An empirical study on computing consensus translations from multiple machine translation systems

Author：Wolfgang Macherey, Franz Josef Och

Professor：陳嘉平

Reporter：陳逸昌

Outline

- Introduction
- Methods for system combination
- Experimental results
- Conclusion

Introduction

- Three questions?
  - To what extent can translation quality benefit from combining systems developed by multiple research labs?
  - How can a set of diverse translation systems be built from a single translation engine?
  - How can an ensemble of translation outputs be selected from a large pool of translation?

Methods for systems combination

- Systems combination via candidate selection
- ROVER-Like combination schemes
- A Two-pass search algorithm

Systems combination via candidate selection

- Simply returns one of the original candidate translation.

- Typically, this selection is made based on translation scores, confidence estimations, language and other models.

The BLEU correlation matrix

Id	01	02	03	04	05	⋯	14	15	16
01	1.00	0.27	0.26	0.23	0.26	⋯	0.15	0.15	0.12
02	0.27	1.00	0.27	0.22	0.25	⋯	0.15	0.15	0.12
03	0.26	0.27	1.00	0.21	0.28	⋯	0.15	0.15	0.10
04	0.23	0.22	0.21	1.00	0.19	⋯	0.14	0.12	0.12
05	0.26	0.25	0.28	0.19	1.00	⋯	0.16	0.17	0.11
06	0.27	0.24	0.25	0.21	0.26	⋯	0.16	0.18	0.13
⋮						⋱			⋮
14	0.15	0.15	0.15	0.14	0.16	⋯	1.00	0.12	0.08
15	0.15	0.15	0.15	0.12	0.17	⋯	0.12	1.00	0.09
16	0.12	0.12	0.10	0.12	0.11	⋯	0.08	0.09	1.00

- (B_{ij}) is defined as the sentence-based BLEU score between a candidate translation e_i and a pseudo-reference translation e_j.
- e_i and e_j is the output of translation systems.
- For example: (B_{12}) use the output from 2nd system as reference to score the output from 1st system.

- Each translation system m is assigned to a system prior weight ω_m.
- use ω and B to chose the best sentence.
- Two useful properties:
  - Doesn't depend on score translation outputs.
  - The system prior weights ω can easily be trained using the Minimum Error Rate Training.
- Call the combination scheme MBR-like system combination.

ROVER-like combination schemes

- Computing a composite translation by voting on confusion networks .
- Use TER to construct the confusion network.

ROVER-like combination schemes

$$\omega_m^{'} = \frac{\left(\omega^T \cdot b_m\right)^l}{\sum_{\tilde{m}} \left(\omega^T \cdot b_{\tilde{m}}\right)^l}, l \in \left[0, +\infty\right)$$

- ω_m :the system prior weight
- b_m :the BLEU matrix m_{th} row vector
- The system which refined system prior weights are denoted by word sausages+.

A two-pass search algorithm

- First pass
  - uses a greedy strategy to determine a bag of words which minimizes the position-independent word error rate (PER).
  - The first pass finishes when putting further constituents into the bag of words does not improve the PER.

A two-pass search algorithm

- Second pass
  - Starts with the empty string and then expends all active hypotheses by systematically inserting the next unused word.
  - The resulting hypotheses are scored with respect to the TER measure with pseudo references.
  - The second pass will finish if either no unused word are left or if expending the set of hypotheses does not improve TER.

Experimental results

- On two corpora for Chinese-English
  - Random selected from MT evaluations up to the year 2005
  - NIST MT evaluations 2006
    - Broadcast news articles(565 sentences)
    - Newswire texts(616 sentences)
    - News group texts(483 sentences)
- 4 reference translation

Table 1: *Corpus statistics for two Chinese-English text translation sets: ZHEN-05 is a random selection of test data used in NIST evaluations prior to 2006; ZHEN-06 comprises the NIST portion of the Chinese-English evaluation data used in the 2006 NIST machine translation evaluation.*

corpus		**Chinese**	**English**
ZHEN-05	sentences	2390	
	chars / words	110647	67737
ZHEN-06	sentences	1664	
	chars / words	64292	41845

Combining multiple research systems

system	BLEU score
1	32.1
2	31.71
3	29.59
4	27.7
5	27.05
6	27.02
7	26.75
8	26.41
9	25.05
10	23.48
11	23.26
12	22.38
13	22.13
14	17.42
15	17.2
16	15.21

- The consensus translation results for the MBR-like candidate selection method (select best).

