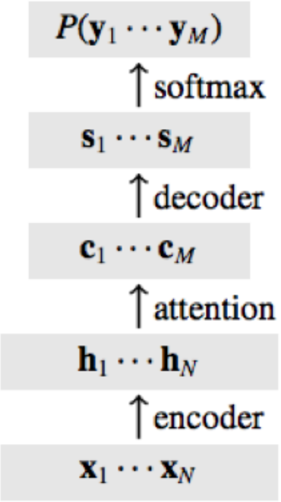
Tied Multitask Learning for Neural Speech Translation

Antonios Anastasopoulos *and* David Chiang

University of Notre Dame

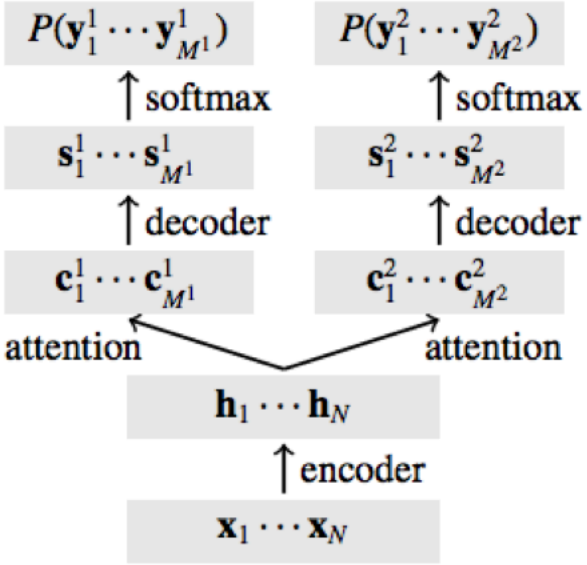
Multitask Learning



$P(\mathbf{y}_1 \cdots \mathbf{y}_M)$
softmax
$\mathbf{s}_1 \cdots \mathbf{s}_M$
decoder
$\mathbf{c}_1 \cdots \mathbf{c}_M$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

$$\begin{aligned} \mathbf{s}_m &= \mathrm{dec}(\mathbf{s}_{m-1}, \mathbf{c}_m, \mathbf{y}_{m-1}) \\ P(\mathbf{y}_m) &= \mathrm{softmax}(\mathbf{s}_m). \end{aligned}$$

single-task



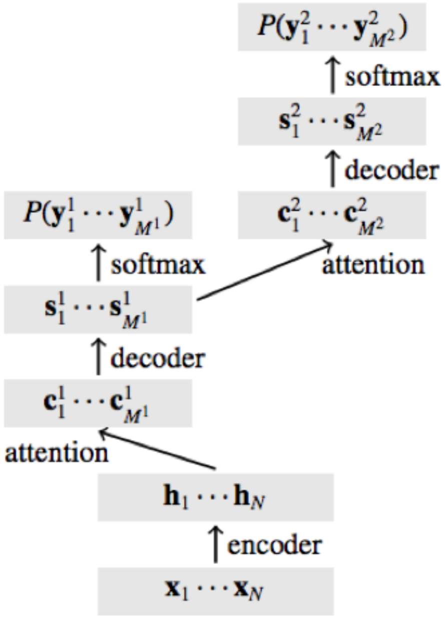
$P(\mathbf{y}^1_1 \cdots \mathbf{y}^1_{M^1})$
$P(\mathbf{y}^2_1 \cdots \mathbf{y}^2_{M^2})$
softmax
softmax
$\mathbf{s}^1_1 \cdots \mathbf{s}^1_{M^1}$
$\mathbf{s}^2_1 \cdots \mathbf{s}^2_{M^2}$
decoder
decoder
$\mathbf{c}^1_1 \cdots \mathbf{c}^1_{M^1}$
$\mathbf{c}^2_1 \cdots \mathbf{c}^2_{M^2}$
attention
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

$$\begin{aligned} \mathbf{c}^1_m &= \sum_n \alpha^1_{mn} \mathbf{h}_n \\ \mathbf{s}^1_m &= \mathrm{dec}^1(\mathbf{s}^1_{m-1}, \mathbf{c}^1_m, \mathbf{y}^1_{m-1}) \\ P(\mathbf{y}^1_m) &= \mathrm{softmax}(\mathbf{s}^1_m) \end{aligned}$$

$$\begin{aligned} \mathbf{c}^2_m &= \sum_n \alpha^2_{mn} \mathbf{h}_n \\ \mathbf{s}^2_m &= \mathrm{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1}) \\ P(\mathbf{y}^2_m) &= \mathrm{softmax}(\mathbf{s}^2_m). \end{aligned}$$

standard multitask

- higher-level intermediate representations should carry information useful for an end task
- e.g. speech->transcription->translation

