Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation

Pavel Pecina[1], Antonio Toral[2], Josef van Genabith[2]

[1]Charles University in Prague, Czech Republic
[2]Dublin City University, Ireland

Common industry scenario:

- A statistical MT system trained and tuned on general domain data needs to be adapted to a specific domain for which no or only very limited in-domain parallel data is available.

In this work, we show how:

1. performance of such systems drops when applied to specific domains
2. perplexity of test data correlates with translation quality
3. system parameters changes when tuned on in-domain data
4. such systems can be adapted successfully by cross/no tuning

Common industry scenario:

- A statistical MT system trained and tuned on general domain data needs to be adapted to a specific domain for which no or only very limited in-domain parallel data is available.

In this work, we show how:

1. performance of such systems drops when applied to specific domains
2. perplexity of test data correlates with translation quality
3. system parameters changes when tuned on in-domain data
4. such systems can be adapted successfully by cross/no tuning

... in context of Panacea and Khresmoi (EU FP7 projects).

Talk overview

1. Introduction
2. Baseline system
3. Measuring domain divergence
4. In-domain parameter tuning
5. Analysis of model parameters
6. Analysis of phrase-length distribution
7. Overfitting reduction
8. Conclusions

Phrase-Based Statistical Machine Translation in Moses

Log-linear model:

- based on the noisy channel model
- input sentence f split into phrases, that are translated, ev. reordered
- translation $\overline{e}$ searched for by maximizing translation probability formulated as log-linear combination of functions h_i and weights λ_i

$$\overline{e} = \arg\max_{e} p(e|f) \qquad p(e|f) = \prod_{i}^{n} h_i(e,f)^{\lambda_i}$$

Components (Moses):

1. phrase translation model ensuring phrase translation adequacy(h_9–h_{12})
2. language model ensuring translations fluency (h_8)
3. reordering model allowing phrases reordering (h_1–h_7)
4. word penalty regulating translation length (h_{14})

Features trained on training data, weights tuned by MERT on dev data.

System description

Baseline system (general-domain):

- trained on Europarl v5
- max phrase length 7; 5-gram LM

	sentences	*tokens*
English–French	1,725K	47M
English–Greek	964K	27M

Development and test sets:

1. General – WPT 2005 test and development sets from Europarl
2. Natural environment – web-crawled within the Panacea project
3. Labour legislation – web-crawled within the Panacea project
4. Medicine – extracted from the EMEA parallel corpus

English–French

	gen	*env*	*lab*	*med*
dev	2,000	1,392	1,411	1,064
test	2,000	2,000	2,000	2,000

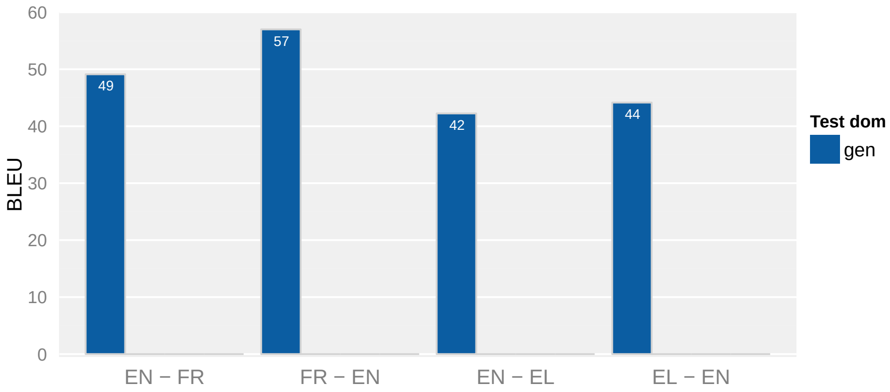
English–Greek

	gen	*env*	*lab*	*med*
dev	2,000	1,000	506	1,064
test	2,000	2,000	2,000	2,000

Baseline system performance – trained and tuned on general domain

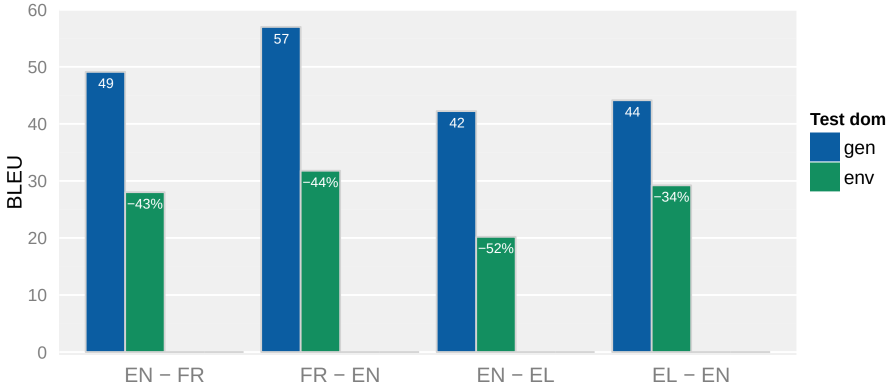
Baseline system performance – trained and tuned on general domain

- is optimal when applied to general domain



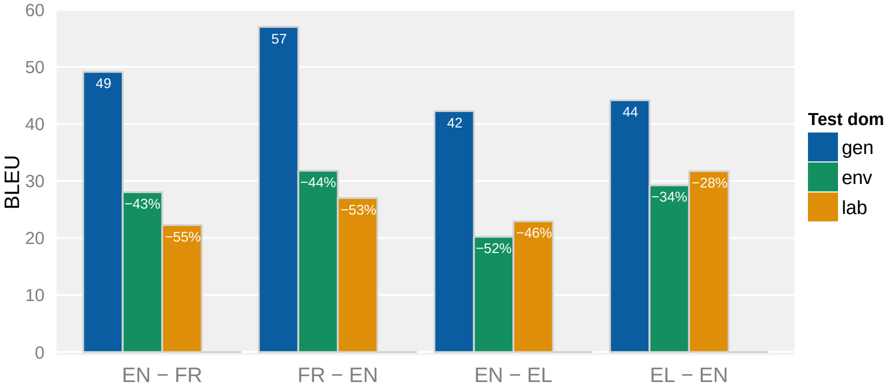
Baseline system performance – trained and tuned on general domain

- is optimal when applied to general domain
- is suboptimal when applied to specific domains



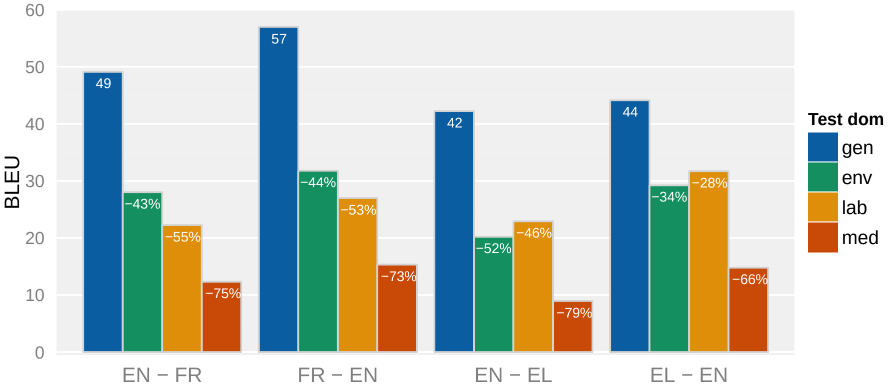
Baseline system performance – trained and tuned on general domain

- is optimal when applied to general domain
- is suboptimal when applied to specific domains



