The TALP–UPC Spanish–English WMT Biomedical Task: Bilingual Embeddings and Char-based Neural Language Model Rescoring in a Phrase-based System

Marta R. Costa-jussà, Cristina España-Bonet, Pranava Madhyastha, Carlos Escolano, José A. R. Fonollosa

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona

ACL 2016
AUGUST 7 - 12 | BERLIN, GERMANY
Presented at First Conference on Machine Translation (WMT16)

Abstract

– Base System: Standard Phrase Based MT system
– Toppings:
  – Vocabulary Expansion
  – Character Based Neural LM
  – Rescoring
– Progressive improvements over several measures.

The Basic System

Phrase Based SMT:

– Goal: Focuses on finding the most probable target sentence given the source sentence.
– Uses a Log-Linear model combining a set of *feature funtions*.
– Feature functions include translation model(s) and *language model*.

Vocabulary Expansion with Bilingual Word-Embeddings

– Key Idea: Mapping Embeddings from two different language to a *Common Subspace*
– We look at this as a bilinear prediction task:

$$\Pr(t|s;W) = \frac{\exp\{\phi_{\tilde{s}}(s)^\top W \phi_{\tilde{t}}(t)\}}{\sum_{t'} \exp\{\phi_{\tilde{s}}(s)^\top W \phi_{\tilde{t}}(t')\}}$$

– This corresponds to:
a) Get the word embeddings over a sufficiently large corpora in both languages;
b) Use a relatively small dictionary as supervision to map the embeddings into a common subspace.
– We learn using spectral regularized model giving us an additional advantage of having compressed bilingual embeddings.

Character-based Neural Language Model

– Motivation: Current LM's including RNN-based ones have limitations:
a) Limited to a finite-size vocabulary for both computational and sparsity reasons;
b) Orthographic representation of the words is completely ignored; is blind to the presence of stems, prefixes, suffixes and any other kind of affixes in words.
– We use recently proposed Char-based NLMs(Kim et al., 2016)
– Basic Process:
a) Get character-based embedding layer that associates each word (sequence of characters) with a sequence of vectors;
b) Process with set of 1D convolution filters of different lengths followed with a max pooling layer and two additional highway layers;
c) The output of the second highway layer replaces the standard source word embedding in the RNN.

Data and OOV%

– Standard Spanish-English BioMedical data + Parallel Copora + Wiki.
– Training + Dev (randomly sampled) and test.
– OOVs lower as training data is indomain.
– Standard preprocessing of the data on both sides using standard pipelines.

		English			Spanish		
	Seg.	Tokens	OOV_{STT}	OOV_{BTT}	Tokens	OOV_{STT}	OOV_{BTT}
devBio	1000	18967	16 (0.08%)	2 (0.01%)	19931	14 (0.07%)	6 (0.03%)
testBio	1000	26105	31 (0.11%)	19 (0.07%)	27651	25 (0.09%)	9 (0.03%)
Biological	4344	115709	434 (0.37%)	333 (0.29%)	126008	415 (0.33%)	254 (0.20%)
Health	5111	125624	133 (0.10%)	98 (0.08%)	146368	160 (0.11%)	40 (0.03%)

System Description

SMT System: Standard phrase-based; in-domain system - we use a 5−gram `SRILM` based SLM on target corpora; extended systems - we use monololingual+parallel copora; WA using `GIZA++` and we use standard MERT.
Vocab Expansion: 300−dimensional CBOW-word embeddings from monolingual copora with `word2vec` for each languages; Supervision using a dictionary obtained from `Apertium`.
OOVs only: We use this module to give us info for OOVs only; We use the probabilistic distribution of target language words given the source language word.
Reranking using Char-NLM: The 1000-best list of translations given by the SMT engine is re-ranked using Char-NLM; the Char-NLM system has been trained with 1D convolutional filters of width [1,2,3,4,5,6,7] and size [50, 100, 150, 200, 200, 200, 200] for a total of 1,100 filters with a tanh ac- tivation, 2 highway layers with a ReLU activation, and 2 LSTM with 650 hidden units.
Systems: {X}TT{X}LM, where X = S: Small and B: Big; oov = with our oov module; reranking = reranking using the char-based NLM.