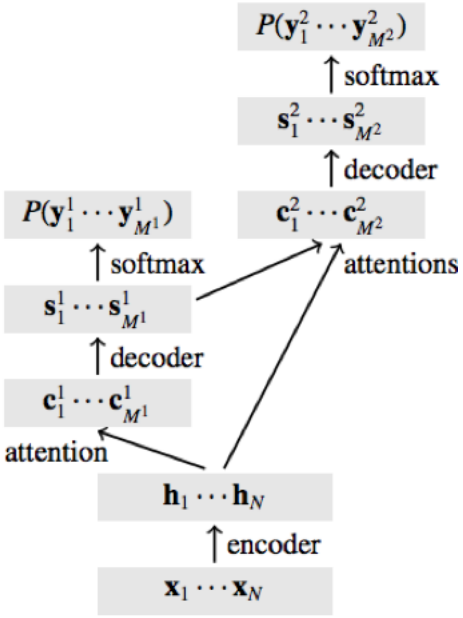

$P(\mathbf{y}^2_1 \cdots \mathbf{y}^2_{M^2})$
softmax
$\mathbf{s}^2_1 \cdots \mathbf{s}^2_{M^2}$
decoder
$P(\mathbf{y}^1_1 \cdots \mathbf{y}^1_{M^1})$
$\mathbf{c}^2_1 \cdots \mathbf{c}^2_{M^2}$
softmax
attention
$\mathbf{s}^1_1 \cdots \mathbf{s}^1_{M^1}$
decoder
$\mathbf{c}^1_1 \cdots \mathbf{c}^1_{M^1}$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

$$\mathbf{c}^2_m = \sum_{m'} \alpha^{12}_{mm'} \mathbf{s}^1_{m'}$$

$$\mathbf{s}^2_m = \mathrm{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1})$$

$$P(\mathbf{y}^2_m) = \mathrm{softmax}(\mathbf{s}^2_m).$$

cascade


$P(\mathbf{y}^2_1 \cdots \mathbf{y}^2_{M^2})$
softmax
$\mathbf{s}^2_1 \cdots \mathbf{s}^2_{M^2}$
decoder
$P(\mathbf{y}^1_1 \cdots \mathbf{y}^1_{M^1})$
$\mathbf{c}^2_1 \cdots \mathbf{c}^2_{M^2}$
softmax
attentions
$\mathbf{s}^1_1 \cdots \mathbf{s}^1_{M^1}$
decoder
$\mathbf{c}^1_1 \cdots \mathbf{c}^1_{M^1}$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

$$\mathbf{c}^2_m = \left[\textstyle\sum_{m'} \alpha^{12}_{mm'} \mathbf{s}^1_{m'} \quad \textstyle\sum_n \alpha^2_{mn} \mathbf{h}_n \right]$$

$$\mathbf{s}^2_m = \mathrm{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1})$$

$$P(\mathbf{y}^2_m) = \mathrm{softmax}(\mathbf{s}^2_m).$$

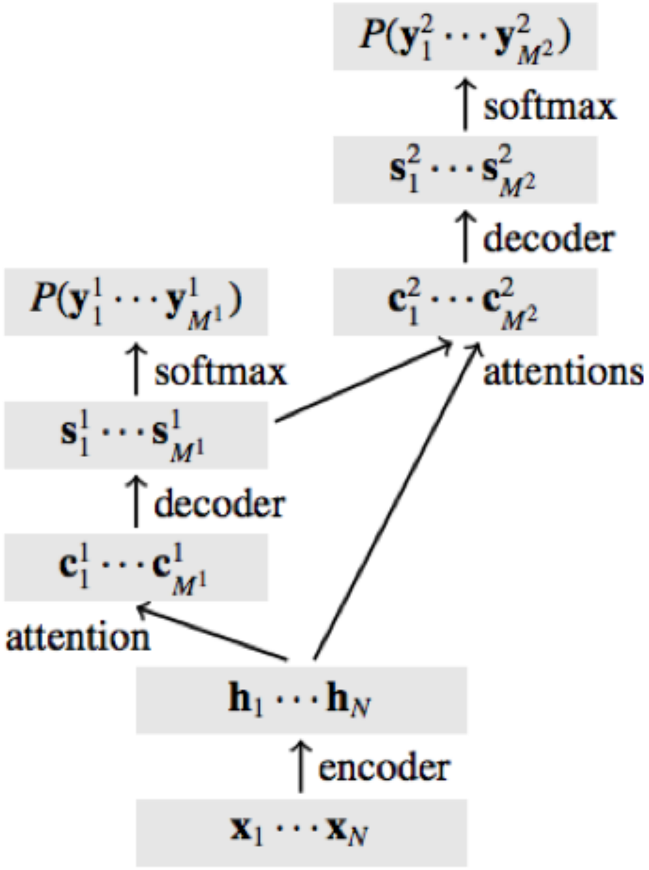
triangle

Objective Function

$$\text{score}(\mathbf{Y}^1, \mathbf{Y}^2 \mid \mathbf{X}; \theta) = \lambda \log P(\mathbf{Y}^1 \mid \mathbf{X}; \theta) + (1 - \lambda) \log P(\mathbf{Y}^2 \mid \mathbf{X}, \mathbf{S}^1; \theta)$$

$$\mathcal{L}(\theta) = \sum \text{score}(\mathbf{Y}^1, \mathbf{Y}^2 \mid \mathbf{X}; \theta)$$

λ is a parameter that controls the importance of each sub-task


$P(\mathbf{y}_1^2 \cdots \mathbf{y}_{M^2}^2)$
softmax
$\mathbf{s}_1^2 \cdots \mathbf{s}_{M^2}^2$
decoder
$P(\mathbf{y}_1^1 \cdots \mathbf{y}_{M^1}^1)$
$\mathbf{c}_1^2 \cdots \mathbf{c}_{M^2}^2$
softmax
attentions
$\mathbf{s}_1^1 \cdots \mathbf{s}_{M^1}^1$
decoder
$\mathbf{c}_1^1 \cdots \mathbf{c}_{M^1}^1$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