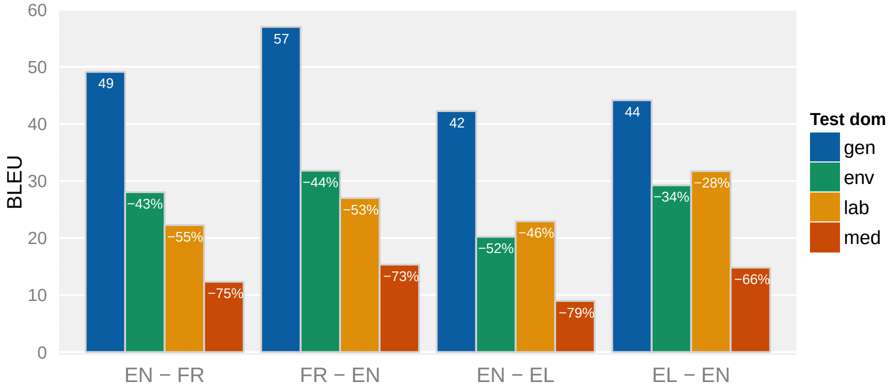
Baseline system performance – trained and tuned on general domain

- is optimal when applied to general domain
- is suboptimal when applied to specific domains



Baseline system performance – trained and tuned on general domain

- is optimal when applied to general domain
- is suboptimal when applied to specific domains



The average decrease over all translation directions/domains is 53.97%.

Domain divergence and its correlation with translation quality

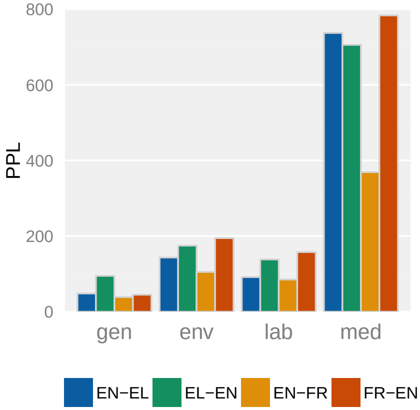
- Translation quality depends on the extent to which the test domain differs from the training domain.

Domain divergence and its correlation with translation quality

- Translation quality depends on the extent to which the test domain differs from the training domain.
- Domain divergence can be quantified by cross perplexity of the test data given the source side of training data.

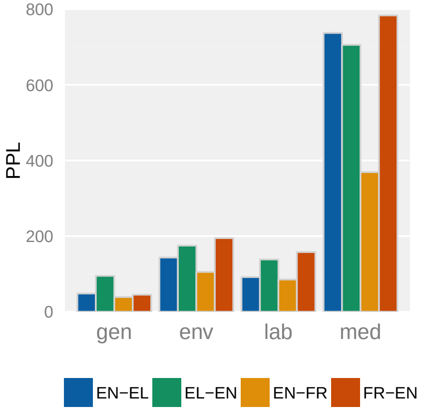
Domain divergence and its correlation with translation quality

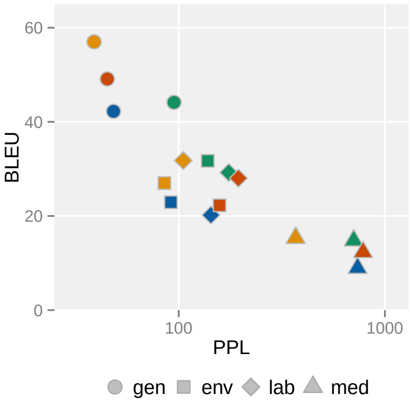
- Translation quality depends on the extent to which the test domain differs from the training domain.
- Domain divergence can be quantified by cross perplexity of the test data given the source side of training data.



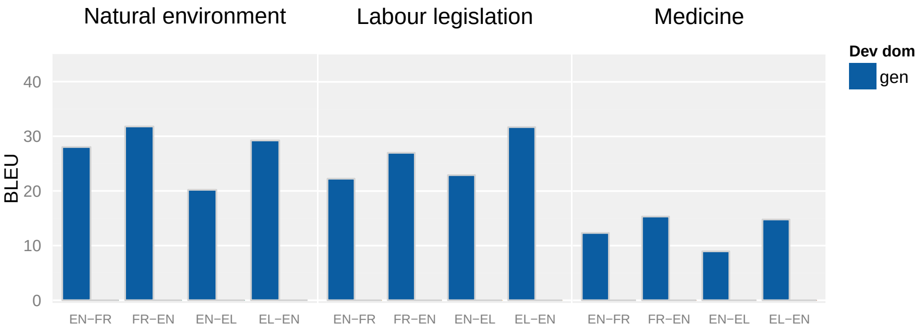
Domain divergence and its correlation with translation quality

- Translation quality depends on the extent to which the test domain differs from the training domain.
- Domain divergence can be quantified by cross perplexity of the test data given the source side of training data.