Evaluation of the in-house test set for the En2Es systems

	WER	PER	TER	BLEU	NIST	GTM-2	MTRst	MTRpa	RG-S*	ULC
BTTBLM.oov	48.45	29.82	44.27	43.84	8.81	36.30	61.58	62.87	49.85	66.03
BTTBLM.oov.reranked	47.58	**29.74**	43.56	44.43	8.90	36.97	62.01	63.25	50.43	67.16
BTTBLM	47.74	30.39	43.72	43.61	8.86	36.51	61.50	62.76	49.98	66.19
BTTBLM.reranked	47.64	29.91	43.52	44.24	8.89	36.90	61.88	63.14	50.29	66.95
STTBLM.oov	48.00	29.60	43.73	44.32	8.87	36.65	62.13	63.32	50.12	66.88
STTBLM.oov.reranked	47.22	29.85	43.11	44.57	8.96	37.21	62.22	63.42	50.44	67.57
STTBLM	**47.01**	29.93	**42.81**	44.51	**8.98**	37.36	62.28	63.47	50.49	67.75
STTBLM.reranked	47.10	29.91	42.96	44.65	8.97	37.40	62.31	63.46	**50.68**	67.78
STTSLM.oov	47.84	29.28	43.61	**44.99**	8.88	37.36	**62.33**	63.44	50.51	67.60
STTSLM.oov.reranked	47.41	29.82	43.25	44.52	8.94	37.29	62.25	63.36	**50.68**	67.54
STTSLM	47.29	29.84	43.16	44.64	8.96	**37.58**	62.27	63.42	50.56	67.71
STTSLM.reranked	47.40	29.93	43.24	44.39	8.94	37.36	62.21	63.30	50.56	67.44
total.reranked	47.06	29.82	43.03	44.75	**8.98**	37.56	**62.33**	**63.53**	50.66	**67.88**

Evaluation of the in-house test set for the Es2En systems

	WER	PER	TER	BLEU	NIST	GTM-2	MTRst	MTRpa	RG-S*	ULC
BTTBLM.oov	50.95	29.98	46.79	40.94	8.59	35.02	35.03	37.28	49.13	65.30
BTTBLM.oov.reranked	50.41	29.75	46.23	41.58	8.65	35.52	35.25	37.48	49.50	66.24
BTTBLM	50.21	29.33	45.98	41.97	8.68	35.88	35.44	37.65	50.01	66.97
BTTBLM.reranked	50.41	29.63	46.28	41.62	8.65	35.51	35.27	37.53	49.50	66.29
STTBLM.oov	50.75	29.95	46.68	40.82	8.61	34.83	35.05	37.12	49.15	65.27
STTBLM.oov.reranked	50.19	29.22	46.04	42.10	8.71	35.72	35.57	37.65	49.95	67.04
STTBLM	50.91	29.74	46.74	41.16	8.62	34.97	35.33	37.40	49.39	65.67
STTBLM.reranked	50.27	**29.08**	46.01	42.19	8.72	35.79	35.62	**37.66**	50.08	67.20
STTSLM.oov	**49.79**	29.45	**45.62**	42.16	**8.75**	**35.94**	35.57	37.60	**50.13**	**67.31**
STTSLM.oov.reranked	50.15	**29.08**	45.99	**42.30**	8.71	35.88	**35.65**	**37.66**	50.10	67.30
STTSLM	50.62	29.53	46.46	41.71	8.65	35.47	35.46	37.48	49.71	66.34
STTSLM.reranked	50.25	29.12	46.04	42.13	8.70	35.76	35.59	37.62	49.97	67.09
total.reranked	50.06	29.42	45.93	42.06	8.71	35.80	35.47	37.65	49.93	67.00

Observations and Conclusions

– Preliminary Results show that the OOV module consistently improves the translations with respect to our baseline specially in the health subdomain as measured by BLEU.
– The re-ranking module is also always better than the in-domain phrase-based baseline.
– On the competition test set - the performance of reranking module is significantly better than the above results.
– In general, the final results indicate that:
a) The system that includes re-ranking with a char-based neural language model is 2 points of BLEU over the average systems in biological subdomain;
b) It is at least 1 point better than the average on the health sub-domain.

Acknowledgements: This work is supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, contract TEC2015-69266-P (MINECO/FEDER, UE).