Optimising Multiple Metrics with MERT

Christophe SERVAN and Holger SCHWENK

05-10/09/2011

Introduction

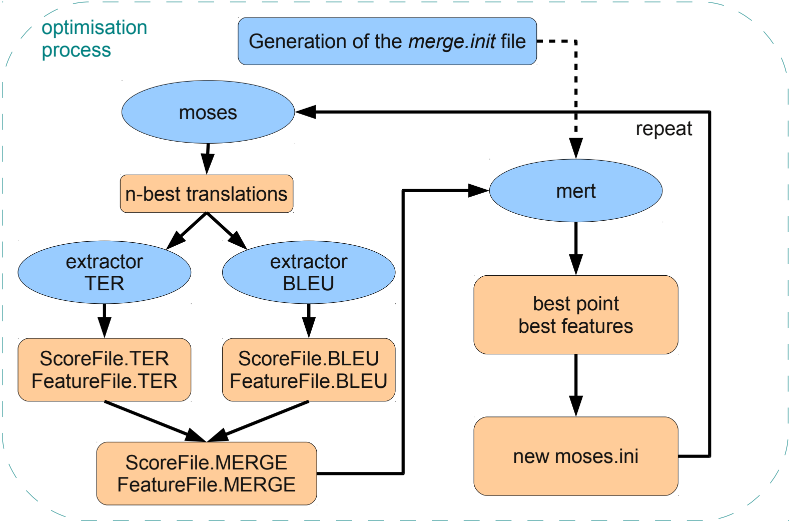
Metric combination & TER optimisation: why?

- difficulty to reproduce some experiments
- GALE evaluation main metric is HTER
  $\Longrightarrow$ need to tune with TER
- former MTMarathon project not achieved

Implementation

- TER scorer
- merge scorer
- *mert-moses.pl* switch

Process description: combination of BLEU and TER



TER scorer

- extension of our TER library (C++ implementation)
- small modification of the MERT implementation
- optimisation of $1 - TER$ (called $negTER$)

$$TER = \frac{nbr\ \ of\ \ edition}{average\ \ length\ of\ \ references} \tag{1}$$

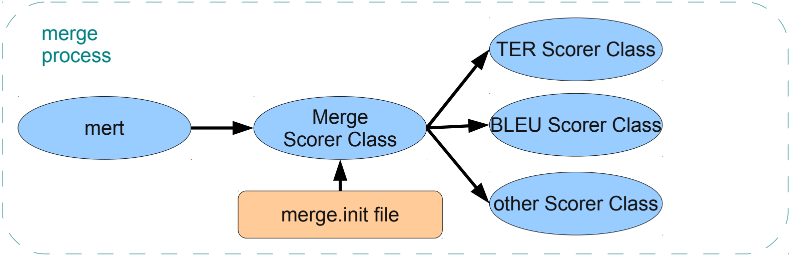
TER scorer

Details

- `extractor` software side:
  - the *setReferenceFiles* function
    $\Longrightarrow$ to load references
  - the *prepareStats* function
    $\Longrightarrow$ returns the number of edition and the average length of the TER score
- `mert` software side:
  - the *calculateScore* function
    $\Longrightarrow$ to calculate the TER scores

Merge Scorer

- a scorer to combine metrics
- uses the other scorer already implemented in `mert`



Merge Scorer

Details

- empty function for the `extractor` software side, the Merge scorer is only used in the `mert` software.
- `mert` software side:
  - the *calculateScore* function
    $\Longrightarrow$ this function allows to call each *calculateScore* function of every scorer implemented.

Other modifications

- the need of a init file that contains weight associated to every metric, feature ans score files:

Metric	weight	feature file name	score file name
BLEU	2	BLEU_FEATURE_FILE	BLEU_SCORING_FILE
TER	1	TER_FEATURE_FILE	TER_SCORING_FILE

- add scorer names to the *scorerFactory* file
- the *calculateScore* function of every scorer in order to accept a vector of float instead of a vector of int
- modification of the script *mert-moses.pl* and add the switch *--sctype* to give the configuration of every metrics used into the init file.