Combining multiple research systems

combination	primary system			consensus				oracle	
	BLEU	CI 95%	BP	BLEU	Δ	BP	pair-CI 95%	BLEU	BP
01, 02, 03	32.10	(±0.88)	0.93	32.97	(+0.87)	0.92	[+0.29, +1.46]	38.54	0.94
01, 15, 16*	32.10	(±0.88)	0.93	23.55	(-8.54)	0.92	[-9.29, -7.80]	33.55	0.95
02, 03, 04	31.71	(±0.90)	0.96	31.55	(-0.16)	0.92	[-0.65, +0.29]	37.23	0.95
03, 04, 05	29.59	(±0.88)	0.87	29.55	(-0.04)	0.88	[-0.53, +0.41]	35.55	0.92
03, 04, 06*	29.59	(±0.88)	0.87	29.83	(+0.24)	0.90	[-0.29, +0.71]	35.69	0.93
04, 05, 06	27.70	(±0.87)	0.94	28.52	(+0.82)	0.91	[+0.15, +1.49]	34.67	0.94
05, 06, 07	27.05	(±0.81)	0.88	28.21	(+1.16)	0.92	[+0.63, +1.66]	33.89	0.94
05, 06, 08*	27.05	(±0.81)	0.88	28.47	(+1.42)	0.91	[+0.95, +1.95]	34.18	0.93
06, 07, 08	27.02	(±0.76)	0.92	28.12	(+1.10)	0.94	[+0.59, +1.59]	33.87	0.95
07, 08, 09	26.75	(±0.79)	0.97	27.79	(+1.04)	0.94	[+0.52, +1.51]	33.54	0.95
08, 09, 10	26.41	(±0.81)	0.92	26.78	(+0.37)	0.94	[-0.07, +0.86]	32.47	0.96
09, 10, 11	25.05	(±0.84)	0.90	24.96	(-0.09)	0.94	[-0.59, +0.46]	30.92	0.97
10, 11, 12	23.48	(±0.68)	1.00	24.24	(+0.76)	0.94	[+0.27, +1.30]	30.08	0.96
11, 12, 13	23.26	(±0.74)	0.95	24.05	(+0.79)	0.92	[+0.40, +1.23]	29.56	0.93
12, 13, 14	22.38	(±0.78)	0.87	22.68	(+0.30)	0.89	[-0.28, +0.95]	28.58	0.91
13, 14, 15	22.13	(±0.72)	0.89	21.29	(-0.84)	0.90	[-1.33, -0.33]	26.61	0.92
14, 15, 16	17.42	(±0.66)	0.93	18.45	(+1.03)	0.92	[+0.45, +1.56]	23.30	0.95
15	17.20	(±0.64)	0.91	—	—	—	—	—	—
16	15.21	(±0.63)	0.96	—	—	—	—	—	—

Combining multiple research systems

- All translation systems achieve nearly equal quality.
- Combing high-performing systems with low-quality translation typically results in clear performance losses.

Non-uniform system prior weights

- Investigate the effect of using non-uniform system weights for the combination.

- Split oracle test set into five random partitions of almost equal size.

