Two Papers on Universal Neural Machine Translation

Presenter: Yong Jiang

Contextual Parameter Generation for Universal Neural Machine Translation

Emmanouil Antonios Platanios[†], Mrinmaya Sachan[†], Graham Neubig[‡], Tom M. Mitchell[†]
[†]Machine Learning Department, [‡]Language Technologies Institute
Carnegie Mellon University
{e.a.platanios,mrinmays,gneubig,tom.mitchell}@cs.cmu.edu

Motivation

- Aim at Multilingual NMT task
- Universal NMT (Google MNMT): oversimplify
- Per-language encoder-decoder: lack of sharing info

Approach

- Contextual Parameter Generator (CPG)
- Learns language embeddings as a context for translation
- Use them to generate the parameters of a shared translation model for ALL language pairs

Approach

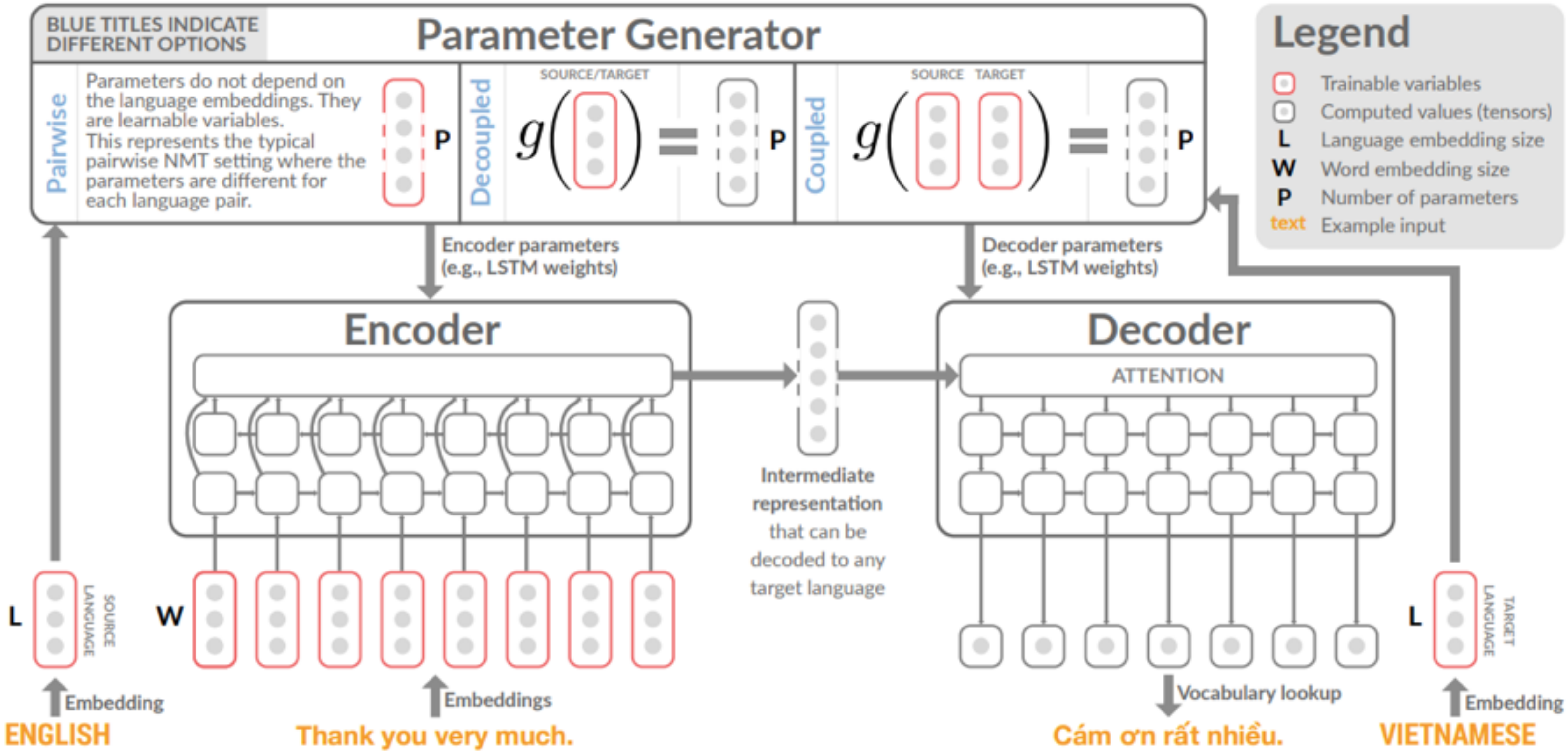

Figure 1: Overview of an NMT system, under our modular framework. Our main contribution lies in the parameter generator module (i.e., coupled or decoupled — each of the boxes with blue titles is a separate option). Note that g denotes a parameter generator network. In our experiments, we consider linear forms for this network. However, our contribution does not depend on the choices made regarding the rest of the modules; we could still use our parameter generator with different architectures for the encoder and the decoder, as well as using different kinds of vocabularies.

- Param Gen: W*E
- Controlled Gen: W*P*E (low rank)

Experiments

Table 1: Comparison of our proposed approach (shaded rows) with the base pairwise NMT model (PNMT) and the Google multilingual NMT model (GML) for the IWSLT-15 dataset. The *Percent Parallel* row shows what portion of the parallel corpus is used while training; the rest is being used only as monolingual data. Results are shown for the BLEU and Meteor metrics. CPG* represents the same model as CPG, but trained without using auto-encoding training examples. The best score in each case is shown in **bold**.

		BLEU				Meteor			
