Recent Developments in Multilingual Modeling

Graham Neubig
4/10/2017

Carnegie Mellon University

Today's Agenda

- What if we want to **translate between languages**?
- What if we want to **embed things in different languages**?
- What if we want to **use language diversity** to help us learn?
- How can we incorporate **multi-modal information** in multi-lingual tasks?

Note: this is a non-exhaustive survey focusing on recent work. Use it as a starting point.

Translation

Statistical Machine Translation

F = *kare wa ringo wo tabeta .*

E = He ate an apple .

Probability model: P(*E*|*F*;*Θ*)

Parameters

Basic Idea: Calculate Probability of Next Word and argmax

F = *watashi wa kouen wo shiteimasu*

$P(e_1=I\|F) = 0.96$	$P(e_1=talk\|F) = 0.03$ $P(e_1=it\|F) = 0.01$	...	$e_1=$ I
$P(e_2=am\|F, e_1) = 0.9$	$P(e_2=was\|F, e_1) = 0.09$	...	$e_2=$ am
$P(e_3=giving\|F, e_{1,2}) = 0.4$	$P(e_3=talking\|F, e_{1,2}) = 0.3$ $P(e_3=presenting\|F, e_{1,2}) = 0.03$	...	$e_3=$ giving
$P(e_4=a\|F, e_{1,3}) = 0.8$	$P(e_4=my\|F, e_{1,3}) = 0.15$	...	$e_4=$ a
$P(e_5=talk\|F, e_{1,4}) = 0.4$ $P(e_5=presentation\|F, e_{1,4}) = 0.3$	$P(e_5=lecture\|F, e_{1,4}) = 0.15$ $P(e_5=discourse\|F, e_{1,4}) = 0.1$	...	$e_5=$ talk
$P(e_6=</s>\|F, e_{1,5}) = 0.8$	$P(e_6=now\|F, e_{1,5}) = 0.1$	...	$e_6=$ </s>

In Other Words, Translation Can be Formulated As:

A Probabilistic Model

$$P(E|F) = \prod_{i=1}^{I+1} P(e_i|F, e_1^{i-1})$$

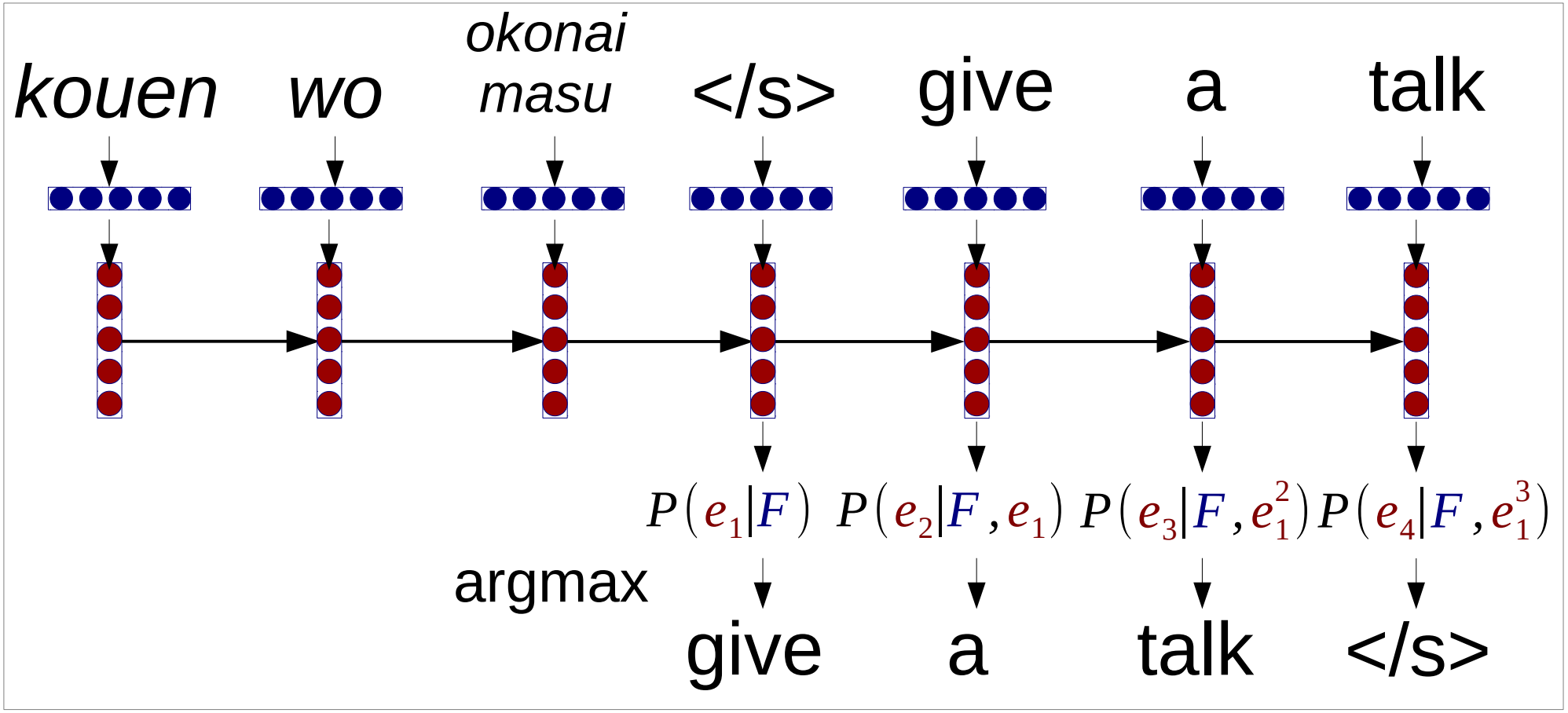
A Translation Algorithm

$i = 0$
while e_i is not equal to "</s>":
  i ← i+1
  e_i ← argmax$_e$ P(e_i|F, $e_{1,i-1}$)

Big question: How to estimate this probability?

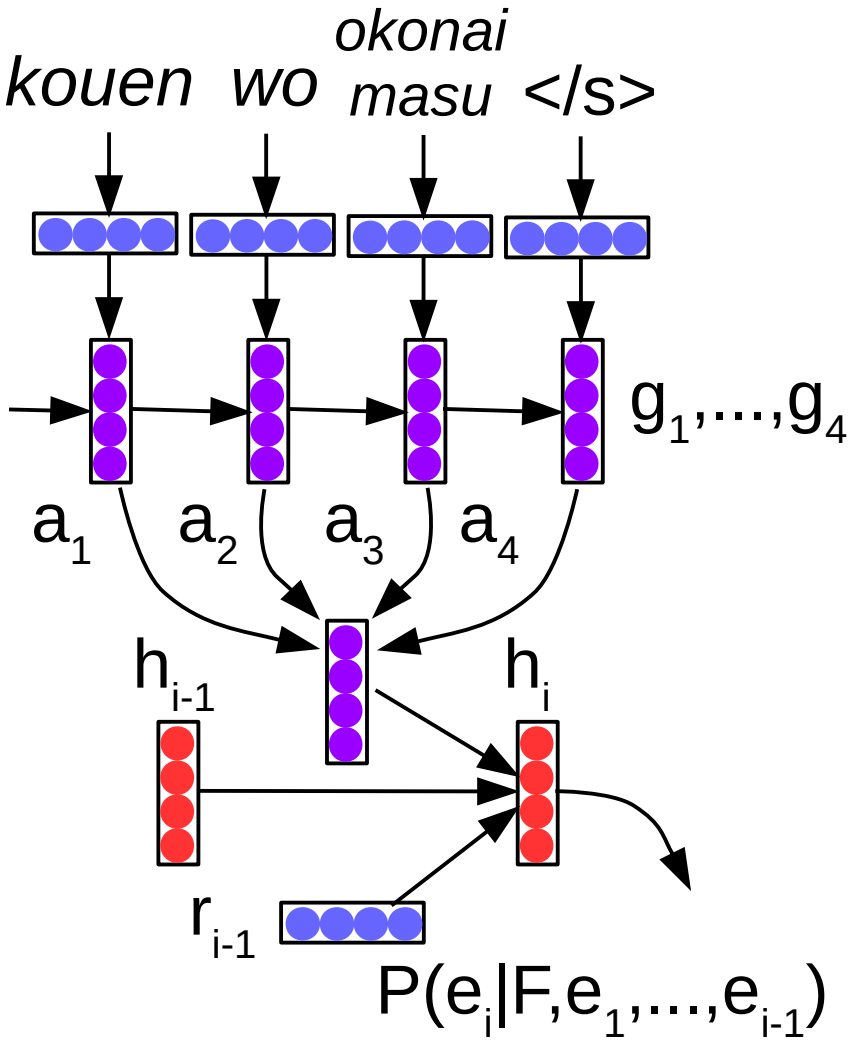
Encoder-Decoder Model [Sutskever+ 14]

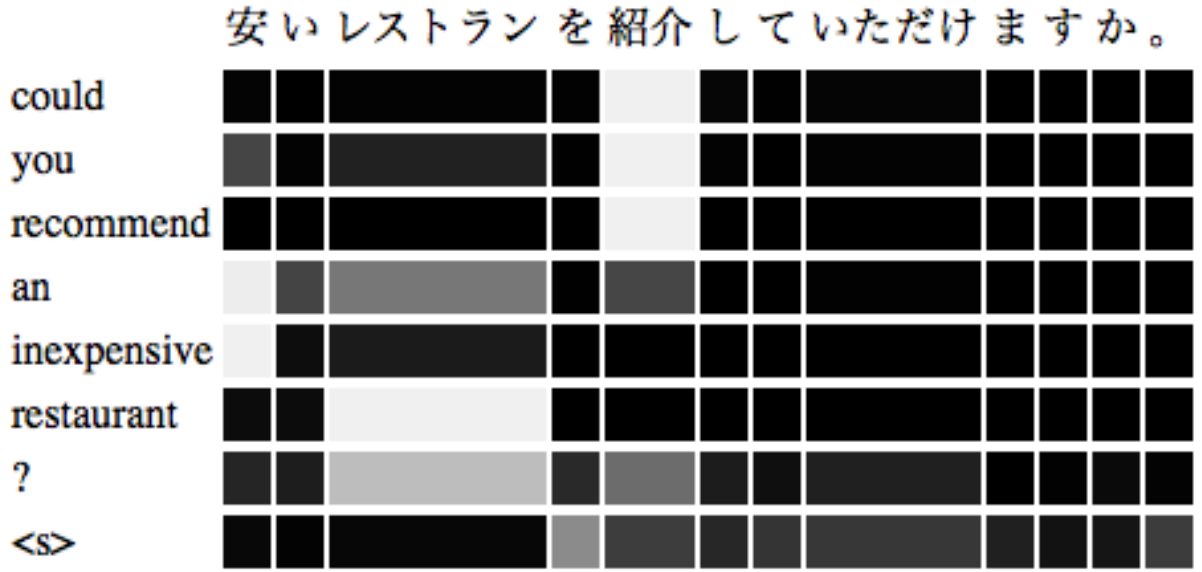
- Estimate $P(e_i|F, e_1^{i-1})$ with long short-term memory (LSTM) recurrent neural nets
- Encoder generates vector representation of source
- Decoder predicts probabilities and takes argmax



Attentional Nets [Bahdanau+ 15]

- While translating, decide which word to "focus" on





Exciting Results!

