The SDL Language Weaver Systems in the WMT12 QE Shared Task

Team : Radu Soricut, Nguyen Bach, Ziyuan Wang

System 1 : M5P-based QE system with 15FFs, directly optimized for DeltaAvg

System 2 : SVR-based QE system with 20FFs, manual FF selection

Outcome : Placed 1st & 2nd on both Ranking and Scoring Tasks

The Feature Set

SDL-LW system submissions created starting from 3 distinct sets of features

- 17 BFs: the baseline feature set
- 8 MFs: the internal features of Moses
- 17 LFs: a set of features that we developed internally

Total: 42 FFs (non-sparse)

The Baseline Features for QE

Systems	**Ranking**		**Scoring**		
	DeltaAvg	Spearman	MAE	RMSE	Interval
17 BFs with M5P	0.53	0.56	0.69	0.83	[2.3-4.9]
17 BFs with SVR	0.55	0.58	0.69	0.82	[2.0-5.0]
best-system	0.63	0.64	0.61	0.75	[1.7-5.0]

Table: Performance of the Baseline Features using M5P and SVR models on the test set.

The Internal Features of Moses for QE

MF1 Distortion cost

MF2 Word penalty cost

MF3 Language-model cost

MF4 Cost of the phrase-probability of source given target $\Phi(s|t)$

MF5 Cost of the word-probability of source given target $\Phi_{lex}(s|t)$

MF6 Cost of the phrase-probability of target given source $\Phi(t|s)$

MF7 Cost of the word-probability of target given source $\Phi_{lex}(t|s)$

MF8 Phrase penalty cost

The Internal Features of Moses for QE

Systems	**Ranking**		**Scoring**		
	DeltaAvg	Spearman	**MAE**	RMSE	Interval
8 MFs with M5P	0.58	0.58	0.65	0.81	[1.8-5.0]
best-system	0.63	0.64	0.61	0.75	[1.7-5.0]

Table: Performance of the Moses-based Features with an M5P model on the test set.

Note: the "8 MFs with M5P" system would have been ranked 4th (out of 17 entries) in the Ranking task, and 5th (out of 19 entries) in the Scoring task. Better than baseline features alone.

