A Systematic Exploration of Diversity in Machine Translation

Kevin Gimpel
Toyota Technological Institute at Chicago

Dhruv Batra
Virginia Tech

Chris Dyer
Carnegie Mellon University

Gregory Shakhnarovich
Toyota Technological Institute at Chicago

Motivation

In structured prediction, M-best lists are not diverse

Packed representations partially address this, but add complexity

Can we inject **diversity** into M-best lists for machine translation?

"They say we're not placing enough emphasis on diversity."

Generating diverse translations

predicted translation ← $\hat{\mathbf{y}}, \hat{\mathbf{h}}$; parameters ← $\mathbf{w}$; feature vector ← $\boldsymbol{\phi}(x, \mathbf{y}, \mathbf{h})$

$$\langle \hat{\mathbf{y}}, \hat{\mathbf{h}} \rangle = \operatorname*{argmax}_{\langle \mathbf{y}, \mathbf{h} \rangle \in \mathcal{T}_x} \mathbf{w}^\top \boldsymbol{\phi}(x, \mathbf{y}, \mathbf{h})$$

latent derivation ← $\langle \mathbf{y}, \mathbf{h} \rangle$; $\mathcal{T}_x$ ← Input sentence

We use the framework of Batra et al. (2012):

best: $\langle \mathbf{y}^1, \mathbf{h}^1 \rangle = \operatorname*{argmax}_{\langle \mathbf{y}, \mathbf{h} \rangle \in \mathcal{T}_x} \mathbf{w}^\top \boldsymbol{\phi}(x, \mathbf{y}, \mathbf{h})$

diverse 2nd-best: $\langle \mathbf{y}^2, \mathbf{h}^2 \rangle = \operatorname*{argmax}_{\langle \mathbf{y}, \mathbf{h} \rangle \in \mathcal{T}_x} \mathbf{w}^\top \boldsymbol{\phi}(x, \mathbf{y}, \mathbf{h}) + \lambda_1 \Delta(\mathbf{y}^1, \mathbf{y})$

…

diverse *m*th-best: $\langle \mathbf{y}^m, \mathbf{h}^m \rangle = \operatorname*{argmax}_{\langle \mathbf{y}, \mathbf{h} \rangle \in \mathcal{T}_x} \mathbf{w}^\top \boldsymbol{\phi}(x, \mathbf{y}, \mathbf{h}) + \sum_{j=1}^{m-1} \lambda_j \Delta(\mathbf{y}^j, \mathbf{y})$

"dissimilarity weight" ← λ_1 ; "dissimilarity function" ← Δ ; compute dissimilarity to all previous translations ← $\sum_{j=1}^{m-1}$

Following Batra et al., we set $\lambda = \lambda_j, \forall j$

λ is tuned via grid search for each task

Dissimilarity functions for MT

Count all *n*-gram matches in the two translations:

$$\Delta_n(\mathbf{y}, \mathbf{y}') = -\sum_{i=1}^{|\mathbf{y}|-q} \sum_{j=1}^{|\mathbf{y}'|-q} [\![\mathbf{y}_{i:i+q} = \mathbf{y}'_{j:j+q}]\!]$$

$q = n - 1$; Iverson bracket ← $[\![\cdot]\!]$

λ and *n* are tuned to maximize oracle BLEU of diverse lists

Can be Implemented as another language model; no change to decoder is required

Experimental setup

language pair	model	# sentence pairs
Arabic → English	phrase-based	4.3M
Chinese → English	hierarchical phrase-based	300K
German → English	phrase-based	1.9M

Moses used for all language pairs

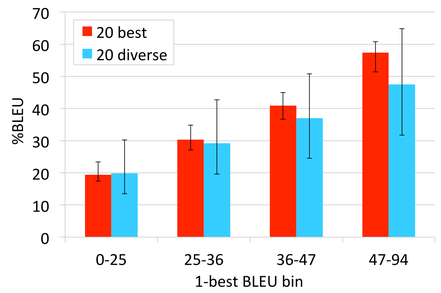
5-gram LMs with an extra 600M tokens from Gigaword

Examples of diverse translations

reference:	The government wants to limit the torture of "witches," a brochure was released
1-best translation:	The government wants the torture of 'witch' and gave out a booklet
unique 5-best list:	The government wants the torture of "witch" and gave out a booklet The government wants the torture of 'witch' and gave out a brochure The government wants the torture of 'witch' and gave out a leaflet The government wants the torture of "witch" and gave out a brochure
diverse list:	The government wants to stop torture of "witch" and issued a leaflet issued The government wants to "stop the torture of" witches and gave out a brochure The government intends to the torture of "witchcraft" and were issued a leaflet The government is the torture of "witches" stamp out and gave a brochure

references:	Kanazawa indicated as Akihito did not bleed much during the operation, the doctors used his own blood for transfusion. Kanazawa said Akihito did not lose a lot of blood during the surgery. Actually, doctors had only used the blood drawn from him for transfusions.
1-best:	Kanazawa College said Akihito bleeding during surgery, doctors, not only in his own blood transfusion.
unique 5-best list:	Kanazawa College said Akihito bleeding during surgery, doctors, not only with his own blood transfusion. Kanazawa College, said Akihito bleeding during surgery, doctors, not only with his own blood transfusion. Kanazawa College said that Akihito bleeding during surgery, doctors, not only in his own blood transfusion. Kanazawa College said that Akihito bleeding during surgery, doctors, not only with his own blood transfusion.
diverse list:	Kanazawa College, said Akihito did not hemorrhage during surgery, the doctor only used his own blood to blood transfusions. Kanazawa College when Akihito hemorrhage, doctors said, but with his own blood transfusions of blood. When Akihito Kanazawa College said that surgical bleeding, doctors only his own blood for transfusion. Kanazawa College that Akihito did not bleeding during surgery, blood transfusion and doctors with his own.

