DCU Confusion Network-based System Combination for ML4HMT

Tsuyoshi Okita (Dublin City University)
Josef van Genabith (Dublin City University)

ML4HMT Workshop at Barcelona

Table Of Contents

System Combination Strategy

- System combination has become a standard tool in MT evaluation campaigns.
- Confusion network-based system combination is dominant.
  - DCU's system is similar with LIUM's MANY [Barrault, 2010] and CMU's MEMT [Heafield, 2010].
- Main components are
  - MBR decoding
  - confusion network

MBR Decoding (1)

- Given such a loss function $L(E, E')$ between an automatic translation E' and the reference E, a set of translation outputs $\mathcal{E}$, and an underlying probability model $P(E|F)$, a MBR decoder is defined as follows [Kumar and Byrne, 02]:

$$\begin{aligned} E &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\ &= \text{argmin}_{E' \in \mathcal{E}} \sum_{E' \in \mathcal{E}} L(E, E') P(E|F) \end{aligned} \tag{1}$$

where $R(E')$ denotes the Bayes risk of candidate translation E' under the loss function L.

MBR Decoding (2)

- MBR decoding [Kumar and Byrne, 02]:

$$\hat{E}_{best}^{MBR} = \text{argmin}_{E' \in \mathcal{E}} R(E')$$

$$= \text{argmin}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \quad (2)$$

$$= \text{argmax}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} BLEU_E(E') P(E|F) \quad (3)$$

$\mathcal{E}$: translation outputs of all the MT systems, E: Reference translation,$L(E, M(F))$: loss function.

- MAP decoding (Noisy Channel Model)

$$\hat{E}_{best}^{MAP} = \text{argmax}_E \sum_A P(E, A|F) \quad (4)$$

MBR Decoding (3)

A	MT output seqs				prob	1-gram expectation
1	a	a	a	c	0.30	$\mathbb{E}_A$(aaac)=0.3*4+0.2*0+0.2*0+0.2*0+0.1*0=1.2
2	b	b	c	d	0.20	$\mathbb{E}_A$(bbcd)=0.3*0+0.2*4+0.2*3+0.2*3+0.1*1=2.1
3	b	b	b	d	0.20	$\mathbb{E}_A$(bbbd)=0.3*0+0.2*3+0.2*4+0.2*2+0.1*2=2.0
4	b	b	c	f	0.20	$\mathbb{E}_A$(bbcf)=0.3*0+0.2*3+0.2*2+0.2*4+0.1*0=1.8
5	f	f	b	d	0.10	$\mathbb{E}_A$(ffbd)=0.3*0+0.2*1+0.2*2+0.2*0+0.1*4=1.0

B	MT output seqs				prob	1-gram expectation
1	a	a	a	c	0.33	$\mathbb{E}_B$(aaac)=0.33*4+0.22*0+0.22*0+0.22*0+0.00*0=1.32
2	b	b	c	d	0.22	$\mathbb{E}_B$(bbcd)=0.33*0+0.22*4+0.22*3+0.22*3+0.00*1=**2.20**
3	b	b	b	d	0.22	$\mathbb{E}_B$(bbbd)=0.33*0+0.22*3+0.22*4+0.22*2+0.00*2=1.98
4	b	b	c	f	0.22	$\mathbb{E}_B$(bbcf)=0.33*0+0.22*3+0.22*2+0.22*4+0.00*0=1.98
5	-	-	-	-	0.00	

Table: MBR decoding can be schematically described as the expectation of the number of matching between the MT output sequence and some sequence, as is described in this table.

MBR Decoding (4)

- Realization of simple question.
  - How about 'abcf'? Note 'abcf' is not an element of $\mathcal{E}$.
  - How about 'aaaf'?
  - How about ...?
  - This becomes confusion network.

System Combination Procedures (1)

- Procedures: For given set of MT outputs,
  1. Choose backbone by a MBR decoder from MT outputs $\mathcal{E}$.
  2. Monolingual word alignment of two word sequences based on the backbone (This becomes a confusion network).
  3. Run the decoding algorithm to choose the best path in the confusion network.

System Combination Procedures (2)

segment 3	
Input 1	they are normally on a week .
Input 2	these are normally made in a week .
Input 3	este himself go normally in a week .
Input 4	these do usually in a week .
Input 5	they are normally in one week .
Backbone(2)	these are normally made in a week .

Backbone(2)	these	are	normally	made	in	a	week .
hyp(1)	$they_S$	are	normally	$*****_D$	on_S	a	week .
hyp(3)	$este_S$	$himself_S$	go_S	$normally_S$	in	a	week .
hyp(4)	these	$*****_D$	do_S	$usually_S$	in	a	week .
hyp(5)	$they_S$	are	normally	$*****_D$	in	one_S	week .
Output	these	are	normally	*****	in	a	week .

System Combination Strategy

- Problems:
  - Good for automatic score. May not good for human evaluation (Not obtaining METEOR gain).
    - No coordination between words in i-th position and j-th position.(as a result , the magnetization of the recording layer 3 of magnetization becomes almost zero .)
    - Strong assumption that testset score reflects devset score.
    - Selection of given MT inputs requires expert knowledge. ⟸ This paper focuses on this!
    - Monolingual alignment relies on TER [Snover, 06].

