Analyzing Optimization for Statistical Machine Translation MERT Learns Verbosity, PRO Learns Length معهد قطر لبحوث الحوسبة

Francisco Guzmán, Preslav Nakov and Stephan Vogel Qatar Computing Research Institute, HBKU

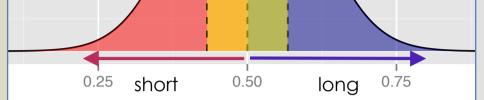
1. Introduction

- MT performance depends on the **tuning set** (Zheng et al., 2010, Liang et al 2010)
- Optimization can be improved by selecting a suitable tuning set.
- PRO has *issues* with **length**:
- generates shorter translations (Nakov et al., 2012)
- is susceptible to produce pathologically long translations (Nakov et al., 2013)

3. Set	PRO			
	Arabic-English	Spanish-	English	VS
Datasets	NIST 04, 05, 06, 09	WMT 08, C	9,10,11	MERT
References	multiple, single	single		medium
Partitions	•	Length: short, mid, long Verbosity: low, mid, high		

Optimizers

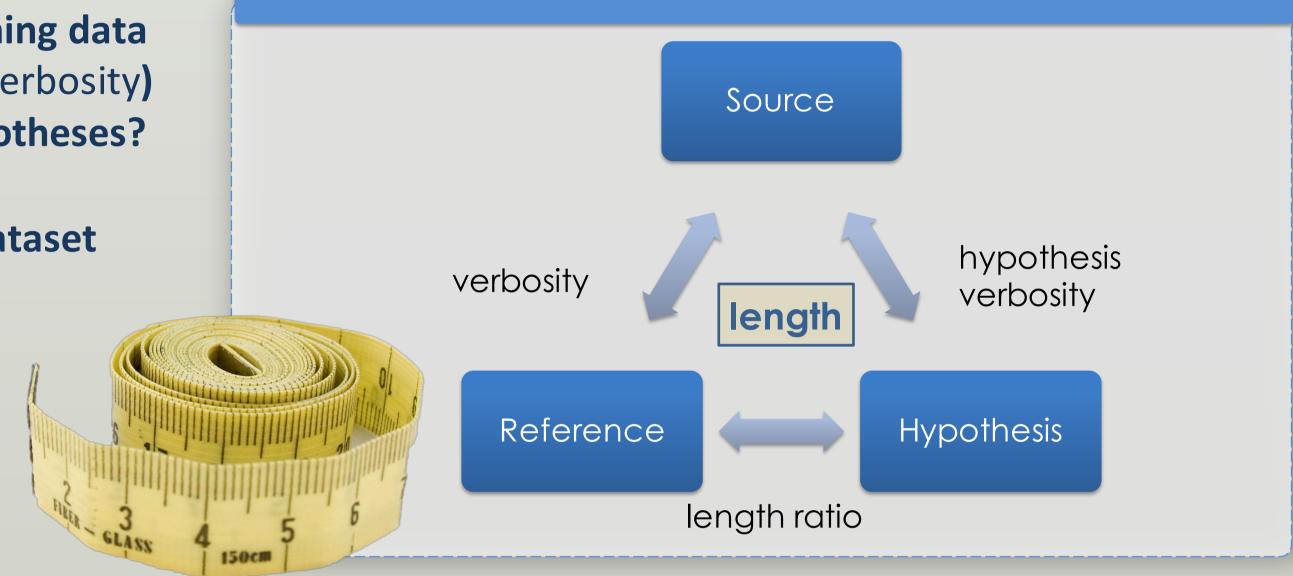
PRO & MERT



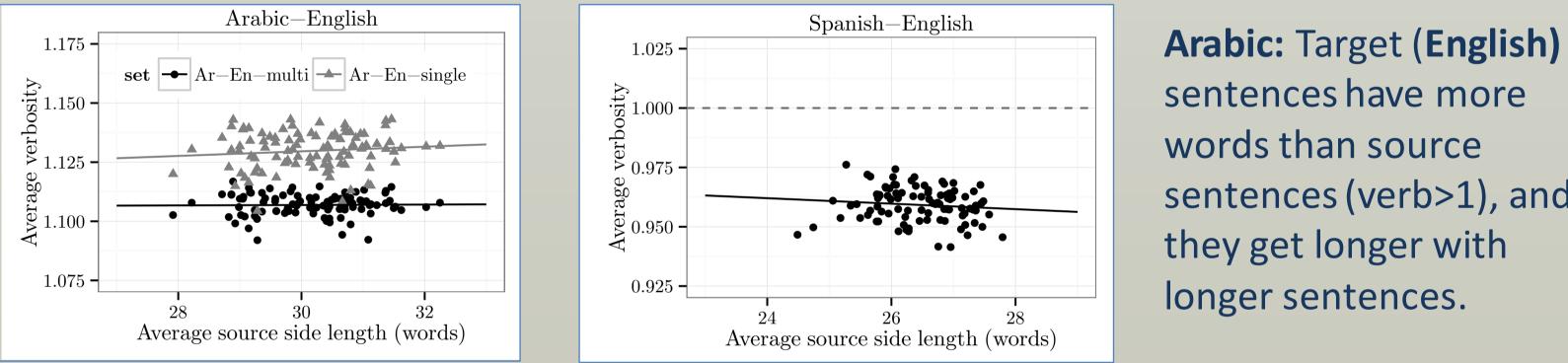
2. Analyzing Length for SMT Optimization

The effect of tuning data (source length, verbosity) on the *final* hypotheses?