Regularization

$\mathbf{A}$: the matrix of attention weights, $\mathbf{A}_{ij} = \alpha_{ij}$

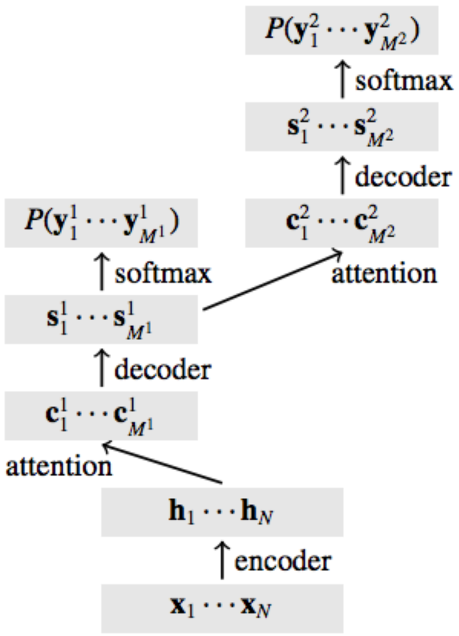
- transitivity

If source word x_i aligns to target word $y_j^1$and y_j^1 aligns to target word y_k^2, then x_i should also probably align to y_k^2.

$$\mathcal{L}_{\text{trans}} = \text{score}(\mathbf{Y}^1, \mathbf{Y}^2) - \lambda_{\text{trans}} \left\|\mathbf{A}^{12}\mathbf{A}^1 - \mathbf{A}^2\right\|_2^2$$

- invertibility

$$\mathcal{L}_{\text{inv}} = \text{score}(\mathbf{Y}^1, \mathbf{Y}^2) - \lambda_{\text{inv}} \left\|\mathbf{A}^1\mathbf{A}^{12} - \mathbf{I}\right\|_2^2$$


$P(\mathbf{y}_1^2 \cdots \mathbf{y}_{M^2}^2)$
softmax
$\mathbf{s}_1^2 \cdots \mathbf{s}_{M^2}^2$
decoder
$P(\mathbf{y}_1^1 \cdots \mathbf{y}_{M^1}^1)$
$\mathbf{c}_1^2 \cdots \mathbf{c}_{M^2}^2$
softmax
attention
$\mathbf{s}_1^1 \cdots \mathbf{s}_{M^1}^1$
decoder
$\mathbf{c}_1^1 \cdots \mathbf{c}_{M^1}^1$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