- IWSLT 2015:
  Best results on de-en
- WMT 2016:
  Best results on most language pairs
- WAT 2016:
  Best results on most language pairs
- Commercial deployment in 2016

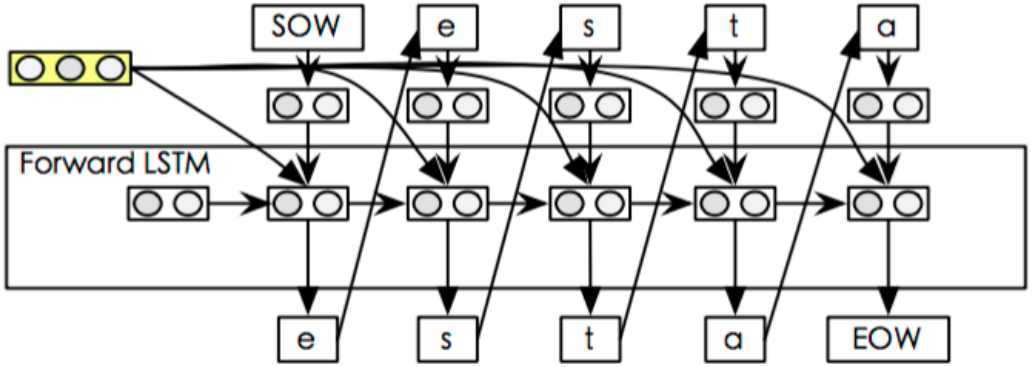
What Do/Don't We Know in 2017?

Modeling

- **Modeling:** How do we define P(E|F;Θ)?
- Attentional nets→new standard? [Bahdanau+15, Luong+15]

Character-based, Subword-based Models

- Character-based word representations [Ling+2015]



- Subword segmentations models [Sennrich+2016]

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
WDict	Forschungsinstitute
C2-50k	Fo|rs|ch|un|gs|in|st|it|ut|io|ne|n
BPE-60k	Gesundheits|forsch|ungsinstitu|ten
BPE-J90k	Gesundheits|forsch|ungsin|stitute

- Purely character-based translation [Chung+2016]

Better Alignment/Representation

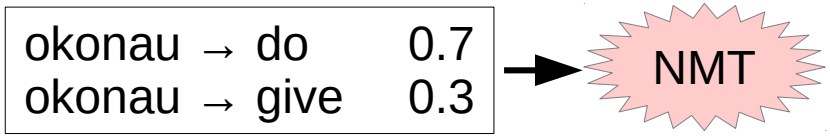
- Models of coverage, reordering, etc. [Cohn+16]

watshi	*wa*	*kouen*	*wo*	*okonau*
OK	OK	TODO	TODO	TODO

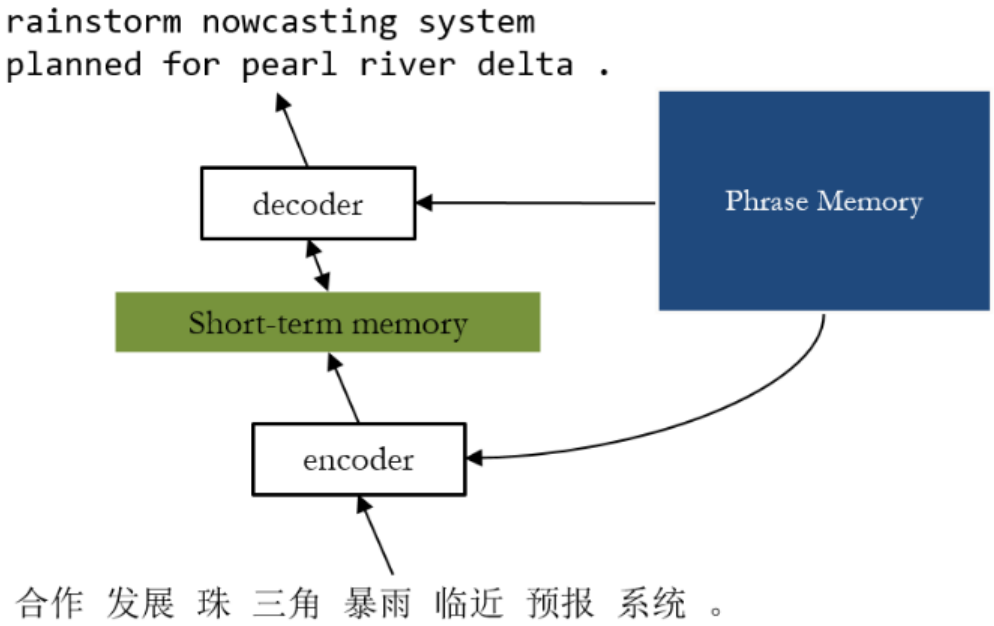
- Better representation
  - Convolutional neural networks [Kalchbrenner+13]
  - Tree-based networks [Eriguchi+16]
  - Multiple time-scale networks [Chung+16,Duong+16]

Incorporating External Knowledge

- External translation lexicons [Arthur+16]



- Phrase tables [Tang+16]

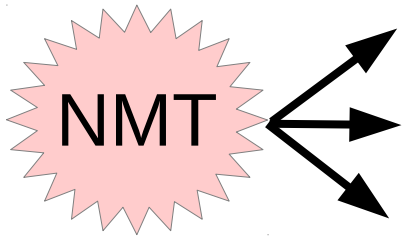


Learning

- **Learning:** How do we learn Θ?
- NMT standard → maximum likelihood over a bilingual corpus

Optimizing Evaluation Metrics

- Directly optimize BLEU, METEOR, etc.



	eval	gradient
giving a talk	0.4	small negative
talk a do	0.1	large negative
I am giving a talk	1.0	large positive

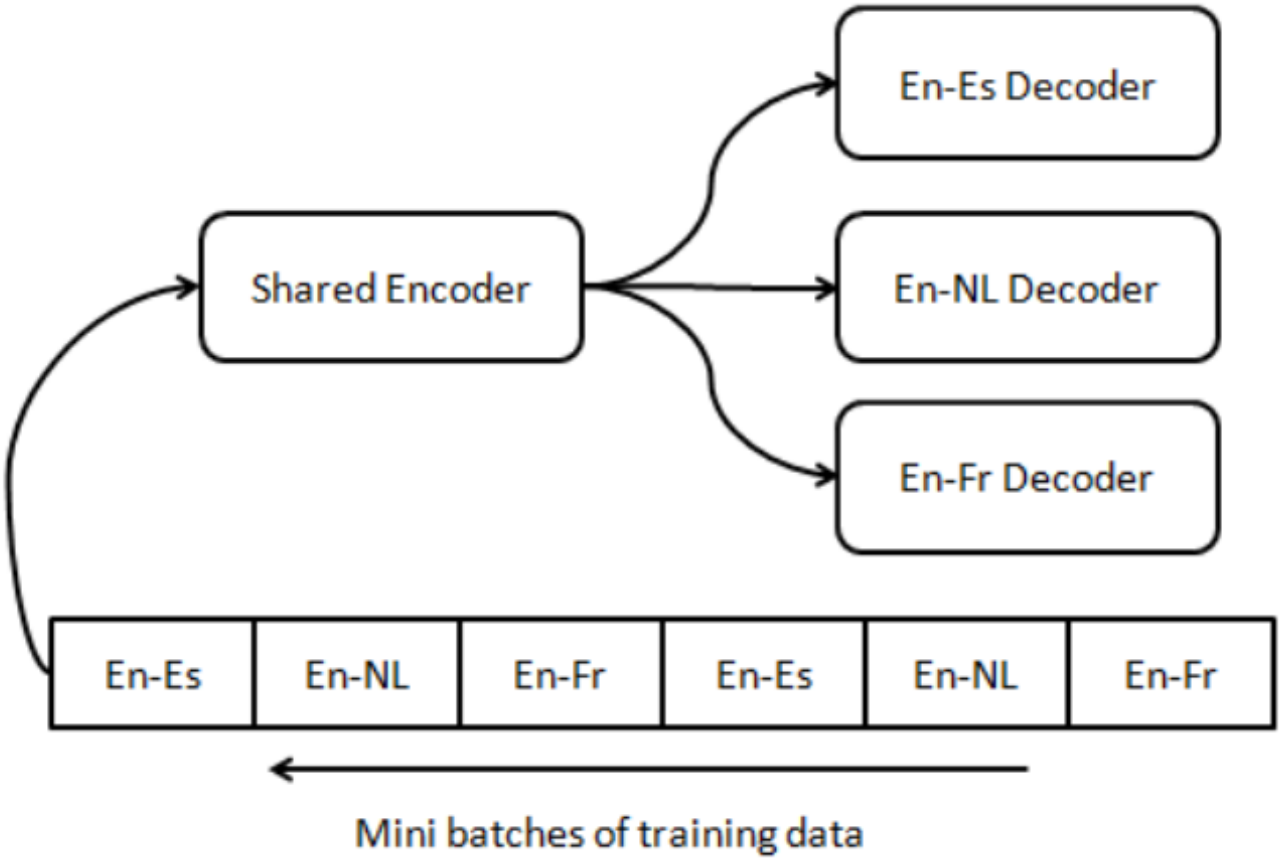
- Methods
  - Reinforcement learning [Ranzato+15]
  - Minimum risk training [Shen+16]
  - Beam search optimization [Wiseman+16]

Multilingual Learning

- How do we utilize information from all of the languages in the world to create better systems?

Separate Encoder/Decoders [Dong+15, Firat+16]

- One encoder or decoder for each language



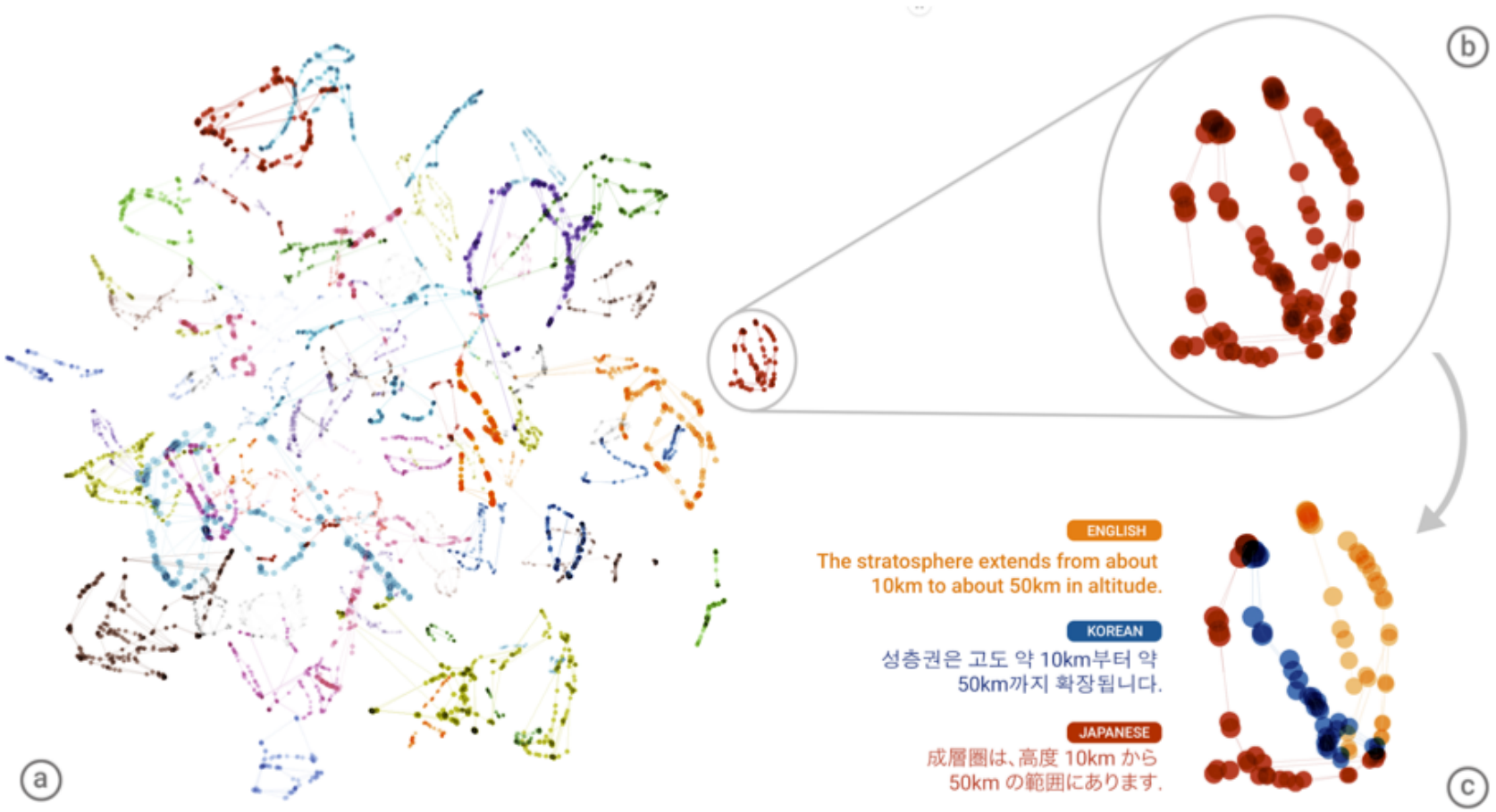
Shared Encoder/Decoder [Johnson+16,Ha+16]