		PNMT	GML	CPG*	CPG	PNMT	GML	CPG*	CPG
100% Parallel Data	En→Cs	14.89	15.92	16.88	**17.22**	19.72	20.93	21.51	**21.72**
	Cs→En	24.43	25.25	26.44	**27.37**	27.29	27.46	28.16	**28.52**
	En→De	25.99	15.92	26.41	**26.77**	44.72	42.97	45.97	**46.30**
	De→En	30.93	29.60	31.24	**31.77**	30.73	29.90	30.95	**31.13**
	En→Fr	38.25	34.40	38.10	**38.32**	57.43	53.86	57.42	**57.68**
	Fr→En	37.40	35.14	37.11	**37.89**	34.83	33.14	34.34	**34.89**
	En→Th	23.62	22.22	26.03	**26.33**	-	-	-	-
	Th→En	15.54	14.03	16.54	**26.77**	21.58	21.02	22.78	**23.05**
	En→Vi	27.47	25.54	28.33	**29.03**	-	-	-	-
	Vi→En	24.03	23.19	25.91	**26.38**	27.59	26.96	28.23	**28.79**
	Mean	26.26	24.12	27.30	**27.80**	32.98	32.03	33.67	**34.01**
10% Parallel Data	En→Cs	5.71	8.18	8.40	**9.49**	12.18	14.97	15.25	**15.90**
	Cs→En	6.64	14.56	14.81	**15.38**	13.02	20.04	19.98	**20.87**
	En→De	11.70	14.60	15.09	**16.03**	29.98	33.74	34.88	**36.19**
	De→En	18.10	19.02	19.77	**20.25**	22.57	23.27	23.65	**24.40**
	En→Fr	24.47	25.15	24.00	**25.79**	44.10	44.84	44.95	**46.22**
	Fr→En	23.79	25.02	24.55	**27.12**	26.28	26.61	26.20	**28.18**
	En→Th	7.86	15.58	**18.41**	17.65	-	-	-	-
	Th→En	7.13	9.11	**10.19**	10.14	13.91	16.32	16.78	**16.92**
	En→Vi	18.01	17.51	**18.92**	18.90	-	-	-	-
	Vi→En	6.69	16.00	16.28	**16.86**	13.39	21.01	21.34	**22.28**
	Mean	13.01	16.47	17.04	**17.76**	21.93	25.10	25.38	**26.37**
1% Parallel Data	En→Cs	0.49	1.25	1.57	**2.38**	4.60	6.24	6.28	**8.38**
	Cs→En	1.10	1.76	1.87	**4.60**	6.29	7.13	7.08	**11.15**
	En→De	1.22	4.13	4.06	**6.46**	12.23	18.29	17.61	**23.83**
	De→En	1.46	3.42	3.86	**7.49**	7.58	8.79	8.95	**13.73**
	En→Fr	2.88	7.74	7.41	**12.45**	13.88	21.29	21.80	**30.36**
	Fr→En	4.05	5.22	5.06	**11.39**	9.58	9.86	9.83	**16.34**
	En→Th	1.22	5.72	8.01	**9.26**	-	-	-	-
	Th→En	1.42	1.66	1.65	**3.37**	6.08	7.22	5.89	**8.74**
	En→Vi	5.35	5.61	5.48	**8.00**	-	-	-	-
	Vi→En	2.01	3.57	3.64	**6.43**	7.86	8.76	8.48	**12.04**
	Mean	2.12	4.01	4.26	**7.18**	8.51	10.95	10.74	**15.58**