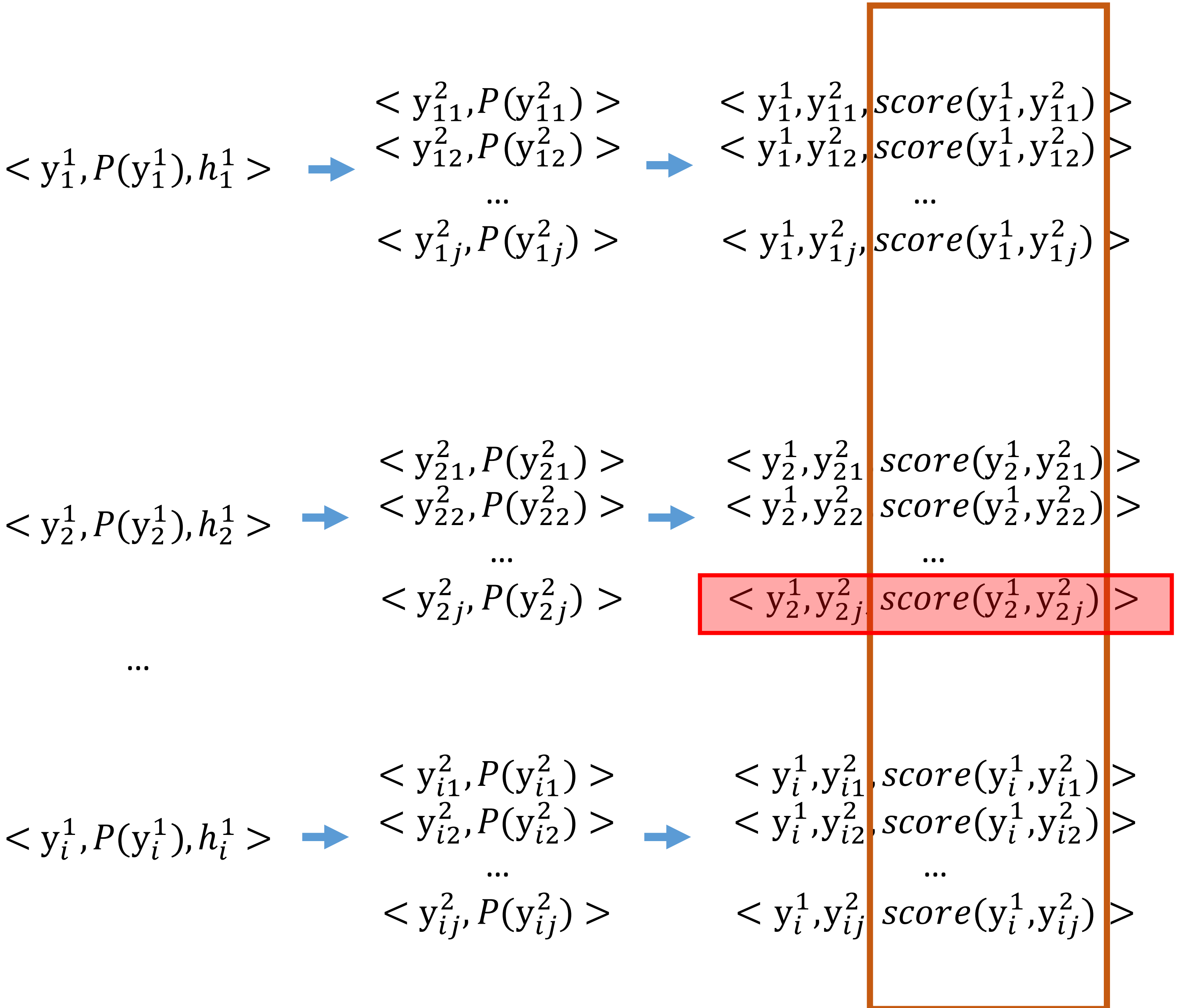
Decoding

- two-phase beam search

1. The first decoder produces a set of triplets consisting of a candidate transcription Y^1, a score $P(Y^1)$ and a hidden state H^1.

2. For each transcription candidate from the first decoder, the second decoder now produces through beam search a set of candidate translations Y^2, each with a score $P(Y^2)$.

3. We then output the combination that yields the highest total $score(Y^1, Y^2)$.



$< y_1^1, P(y_1^1), h_1^1 >$ → $< y_{11}^2, P(y_{11}^2) >$, $< y_{12}^2, P(y_{12}^2) >$, …, $< y_{1j}^2, P(y_{1j}^2) >$ → $< y_1^1, y_{11}^2, score(y_1^1, y_{11}^2) >$, $< y_1^1, y_{12}^2, score(y_1^1, y_{12}^2) >$, …, $< y_1^1, y_{1j}^2, score(y_1^1, y_{1j}^2) >$

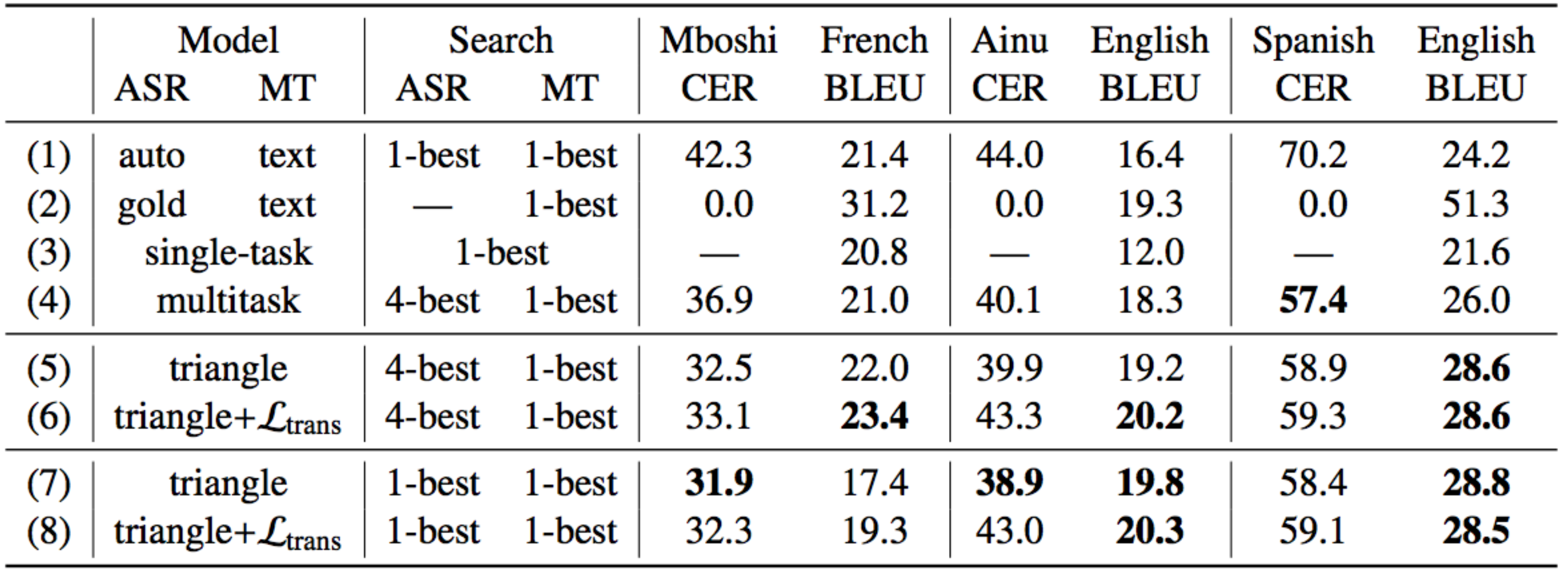
$< y_2^1, P(y_2^1), h_2^1 >$ → $< y_{21}^2, P(y_{21}^2) >$, $< y_{22}^2, P(y_{22}^2) >$, …, $< y_{2j}^2, P(y_{2j}^2) >$ → $< y_2^1, y_{21}^2, score(y_2^1, y_{21}^2) >$, $< y_2^1, y_{22}^2, score(y_2^1, y_{22}^2) >$, …, $< y_2^1, y_{2j}^2, score(y_2^1, y_{2j}^2) >$