- Share the encoders and decoders for each, add a symbol for the target language

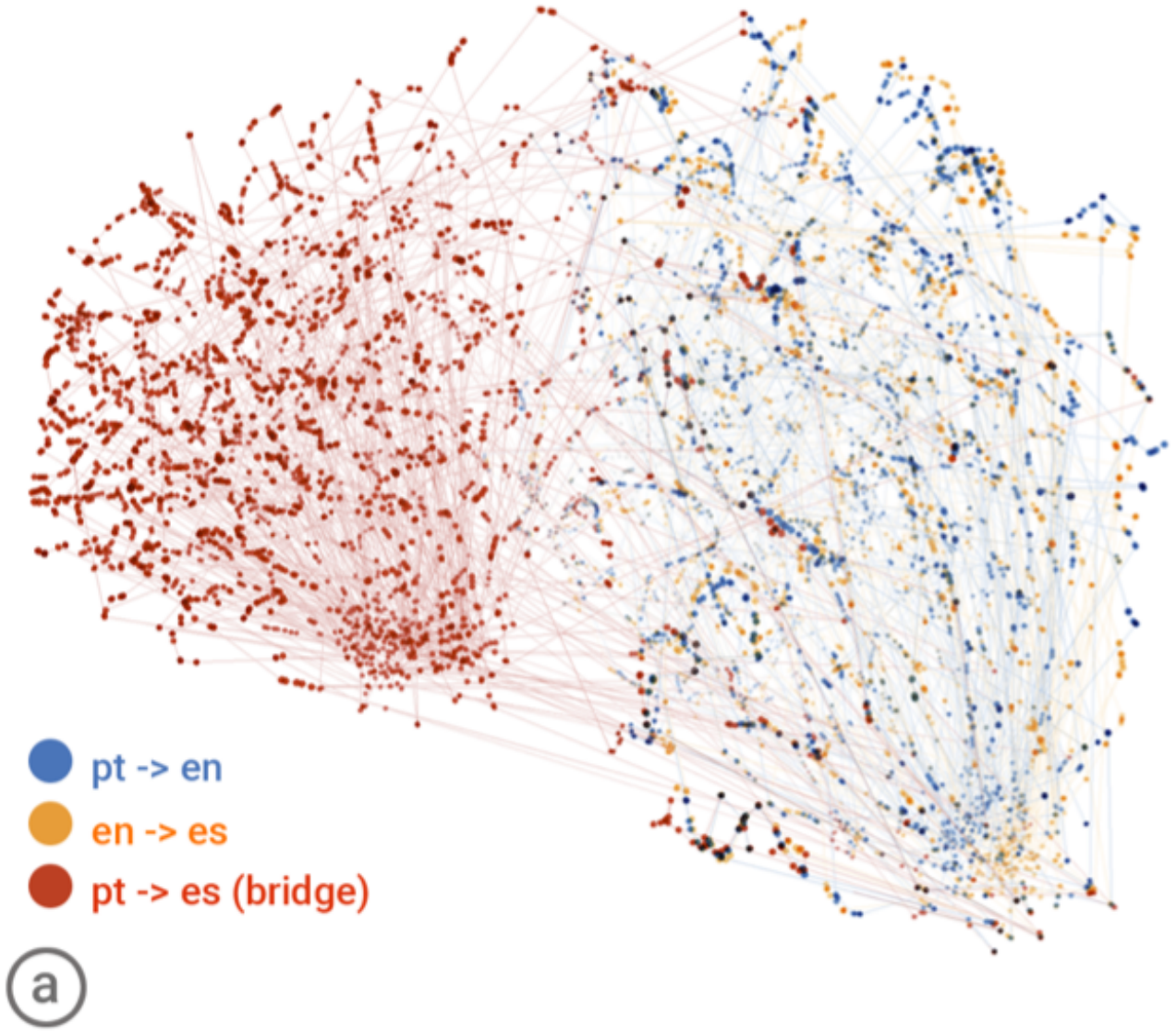
Hello, how are you? -> ¿Hola como estás?

<es> Hello, how are you? -> ¿Hola como estás?

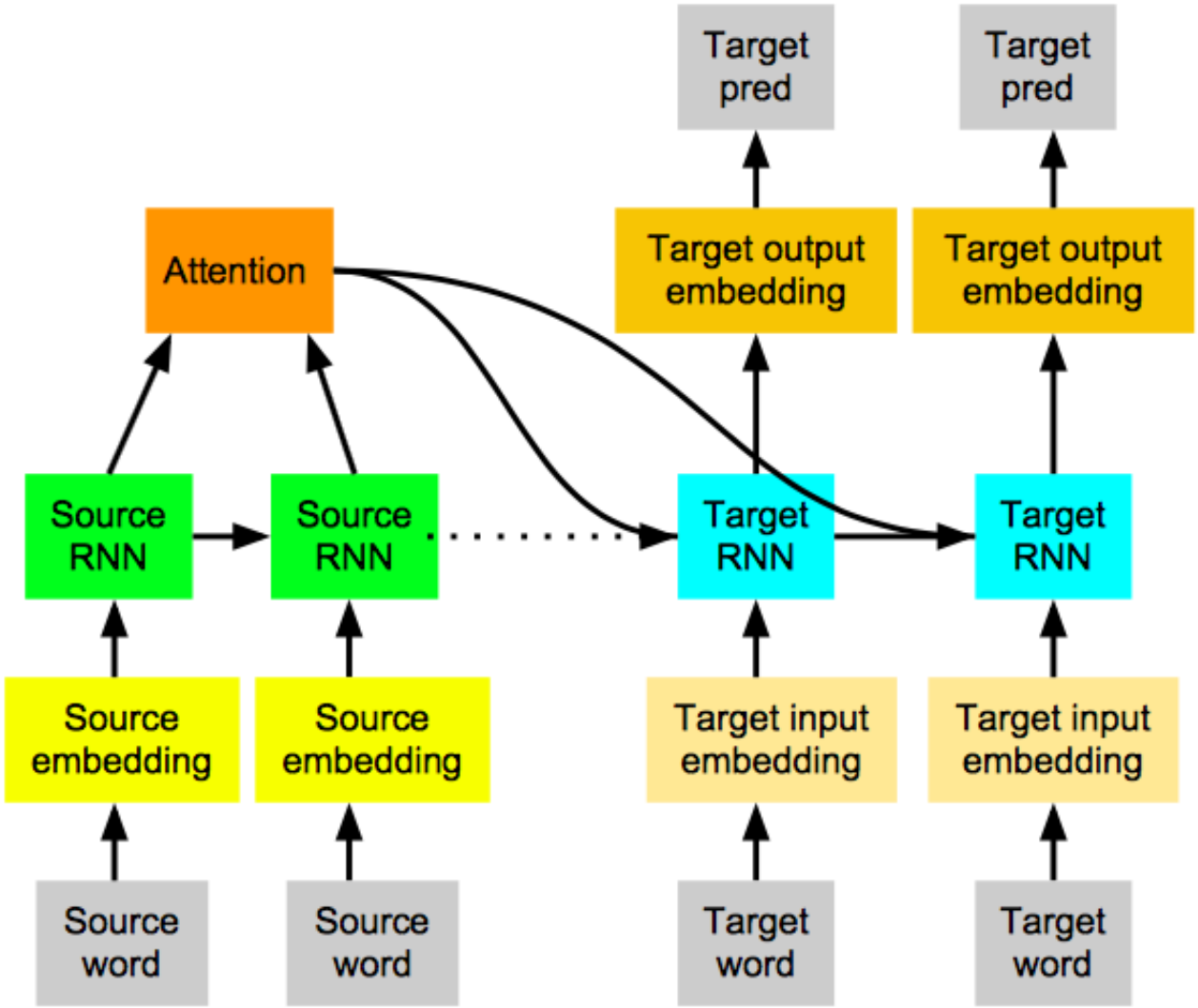
Shared Semantic Space?
[Johnson+ 16]



Or Not...



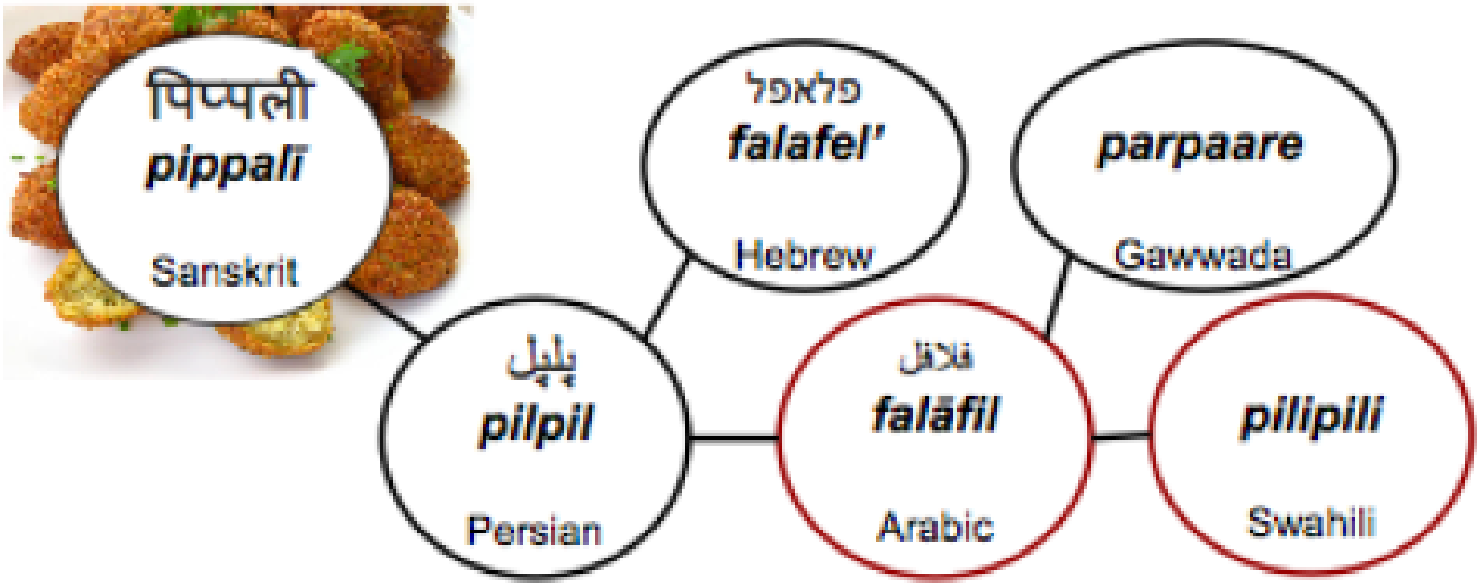
Transfer Learning [Zoph+16]