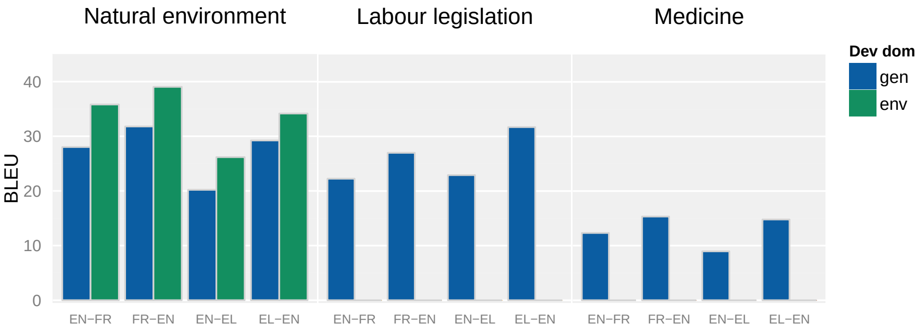




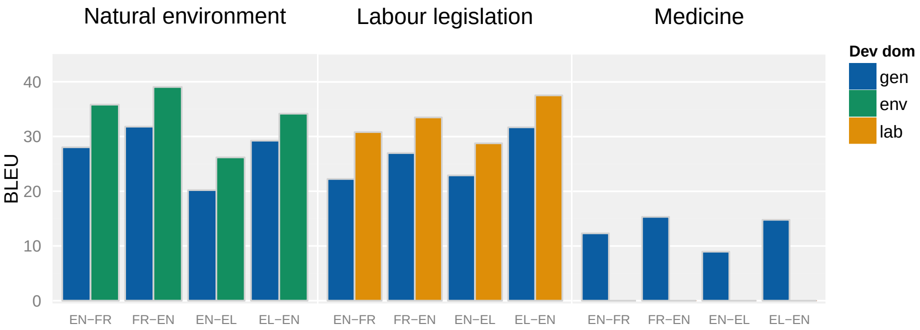
Parameter tuning on in-domain data



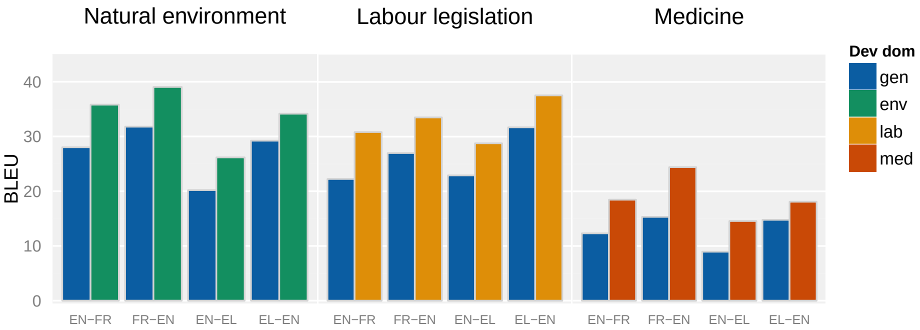
Parameter tuning on in-domain data



Parameter tuning on in-domain data

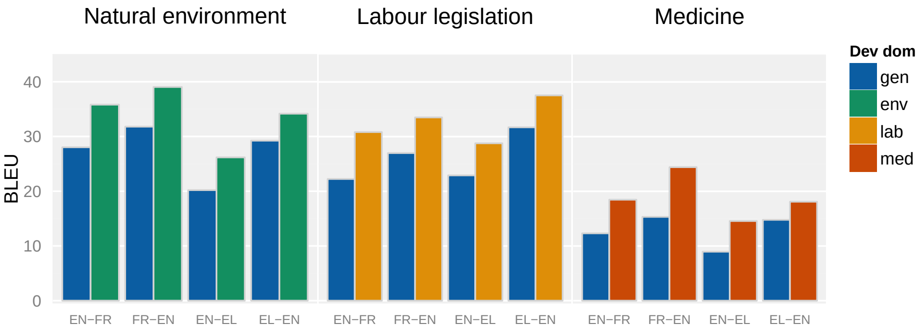


Parameter tuning on in-domain data



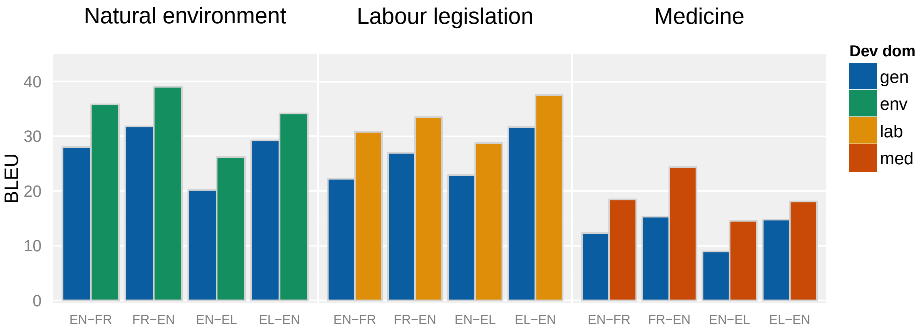
- The overall average increase of BLEU is 33.16% relative.

Parameter tuning on in-domain data



- The overall average increase of BLEU is 33.16% relative.
- Development sets contain only several hundred sentence pairs each.

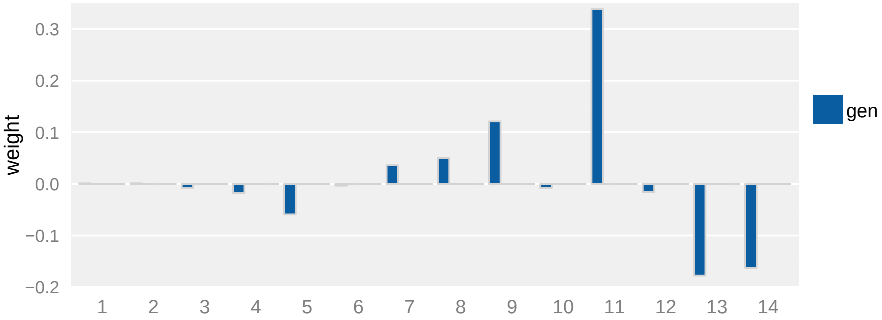
Parameter tuning on in-domain data

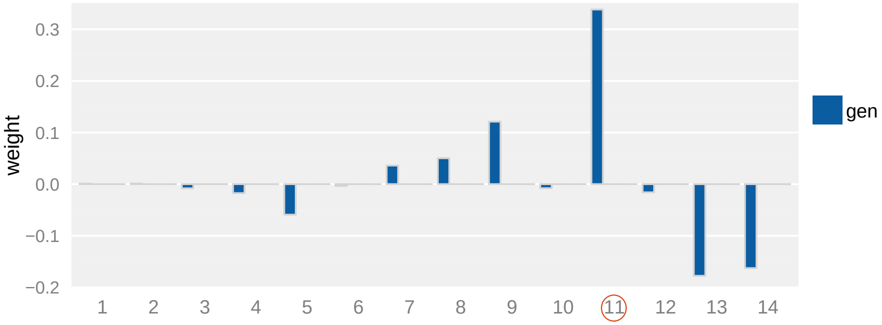


- The overall average increase of BLEU is 33.16% relative.
- Development sets contain only several hundred sentence pairs each.
- The main models remain the same, only the weigh vector changes.

Moses weights optimized on **general domain** (EN-FR)

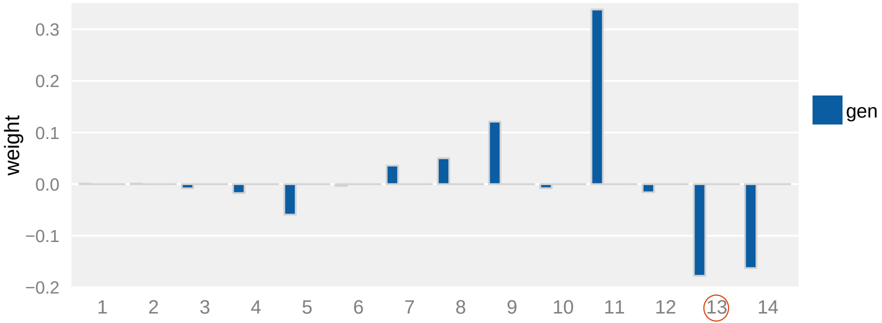
Moses weights optimized on **general domain** (EN-FR)





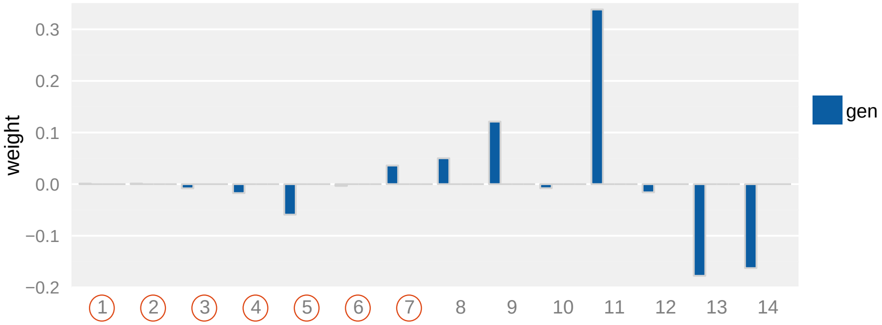
1. *Direct phrase translation probability* (h_{11}): high weight $\rightarrow$ high reward for hypotheses consisting of phrases with high translation probability

Moses weights optimized on **general domain** (EN-FR)



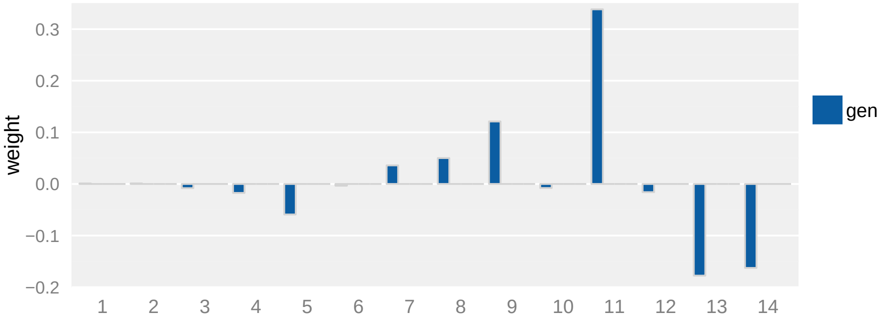
1. *Direct phrase translation probability* (h_{11}): high weight → high reward for hypotheses consisting of phrases with high translation probability
2. *Phrase penalty* (h_{13}): low negative weights → the systems prefer hypotheses consisting of fewer but longer phrases.