Experiments

Table 2: Comparison of our proposed approach (shaded rows) with the base pairwise NMT model (PNMT) and the Google multilingual NMT model (GML) for the IWSLT-17 dataset. Results are shown for the BLEU metric only because Meteor does not support `It`, `Nl`, and `Ro`. CPG^8 represents CPG using language embeddings of size 8. The "$_{C4}$" subscript represents the low-rank version of CPG for controlled parameter sharing (see Section 3.1), using rank 4, etc. The best score in each case is shown in **bold**.

		BLEU							
		PNMT	GML	CPG^8	CPG^8_{C4}	CPG^8_{C2}	CPG^8_{C1}	CPG^{64}_{C8}	CPG^{512}_{C8}
Supervised	De→En	21.78	21.25	**22.56**	20.78	22.09	21.23	21.50	22.38
	De→It	13.16	13.84	**14.73**	14.34	14.43	13.84	14.34	14.11
	De→Ro	10.85	11.95	12.24	12.37	**12.72**	10.37	11.32	11.94
	En→De	**19.75**	17.06	19.41	19.04	18.42	17.04	17.46	19.29
	En→It	27.70	25.74	27.57	27.11	**28.21**	26.26	27.26	27.48
	En→Nl	24.41	22.46	24.47	**25.15**	24.64	23.94	24.48	24.50
	En→Ro	19.23	18.60	20.83	**20.96**	18.69	17.23	20.20	20.86
	It→De	14.39	12.76	14.61	**15.06**	14.15	13.12	14.18	14.69
	It→En	29.84	27.96	**30.62**	30.10	29.44	29.22	29.56	30.18
	It→Nl	16.74	16.27	17.99	**18.11**	18.05	17.13	17.71	17.99
	Nl→En	26.30	24.78	26.31	26.17	25.74	26.15	**26.33**	26.20
	Nl→It	16.03	16.10	16.81	**17.50**	17.03	16.81	16.89	17.09
	Nl→Ro	12.84	12.48	14.01	**14.44**	12.56	11.79	12.38	13.66
	Ro→De	12.75	12.21	13.58	**13.66**	13.02	12.62	12.96	13.63
	Ro→En	24.33	22.88	23.83	23.88	24.20	23.58	**24.65**	23.57
	Ro→Nl	13.70	14.11	15.34	**15.51**	15.11	14.65	15.29	15.19
	Mean	18.99	18.15	19.68	**19.75**	19.28	18.44	19.16	19.74
Zero-Shot	De→Nl	12.75	12.50	12.74	**12.80**	11.65	12.41	12.67	12.75
	It→Ro	9.97	9.57	10.57	10.17	10.42	9.65	**10.69**	10.32
	Nl→De	11.32	10.47	11.52	11.20	11.28	10.89	**11.63**	11.45
	Ro→It	11.69	10.82	11.51	11.40	11.66	11.42	**11.78**	11.27
	Mean	11.43	10.84	11.59	11.39	11.25	11.09	**11.69**	11.44

