معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Pairwise Neural Machine Translation Evaluation

Francisco Guzmán, Shafiq Joty, Lluís Màrquez and **Preslav Nakov**
Qatar Computing Research Institute, HBKU

Motivation

Pairwise MT Evaluation

- *Learn* to differentiate *better* from *worse* translations
- State-of-the-art: structured input and preference-kernel learning (Guzmán et al., EMNLP 2014)
- Inspired by human ranking-based MT evaluation. Evaluators compare pairs of hypotheses

Input: (*Translation1, Translation2, Reference*)

Question: Is *T1* a better translation than *T2*, given *R*?

Why Neural Networks?

- State-of-the-art uses computationally expensive tree kernels (esp. at test time). NNs provide fast inference
- NNs can learn effectively from compact *semantic* and *syntactic* distributed representations
- They are highly competitive

Setting

Learning Task

- Binary classification: $y = \begin{cases} 1 & \text{if } t_1 \text{ is better than } t_2 \text{ given } r \\ 0 & \text{if } t_1 \text{ is worse than } t_2 \text{ given } r \end{cases}$
- Model: $p(y|t_1, t_2, r) = \mathrm{Ber}(y|f(t_1, t_2, r))$

$$\hat{y}_{n\theta} = f(t_1, t_2, r) = \mathrm{sig}(\mathbf{w}_{\mathbf{v}}^{\mathbf{T}} \phi(t_1, t_2, r) + b_v)$$

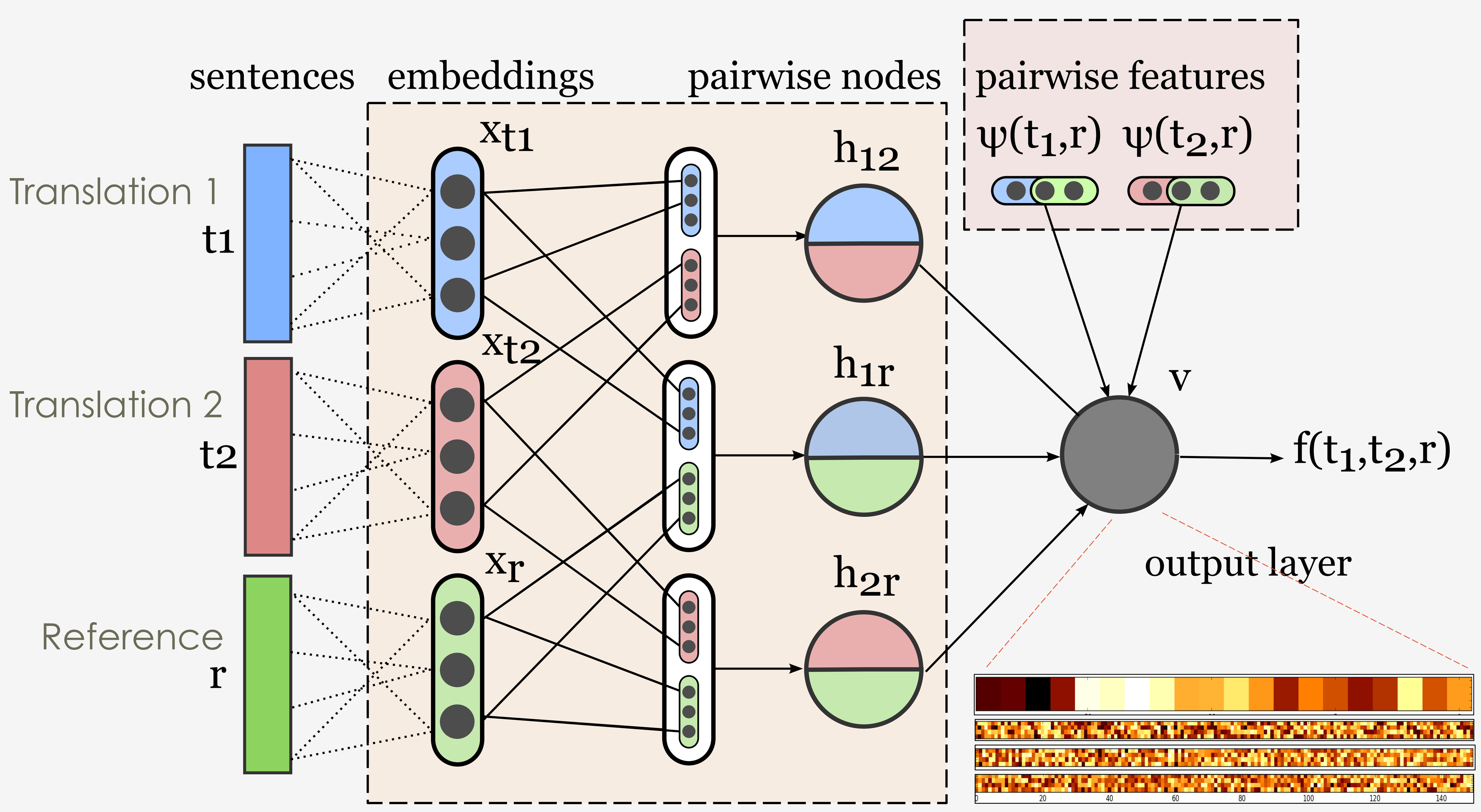
- Cost function:
  - Negative log-likelihood: $J_\theta = -\sum_n y_n \log \hat{y}_{n\theta} + (1 - y_n) \log (1 - \hat{y}_{n\theta})$
  - Kendall's-tau: $J_\theta = -\sum_n y_n \,\mathrm{sig}(-\gamma \Delta_n) + (1 - y_n)\,\mathrm{sig}(\gamma \Delta_n)$