Transfer Learning [Zoph+16]

Source Embeddings	Source RNN	Target RNN	Attention	Target Input Embeddings	Target Output Embeddings	Dev BLEU ↑	Dev PPL ↓
🔒	🔒	🔒	🔒	🔒	🔒	0.0	112.6
🔓	🔒	🔒	🔒	🔒	🔒	7.7	24.7
🔓	🔓	🔒	🔒	🔒	🔒	11.8	17.0
🔓	🔓	🔓	🔒	🔒	🔒	14.2	14.5
🔓	🔓	🔓	🔓	🔒	🔒	**15.0**	13.9
🔓	🔓	🔓	🔓	🔓	🔒	14.7	**13.8**
🔓	🔓	🔓	🔓	🔓	🔓	13.7	14.4

Cross-lingual Bridging [Tsvetkov+16]

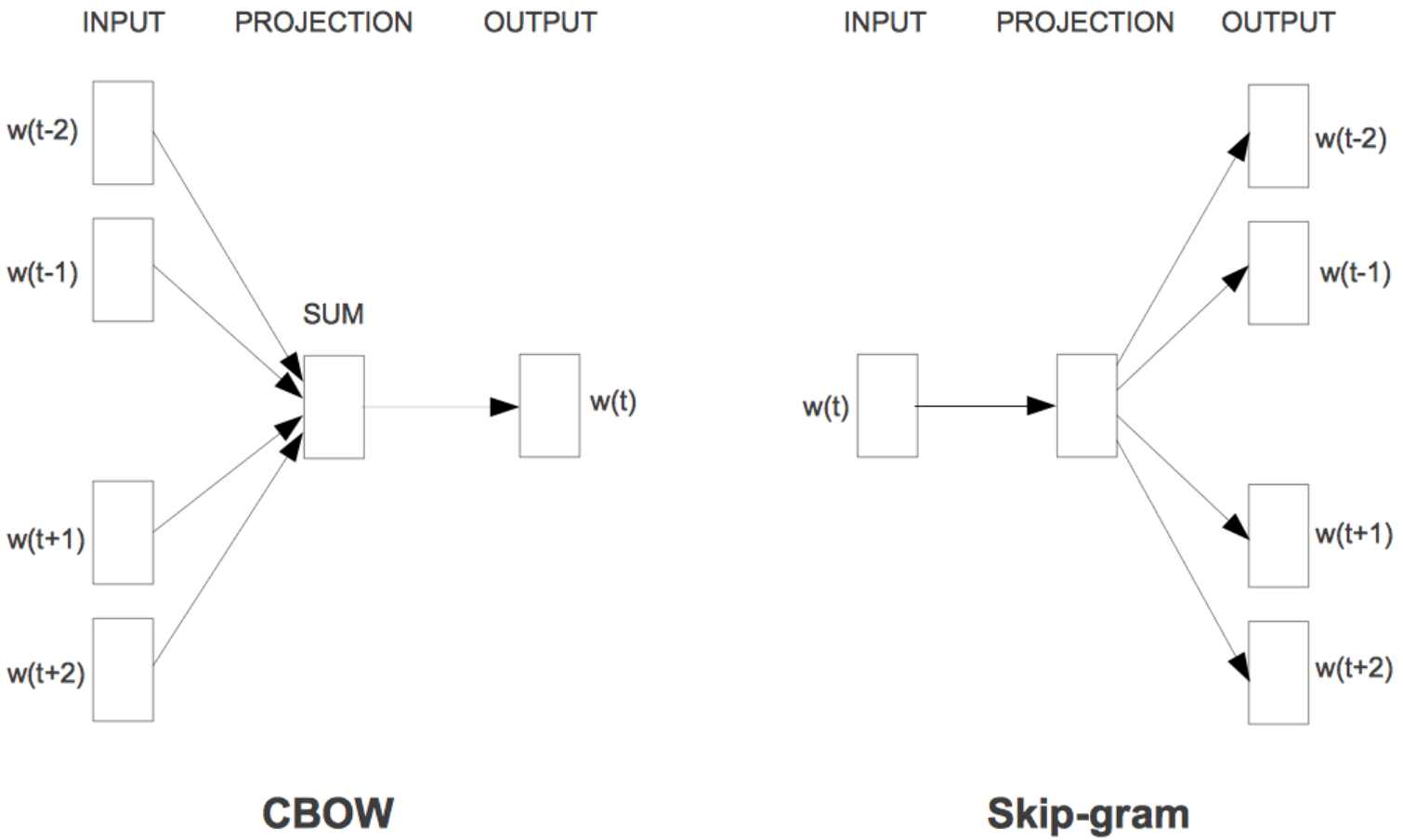


- translating out-of-vocabulary words using cross-lingual bridges
- modeling cross-lingual phonology, morphology, syntax, semantics
- transfer learning for resource-poor languages

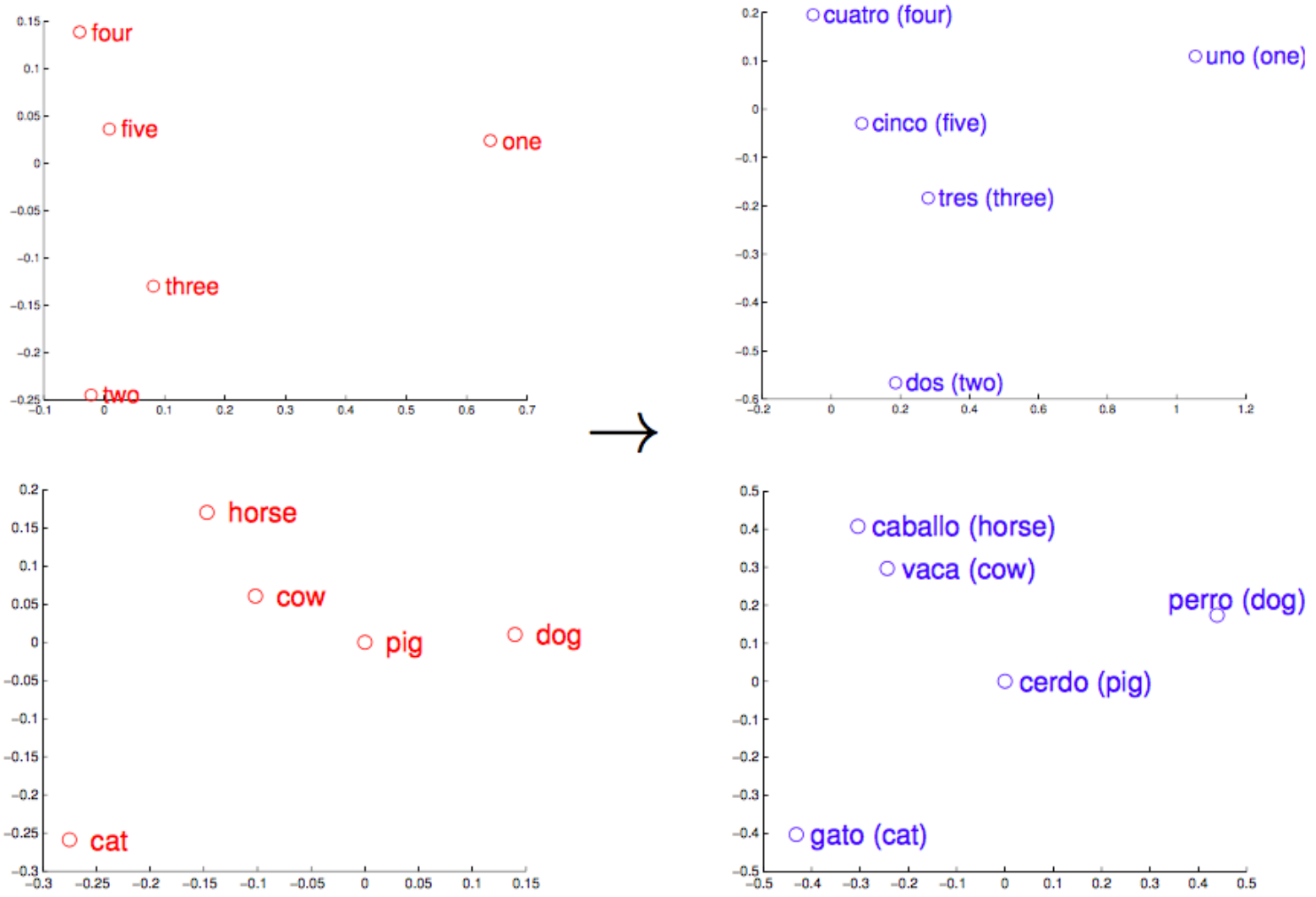
Multi-lingual Embedding

Word Embedding

- e.g. skip-gram, CBOW model [Mikolov+13]

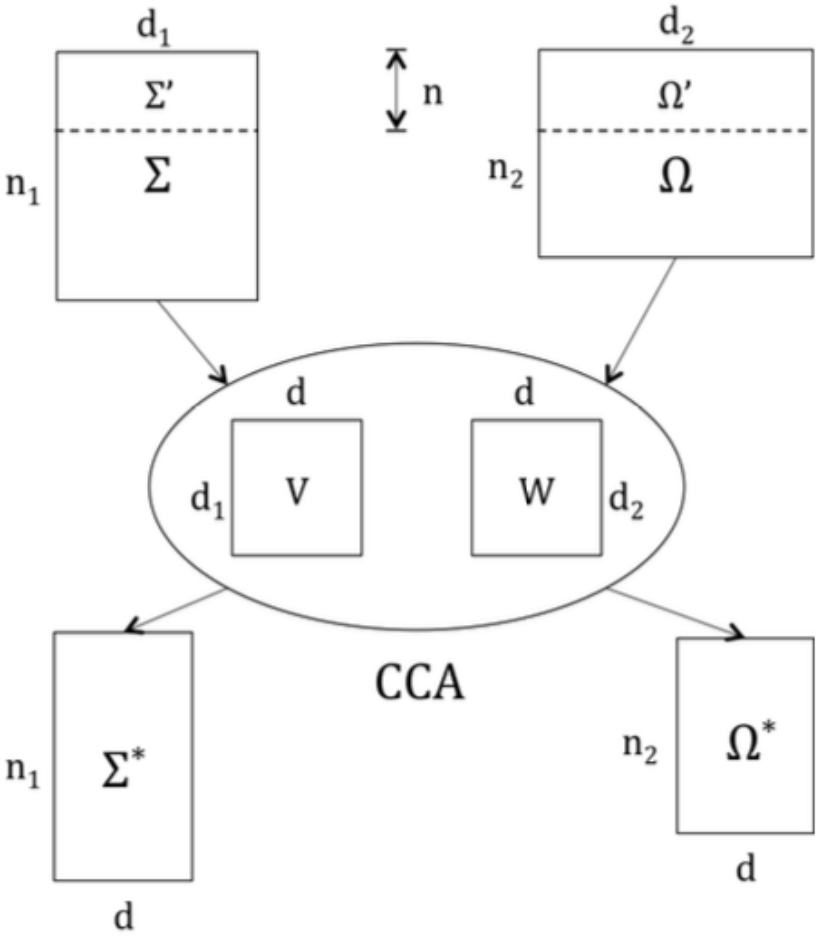


Cross-lingual Consistency in Word Embeddings [Mikolov+13]



Bilingual Embeddings: Maximizing Correlation [Faruqui+14]

- Take existing embeddings, try to project them onto a space that matches bilingual dictionary



Translation-invariant Matrix Decomposition [Huang+15]

- Perform matrix decomposition to minimize word-context and cross-lingual objectives