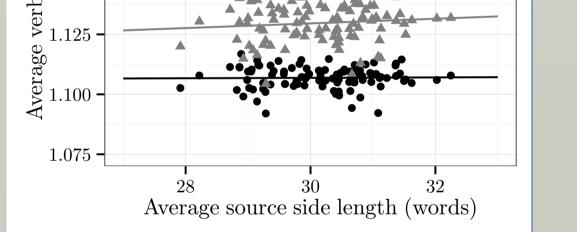
What type of dataset works best?



Tuning set verbosity depends on source length



4. Results: Length MERT PRO **Correlation:** Tuning 98% verbosity vs. hypothesis verbosity **Correlation:** Tuning **Correlation:** Tuning verbosity vs. source length vs. length hypothesis verbosity ratio MERT-tuning verbosity vs. test set length ratio Ar–En–multi Ar-En-single Es-En sity verbo 1.2 hypothesis 11 1:5 - 0.1 set



words than source sentences (verb>1), and they get longer with longer sentences.

Spanish: Target sentences have fewer words. They get shorter with longer source sentences.

verbosity tuning sets!

Avoid low-verbosity

Prefer verbosity as a

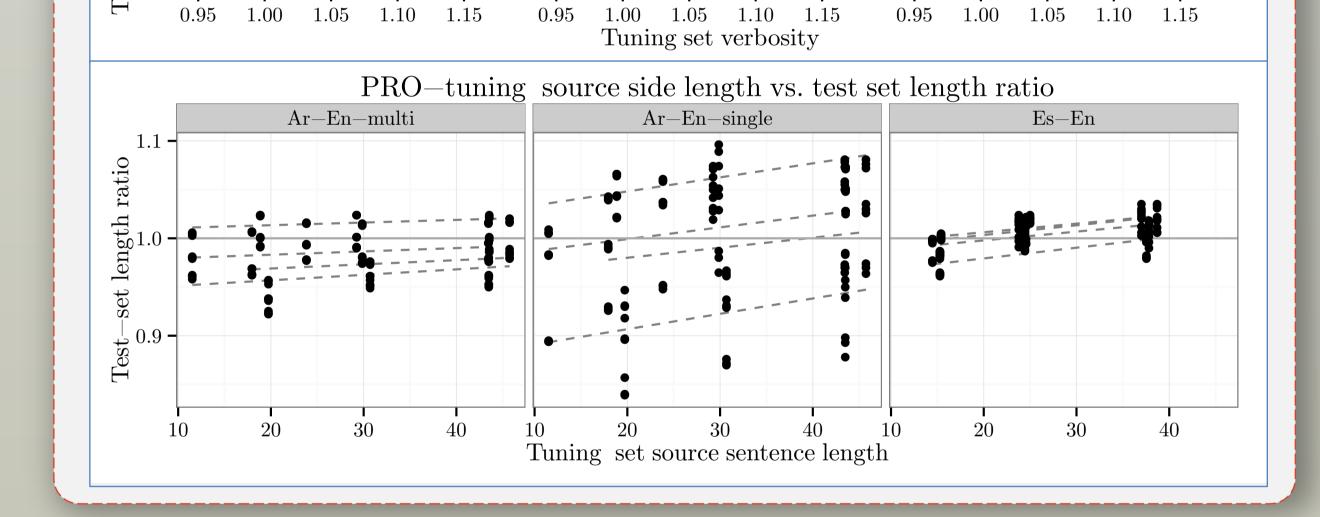
selection criteria

sets

Figure 1: Average source sentence length (x axis) vs. average verbosity (y axis) for 100 random samples, each with 500 sentence pairs extracted from NIST (Left: Arabic-English, multi- and single-reference) and from WMT (Right: Spanish-English, single-reference) data.

What we discovered

Source length <-> verbosity We can use source length to control verbosity



5. Results: BLEU

	Arabic-English (1-ref)				
tuning	short	mid	long		
MERT					
short	26.69 *	28.14	27.49		
mid	26.22	28.39 *	27.96		
long	25.80	28.20	28.27*		
PRO-fix	25.05	27.((07.00		
short	25.95	27.66	27.28		
mid	25.98	28.23	28.19		
long	25.87	28.11	28.05		

Cross-testing (BLEU)

MERT: best BLEU when cuning on similar-to-test

PRO: learned parameters re independent of testset length.

6. Choosing a Tuning Set: verbosity or length?

	test				MERT: Choose long
•	Arabic-English (multi-ref)	Arabic-English (1-ref)	WMT Spanish-English		source tuning sets!
tuning	MERT PRO-fix	MERT PRO-fix	MERT PRO-fix		PRO: Choose high-

length							
short	48.71	49.12	26.74	27.35	26.79	27.07	
mid	49.27	49.59	26.97	27.23	26.99	26.88	
long	49.35	49.20	27.23	27.28	27.02	26.84	
verbosity							
low-verb	47.90	47.60	25.89	25.88	26.70	26.61	
mid-verb	49.16	49.52	27.69	27.95	27.09	26.81	•
high-verb	50.28	50.79 *	27.36	28.03 *	27.01	27.38*	
e							

	PRO	MERT		
Likes	Length	Verbosity	30.0 BLEU BP	
Best strategy	High verbosity tuning set	Mixed. High verbosity tuning set	29.5- DETER	
Worst strategy	Select low verbosity tuning	Select low verbosity tuning sets		
			28.5 all top90 top80 top70 top60 top50	

7. Conclusion

- Know your tuning datasets: Different language pairs and translation directions may have different *source-side length* - verbosity dependencies.
- When optimizing with PRO: select or construct a highverbosity dataset as this could potentially compensate for PROs tendency to yield too short translations.
- When optimizing with MERT: If you know beforehand the test set, select the *closest* tuning set. Otherwise, tune on longer sentences.