Moses weights optimized on **general domain** (EN-FR)

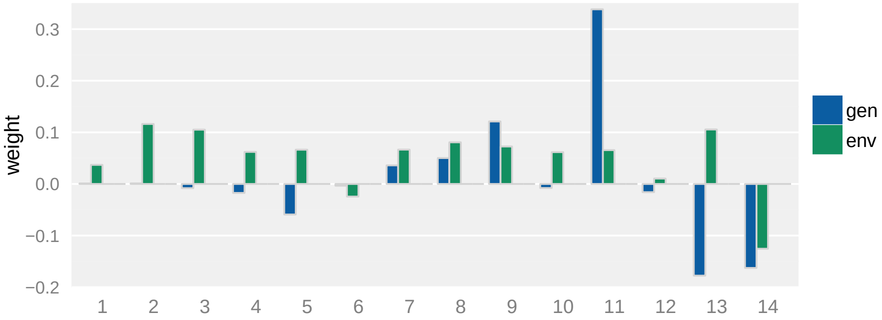


1. *Direct phrase translation probability (h_{11})*: high weight → high reward for hypotheses consisting of phrases with high translation probability
2. *Phrase penalty (h_{13})*: low negative weights → the systems prefer hypotheses consisting of fewer but longer phrases.
3. *Reordering model (h_{1-7})*: weights around zero, reordering not preferred

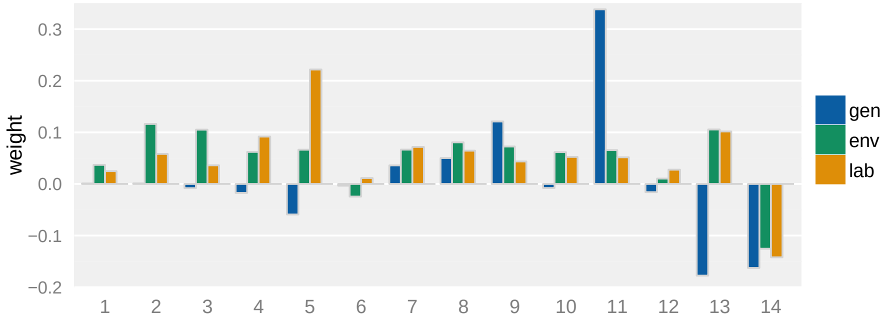
Moses weights optimized on **general** vs. **specific domain** I. (EN-FR)



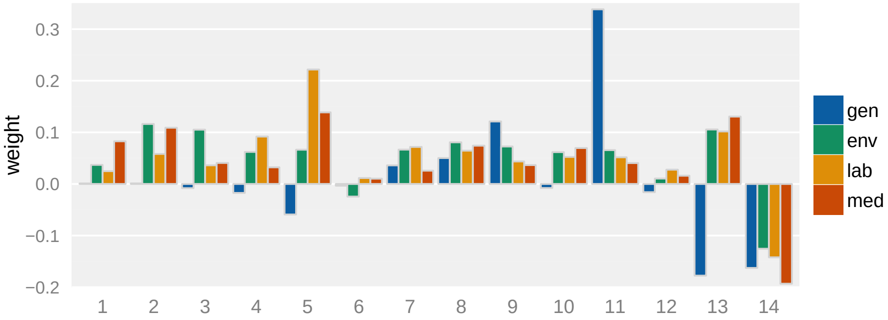
Moses weights optimized on **general** vs. **specific domain** I. (EN-FR)

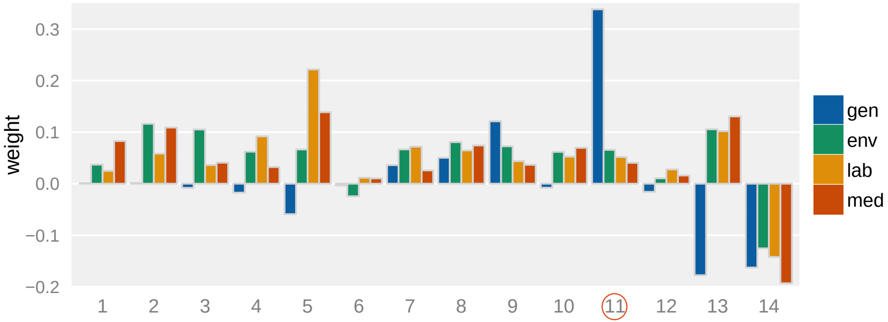


Moses weights optimized on **general** vs. **specific domain** I. (EN-FR)



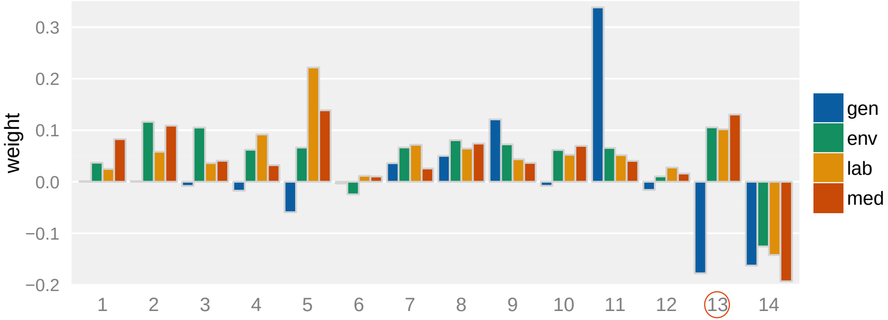
Moses weights optimized on **general** vs. **specific domain** I. (EN-FR)



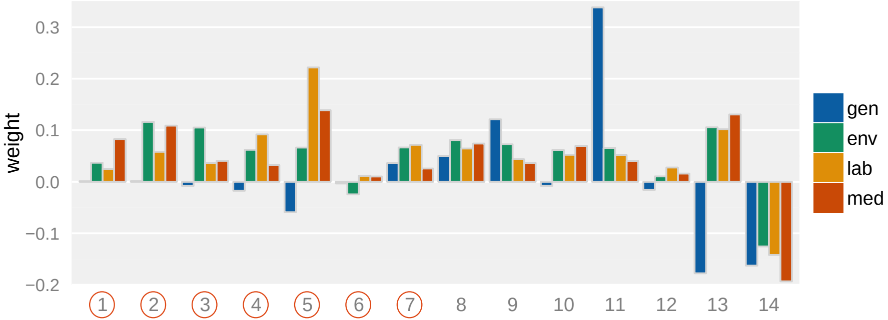


- *Direct phrase translation probability* (h_{11}): weights decrease rapidly
- The translation tables do not provide enough good quality translations for the specific domains
- The best translation hypotheses consist of phrases with varying translation probability scores.