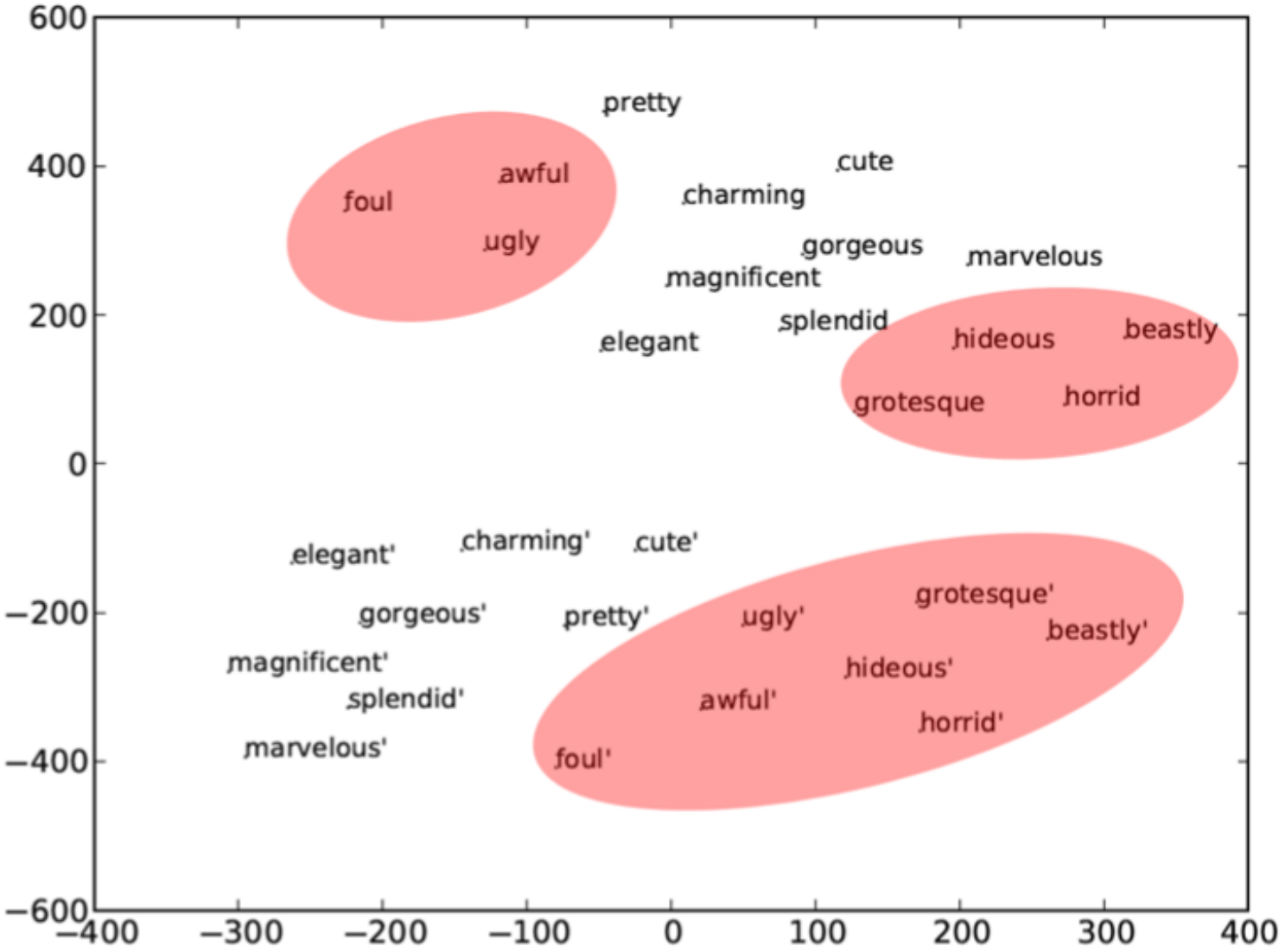
- $\mathbf{X}$: a single multilingual cooccurrence matrix (with all the $M_1 + M_2$ words as the rows, and $N_1 + N_2$ contexts as columns). Entries in this matrix specify the cooccurrence between a word in any language and a context in any language.
- $\mathbf{D}_1$: a word dictionary matrix (with all the $M_1 + M_2$ English and Spanish words as both rows and columns). Entries in this matrix specify which words are translations of which other words, and is generally block-normalized, so that (e.g.) each Spanish word has a probability distribution over English words.
- $\mathbf{D}_2$: a context dictionary matrix (with all the $N_1 + N_2$ English and Spanish contexts as both rows and columns). This is similar to $\mathbf{D}_1$ in its construction.

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \|\mathbf{D}_1\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \quad (1)$$
$$\|\mathbf{X}\mathbf{D}_2^T - \mathbf{U}\mathbf{V}^T\|_F^2 + \|\mathbf{D}_1\mathbf{X}\mathbf{D}_2^T - \mathbf{U}\mathbf{V}^T\|_F^2.$$

Massively Multi-lingual Embedding [Ammar+16]

- Not just 2, but many languages
- **MultiCluster:** cluster together words in different languages, learn over clusters
- **MultiCCA:** use English as an anchor language, project all language vectors to English

Benefit of Learning Embeddings Multilingually [Faruqui+14]

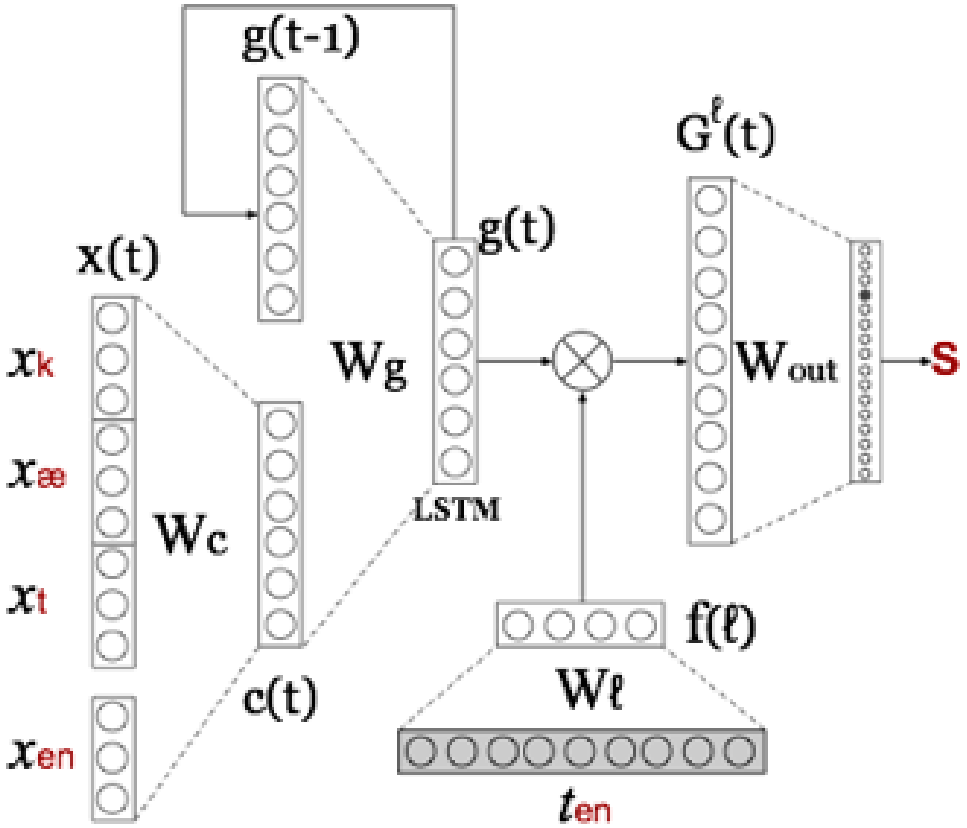


Quantitative Results/Applications

	Task	multiCluster	multiCCA	multiSkip	invariance
extrinsic metrics	dependency parsing	**61.0** [70.9]	58.7 [69.3]	57.7 [68.9]	59.8 [68.6]
	document classification	**92.1** [48.1]	**92.1** [62.8]	90.4 [45.7]	91.1 [31.3]
intrinsic metrics	monolingual word similarity	38.0 [57.5]	43.0 [71.0]	33.9 [55.4]	**51.0** [23.0]
	multilingual word similarity	58.1 [74.1]	**66.6** [78.2]	59.5 [67.5]	58.7 [63.0]
	word translation	43.7 [45.2]	35.7 [53.2]	46.7 [39.5]	**63.9** [30.3]
	monolingual QVEC	10.3 [98.6]	**10.7** [99.0]	8.4 [98.0]	8.1 [91.7]
	multiQVEC	**9.3** [82.0]	8.7 [87.0]	8.7 [87.0]	5.3 [74.7]
	monolingual QVEC-CCA	62.4 [98.6]	63.4 [99.0]	58.9 [98.0]	**65.8** [91.7]
	multiQVEC-CCA	43.3 [82.0]	41.5 [87.0]	36.3 [75.6]	**46.2** [74.7]

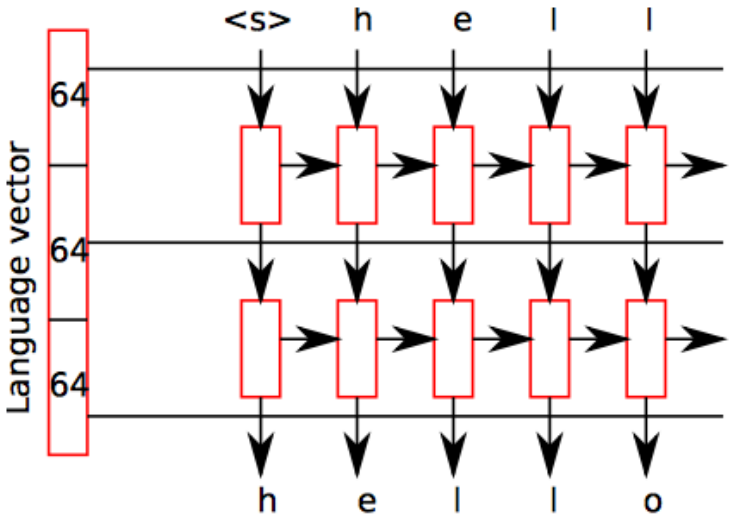
Learning from Multiple Languages for NLP

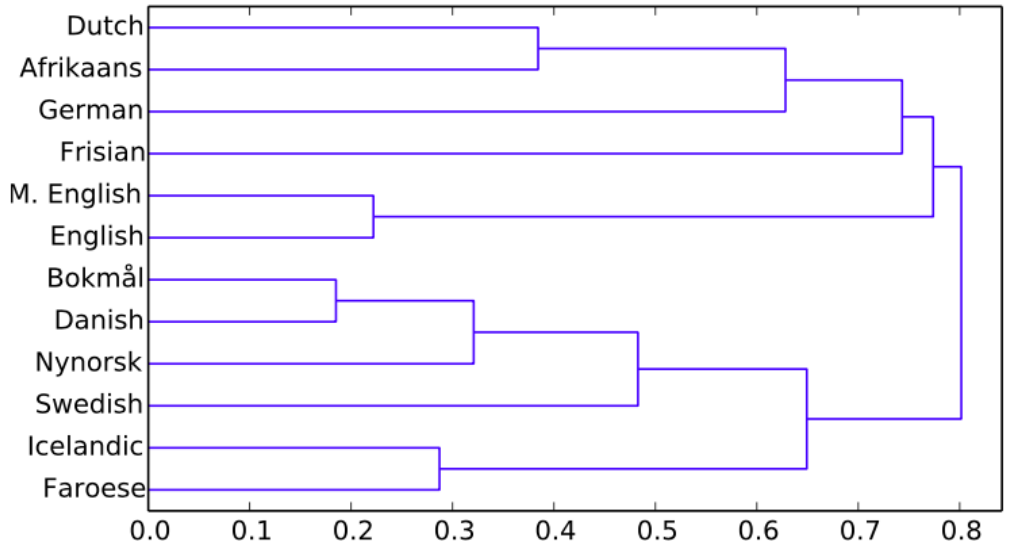
Polyglot Language Models [Tsvetkov+16]