Analysis on Language Embedding

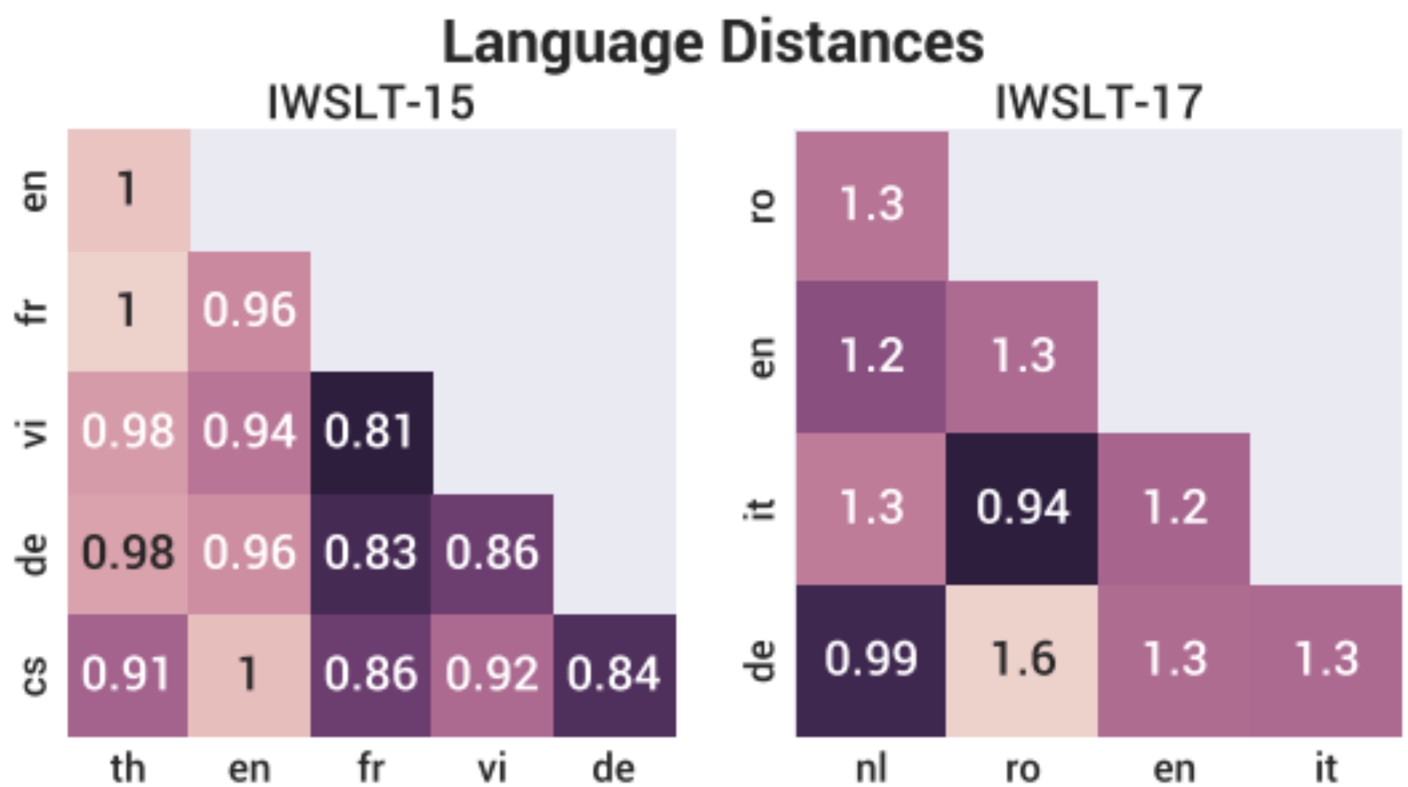

Figure 2: Pairwise cosine distance for all language pairs in the IWSLT-15 and IWSLT-17 datasets. **Darker** colors represent more similar languages.