Moses weights optimized on **general** vs. **specific domain** II. (EN-FR)

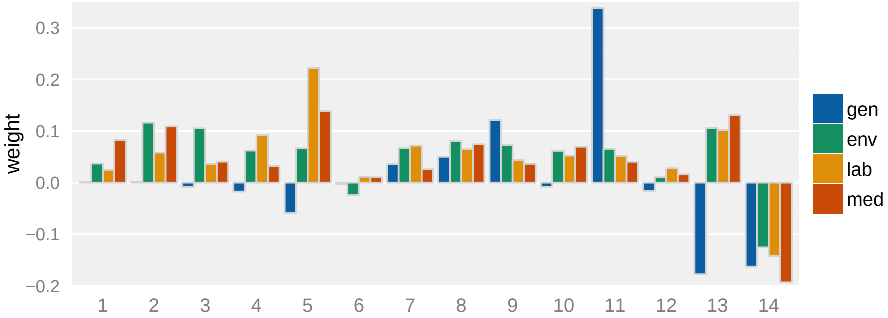


- *Phrase penalty (h_{13})*: weights increase from negative to positive
- Hypotheses consisting of few and long phrases not rewarded
- In most cases such hypotheses are penalized and hypotheses consisting of more (and short) phrases are preferred.



- *Reordering model* (h_1–h_7): weights increased substantially
- For specific-domain data the model significantly prefers hypotheses with altered phrase/word order.

Moses weights optimized on **general** vs. **specific domain** IV. (EN-FR)



- Weights of other features do not change substantially
- Their importance is similar on general- and specific-domain data.

Analysis of phrase-length distribution in test translations (EN-FR)

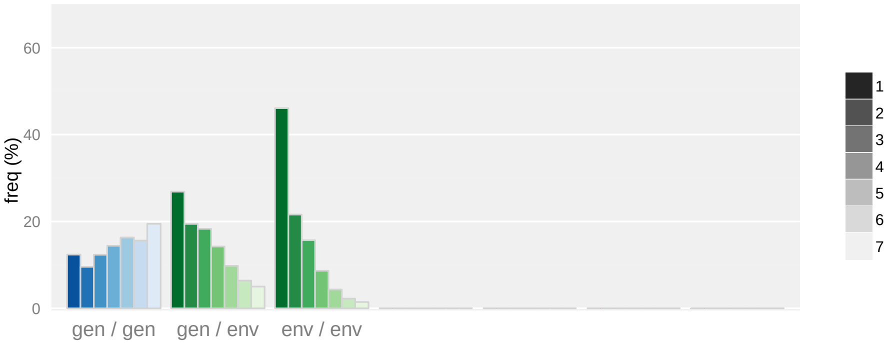
Analysis of phrase-length distribution in test translations (EN-FR)

mean: 4.37

- *gen/gen* – system uses long phrases → optimal translation quality

mean: 4.37 3.00

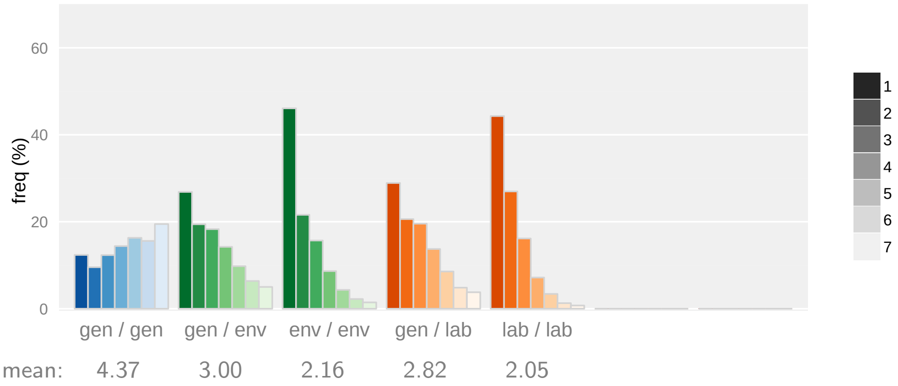
- *gen/gen* – system uses long phrases → optimal translation quality
- *gen/spec* – system trained to use long phrases → sub-optimal quality



mean: 4.37 3.00 2.16

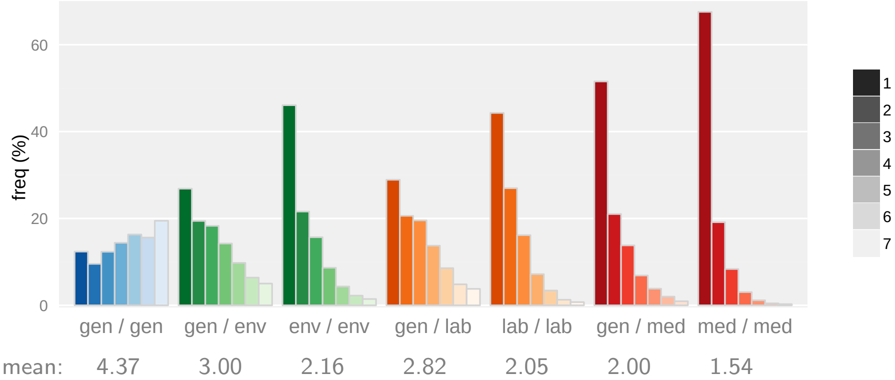
- *gen/gen* – system uses long phrases → optimal translation quality
- *gen/spec* – system trained to use long phrases → sub-optimal quality
- *spec/spec* – system trained to use shorter phrases → improved quality

Analysis of phrase-length distribution in test translations (EN-FR)



- *gen/gen* – system uses long phrases → optimal translation quality
- *gen/spec* – system trained to use long phrases → sub-optimal quality
- *spec/spec* – system trained to use shorter phrases → improved quality

Analysis of phrase-length distribution in test translations (EN-FR)



- *gen/gen* – system uses long phrases → optimal translation quality
- *gen/spec* – system trained to use long phrases → sub-optimal quality
- *spec/spec* – system trained to use shorter phrases → improved quality

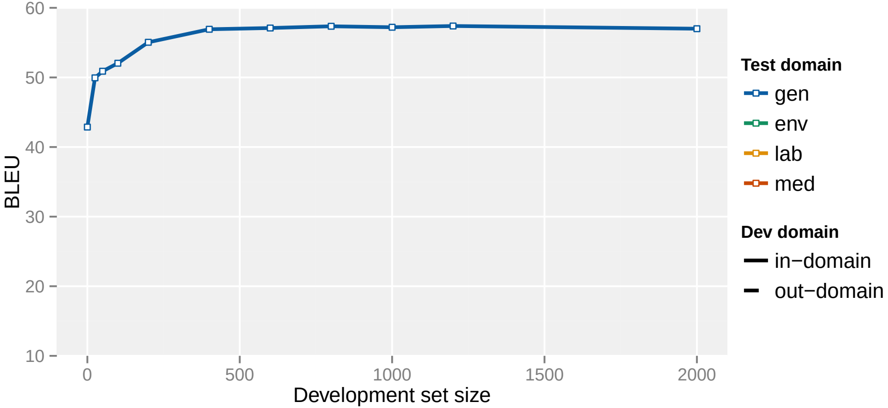
SMT system overtraining

This situation can be interpreted as overtraining: the model overfits the training/tuning data and on different domain it fails to find the best possible translations.