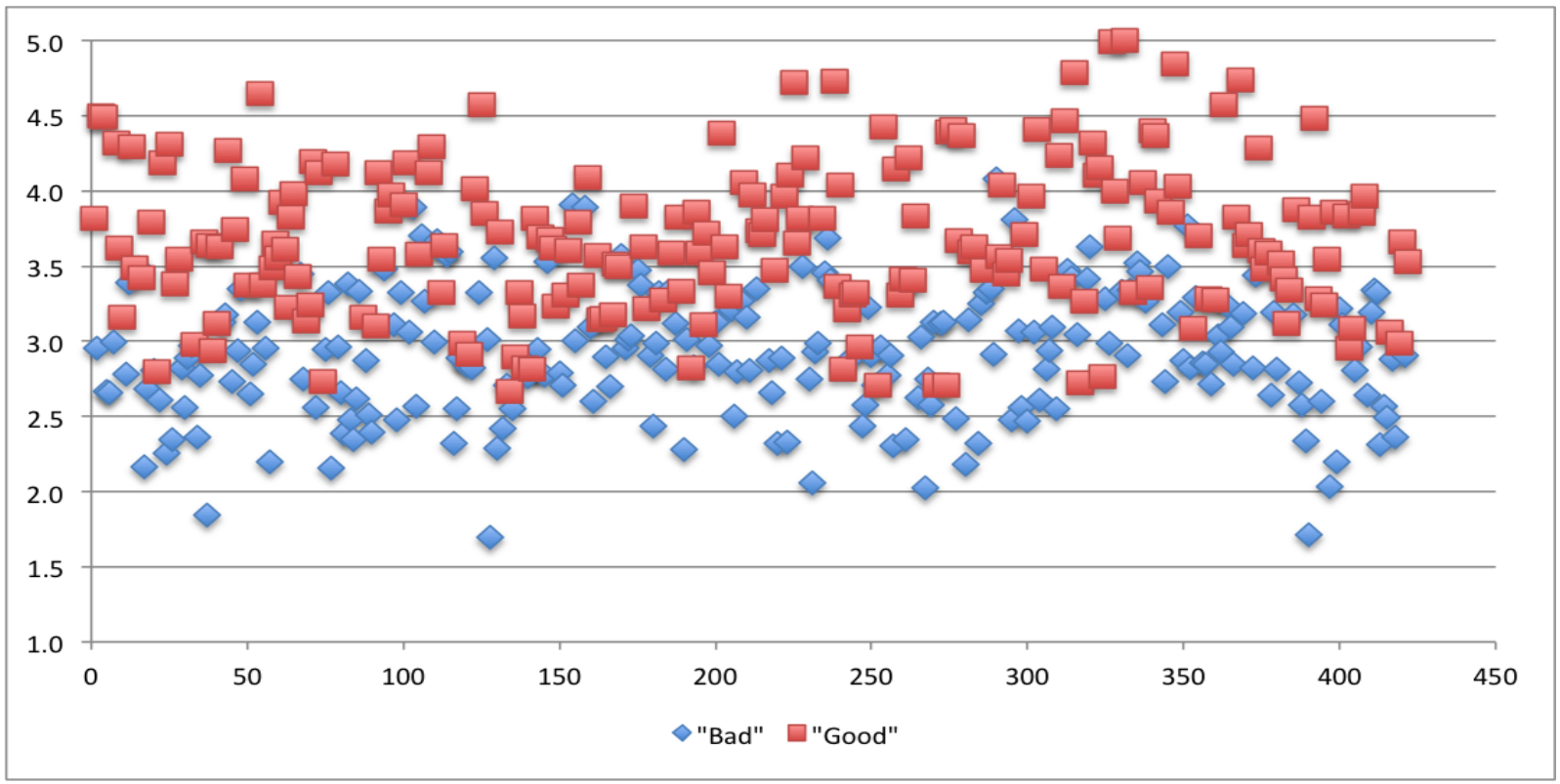
The Need for Feature Selection

Systems	#L.Eq.	**Dev Set**		**Test Set**	
		DeltaAvg	MAE	DeltaAvg	MAE
42 FFs with M5P	10	0.60	0.58	0.56	0.64
15 FFs with M5P	**2**	**0.63**	**0.52**	**0.63**	**0.61**
14 FFs with M5P	6	0.62	**0.50**	0.61	0.62

Table: M5P-model performance for different feature-function sets (15-FFs $\in$ 42-FFs; 14-FFs $\in$ 42-FFs).

The "Winning" M5P-based Submission

Regression-tree model with only 2 equations, for "Bad"/"Good".



The "Winning" Feature Functions (BFs & MFs)

- BF1 number of tokens in the source sentence
- BF3 average source token length
- BF4 LM probability of source sentence
- BF6 average number of occurrences of the target word within the target translation
- BF12 percentage of source-word bigrams in highest-frequency quartile in SMT_{src}
- BF13 percentage of source-word trigrams in lowest-frequency quartile in SMT_{src}
- BF14 percentage of source-word trigrams in highest-frequency quartile in SMT_{src}
- MF3 LM cost of target translation
- MF4 Cost of the phrase-probability of source given target $\Phi(s|t)$
- MF6 Cost of the phrase-probability of target given source $\Phi(t|s)$

The "Winning" Feature Functions (LFs)

LF1 number of out-of-vocabulary tokens in the source sentence

LF10 geometric mean (λ-smoothed) of 1-to-4–gram precision scores of target translation against a pseudo-reference produced by a second MT Eng-Spa system

LF14 count of O2O alignments with Part-of-Speech–agreement

LF15 ratio of O2O alignments with Part-of-Speech–agreement over O2O alignments

LF16 ratio of O2O alignments with Part-of-Speech–agreement over source

The SVR-based Submission

SVR Model (C;γ;ϵ)	#S.V.	**Dev Set** DeltaAvg	**Dev Set** MAE	**Test Set** DeltaAvg	**Test Set** MAE
1.0 ; .0078; 0.50	695	0.62	0.52	0.60	0.66
1.7; .0026; 0.33	**952**	**0.63**	**0.51**	**0.61**	**0.64**
8.0 ; .0019; 0.01	1509	0.64	0.50	0.60	0.68
16.0; .0014; 0.09	1359	0.63	0.51	0.59	0.70

Table: SVR-model performance for dev and test sets.

SVRs are easy to overfit on "exposed" test sets, leading to suboptimal performance on blind tests.

Conclusions

- The **decoder-internal FFs** as QE FFs **help a lot**.
- For **feature-selection**, brute-force techniques **directly optimizing** the evaluation metrics work **under M5P models** (winning submission took 60 hours on 800 machines)
- **Overfitting with SVR** models: too flexible for their own good in current set-up