Better Evaluation of ASR in Speech Translation Context Using Word Embeddings

Ngoc-Tien Le, Christophe Servan, Benjamin Lecouteux
and
Laurent Besacier

LIG–GETALP, University of Grenoble Alpes

September 11^{th} 2016

Spoken Language Translation (SLT)

Two main technologies involved:

- Automatic Speech Recognition (ASR)
- Machine Translation (MT)

More and more end-to-end speech-to-text approaches but

- ASR and MT remain independent in many SLT systems
- Cannot always jointly optimize full SLT pipeline

Problem addressed

- Propose a more "suitable" optimization metric for ASR in a SLT framework
  - Beyond WER

Related Works

- [Dixon et al.2011] investigated how SLT performs as speech decoder parameters change
  - sub-optimal WERs give comparable BLEU scores at faster decoding speeds
- [Bechet et al.2015] analyzed ASR error segments with high negative impact on SLT
  - removing such segments prior to translation can improve SLT
- [Ruiz and Federico2015] proposed a Phonetically-Oriented Word Error Rate (POWER)
  - incorporates alignment of phonemes to better trace the impact of error types in ASR on downstream tasks

Our approach

- Inspired from [Vilar et al.2006] who noticed that many ASR substitution errors are due to slight morphological changes, limiting the impact on SLT performance.

Word representation into a continuous space

- We need different substitution scores according to the syntactic or semantic similarity between words
- Use the representation proposed by [Mikolov et al.2013] and implemented in our toolkit MultiVec [Berard et al.2016]
- (Optionally) recompute the best alignment path between *hyp* and *ref*

The *substitution* score (0 or 1) is replaced by the cosine distance between two words (continuous value in [0,2]).

$$D_c(W_1, W_2) = 1 - S_c(W_1, W_2) \quad (1)$$

WER

WER=70%

		un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
		0	1	2	3	4	5	6	7	8	9
ordre		1	1	2	3	4	5	6	7	8	9
westphalien		2	2	2	3	4	5	6	7	8	9
d'		3	3	3	3	3	4	5	6	7	8
engagements		4	4	4	4	4	4	5	6	7	8
parmi		5	5	5	5	5	5	4	5	6	7
des		6	6	6	6	6	6	5	5	6	7
nations		7	7	7	7	7	7	6	6	6	7
souveraines		8	8	8	8	8	8	7	7	7	7
Alignment:		A	I	S	S	A	S	A	S	S	S
Cost:		0	1	1	1	0	1	0	1	1	1

WER-E

WER-E=48.5%

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1.01	2.07	2.93	4.15	4.89	6.07	7.03	8.05	9.01
westphalien	2	1.79	1.73	2.83	3.93	5.38	5.80	6.90	7.75	8.85
d'	3	3.05	2.97	2.21	2.83	3.83	4.83	5.83	6.83	7.83
engagements	4	3.94	4.02	4.15	3.41	3.30	5.01	5.91	6.92	7.81
parmi	5	4.77	4.80	5.13	5.15	4.61	3.30	4.30	5.30	6.30
des	6	6.04	5.85	5.80	5.61	6.24	4.30	3.64	5.49	6.12
nations	7	6.87	6.83	6.77	6.85	6.55	5.30	5.26	4.42	6.43
souveraines	8	7.92	7.71	7.99	7.71	7.82	6.30	6.15	6.10	**4.85**

Alignment:	A	I	S	S	A	S	A	S	S	S
Cost:	0	1	1.07	0.75	0	0.47	0	0.35	0.78	0.43

WER-S

WER-S=47.6%

		un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
	un	0	1	2	3	4	5	6	7	8	9
ordre		1	1.01	2.01	2.93	3.93	4.89	5.89	6.89	7.89	8.89
westphalien		2	1.79	1.74	2.74	3.74	4.74	5.74	6.72	7.61	8.61
d'		3	2.79	2.74	2.21	2.74	3.74	4.74	5.74	6.74	7.74
engagements		4	3.79	3.74	3.21	3.42	3.21	4.21	5.21	6.21	7.21
parmi		5	4.77	4.65	4.21	4.21	4.21	3.21	4.21	5.21	6.21
des		6	5.77	5.65	5.21	4.68	5.21	4.21	3.55	4.55	5.55
nations		7	6.77	6.57	6.21	5.68	5.63	5.21	4.55	4.34	5.34
souveraines		8	7.77	7.57	7.21	6.68	6.63	6.21	5.55	5.34	4.76

Alignment:	A	S	S	I	A	S	A	S	S	S
Cost:	0	1.01	0.73	1	0	0.47	0	0.35	0.78	0.43

Questions for next SLT experiments

- Do WER-E and WER-S better correlate with BLEU (or other MT metrics) ?
- Do we improve BLEU (or other MT metrics) by selecting the best ASR hypothesis according to WER-E and WER-S ?
- (Preliminary) What do we get when we optimize ASR using WER-E or WER-S instead of WER ?

ASR & SMT systems

- French to English SLT
- Hybrid HMM-DNN ASR
  - KALDI toolkit [Povey et al.2011]
  - trained on more than 600 hours (ESTER, REPERE, ETAPE and BREF 120)
  - CD-DNN-HMM acoustic model
  - 3gram LM (ESTER)
  - N-Best output size: 1000
- Phrase-Based SMT
  - Moses toolkit [Koehn et al.2007]
  - Europarl, Ted and News-Commentary corpora (60M words)

SLT Evaluation Corpus

- French speech utterances (news domain) [Besacier et al.2014]
- *dev*: 2643 utt and *test*: 4050 utt
- online `https://github.com/besacier/WCE-SLT-LIG/`

Tasks	*metrics*	ASR Ref.	ASR 1-best
dev	*WER*	–	21.92
	TER	38.84	55.64
	BLEU	43.05	30.81
	METEOR	40.73	34.02
test	*WER*	–	17.46
	TER	45.64	58.70
	BLEU	44.71	34.27
	METEOR	39.10	34.27

ASR Metrics: WER, WER-E and WER-S

Tasks	*metrics*	ASR 1-best	Oracle from N-best WER	WER-E	WER-S
dev	*WER*	21.92	12.01	12.16	12.15
	WER-E	18.10	10.45	9.98	10.04
	WER-S	17.41	10.19	9.79	9.75
test	*WER*	17.46	7.38	7.53	7.52
	WER-E	13.13	5.86	5.43	5.48
	WER-S	12.53	5.65	5.29	5.25

SLT Correlation scores

Do WER-E and WER-S better correlate with BLEU (or other MT metrics) ?