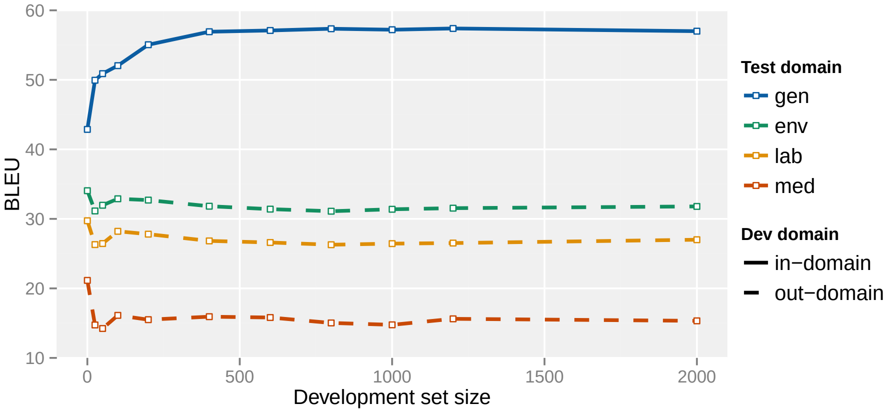
Solutions:

- In-domain parameter tuning – already discussed
- Reducing development data size – how much data we need
- No tuning at all – using default parameter values
- Cross-domain tuning – tuning on different domains

Reducing development data size – how much data we need? (EN-FR)

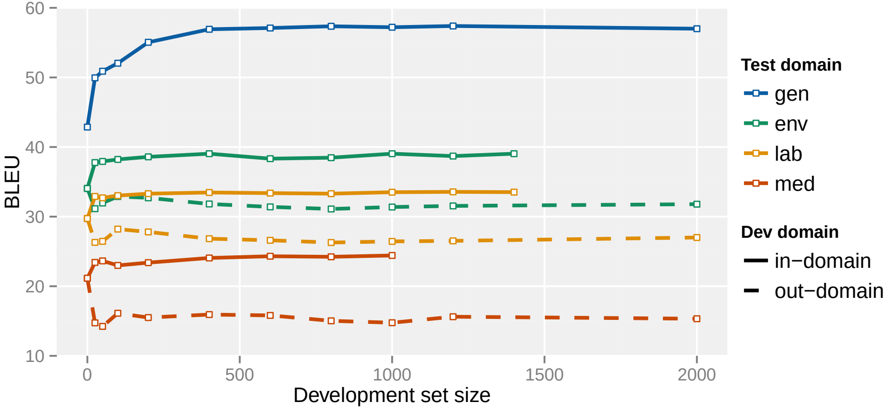


- *gen/gen* – increasing dev set size is beneficial up to 600 sentences



- *gen/gen* – increasing dev set size is beneficial up to 600 sentences
- *gen/spec* – no benefit from using general-domain dev data at all

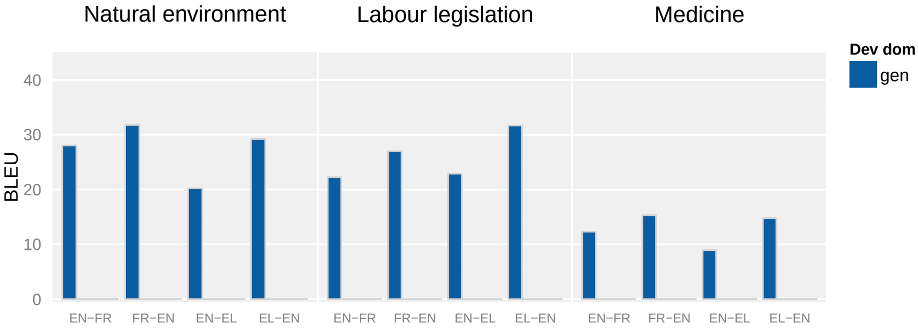
Reducing development data size – how much data we need? (EN-FR)



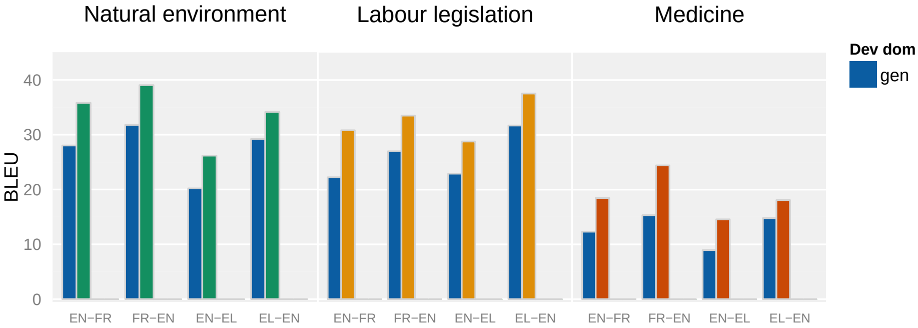
- *gen/gen* – increasing dev set size is beneficial up to 600 sentences
- *gen/spec* – no benefit from using general-domain dev data at all
- *spec/spec* – the plateau is reached much earlier, as few as 200-300 sentence pairs are usually enough to get reasonable results.

No parameter tuning – using default parameter values

No parameter tuning – using default parameter values

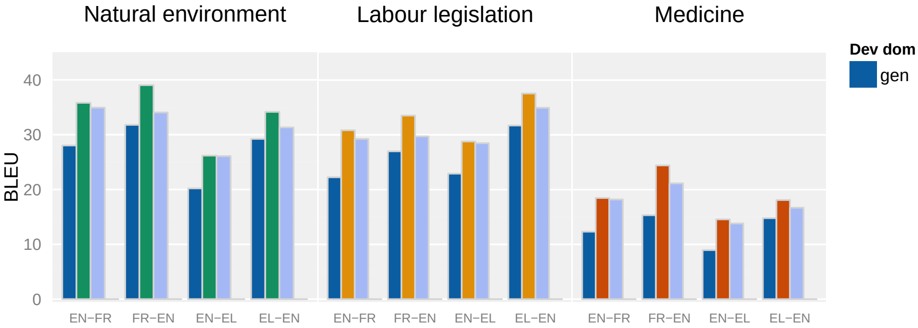


No parameter tuning – using default parameter values



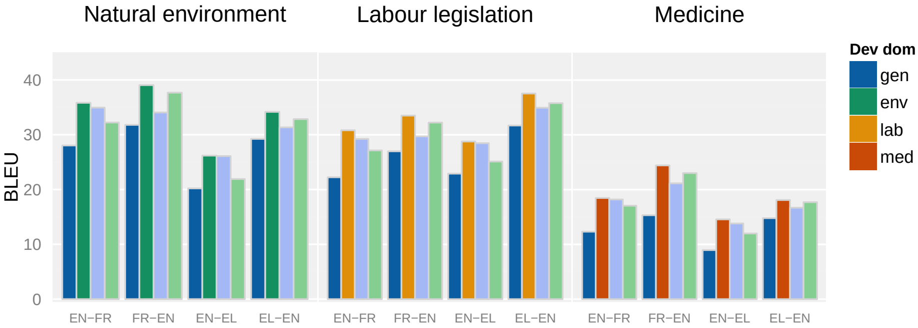
- In-domain tuning: +33.16%

No parameter tuning – using default parameter values



- In-domain tuning: +33.16%
- Using default vector weight: +24.75%

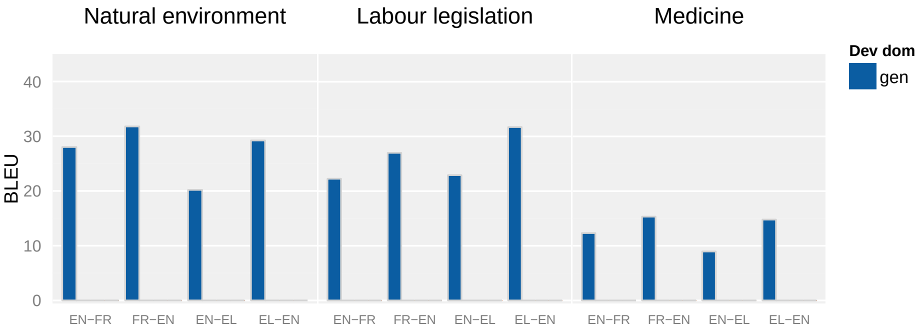
No parameter tuning – using default parameter values