Experiments

- GALE Evaluation
  $\Longrightarrow$ translation from Arabic to English:
  - news
  - speech
  - web
- WMT'11 Evaluation
  $\Longrightarrow$ translation from French to English

experimental protocol

- reference: BLEU tuning
- seeds are fixed
- main metric: $\frac{TER-BLEU}{2}$

WMT'11

Training details

- same system used for WMT'11 evaluation
- 435M Words (europarl 6, news commentary 6, filtered 10^9 corpus, unsupervised data)
- 7G words for target LM (English)
- tuning and tests:
  - tuning corpus: newstest2009
  - internal test: newstest2010
  - evaluation test: newstest2011

WMT'11

- Results:

	newstest2009 (Dev)			newstest2010 (Internal test)			newstest2011 (Evaluation test)		
Optimisation	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$
BLEU	29.14	53.98	12.42	29.65	52.78	11.57	30.19	51.61	10.71
TER	27.65	52.91	12.63	28.79	51.56	11.39	29.36	50.57	10.61
1xBLEU-TER	29.15	53.58	**12.22**	29.95	52.42	**11.24**	30.37	51.36	**10.50**
2xBLEU-TER	29.10	53.88	12.39	29.93	52.55	11.31	30.15	51.56	10.71
3xBLEU-TER	29.19	53.83	12.32	29.99	52.46	*11.24*	30.14	51.56	10.71
4xBLEU-TER	29.21	54.01	12.40	29.98	52.60	11.31	30.08	51.75	10.84
5xBLEU-TER	29.33	53.84	12.26	29.89	52.53	11.32	30.21	51.56	10.68

WMT'11

- Results:

	newstest2009 (Dev)			newstest2010 (Internal test)			newstest2011 (Evaluation test)		
Optimisation	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$
BLEU	29.14	53.98	12.42	29.65	52.78	11.57	30.19	51.61	10.71
TER	27.65	52.91	12.63	28.79	51.56	11.39	29.36	50.57	10.61
1xBLEU-TER	29.15	53.58	**12.22**	29.95	52.42	**11.24**	30.37	51.36	**10.50**
2xBLEU-TER	29.10	53.88	12.39	29.93	52.55	11.31	30.15	51.56	10.71
3xBLEU-TER	29.19	53.83	12.32	29.99	52.46	*11.24*	30.14	51.56	10.71
4xBLEU-TER	29.21	54.01	12.40	29.98	52.60	11.31	30.08	51.75	10.84
5xBLEU-TER	29.33	53.84	12.26	29.89	52.53	11.32	30.21	51.56	10.68

$\Longrightarrow$ TER optimisation: improvement of TER with degrading BLEU.

WMT'11

- Results:

	newstest2009 (Dev)			newstest2010 (Internal test)			newstest2011 (Evaluation test)		
Optimisation	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$	BLEU	TER	$\frac{TER-BLEU}{2}$
BLEU	29.14	53.98	12.42	29.65	52.78	11.57	30.19	51.61	10.71
TER	27.65	52.91	12.63	28.79	51.56	11.39	29.36	50.57	10.61
1xBLEU-TER	29.15	53.58	**12.22**	29.95	52.42	**11.24**	30.37	51.36	**10.50**
2xBLEU-TER	29.10	53.88	12.39	29.93	52.55	11.31	30.15	51.56	10.71
3xBLEU-TER	29.19	53.83	12.32	29.99	52.46	*11.24*	30.14	51.56	10.71
4xBLEU-TER	29.21	54.01	12.40	29.98	52.60	11.31	30.08	51.75	10.84
5xBLEU-TER	29.33	53.84	12.26	29.89	52.53	11.32	30.21	51.56	10.68

$\Longrightarrow$ TER optimisation: improvement of TER with degrading BLEU.
$\Longrightarrow$ combined optimisation: improvement of TER without degrading BLEU.

GALE

Training details

- different genre of corpus: news, web and speech
- size of bitext:

Genre	# lines	# words AR	# words EN
news	3M	72.8M	76.9M
web	2.2M	46.6M	48.3M
speech	2.4M	54.4M	57.3M

- for each genre
  - about 4G words for target LM
  - tuning and tests about 50K words for news and web, 100K for speech
  - 3 references for web and speech, 1 for news

GALE

- Results:

Corpus name	Optimisation	Dev			Test		
		BLEU	TER	(TER-BLEU)/2	BLEU	TER	(TER-BLEU)/2
news	BLEU	33.56	43.80	5.12	32.98	44.25	5.63
	TER	34.07	42.81	**4.37**	33.55	43.18	**4.82**
	1xBLEU-TER	33.55	43.67	*5.06*	33.18	44.00	*5.41*
	2xBLEU-TER	33.47	43.66	5.09	33.10	44.05	5.47
	3xBLEU-TER	33.66	43.45	4.89	33.19	43.91	5.36
web	BLEU	40.78	61.20	10.96	39.27	61.86	11.29
	TER	40.46	60.59	**10.68**	39.24	61.43	**11.10**
	1xBLEU-TER	40.76	61.09	10.79	39.52	61.72	11.10
	2xBLEU-TER	40.62	61.01	10.87	39.28	61.56	11.14
	3xBLEU-TER	40.72	60.86	*10.72*	39.42	61.56	*11.07*
speech	BLEU	33.73	58.03	12.15	33.94	58.03	12.04
	TER	33.30	55.92	**11.31**	33.39	56.34	**11.47**
	1xBLEU-TER	34.04	56.98	*11.47*	34.13	57.17	*11.52*
	2xBLEU-TER	33.97	57.21	11.62	34.12	57.28	11.58
	3xBLEU-TER	33.86	57.97	12.05	33.88	58.13	12.12