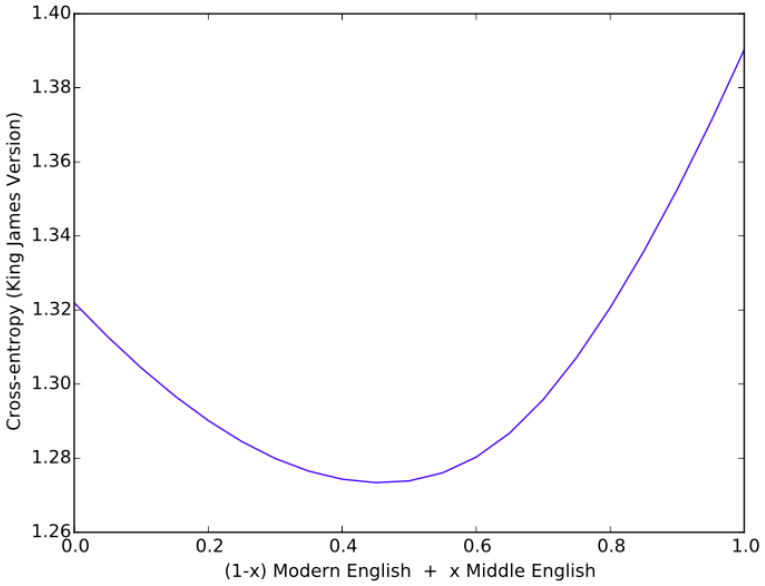


- multilingual phonetic language models
- multilingual distributed representations of words and phrases

Investigating Language Vectors [Östling+17]

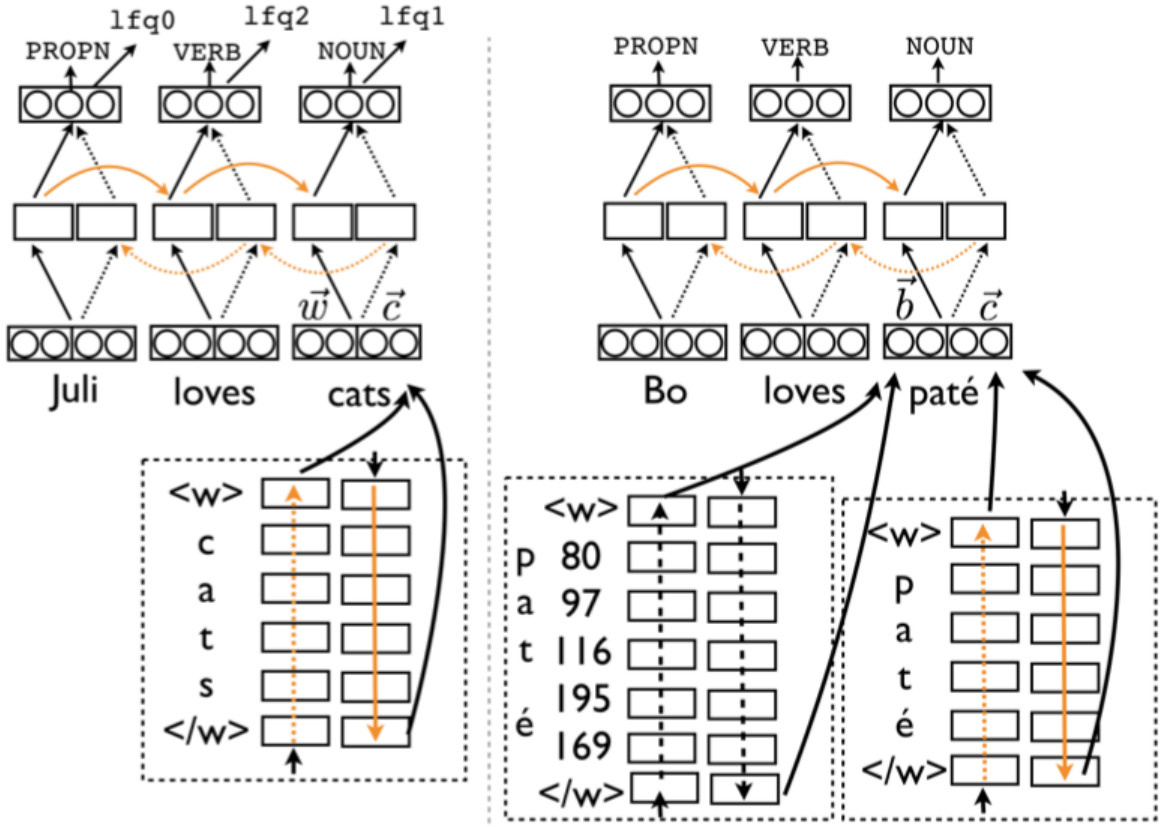






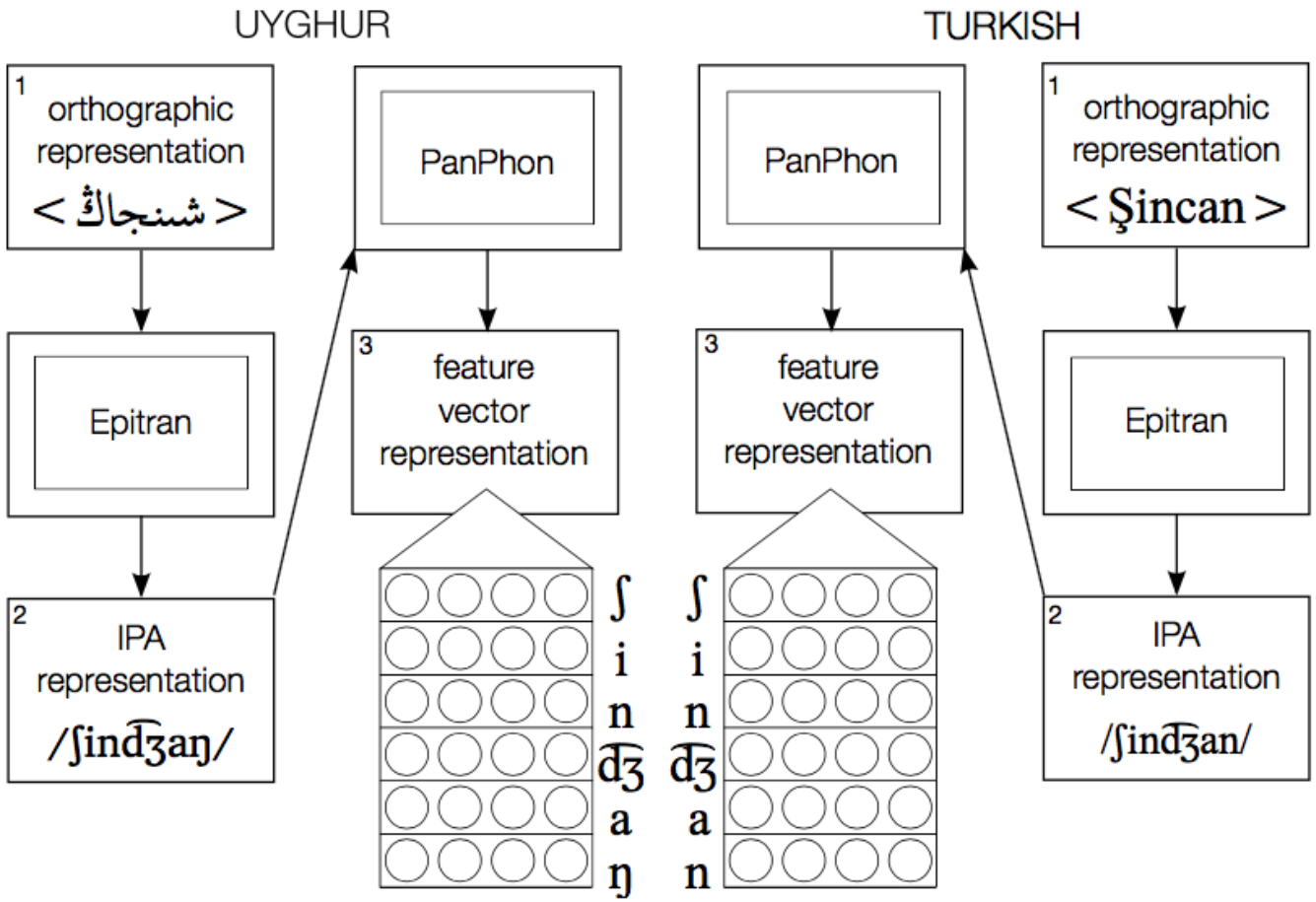
Multi-lingual POS Tagging [Plank+16]

- Multilingual training with byte and character inputs
- Predict frequency bin as auxiliary loss



Multilingual Named Entity Recognition [Bharadwaj+16]

- Use how words sound to map between languages



Multi-lingual Parsing [Ammar+16]

- Train a single parser for many languages
- Use multi-lingual word embeddings

LAS	target language							average
	de	en	es	fr	it	pt	sv	
monolingual	**79.3**	**85.9**	83.7	81.7	88.7	85.7	83.5	84.0
MALOPA	70.4	69.3	72.4	71.1	78.0	74.1	65.4	71.5
+lexical	76.7	82.0	82.7	81.2	87.6	82.1	81.2	81.9
+language ID	78.6	84.2	83.4	**82.4**	**89.1**	84.2	82.6	83.5
+fine-grained POS	78.9	85.4	**84.3**	**82.4**	89.0	**86.2**	**84.5**	**84.3**