…

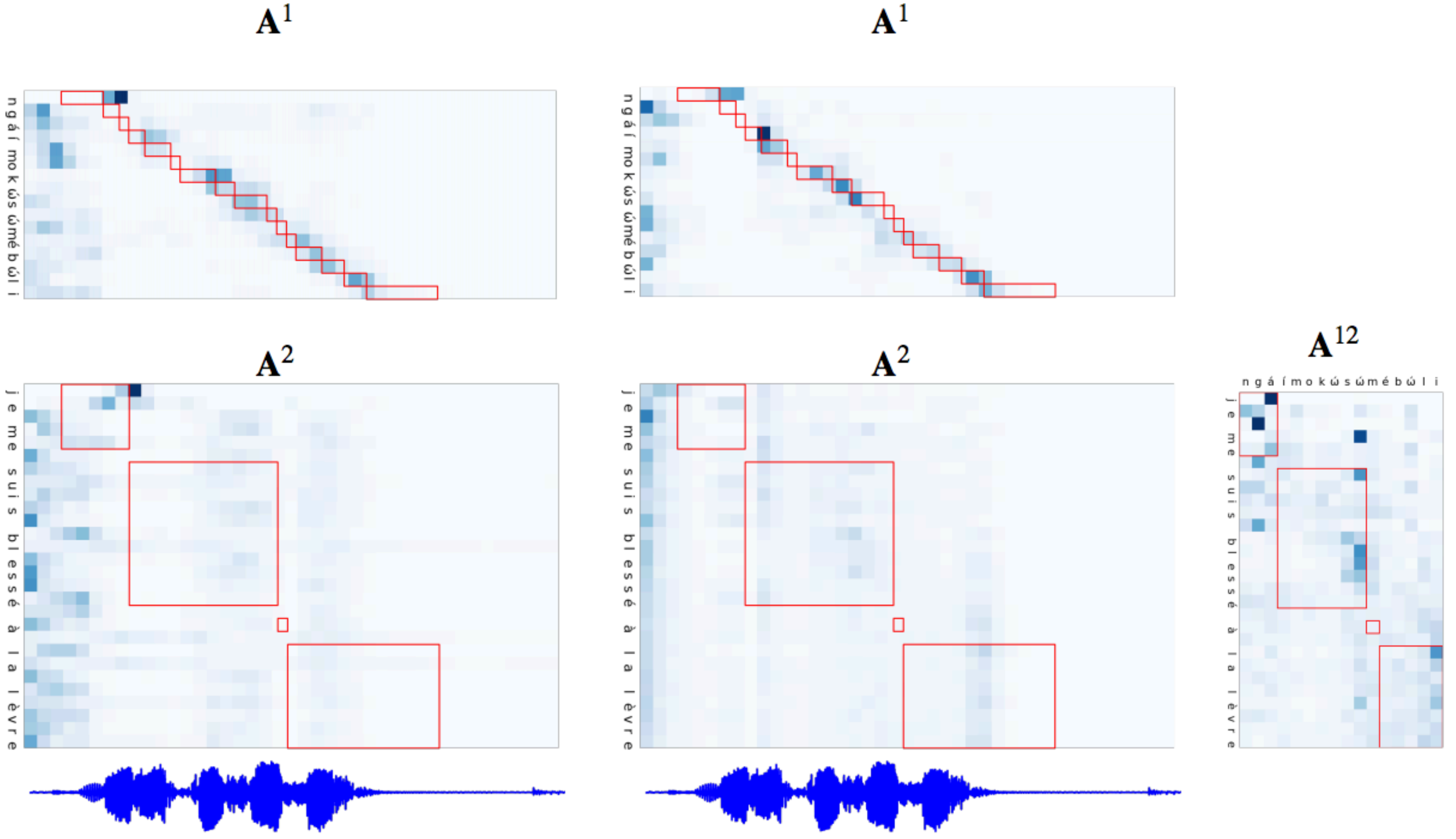
$< y_i^1, P(y_i^1), h_i^1 >$ → $< y_{i1}^2, P(y_{i1}^2) >$, $< y_{i2}^2, P(y_{i2}^2) >$, …, $< y_{ij}^2, P(y_{ij}^2) >$ → $< y_i^1, y_{i1}^2, score(y_i^1, y_{i1}^2) >$, $< y_i^1, y_{i2}^2, score(y_i^1, y_{i2}^2) >$, …, $< y_i^1, y_{ij}^2, score(y_i^1, y_{ij}^2) >$