# systems	combination	uniform		ω opt. on dev.			ω opt. on test	
		BLEU	BP	BLEU	BP	pair-CI 95%	BLEU	BP
3	01 – 03	32.98	0.92	33.03	0.93	[-0.23, +0.34]	33.60	0.93
4	01 – 04	**33.44**	0.93	33.46	0.93	[-0.26, +0.29]	34.97	0.94
5	01 – 05	33.07	0.92	33.14	0.93	[-0.29, +0.43]	34.33	0.93
6	01 – 06	32.86	0.92	33.53	0.93	[+0.26, +1.08]	34.43	0.93
7	01 – 07	33.08	0.93	33.51	0.93	[+0.04, +0.82]	34.49	0.93
8	01 – 08	33.12	0.93	33.47	0.93	[-0.06, +0.75]	34.50	0.94
9	01 – 09	33.15	0.93	33.22	0.93	[-0.35, +0.51]	34.68	0.93
10	01 – 10	33.01	0.93	33.59	0.94	[+0.18, +0.96]	34.79	0.94
11	01 – 11	32.84	0.94	33.40	0.94	[+0.13, +0.98]	34.76	0.94
12	01 – 12	32.73	0.93	33.49	0.94	[+0.34, +1.18]	34.83	0.94
13	01 – 13	32.71	0.93	33.54	0.94	[+0.39, +1.26]	34.91	0.94
14	01 – 14	32.66	0.93	**33.69**	0.94	[+0.58, +1.47]	34.97	0.94
15	01 – 15	32.47	0.93	33.57	0.94	[+0.63, +1.57]	34.99	0.94
16	01 – 16	32.51	0.93	33.62	0.94	[+0.62, +1.59]	35.00	0.94

- The best combined system obtained with trained system weights(01-14)

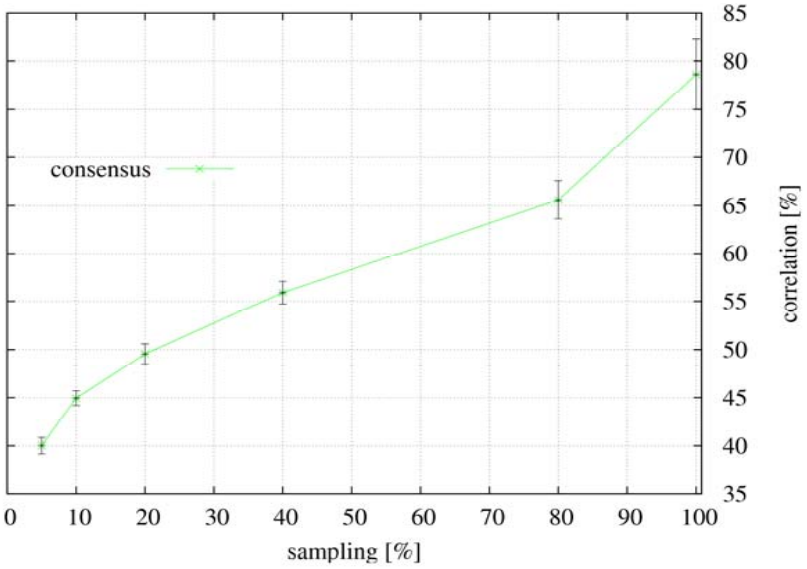
- Not significantly better than the best combined system using uniform weights(01-04)

Effect of correlation on system combination

- The correlation is a key factor which decides the overall system performance improves.
- If the correlation is too large, there will be insufficient diversity among the input.
- If the correlation is too low, there might be no consensus among the input.

- Built a large number of systems trained on randomly sampled portion of the training data.
- Sample sizes ranged between 5% and 10 % with each larger data set doubling the size of the next smaller collection.
- Created 10 data sets for each sample size, thus resulting in a total of 6X10 training corpora.