Multi-modal Translation Models

Speech Translation

ASR

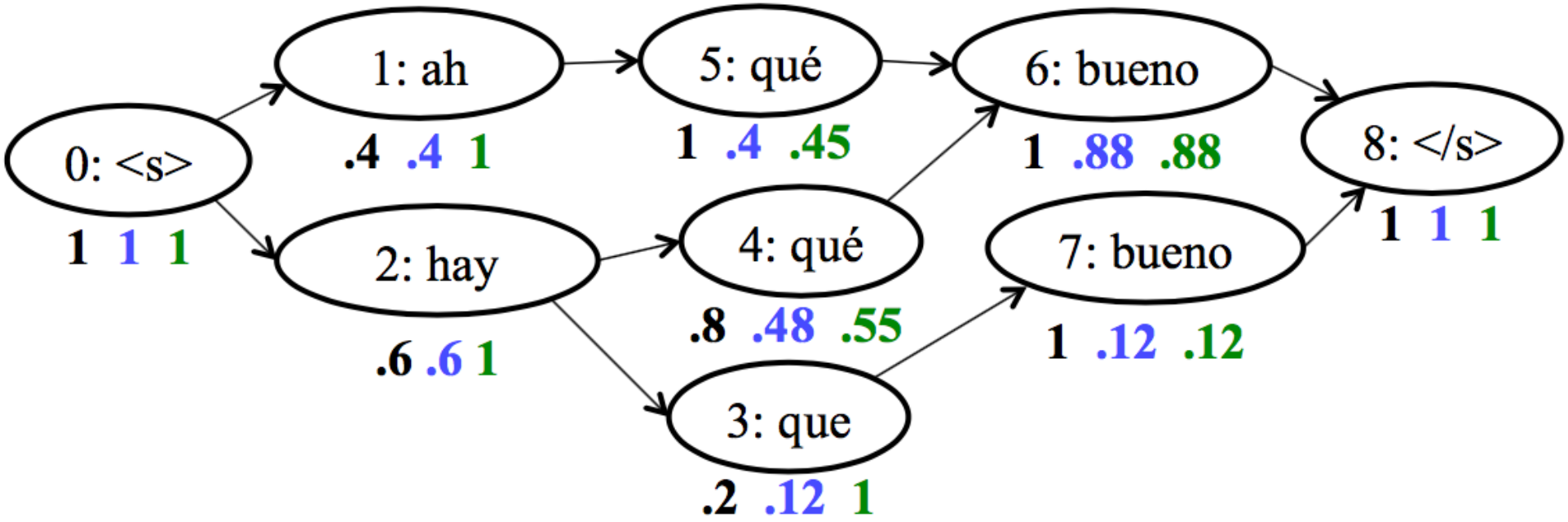
こんにちは、駅はどこですか？

MT

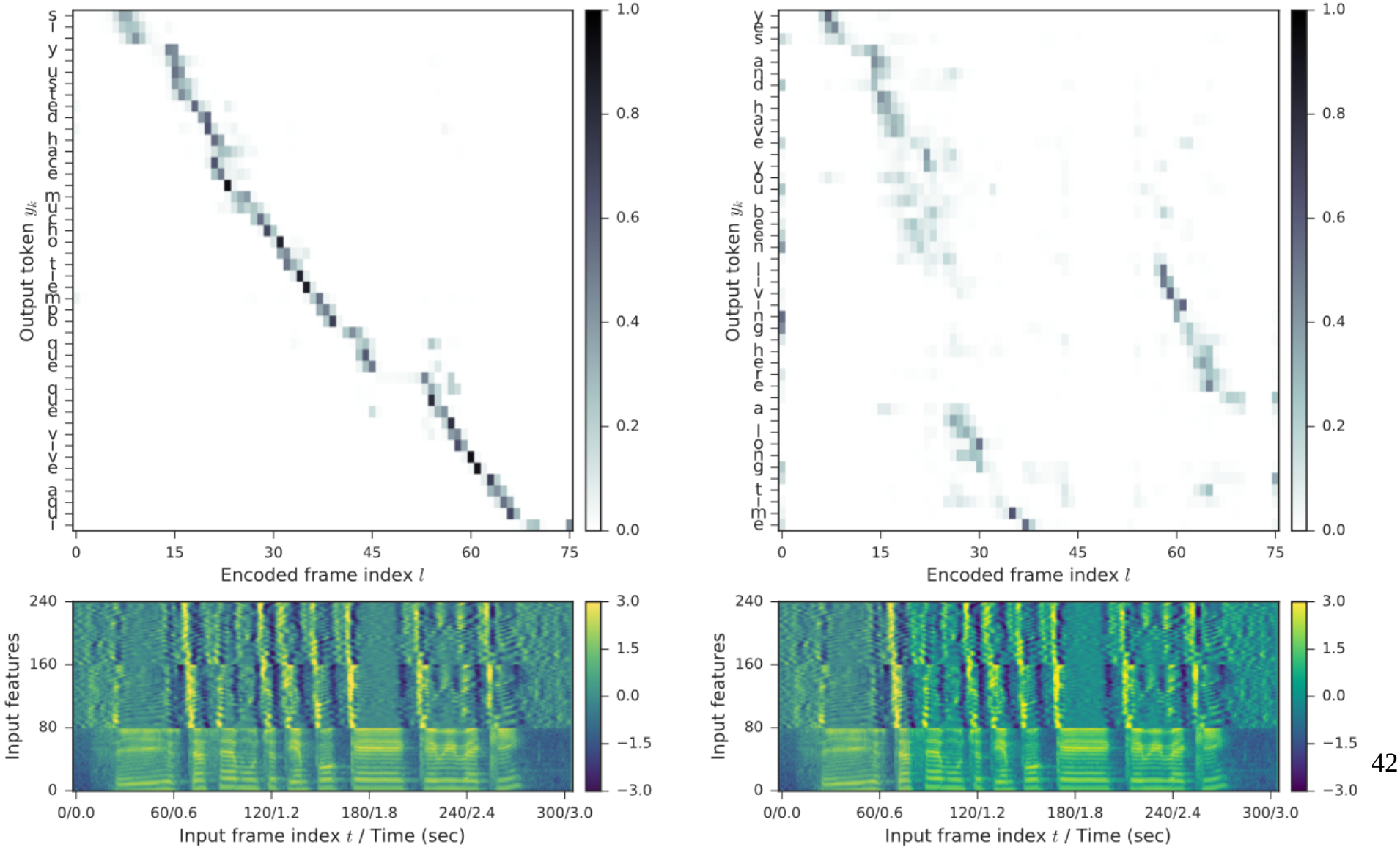
Hello, where is the station?

TTS

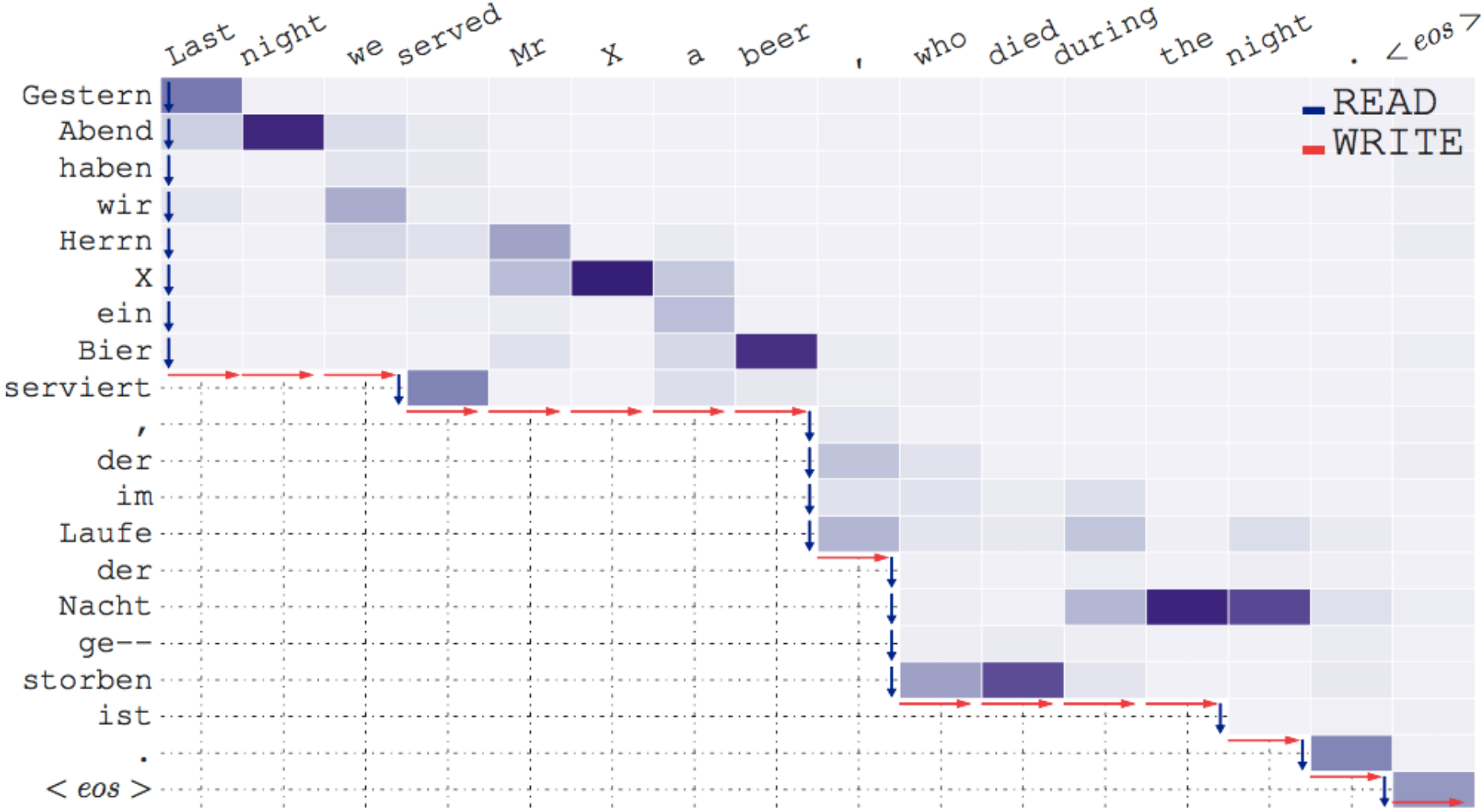
Incorporating Uncertainty of Speech [Sperber+17]



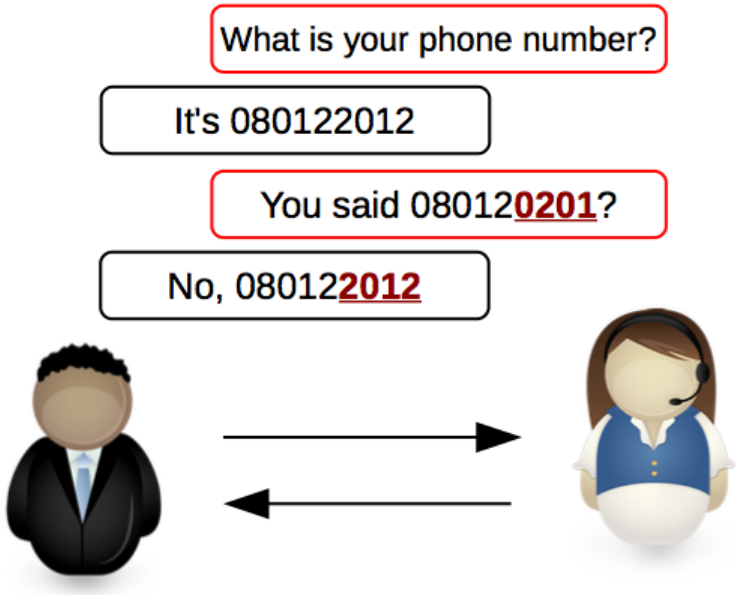
Direct Speech-to-text Translation [Weiss+17]



Real-time Simultaneous Translation [Gu+17]



Translating Prosodic Features [Do+17]