(Self-Attentive) Autoencoder-based Universal Language Representation for Machine Translation

Carlos Escolano, Marta R. Costa-jussà and José A. R. Fonollosa
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
{carlos.escolano,marta.ruiz,jose.fonollosa}@upc.edu

Motivation

- Learning interlingual embeddings is useful

Approach

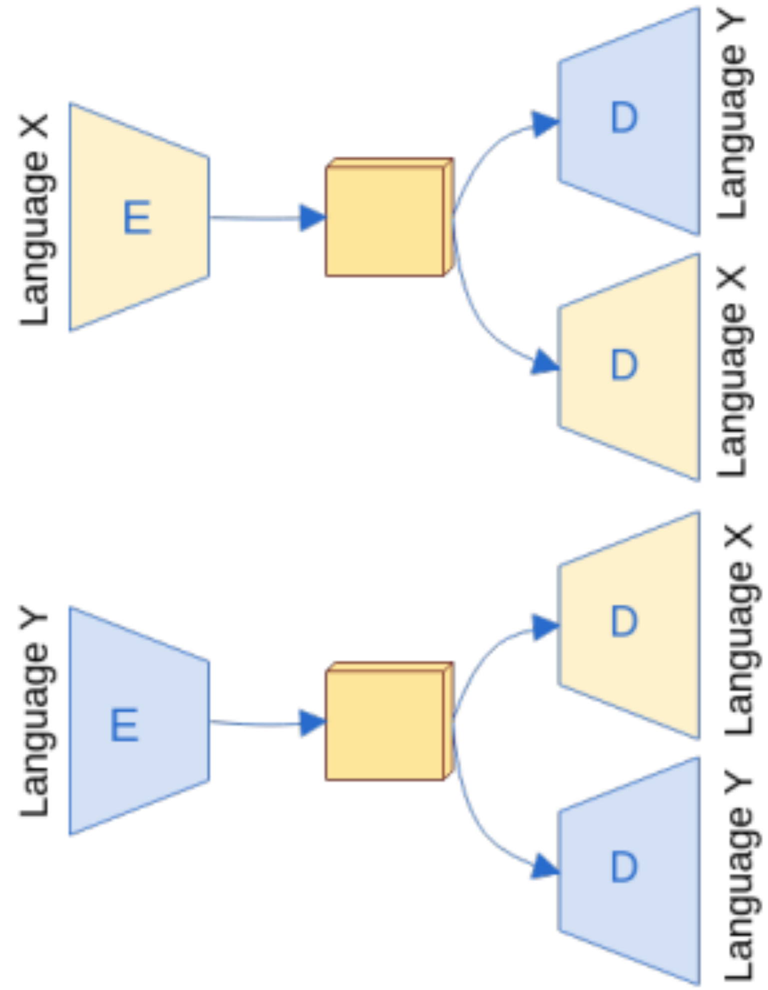

Figure 1: Architecture example. Every module is compatible with the intermediate representation.