Experiments

- Speech Transcription and Translation

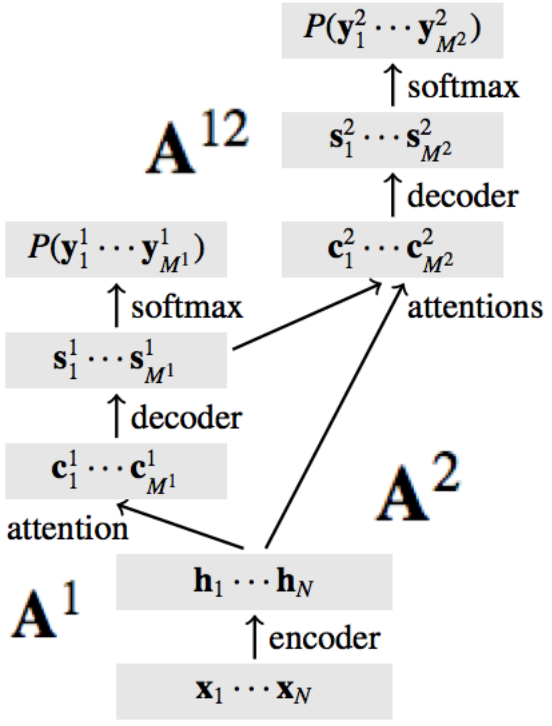


	Model		Search		Mboshi	French	Ainu	English	Spanish	English
	ASR	MT	ASR	MT	CER	BLEU	CER	BLEU	CER	BLEU
(1)	auto	text	1-best	1-best	42.3	21.4	44.0	16.4	70.2	24.2
(2)	gold	text	—	1-best	0.0	31.2	0.0	19.3	0.0	51.3
(3)	single-task		1-best		—	20.8	—	12.0	—	21.6
(4)	multitask		4-best	1-best	36.9	21.0	40.1	18.3	**57.4**	26.0
(5)	triangle		4-best	1-best	32.5	22.0	39.9	19.2	58.9	**28.6**
(6)	triangle+$\mathcal{L}_{\text{trans}}$		4-best	1-best	33.1	**23.4**	43.3	**20.2**	59.3	**28.6**
(7)	triangle		1-best	1-best	**31.9**	17.4	**38.9**	**19.8**	58.4	**28.8**
(8)	triangle+$\mathcal{L}_{\text{trans}}$		1-best	1-best	32.3	19.3	43.0	**20.3**	59.1	**28.5**



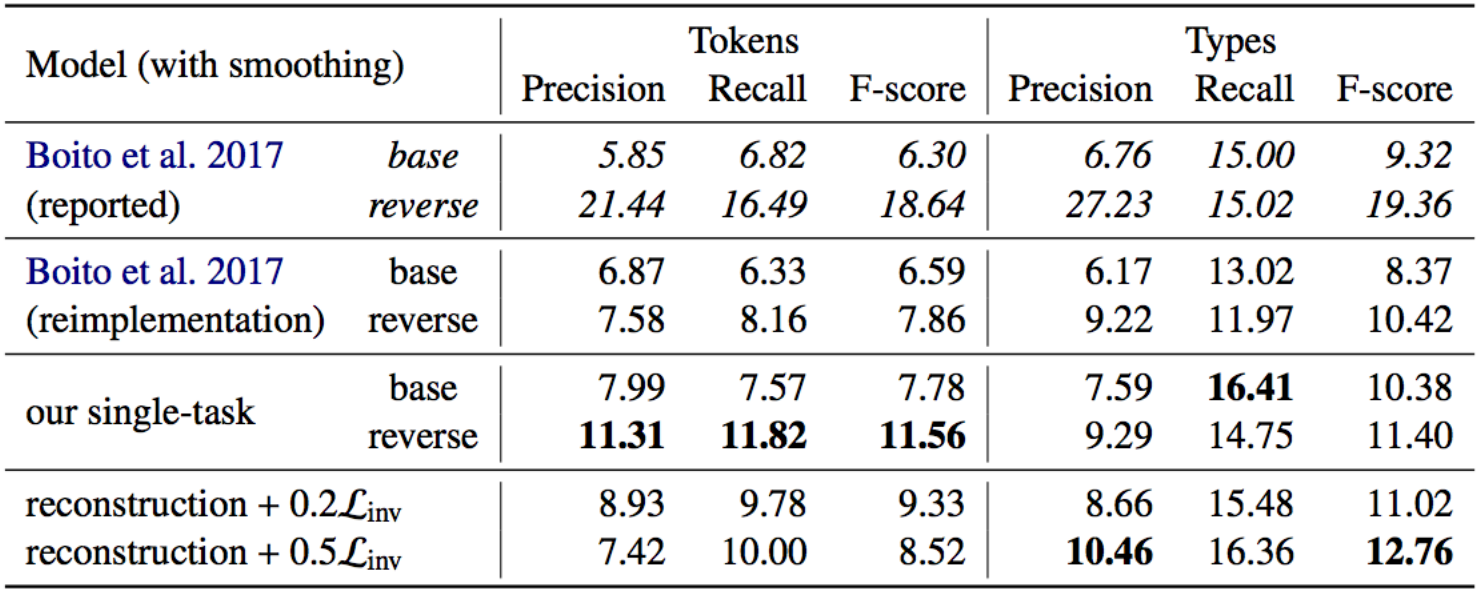
(a) multitask

(b) triangle + transitivity


$P(\mathbf{y}^2_1 \cdots \mathbf{y}^2_{M^2})$
softmax
$\mathbf{A}^{12}$
$\mathbf{s}^2_1 \cdots \mathbf{s}^2_{M^2}$
decoder
$P(\mathbf{y}^1_1 \cdots \mathbf{y}^1_{M^1})$
$\mathbf{c}^2_1 \cdots \mathbf{c}^2_{M^2}$
softmax
attentions
$\mathbf{s}^1_1 \cdots \mathbf{s}^1_{M^1}$
decoder
$\mathbf{c}^1_1 \cdots \mathbf{c}^1_{M^1}$
$\mathbf{A}^2$
attention
$\mathbf{A}^1$
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