GALE

- Results:

Corpus name	Optimisation	Dev			Test		
		BLEU	TER	(TER-BLEU)/2	BLEU	TER	(TER-BLEU)/2
news	BLEU	33.56	43.80	5.12	32.98	44.25	5.63
	TER	34.07	42.81	**4.37**	33.55	43.18	**4.82**
	1xBLEU-TER	33.55	43.67	*5.06*	33.18	44.00	*5.41*
	2xBLEU-TER	33.47	43.66	5.09	33.10	44.05	5.47
	3xBLEU-TER	33.66	43.45	4.89	33.19	43.91	5.36
web	BLEU	40.78	61.20	10.96	39.27	61.86	11.29
	TER	40.46	60.59	**10.68**	39.24	61.43	**11.10**
	1xBLEU-TER	40.76	61.09	10.79	39.52	61.72	11.10
	2xBLEU-TER	40.62	61.01	10.87	39.28	61.56	11.14
	3xBLEU-TER	40.72	60.86	*10.72*	39.42	61.56	*11.07*
speech	BLEU	33.73	58.03	12.15	33.94	58.03	12.04
	TER	33.30	55.92	**11.31**	33.39	56.34	**11.47**
	1xBLEU-TER	34.04	56.98	*11.47*	34.13	57.17	*11.52*
	2xBLEU-TER	33.97	57.21	11.62	34.12	57.28	11.58
	3xBLEU-TER	33.86	57.97	12.05	33.88	58.13	12.12

⟹ TER optimisation: improvement of TER with degrading BLEU for web and speech, improvement of both metrics with news.

GALE

- Results:

Corpus name	Optimisation	Dev			Test		
		BLEU	TER	(TER-BLEU)/2	BLEU	TER	(TER-BLEU)/2
news	BLEU	33.56	43.80	5.12	32.98	44.25	5.63
	TER	34.07	42.81	**4.37**	33.55	43.18	**4.82**
	1xBLEU-TER	33.55	43.67	*5.06*	33.18	44.00	*5.41*
	2xBLEU-TER	33.47	43.66	5.09	33.10	44.05	5.47
	3xBLEU-TER	33.66	43.45	4.89	33.19	43.91	5.36
web	BLEU	40.78	61.20	10.96	39.27	61.86	11.29
	TER	40.46	60.59	**10.68**	39.24	61.43	**11.10**
	1xBLEU-TER	40.76	61.09	10.79	39.52	61.72	11.10
	2xBLEU-TER	40.62	61.01	10.87	39.28	61.56	11.14
	3xBLEU-TER	40.72	60.86	*10.72*	39.42	61.56	*11.07*
speech	BLEU	33.73	58.03	12.15	33.94	58.03	12.04
	TER	33.30	55.92	**11.31**	33.39	56.34	**11.47**
	1xBLEU-TER	34.04	56.98	*11.47*	34.13	57.17	*11.52*
	2xBLEU-TER	33.97	57.21	11.62	34.12	57.28	11.58
	3xBLEU-TER	33.86	57.97	12.05	33.88	58.13	12.12

⟹ TER optimisation: improvement of TER with degrading BLEU for web and speech, improvement of both metrics with news.

⟹ combined optimisation: improvement of TER without degrading BLEU but not the best result (TER-BLEU)/2.

Conclusions

negTER optimisation

- improvement of TER with degrading BLEU with the Fr→En task of WMT'11.
- improvement of TER with degrading BLEU with the Ar→En task of GALE'11 except for the news eval (improvement of both metrics).

Combined optimisation

- improvement of TER and BLEU with both task and language pair compared to the BLEU optimisation only.

⟹ investigations about theses performances (language pair, language type... ?)

Conclusions & outlooks

- new scorer tested successfully on two language pair : Fr→En (news) and Ar→En (news, web and speech)
- find the best metric combination related to human judgements
- add further metrics (TERp, METEOR...)
- software available at: `https://mosesdecoder.svn.sourceforge.net/svnroot/mosesdecoder/branches/mert-other_metrics` and packaged at for the latest moses (soon into the main trunk): `http://www-lium.univ-lemans.fr/~servan/package_multi_scorer.rev4138.tgz`
- usage: `mert-moses.multi.pl` with the switch: `--sctype=BLEU:2,TER:1`

Thank you!