sampling	primary			mbr-like		sausages		sausages+		two-pass		two-pass+	
[%]	BLEU	CI 95%	BP	BLEU	BP	BLEU	BP	BLEU	BP	BLEU	BP	BLEU	BP
5	27.82	(±0.65)	1.00	29.51	1.00	29.00	0.97	30.25	0.99	29.58	0.94	29.93	0.96
10	29.70	(±0.69)	1.00	31.42	1.00	30.74	0.98	31.99	0.99	31.30	0.95	31.75	0.97
20	31.37	(±0.69)	1.00	32.56	1.00	32.64	1.00	33.17	0.99	32.60	0.96	32.76	0.98
40	32.66	(±0.66)	1.00	33.52	1.00	33.23	0.99	33.98	1.00	33.65	0.97	33.88	0.99
80	33.67	(±0.66)	1.00	**34.17**	1.00	33.93	0.99	**34.38**	1.00	**34.20**	0.99	**34.35**	1.00
100	**33.90**	(±0.67)	1.00	34.03	1.00	**33.98**	1.00	34.02	1.00	33.90	1.00	34.08	1.00



Effect of correlation on system combination

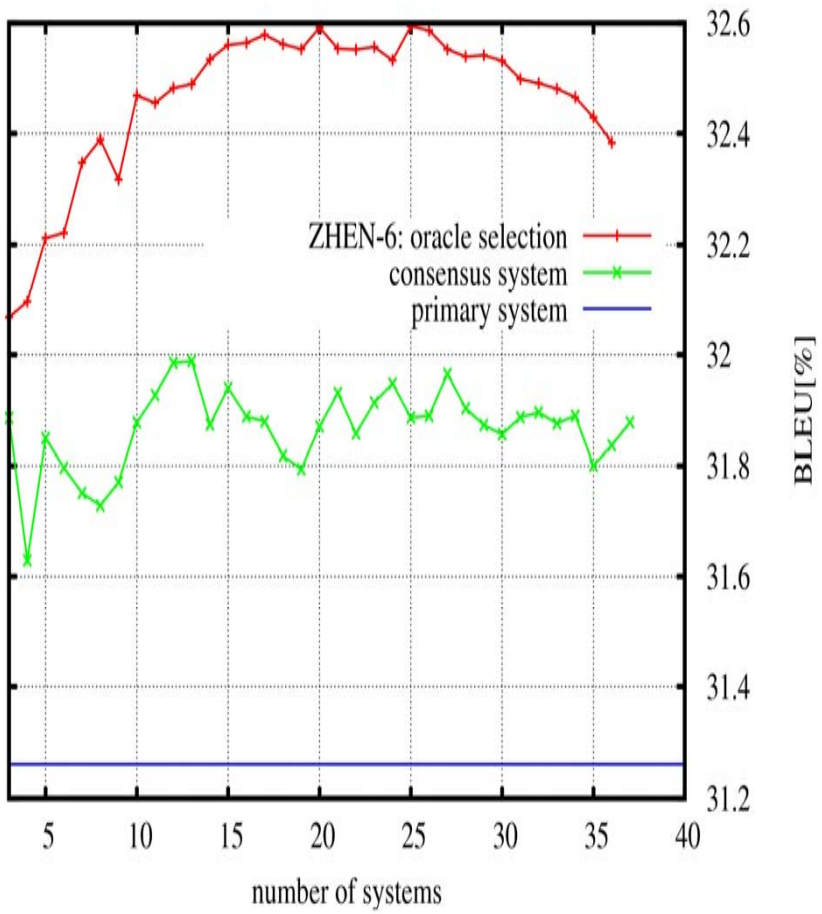
- Increasing the correlation substantially reduces the potential of a consensus system to outperform the primary translation system.

- For two-pass and sausage, as soon as the correlation increases and their outputs produced by the individual systems become more similar.

Toward automatic system generation and selection

- A method to reduce correlation without sacrificing system performance.
- Change few parameters may have a strong impact on system correlation.
- Parameters that were changed include
  - the maximum jump width in word re-ordering
  - The choice of feature function weights for the log-linear translation models.
  - The set of language models.

Toward automatic system generation and selection



- Used a greedy strategy to rank the systems with respect to their ability to improve system combination.
- Adding those systems, which gave the highest gain in terms of BLEU score according to the MBR-like method.

Conclusion

- Both high diversity and similar translation quality are essential.
- Trained a large pool of translation systems from a single translation engine with low correlation.