Our MBR Decoding Strategy (1)

- How about enlarging hypothesis space $\mathcal{E}_H$? (Tromble et al. enlarges evidence space $\mathcal{E}_E$: Lattice MBR).
- We consider all the possible subsets of the full set of MT outputs, as is shown in (6):

$$\hat{E} = \text{argmin}_{\mathcal{E}_i \subseteq \mathcal{E}} \sum_{E' \in \mathcal{E}_i} L(E, E')P(E|F) \quad (5)$$

$$= \text{argmin}_{E' \in \mathcal{E}_{H_i}, \mathcal{E}_{H_i} \subseteq \mathcal{E}} \sum_{E' \in \mathcal{E}_{E_i}} L(E, E')P(E|F) \quad (6)$$

  where $\mathcal{E}_{H_i} \subseteq \mathcal{E}$ indicates that we choose $\mathcal{E}_{H_i}$ from all the possible subsets of $\mathcal{E}$ (or a power set of $\mathcal{E}$), $\mathcal{E}_{H_i}$ denote a i-th hypothesis space, $\mathcal{E}_{E_i}$ denote a i-th evidence space, and $\mathcal{E}_{H_i}$ has relation with $\mathcal{E}_{E_i}$.
- A power set of $\mathcal{E} = \{1, 2\}$ is $\{\{1, 2\}, \{1\}, \{2\}, \emptyset\}$.

Our MBR Decoding Strategy (2)

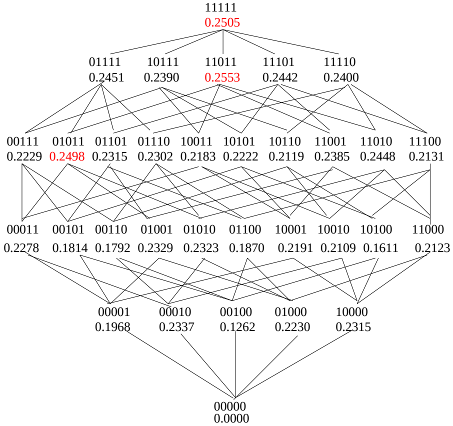


Figure: Figure shows the lattice of five MT output sequences encoded as binary sequences ('11111', '01111', etc) and BLEU scores ('0.2505', '0.2451', etc).

Experiments (ES-EN)

	NIST	BLEU	METEOR
system t1	6.3934	0.1968/0.1289*	0.5022487
system t2	6.3818	0.2337/0.1498*	**0.5732194**
system t3	4.5648	0.1262/0.0837*	0.4073446
system t4	6.2136	0.2230/0.1343*	0.5544878
system t5	6.7082	0.2315/0.1453*	0.5412563
our algorithm	**6.8419**	**0.2553**	0.5683086
combination 1 (t1,t2,t3,t4,t5)	6.7151	0.2505	0.5701207
combination 2 (t1,t2,t4,t5)	6.8419	0.2553	0.5683086
combination 3 (t2,t4,t5)	6.7722	0.2498	0.5687383

Table: All the scores are on testset except those marked * (which are on devset). On comparison, we did sampling of three combinations of the single systems, which shows that our results are equivalent to the combination 2.

Experiments (JP-EN)

JP-EN	NIST	BLEU	METEOR
system t1	7.0374	0.2532	0.6083487
system t2	7.2992	0.2775	0.6223682
system t3	5.1474	0.1243	0.4527874
system t4	6.6323	0.1913	0.5590906
system t5	6.6682	0.2165	0.5827379
system t6	6.8597	0.2428	0.5909936
system t7	7.2555	0.2755	0.6193990
system t8	6.1250	0.1946	0.6090198
system t9	7.2182	0.2529	0.6062563
system t10	5.6288	0.1727	0.5141809
system t11	7.2625	0.2529	0.6105696
baseline	7.2992	0.2775	0.6223682
heuristic 1	7.4292	0.2750	0.6228906
heuristic 2	7.2992	0.2775	0.6223682
our algorithm ('11100010101')	**7.5161**	**0.2869**	**0.6305818**

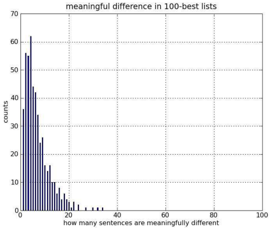
Conclusion

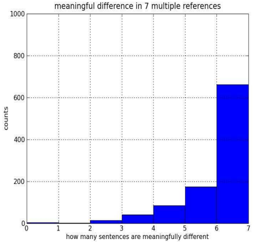
- System combination module in MaTrEx at Dublin City University.
    - We introduce a new input selection mechanism which removes some radically bad systems for the sake of achieving final better overall performance.
    - Automatic evaluation in terms of BLEU, NIST, WER, PER was the first among four participants. Improvement was 2.16 BLEU points absolute and 9.2% relative compared to the best single system.
    - Note that we observed this between JP-EN [Okita et al.,10] as well.

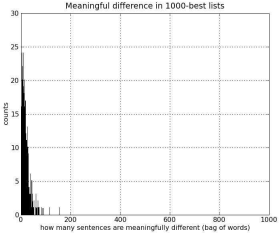
Further Studies

- Effect of bad translation inputs in system combination. (Our implementation of Eq (2) was naive).
- Case when our inputs are the 1000-best list? [Tromble et al.,08;DeNero et al.,08]. Their improvements tend to be quite small compared to the confusion network-based approach. The 100-best list and the 1000-best list produced by Moses [Koehn et al.,07] tend not to be sufficiently different. How about carefully add only some useful outputs rather than employed as is in system combination?

Further Studies







Acknowledgement

Thank you for your attention.

- We would like to thank Jinhua Du.
- This research is supported by the the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME project (Grant agreement No. 249119).
- This research is supported by the 7th Framework Programme of the European Commission through the Euro Matrix Plus project (Grant No. 231720).
- Irish Centre for High-End Computing.