$$\Delta = f(t_1, t_2, r) - f(t_2, t_1, r)$$

Features

- Pairwise lexical features: BLEU, METEOR, NIST, TER
- Word embeddings:
  - *Syntactic* embeddings from an RNN parser (Socher et al. 2013)
  - *Semantic* embeddings from word2vec, GloVE, COMPOSES

Neural Architecture



Trained network

Experimental Setup

- **Data (human pairwise judgments):**
  - Train: WMT11 (11,160 pairs)
  - Dev: WMT13 (5,000 pairs)
  - Test: WMT12 (3,798 pairs)
  - Features were normalized using min-max
- **Training:**
  - *Optimization:* SGD+adagrad for 10k epochs with *early stopping* and *L2 regularization*
  - *Learning rate:* 0.01
  - *Mini batch size:* 30
  - *Weight initialization:* uniform [-0.01, 0.01]
  - *Hidden layer size:* 4 with *tanh* activations
- **Evaluation:** WMT12 version of Kendall's tau

Results (Kendall Tau)

NN with Different Features

Lexical	27.06
Lex+Syntax	28.51
Lex+Semantics	29.07
Lex+Syn+Semantics	**29.70**

Other Metrics

BLEU	18.46
METEOR	23.56
DiscoTK	*30.50*
Kernel Approach	23.70

Different Semantic Embeddings

Source	Alone	Comb.
GW25	10.01	29.70
GW300	9.66	**29.90**
CC-300-42B	12.16	29.68
CC-300-840B	11.41	**29.88**
Word2Vec300	7.72	29.13
COMPOSES400	12.35	28.54

BLEU Components + Embeddings

BLEU	18.46
BLEUCOMP	19.75
+SYN25	23.70
+GW25	24.92
+SYN25+GW25	**26.15**

Deep vs. Flat NN

Single-layer	29.10
Multi-layer	**29.70**

Logistic vs. Kendall Cost

Logistic	29.70
Kendall	29.53
Logistic + Kendall	**29.92**

Conclusion and Future Work

- **Proposed a novel NN framework for MT evaluation:**
  - Flexible in incorporating different sources of information
  - Results are additive w.r.t. the sources of information
  - Enables fast inference
  - Achieves state-of-the-art results
- **Future work:**
  - Add source-sentence information
  - Use the NN framework for:
    - re-ranking
    - quality estimation
    - system combination