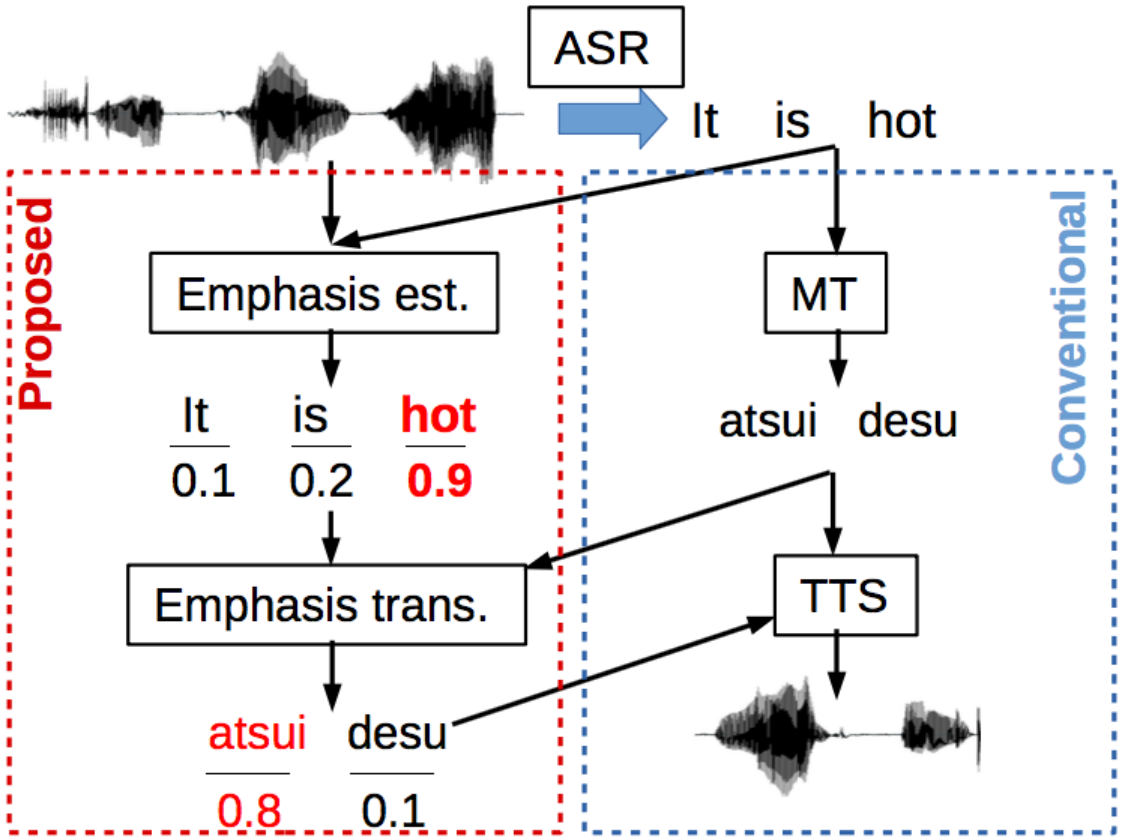
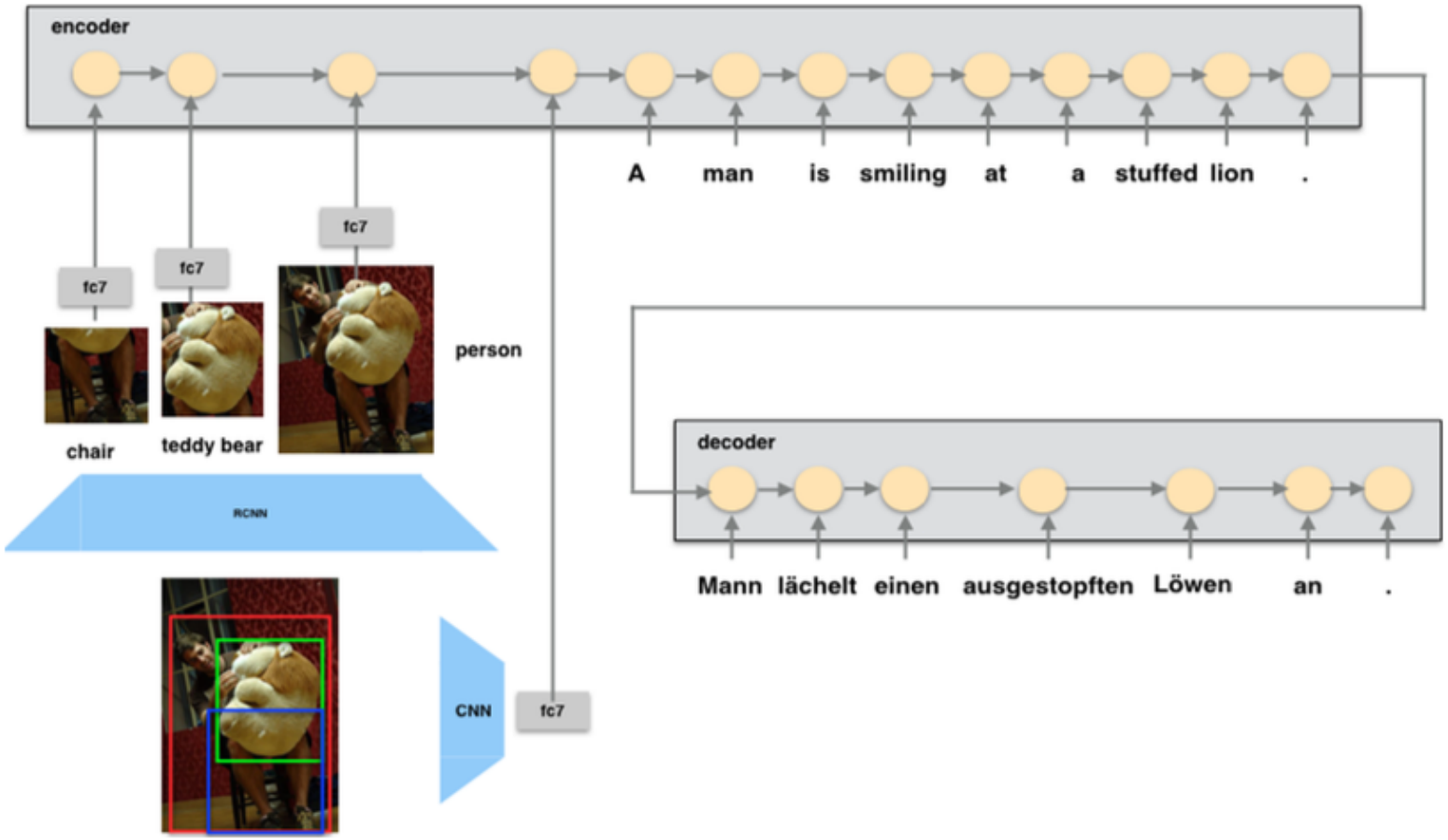
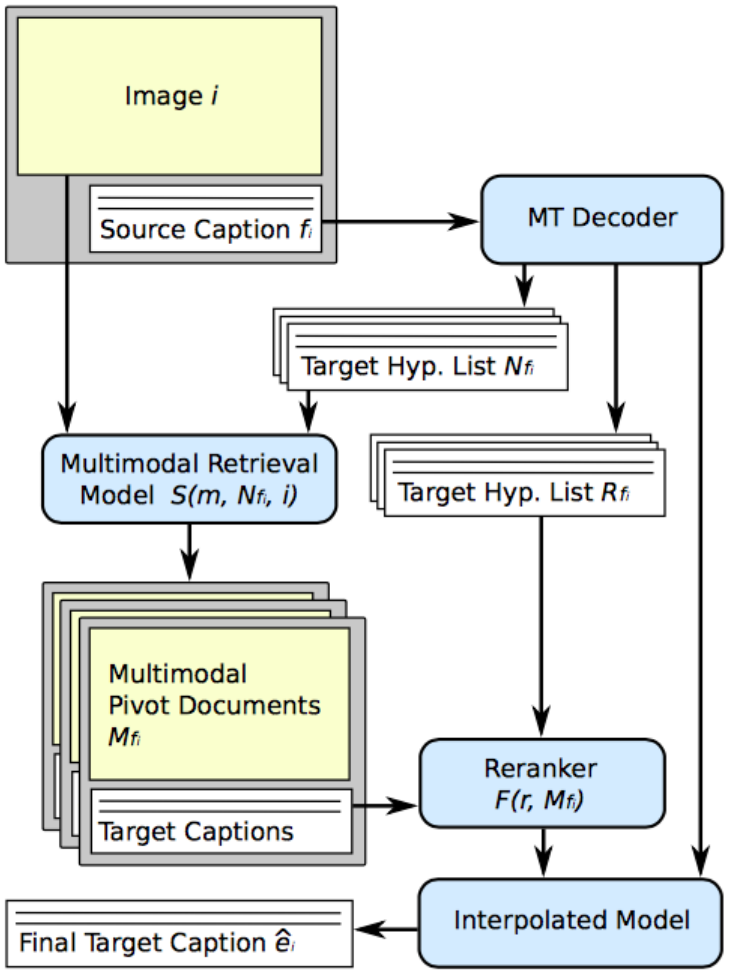


Image-augmented Translation [Huang+16]



Multi-modal Pivots [Hitschler+16]



References

References (1)

- W. Ammar, G. Mulcaire, M. Ballesteros, C. Dyer, and N. Smith. Many languages, one parser. Transactions of the Association for Computational Linguistics, 4:431–444, 2016.
- W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925, 2016.
- P. Arthur, G. Neubig, and S. Nakamura. Incorporating discrete translation lexicons into neural machine translation. In Proc. EMNLP, 2016.
- D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Proc. ICLR, 2015.
- J. Chung, K. Cho, and Y. Bengio. A character-level decoder without explicit segmentation for neural machine translation. In Proc. ACL, pages 1693–1703, 2016.
- T. Cohn, C. D. V. Hoang, E. Vymolova, K. Yao, C. Dyer, and G. Haffari. Incorporating structural alignment biases into an attentional neural translation model. In Proc. NAACL, pages 876–885, 2016.
- D. Dong, H. Wu, W. He, D. Yu, and H. Wang. Multi-task learning for multiple language translation. In Proc. ACL, pages 1723–1732, 2015.
- L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn. An attentional model for speech translation without transcription. In Proc. NAACL, pages 959, 2016.
- A. Eriguchi, K. Hashimoto, and Y. Tsuruoka. Tree-to-sequence attentional neural machine translation. In Proc. ACL, pages 823–833, 2016.
- M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In Proc. EACL, pages 462–471, 2014.
- O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proc. NAACL, pages 866–875, 2016.
- J. Gu, G. Neubig, K. Cho, and V. O. Li. Learning to translate in real-time with neural machine translation. In Proc. EACL, 2017.
- T.-L. Ha, J. Niehues, and A. Waibel. Toward multilingual neural machine translation with universal encoder and decoder. arXiv preprint arXiv:1611.04798, 2016.
- J. Hitschler, S. Schamoni, and S. Riezler. Multimodal pivots for image caption translation. In Proc. ACL, pages 2399–2409, 2016.
- K. Huang, M. Gardner, E. Papalexakis, C. Faloutsos, N. Sidiropoulos, T. Mitchell, P. P. Talukdar, and X. Fu. Translation invariant word embeddings. In EMNLP, pages 1084–1088, 2015.
- P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer. Attention-based multimodal neural machine translation. In Proc. WMT, pages 639–645, 2016.
- M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vi´egas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. arXiv preprint arXiv:1611.04558, 2016.

References (2)

- N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In Proc. EMNLP, pages 1700–1709, 2013.
- T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In Proc. EMNLP, pages 1412–1421, 2015.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168, 2013.
- R. O ̈stlingandJ.Tiedemann.Continuousmultilingualitywithlanguagevectors.InProc.EACL,pages644–649,2017.
- B. Plank, A. Søgaard, and Y. Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In Proc. ACL, pages 412–418, 2016.
- M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. Proc. ICLR, 2016.
- S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu. Minimum risk training for neural machine translation. In Proc. ACL, pages 1683–1692, 2016.
- M. Sperber, G. Neubig, J. Niehues, and A. Waibel. Neural lattice-to-sequence models for uncertain inputs. arXiv preprint arXiv:1704.00559, 2017.
- I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Proc. NIPS, pages 3104–3112, 2014.
- Y. Tang, F. Meng, Z. Lu, H. Li, and P. L. H. Yu. Neural machine translation with external phrase memory. CoRR, abs/1606.01792, 2016.
- Y. Tsvetkov and C. Dyer. Cross-lingual bridges with models of lexical borrowing. Journal of Artificial Intelligence Research, 55:63–93, 2016.
- Y. Tsvetkov, S. Sitaram, M. Faruqui, G. Lample, P. Littell, D. Mortensen, A. W. Black, L. Levin, and C. Dyer. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In Proc. NAACL, pages 1357–1366, 2016.
- R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen. Sequence-to-sequence models can directly transcribe foreign speech. arXiv preprint arXiv:1703.08581, 2017.
- S. Wiseman and A. M. Rush. Sequence-to-sequence learning as beam-search optimization. In Proc. EMNLP, pages 1296–1306, 2016.
- B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. In Proc. EMNLP, pages 1568–1575, 2016.