Pearson Correlation between ASR metrics (WER, WER-E or WER-S) and SLT performances (TER, BLEU, METEOR) - each point measured on blocks of 100 sentences:

Tasks	*metrics*	Pearson Correlation WER	WER-E	WER-S
dev	*TER*	0.732	0.767	**0.773**
	BLEU	-0.677	-0.708	**-0.710**
	METEOR	-0.753	**-0.799**	-0.797
tst	*TER*	**0.457**	**0.457**	0.441
	BLEU	-0.624	**-0.661**	-0.606
	METEOR	-0.672	**-0.692**	-0.678

Spearman Correlation scores:

Tasks	*metrics*	Spearman Correlation WER	WER-E	WER-S
dev	*TER*	0.768	**0.843**	0.841
	BLEU	-0.703	**-0.774**	-0.769
	METEOR	-0.783	**-0.879**	-0.873
tst	*TER*	0.511	**0.525**	0.518
	BLEU	**-0.627**	-0.578	-0.586
	METEOR	-0.602	**-0.643**	-0.638

Oracle analysis

Do we improve BLEU (or other MT metrics) by selecting the best ASR hypothesis according to WER-E and WER-S ?

Speech Translation (SLT) performances:

Tasks	*metrics*	ASR 1-best	Oracle from N-best WER	WER-E	WER-S
dev	*TER*	55.64	50.62	50.52	**50.45**
	BLEU	30.81	35.29	35.37	**35.41**
	METEOR	34.02	36.37	36.42	**36.44**
test	*TER*	58.70	54.13	**54.01**	54.03
	BLEU	34.27	39.34	**39.43**	39.42
	METEOR	34.27	36.55	**36.64**	**36.64**

Comparison of SLT perf of the *Oracle WER* vs. the *Oracle WER-E*:

Tasks	Comparison	TER	BLEU	METEOR
Dev	O. WER-E best	**255**	**310**	**321**
	O. WER best	190	271	315
	Ties	2198	2062	2007
Test	O. WER-E best	**341**	**451**	**510**
	O. WER best	264	381	399
	Ties	3445	3218	3141

WER *versus* WER-E ASR optimization

(Preliminary) What do we get when we optimize ASR using WER-E or WER-S instead of WER ?

Tasks	*metrics*	ASR optimized with WER	ASR optimized with WER-E
dev	*TER*	55.64	**55.52**
	BLEU	30.81	**30.84**
	METEOR	**34.02**	34.00
test	*TER*	58.71	**58.56**
	BLEU	34.27	**34.38**
	METEOR	**34.27**	34.26

Output example

ASR example

REF ASR	ce serait intéressant de voir un ordinateur ...	WER	WER-E	WER-S
OptWER	ce **sera** intéressant de voir un ordinateur ...	9.09	2.43	2.43
OptWER-E	ce **serait** intéressant de voir un ordinateur ...	0.00	0.00	0.00

ASR optimization done on the whole dev corpus

SLT example

REF SLT	it would be interesting to see a computer ...	TER	SentBLEU	METEOR
OptWER - SLT	this **will** be interesting to see a computer ...	33.33	62.63	49.33
OptWER-E - SLT	it **would** be interesting to see a computer ...	16.67	79.11	92.73

Conclusion

- Extension of WER to penalize differently substitution errors using word embeddings
- WER-E and WER-S better correlated with SLT performances
- Opens possibilities to optimize ASR using metrics clever than WER
- Reproductible research
  - data of our experiments available on *github*[1]
  - modified WER code available on *github*[2]

[1] https://github.com/besacier/WCE-SLT-LIG/tree/master/IS2016
[2] https://github.com/cservan/tercpp-embeddings

References I

Frederic Bechet, Benoit Favre, and Mickael Rouvier.
2015.
"speech is silver, but silence is golden": improving speech-to-speech translation performance by slashing users input.
In *Proceedings of Interspeech 2015*, Dresden, Germany, September.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier.
2016.
MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP.
In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, May.

Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Luong, Kaing Hour, and Marwa Hadjsalah.
2014.
Word confidence estimation for speech translation.
In *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, USA, December.

Paul R. Dixon, Andrew Finch, Chiori Hori, and Hideki Kashioka.
2011.
Investigation on the effects of ASR tuning on speech translation performance.
In *The proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, CA, USA, December.

References II


P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst.
2007.
Moses: Open Source Toolkit for Statistical Machine Translation.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, Juin.


Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
2013.
Efficient estimation of word representations in vector space.
In *The Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*, Scottsdale, AR, USA, May.


Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely.
2011.
The kaldi speech recognition toolkit.
In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, December.


Nicholas Ruiz and Marcello Federico.
2015.
Phonetically-oriented word error alignment for speech recognition error analysis in speech translation.
In *IEEE 2015 Workshop on Automatic Speech Recognition and Understanding*, December.

References III

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney.
2006.
Error analysis of statistical machine translation output.
In *Proceedings of LREC 2006*, Genoa, Italy, May.