ML4HMT: DCU Teams Overview

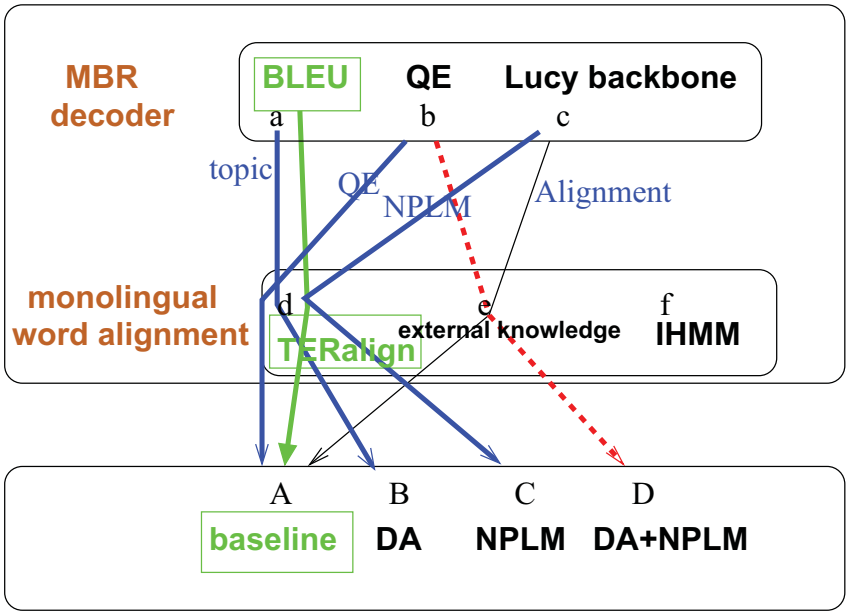
Tsuyoshi Okita
Dublin City University

CNGL Centre for Next Generation Localisation
NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD DUBLIN University College Dublin
University of Limerick
Trinity College Dublin

DCU Teams Overview

- Meta information
  - DCU-Alignment: alignment information
  - DCU-QE: quality information
  - DCU-DA: domain ID information
  - DCU-NPLM: latent variable information

Our Strategies



confusion network construction

monotonic consensus decoding

Standard system combination (green)

This presentation shows tuning results of blue lines.

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
     - Run monolingual word aligner

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

Centre for Next Generation Localisation

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
     - Run monolingual word aligner
  2. Run monotonic (consensus) decoder (with MERT tuning)

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

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 .

⇓ **1. MBR decoding**

Backbone(2)	these are normally made in a week .

⇓ **2. monolingual word alignment**

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 .

⇓ **3. monotonic consensus decoding**

Output	these	are	normally	*****	in	a	week .

1. MBR Decoding

1. Given MT outputs, choose 1 sentence.

$$\begin{aligned}
\hat{E}_{best}^{MBR} &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\
&= \text{argmin}_{E' \in \mathcal{E}} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \\
&= \text{argmin}_{E' \in \mathcal{E}} \sum_{E' \in \mathcal{E}_E} (1 - BLEU_E(E')) P(E|F) \\
&= \text{argmin}_{E' \in \mathcal{E}} \\
&\left[\mathbf{1} - \begin{bmatrix} B_{E_1}(E_1) & B_{E_2}(E_1) & B_{E_3}(E_1) & B_{E_4}(E_1) \\ B_{E_1}(E_2) & B_{E_2}(E_2) & B_{E_3}(E_2) & B_{E_4}(E_2) \\ & \dots & \dots & \\ B_{E_1}(E_4) & B_{E_2}(E_4) & B_{E_3}(E_4) & B_{E_4}(E_4) \end{bmatrix} \right] \begin{bmatrix} P(E_1|F) \\ P(E_2|F) \\ P(E_3|F) \\ P(E_4|F) \end{bmatrix}
\end{aligned}$$

1. MBR Decoding

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 .

$$
\begin{aligned}
&= \text{argmin} \left[\mathbf{1} - \begin{bmatrix} 1.0 & 0.259 & 0.221 & 0.245 \\ 0.267 & 1.0 & 0.366 & 0.377 \\ & \dots & \dots & \\ 0.245 & 0.366 & 0.346 & 1.0 \end{bmatrix} \right] \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} \\
&= \text{argmin}\,[0.565, 0.502, 0.517, 0.506] \\
&= (\mathit{Input}2)
\end{aligned}
$$

Backbone(2)	these are normally made in a week .

2. Monolingual Word Alignment

- TER-based monolingual word alignment
  - Same words in different sentence are aligned
  - Proceeded in a pairwise manner: Input 1 and backbone, Input 3 and backbone, Input 4 and backbone.

Backbone(2)	these	are	normally	made	in	a	week .
hyp(1)	they$_S$	are	normally	*****$_D$	on$_S$	a	week .
Backbone(2)	these	are	normally	made	in	a	week .
hyp(3)	este$_S$	himself$_S$	go$_S$	normally$_S$	in	a	week .
Backbone(2)	these	are	normally	made	in	a	week .
hyp(4)	these	*****$_D$	do$_S$	usually$_S$	in	a	week .