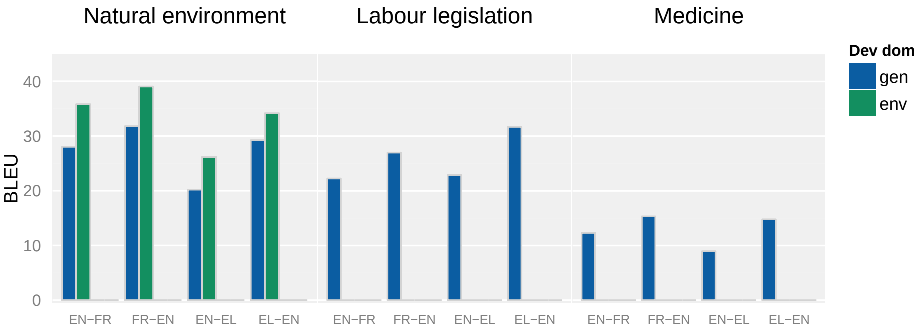
- In-domain tuning: +33.16%
- Using default vector weight: +24.75%
- Using flat vector weight: +21.10%

Cross-domain tuning – tuning on different domains

Cross-domain tuning – tuning on different domains

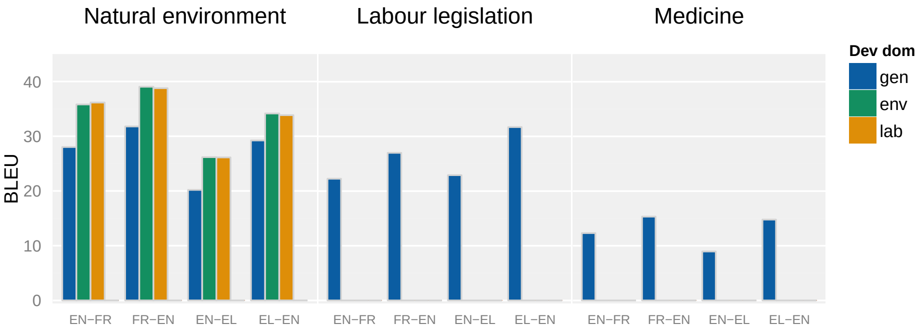


Cross-domain tuning – tuning on different domains



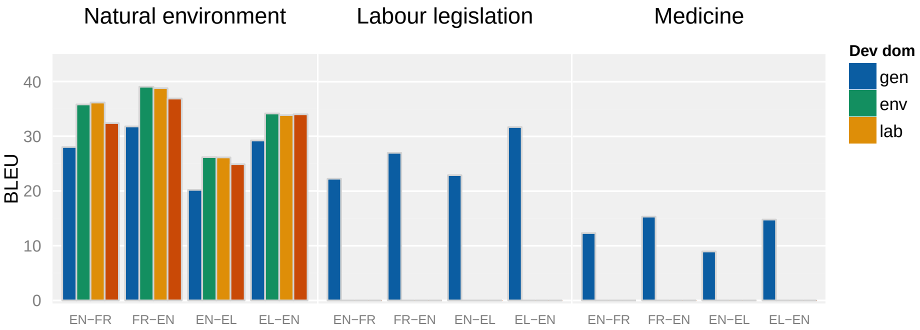
- In-domain tuning: +33.16%

Cross-domain tuning – tuning on different domains



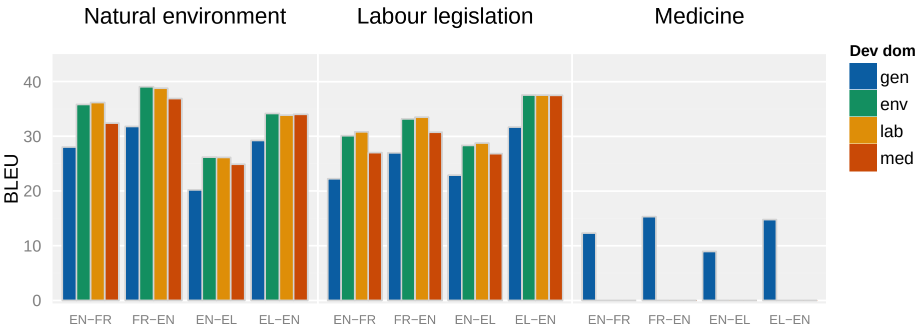
- In-domain tuning: +33.16%

Cross-domain tuning – tuning on different domains



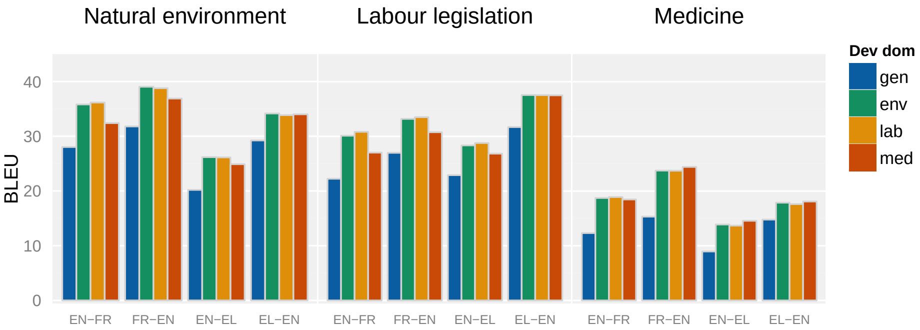
- In-domain tuning: +33.16%

Cross-domain tuning – tuning on different domains



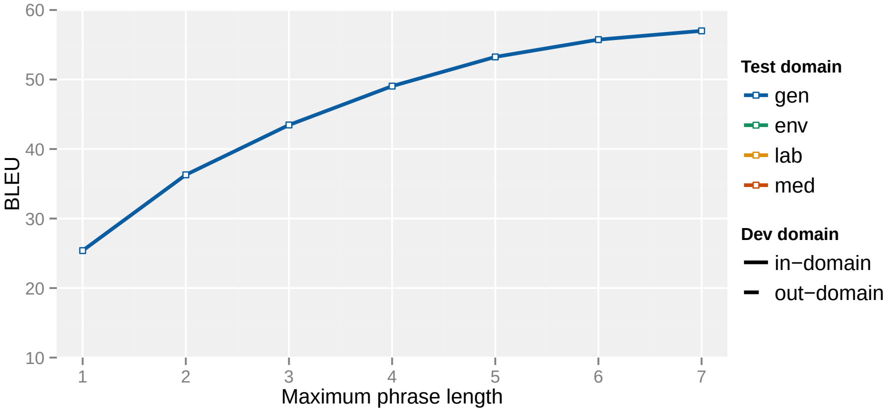
- In-domain tuning: +33.16%

Cross-domain tuning – tuning on different domains



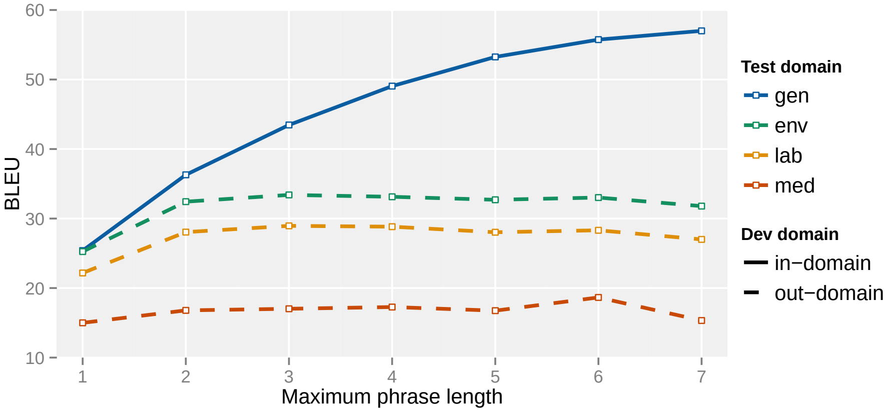
- In-domain tuning: +33.16%
- Cross-domain tuning: +29.25%