Experiments

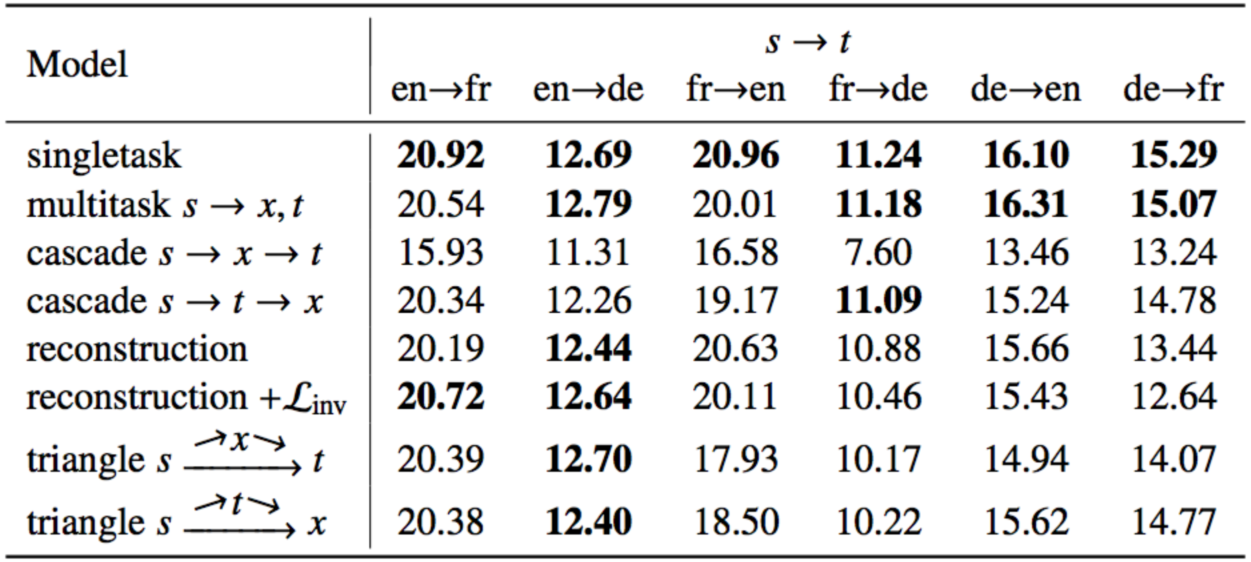
- Word Discovery



Model (with smoothing)		Tokens			Types		
		Precision	Recall	F-score	Precision	Recall	F-score
Boito et al. 2017 (reported)	*base*	*5.85*	*6.82*	*6.30*	*6.76*	*15.00*	*9.32*
	reverse	*21.44*	*16.49*	*18.64*	*27.23*	*15.02*	*19.36*
Boito et al. 2017 (reimplementation)	base	6.87	6.33	6.59	6.17	13.02	8.37
	reverse	7.58	8.16	7.86	9.22	11.97	10.42
our single-task	base	7.99	7.57	7.78	7.59	**16.41**	10.38
	reverse	**11.31**	**11.82**	**11.56**	9.29	14.75	11.40
reconstruction + $0.2\mathcal{L}_{\text{inv}}$		8.93	9.78	9.33	8.66	15.48	11.02
reconstruction + $0.5\mathcal{L}_{\text{inv}}$		7.42	10.00	8.52	**10.46**	16.36	**12.76**

Experiments

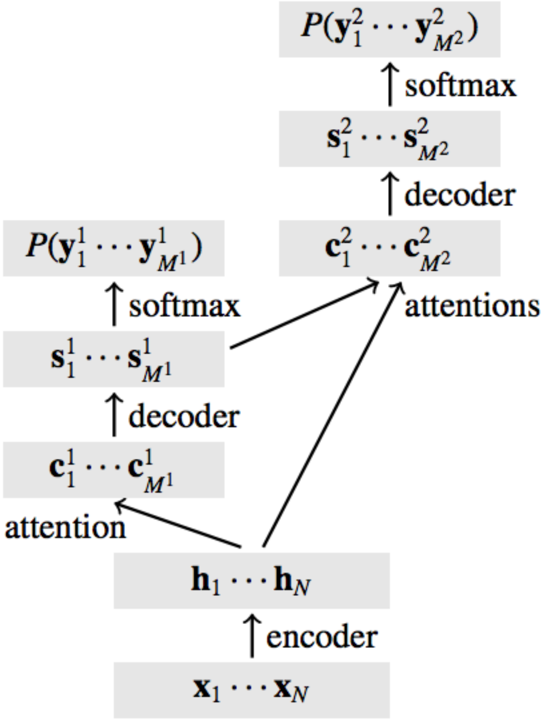
- Negative Results: High-Resource Text Translation
- in the case of text translation between so linguistically close languages, the lower level representations (the output of the encoder) provide as much information as the higher level ones, without the search errors that are introduced during inference.



Model	$s \rightarrow t$					
	en→fr	en→de	fr→en	fr→de	de→en	de→fr
singletask	**20.92**	**12.69**	**20.96**	**11.24**	**16.10**	**15.29**
multitask $s \rightarrow x, t$	20.54	**12.79**	20.01	**11.18**	**16.31**	**15.07**
cascade $s \rightarrow x \rightarrow t$	15.93	11.31	16.58	7.60	13.46	13.24
cascade $s \rightarrow t \rightarrow x$	20.34	12.26	19.17	**11.09**	15.24	14.78
reconstruction	20.19	**12.44**	20.63	10.88	15.66	13.44
reconstruction $+\mathcal{L}_{\text{inv}}$	**20.72**	**12.64**	20.11	10.46	15.43	12.64
triangle $s \xrightarrow{\nearrow x \searrow} t$	20.39	**12.70**	17.93	10.17	14.94	14.07
triangle $s \xrightarrow{\nearrow t \searrow} x$	20.38	**12.40**	18.50	10.22	15.62	14.77

