WER on single systems best hypotheses.

21 Official Workshop Data

Official Workshop Data: Epps 2005 Spanish

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	RWTH	12.6	13.2
	Limsi	11.2	12.3
Oracle*		7.6	8.5
ROVER		11.8	12.2
	+ conf. scores	11.6	12.0
	+ linear weighted conf. scores	10.5	11.6
	+ exp. weighted conf. scores	10.5	11.6
CNC		11.0	11.8
	+ weights	10.7	11.6

* ROVER's oracle WER on single systems best hypotheses.

22 Eval06en: RWTH Internal Systems

Baseline results: Epps 2006 English

Baseline system: fastVTN, voicedness features, 4-gram LM

	Viterbi WER[%]		CN WER[%]		fWER [%]		graph WER[%]		avg. density	
	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.
+ CMLLR/MLLR	14.1	11.8	14.1	11.8	13.9	11.8	2.5	0.1	71	46
+ MMI	13.7	11.7	13.7	11.7	13.5	11.5	2.5	0.1	70	46
+ SAT	13.3	10.8	13.4	10.9	13.1	10.8	3.1	1.2	66	46
+ improved lexicon and LM	12.9	10.3	13.0	10.3	12.7	10.3	3.7	1.7	62	44

23 Eval06en: RWTH Internal Systems

Internal System Combination: Epps 2006 English

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	+CMLLR/MLLR	14.1	11.8
	+MMI	13.7	11.7
	+SAT	13.3	10.8
	+improved lexicon and LM	12.9	10.3
Oracle*		10.8	8.6
ROVER	w/o +CMLLR/MLLR	13.0	10.5
	+ conf. scores	12.6	10.5
	+ linear weighted conf. scores	12.5	10.4
	+ exp. weighted conf. scores	12.6	10.5
CNC		13.1	10.6
	+ weights	12.9	10.2
Frame Based		12.8	10.7
	weights	12.5	10.3

* ROVER's oracle WER on single systems best hypotheses.

24 Eval06en: External Systems

External System Combination: Epps 2006 English

combination method	systems	WER[%] dev.	WER[%] eval.
single systems	Limsi	10.2	8.3
	IBM*	10.7	8.7
	UKA*	12.8	9.9
	RWTH	12.9	10.3
	IRST*	13.1	11.0
Oracle**		4.8	4.2
ROVER	w/o UKA	8.8	7.0
	+ conf. scores	8.7	6.9
	+ linear weighted conf. scores	8.4	6.8
	+ exp. weighted conf. scores	8.5	6.9

* CTM files without confidence scores.

** ROVER's oracle WER on single systems best hypotheses.