Reducing maximum phrase length (EN-FR)



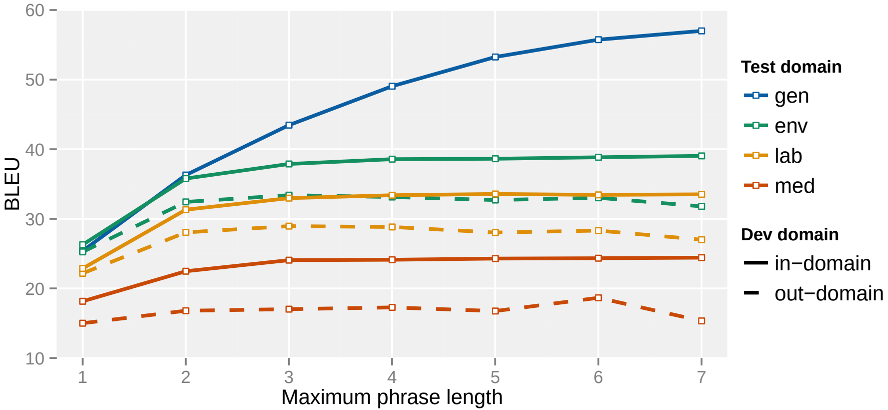
- *gen/gen* – increasing max phrase length is beneficial even beyond 7

Reducing maximum phrase length (EN-FR)



- *gen/gen* – increasing max phrase length is beneficial even beyond 7
- *gen/spec* – optimum reached at 2-3, for higher values BLEU decreases

Reducing maximum phrase length (EN-FR)



- *gen/gen* – increasing max phrase length is beneficial even beyond 7
- *gen/spec* – optimum reached at 2-3, for higher values BLEU decreases
- *spec/spec* – optimum reached at 3-4, longer phrases not needed.

Conclusions

1. Systems trained and tuned on general domain perform poorly on specific domains
2. Perplexity of the source side of the test data given the source side of the training nicely correlates with the translation quality
3. Tuning the systems trained on general domain on specific target domain data recovers a large amount of the loss
4. In-domain tuning requires about 100–200 sentence pairs to achieve decent translation quality
5. Using the default model parameters, performs surprisingly well and always outperforms systems tuned on general domain.
6. Cross-domain tuning offers a good solution when no in-domain development data is available

Thank you for your attention

This research was supported by:

- EU FP7 projects PANACEA and KHRESMOI
- Czech Science Foundation Center of Excellence CEMI
- Science Foundation Ireland project CNGL

Complete data overview

	dom	*set*	*sentences*	*L1 tokens* /	*voc*	*L2 tokens* /	*voc*
English – French	*gen*	train	1,725,096	47,956,886	73,645	53,262,628	103,436
		dev	2,000	58,655	5,734	67,295	6,913
		test	2,000	57,951	5,649	66,200	6,876
	env	dev	1,392	41,382	4,660	49,657	5,542
		test	2,000	58,865	5,483	70,740	6,617
	lab	dev	1,411	52,156	4,478	61,191	5,535
		test	2,000	71,688	5,277	84,397	6,630
	med	dev	1,064	16,807	3,484	18,932	4,865
		test	2,000	31,725	5,268	34,884	7,331
English – Greek	*gen*	train	964,242	27,446,726	61,497	27,537,853	173,435
		dev	2,000	58,655	5,734	63,349	9,191
		test	2,000	57,951	5,649	62,332	9,037
	env	dev	1,000	27,865	3,586	30,510	5,467
		test	2,000	58,073	4,893	63,551	8,229
	lab	dev	506	15,129	2,227	16,089	3,333
		test	2,000	62,953	4,022	66,770	7,056
	med	dev	1,064	16,807	3,484	20,625	3,893
		test	2,000	31,725	5,268	38,614	5,754

Complete results (BLEU)

test	*dev*	*English–French*		*French–English*		*English–Greek*		*Greek–English*	
gen	**gen**	**49.12**	*0.00*	**57.00**	*0.00*	**42.24**	*0.00*	**44.15**	*0.00*
	env	41.51	*–15.49*	41.63	*–26.96*	30.82	*–27.04*	33.99	*–23.01*
	lab	38.65	*–21.32*	44.73	*–21.53*	29.75	*–29.57*	37.01	*–16.17*
	med	34.40	*–29.97*	37.52	*–34.18*	31.02	*–26.56*	34.43	*–22.02*
	def	39.53	*–19.52*	42.87	*–24.79*	30.93	*–26.78*	32.88	*–25.53*
env	*gen*	28.03	*0.00*	31.79	*0.00*	20.20	*0.00*	29.23	*0.00*
	env	35.81	*+27.76*	**39.04**	**+22.81**	**26.18**	**+29.60**	**34.16**	**+16.87**
	lab	**36.16**	**+29.00**	38.78	*+21.99*	**26.13**	**+29.36**	33.85	*+15.81*
	med	32.40	*+15.59*	36.89	*+16.04*	24.89	*+23.22*	**34.01**	**+16.35**
	def	34.94	*+24.65*	34.05	*+7.11*	**26.09**	**+29.16**	31.33	*+7.18*
	flat	32.22	*+14.95*	37.66	*+18.46*	21.91	*+8.47*	32.84	*+12.35*
lab	*gen*	22.26	*0.00*	27.00	*0.00*	22.92	*0.00*	31.71	*0.00*
	env	30.13	*+35.35*	**33.21**	**+23.00**	28.36	*+23.73*	**37.57**	**+18.48**
	lab	**30.84**	**+38.54**	**33.52**	**+24.15**	**28.79**	**+25.61**	**37.55**	**+18.42**
	med	27.04	*+21.47*	30.77	*+13.96*	26.85	*+17.15*	**37.52**	**+18.32**
	def	29.26	*+31.45*	29.73	*+10.11*	28.48	*+24.26*	34.95	*+10.22*
	flat	27.16	*+22.01*	32.24	*+19.41*	25.13	*+9.64*	35.79	*+12.87*
med	*gen*	12.32	*0.00*	15.33	*0.00*	8.96	*0.00*	14.79	*0.00*
	env	**18.74**	**+52.11**	23.75	*+54.92*	13.89	*+55.02*	**17.88**	**+20.89**
	lab	**18.91**	**+53.49**	23.73	*+54.79*	13.69	*+52.79*	17.62	*+19.13*
	med	18.47	*+49.92*	**24.42**	**+59.30**	**14.57**	**+62.61**	**18.10**	**+22.38**
	def	18.20	*+47.73*	21.15	*+37.96*	13.82	*+54.24*	16.70	*+12.91*
	flat	17.06	*+38.47*	23.02	*+50.16*	11.99	*+33.82*	17.71	*+19.74*