Merits

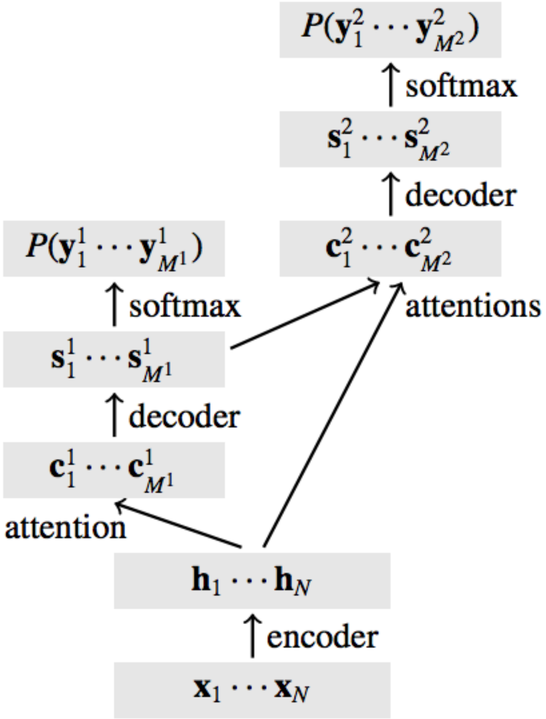
- General Framework
- Transitivity and invertibility attention regularizer

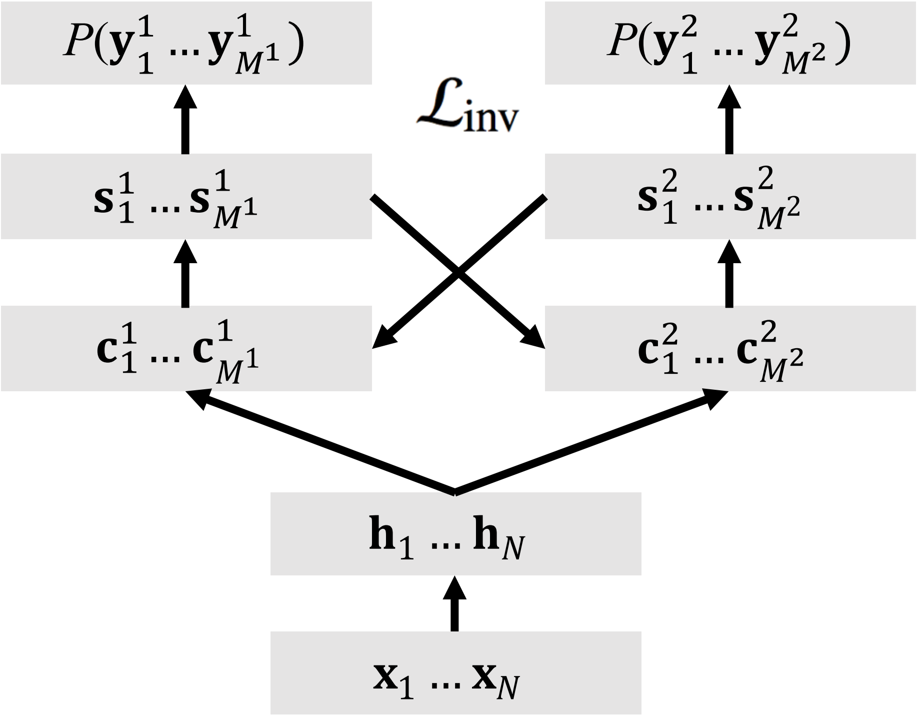


$P(\mathbf{y}_1^2 \cdots \mathbf{y}_{M^2}^2)$
softmax
$\mathbf{s}_1^2 \cdots \mathbf{s}_{M^2}^2$
decoder
$\mathbf{c}_1^2 \cdots \mathbf{c}_{M^2}^2$
attentions
$P(\mathbf{y}_1^1 \cdots \mathbf{y}_{M^1}^1)$
softmax
$\mathbf{s}_1^1 \cdots \mathbf{s}_{M^1}^1$
decoder
$\mathbf{c}_1^1 \cdots \mathbf{c}_{M^1}^1$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

Limitation

- imbalanced structure


$P(\mathbf{y}_1^2 \cdots \mathbf{y}_{M^2}^2)$
softmax
$\mathbf{s}_1^2 \cdots \mathbf{s}_{M^2}^2$
decoder
$\mathbf{c}_1^2 \cdots \mathbf{c}_{M^2}^2$
attentions
$P(\mathbf{y}_1^1 \cdots \mathbf{y}_{M^1}^1)$
softmax
$\mathbf{s}_1^1 \cdots \mathbf{s}_{M^1}^1$
decoder
$\mathbf{c}_1^1 \cdots \mathbf{c}_{M^1}^1$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$


$P(\mathbf{y}_1^1 \dots \mathbf{y}_{M^1}^1)$
$\mathcal{L}_{\text{inv}}$
$P(\mathbf{y}_1^2 \dots \mathbf{y}_{M^2}^2)$
$\mathbf{s}_1^1 \dots \mathbf{s}_{M^1}^1$
$\mathbf{s}_1^2 \dots \mathbf{s}_{M^2}^2$
$\mathbf{c}_1^1 \dots \mathbf{c}_{M^1}^1$
$\mathbf{c}_1^2 \dots \mathbf{c}_{M^2}^2$
$\mathbf{h}_1 \dots \mathbf{h}_N$
$\mathbf{x}_1 \dots \mathbf{x}_N$