Details

- Objective function: $Loss = L_{XX} + L_{YY} + L_{XY} + L_{YX} + d(h(X), h(Y))$
- distance measure:
  - Correlation distance:
    - $d(h(X), h(Y)) = 1 - c(h(X), h(Y))$
  - Maximum distance
    - $d(h(X), h(Y)) = \max(|h(X) - h(Y)|)$

$$c(h(X), h(Y)) = \frac{\sum_{i=1}^{n}(h(x_i - \overline{h(X)}))(h(y_i - \overline{h(Y)}))}{\sqrt{\sum_i^n (h(x_i) - \overline{h(X)})^2 \sum_i^n (h(y_i) - \overline{h(Y)})^2}}$$

Evaluation

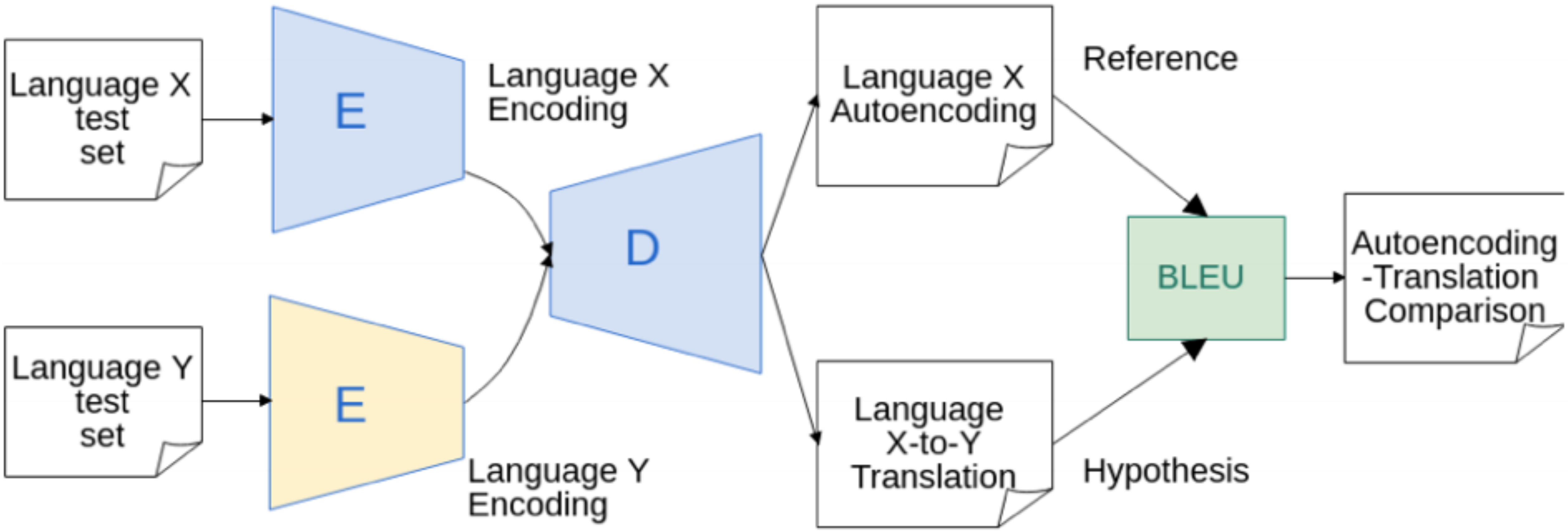

Figure 2: Pipeline of the Interlingua BLEU measure.

Experiments on Translation

Table 1: BLEU results for the different system alternatives, Transformer and different configurations of our architecture, Universal (Univ) with and without decomposed vector quatization (dvq), and correlation distance(corr) and maximum of difference(max)

	EN-TR	**TR-EN**
Transformer	8.32	12.03
Transformer dvq	2.89	8.14
Univ + corr	8.11	12.00
Univ + max	6.19	10.38
Univ + dvq + corr	7.45	7.56
Univ + dvq + max	2.40	5.24

Experiments on Embeddings

Table 2: Comparison of BLEU scores on the *univ+corr* architecture when performing as autoencoder and MT. The third column is the BLEU between autoencoder and translation outputs

Decoder	**Autoencoder**	**MT**	**A-T**
EN	63.32	12.00	11.90
TR	59.33	8.11	6.02