Analysis of diverse lists



Bars show median BLEU+1 on 20-best and 20-diverse lists, error bars show min and max

Stats are averages across sentences in BLEU quartiles chosen based on BLEU+1 of 1-best translation

Ranges of diverse lists always subsume those of M-best lists, but median drops for high BLEU

Plot shows Chinese → English; other language pairs look similar

System combination experiments

Pass diverse translations to system combination framework of Heafield and Lavie (2010)

Baselines are standard M-best lists and unique M-best lists

Dissimilarity parameters for diverse translations tuned to maximize oracle BLEU at M, for each M

		Arabic → English			Chinese → English			German → English		
		M=10	15	20	10	15	20	10	15	20
	baseline		50.1			36.9			21.8	
system combination	M-best	50.2	50.1	50.0	36.7	36.9	37.0	21.7	21.7	21.8
	unique M-best	50.6	50.0	50.8	37.1	36.9	37.1	21.8	**21.9**	**21.9**
	M-diverse	**51.4**	**51.2**	**51.2**	**37.6**	**37.6**	**37.5**	**22.0**	21.8	21.6

Breakdown by BLEU quartile (M = 15):

system combination (and diversity) work best for low-BLEU translations

		Arabic → English				Chinese → English				German → English			
		0-39	39-49	49-61	61+	0-25	25-36	36-47	47+	0-14.5	14.5-21.1	21.1-30.3	30.3+
	baseline	30.1	44.1	55.1	**70.0**	15.2	28.9	41.0	**57.5**	5.3	14.4	23.7	**40.9**
system combination	M-best	30.1	44.6	55.5	68.8	15.9	29.2	40.5	56.8	6.0	15.0	23.6	40.0
	unique M-best	30.4	44.7	55.2	68.4	16.7	29.0	41.2	56.6	5.9	14.9	**23.8**	40.6
	M-diverse	**31.3**	**45.3**	**57.8**	69.1	**17.7**	**30.6**	**41.7**	56.9	**7.6**	**15.2**	23.4	39.6

Reranking experiments

Compared M-best lists to diverse lists in discriminative reranking

Used structured SVM with slack rescaling for training the reranker (Yadollahpour et al., 2013), which worked better than MERT, PRO, and Rampion

Features included inverse IBM Model 1 , syntactic LM of Pauls and Klein (2012), finite/non-finite verbs, discriminative word/POS tag LMs, Google 5-grams, and 7-gram Brown cluster LMs

Comparing list types for Arabic → English:

	features	
list type	**none**	**all**
20 best	50.3	50.6
100 best	50.6	50.8
200 best	50.4	51.2
1000 best	50.5	51.2
unique 20 best	50.5	51.2
unique 100 best	50.6	51.2
unique 200 best	50.4	51.3
diverse 20	50.5	51.1
diverse 20 x 5 best	50.6	51.4
diverse 20 x 10 best	**50.7**	51.3
diverse 20 x 50 best	**50.7**	**51.8**

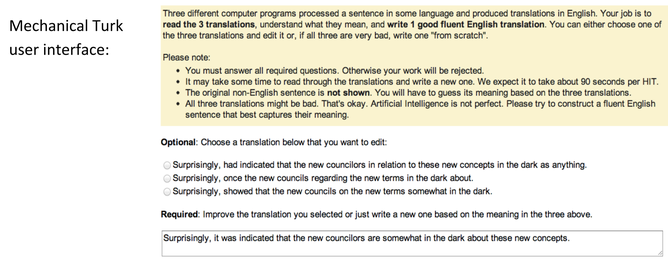
Comparing list types for training/testing:

test \ train	best	diverse
best	51.2	51.7
diverse	50.5	51.8

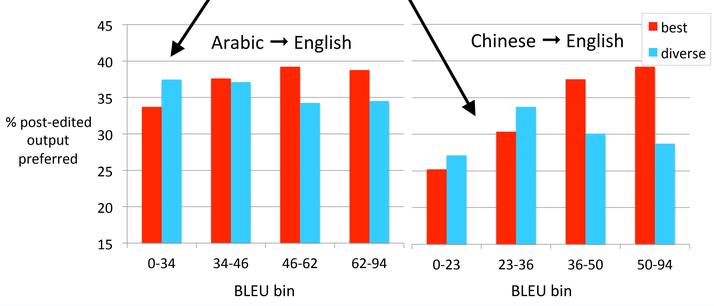
diversity helps during training, even when using M-best lists for testing

Human post-editing experiments

Mechanical Turk user interface:



diversity improves understanding for difficult sentences



References

D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. 2012. Diverse M-best solutions in Markov random fields. In *Proc. of ECCV*.

K. Heafield and A. Lavie. 2010. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93.

A. Pauls and D. Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proc. of ACL*.

P. Yadollahpour, D. Batra, and G. Shakhnarovich. 2013. Discriminative re-ranking of diverse segmentations. In *Proc. of CVPR*.