3. Monotonic Consensus Decoding

- Monotonic consensus decoding is limited version of MAP decoding
  - monotonic (position dependent)
  - phrase selection depends on the position (local TMs + global LM)

$$
\begin{aligned}
e_{best} &= \arg\max_{e} \prod_{i=1}^{I} \phi(i|\bar{e}_i) p_{LM}(e) \\
&= \arg\max_{e} \{\phi(1|\text{these})\phi(2|\text{are})\phi(3|\text{normally})\phi(4|\emptyset)\phi(5|\text{in}) \\
&\quad \phi(6|\text{a})\phi(7|\text{week}) p_{LM}(e), \ldots\} \\
&= \text{these are normally in a week} \qquad (1)
\end{aligned}
$$

1 ||| these ||| 0.50	2 ||| are ||| 0.50	3 ||| normally ||| 0.50
1 ||| they ||| 0.25	2 ||| himself ||| 0.25	...
1 ||| este ||| 0.25	2 ||| $\emptyset$ ||| 0.25	...

System Combination with Extra Alignment Information

Xiaofeng Wu, Tsuyoshi Okita, Josef van Genabith, Qun Liu
Dublin City University

CNGL Centre for Next Generation Localisation
NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD Dublin
University College Dublin
University of Limerick
Trinity College Dublin

Centre for Next Generation Localisation

Table Of Contents

Objective

- Meta information
  - Alignment information
- ML4HMT dataset includes alignment information when MT systems decode.
- Usual monolingual alignment in system combination do not use such external alignment information.

Standard System Combination Procedures

- Procedures: For given set of MT outputs,
  1. (Standard approach) Choose backbone by a MBR decoder from MT outputs $\mathcal{E}$.

$$\begin{aligned} \hat{E}_{best}^{MBR} &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\ &= \text{argmin}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \quad (2) \\ &= \text{argmax}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} BLEU_E(E') P(E|F) \quad (3) \end{aligned}$$

  2. Monolingual word alignment between the backbone and translation outputs in a pairwise manner(This becomes a confusion network).
     - TER alignment [Sim et al., 06]
     - IHMM alignment [He et al., 08]
  3. Run the (monotonic) consensus decoding algorithm to choose the best path in the confusion network.

Our System Combination Procedures

- Procedures: For given set of MT outputs,
  1. (Standard approach) Choose backbone by a MBR decoder from MT outputs $\mathcal{E}$.

$$\begin{aligned} \hat{E}_{best}^{MBR} &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\ &= \text{argmin}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \quad (4) \\ &= \text{argmax}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} BLEU_E(E') P(E|F) \quad (5) \end{aligned}$$

  2. Monolingual word alignment with prior knowledge (about alignment links) between the backbone and translation outputs in a pairwise manner (This becomes a confusion network).
  3. Run the (monotonic) consensus decoding algorithm to choose the best path in the confusion network.

IHMM Alignment [He et al., 08]

- Same as conventional HMM alignment [Vogel et al., 96] except
- Word semantic similarity and word surface similarity
  - word semantic similarity: source word seq = hidden word seq

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

  - exact match, longest matched prefix, longest common subsequences
    - "week" and "week" (exact match).
    - "week" and "weeks" (longest matched prefix).
    - "week" and "biweekly" (longest common subsequences)
- Distance-based distortion penalty.

Alignment Bias

- In (monotonic) consensus decoding,
  - big weight for Lucy alignment and
  - low weight for conflicting alignment with Lucy.
- This can be expressed as

$$p(E_\psi) \quad = \quad \theta_\psi \, logp(E_\psi|F) \tag{6}$$

where $\psi = 1, \ldots, N_{nodes}$ denotes the current node at which the beam search arrived. $\theta_\psi > 1$ if a current node is Lucy alignment and $\theta_\psi = 1$ if a current node is not Lucy alignment.

Lucy Backbone

- We used the Lucy backbone since it seems better than other backbone.

	Devset(1000)		Testset(3003)	
TER Backbone	8.1168	0.3351	7.1092	0.2596
Lucy Backbone	8.1328	0.3376	7.4546	0.2607

Table: TER Backbone selection results.

Extra Alignment Information Experiments

θ_ψ	Devset(1000)		Testset(3003)	
	NIST	BLEU	NIST	BLEU
1	8.1328	0.3376	7.4546	0.2607
1.2	8.1179	0.3355	7.2109	0.2597
1.5	8.1171	0.3355	7.4512	0.2578
2	8.1252	0.3360	7.4532	0.2558
4	8.1180	0.3354	7.3540	0.2569
10	8.1190	0.3354	7.1026	0.2557

Table: The Lucy backbone with tuning of θ_ψ.

Discussion: HMM-MAP (Bayesian HMM) Alignment

- Hidden Markov Model

$$p(s_{1:T}, y_{1:T}) = p(s_1)p(y_1|s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(y_t|s_t) \quad (7)$$

  - $p(s_t|s_{t-1})$: transition matrix
  - $p(y_t|s_t)$: emission matrix
- HMM-MAP (Bayesian HMM)
  - Prior on transition matrix and emission matrix
- IHMM-MAP
  - Prior on transition matrix and emission matrix
  - Word semantic similarity and word surface similarity
  - Distance-based distortion penalty

Conclusion

- We focus on adding extra alignment information on consensus decoding.
- Our results show that with choosing Lucy, which is an RBMT system, as a backbone the result is slightly better (0.11% improvment by BLEU) than the traditional TER backbone selection method.
- Extra alignment information we added in the decoding part does not improve the performance.

Acknowledgement

Thank you for your attention.

- 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 Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.