Augmenting NMT Architecture with Linguistic Features

Part of the EAMT 2024 Tutorial
Linguistically Motivated Neural Machine Translation

Hour Kaing
https://hour.github.io

European Association for Machine Translation

National Institute of Information and Communications Technology

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

Linguistic Features

- Lemma
  - A lemma is the canonical form, or dictionary form of a set of word forms. [Wikipedia]

- For examples
  - In English, *break*, *breaks*, *broke*, *broken* and *breaking* have the same lemma as *break*.
  - In French*, aller, vais, va, allait,* and *ira* have the same lemma as *aller (go)*.

Linguistic Features

- Part-Of-Speech (POS)
  - A POS is a word class or grammatical category of words that have similar grammatical properties. [Wikipedia]

- For examples
  - Noun : *home, house*, and *television*.
  - Verb : *walk, happen*, or *to be*.
  - Universal POS tags used by the Universal Dependency project.

    - ADJ: adjective
    - ADP: adposition
    - ADV: adverb
    - AUX: auxiliary
    - CCONJ: coordinating conjunction
    - DET: determiner
    - INTJ: interjection
    - NOUN: noun
    - NUM: numeral
    - PART: particle
    - PRON: pronoun
    - PROPN: proper noun
    - PUNCT: punctuation
    - SCONJ: subordinating conjunction
    - SYM: symbol
    - VERB: verb
    - X: other

Linguistic Features

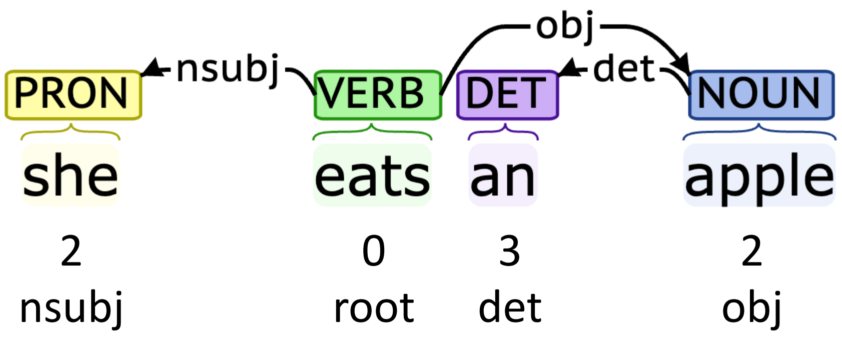
- More specific grammatical properties or morphological features
  - Pronominal type (PronType)
    - Personal/possessive pronoun (Prs) : *he, she, his, her*, …
    - Article (Art) : *a, an, the*
  - Tense (Tense)
    - Past (Past) : went, gone, …
    - Present/non-past tense (Pre) : goes, going, …

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Deixis	Polarity
	DeixisRef	Person
	Degree	Polite
		Clusivity

From Universal Dependency

Linguistic Features

- Dependency grammars
  - The grammars are based on the dependency relation between words.
  - A dependency is a directed link from a head word to a dependent word.



Example generated by Stanza

Linguistic Features

- Constituency/phrase structure grammars
  - The grammars are based on the constituency relation between words
  - A constituent is a word or a group of words that function as a single unit within a hierarchical structure. [Wikipedia]
- Constituency is a hypergraph problem
  - An edge can join any number of vertices

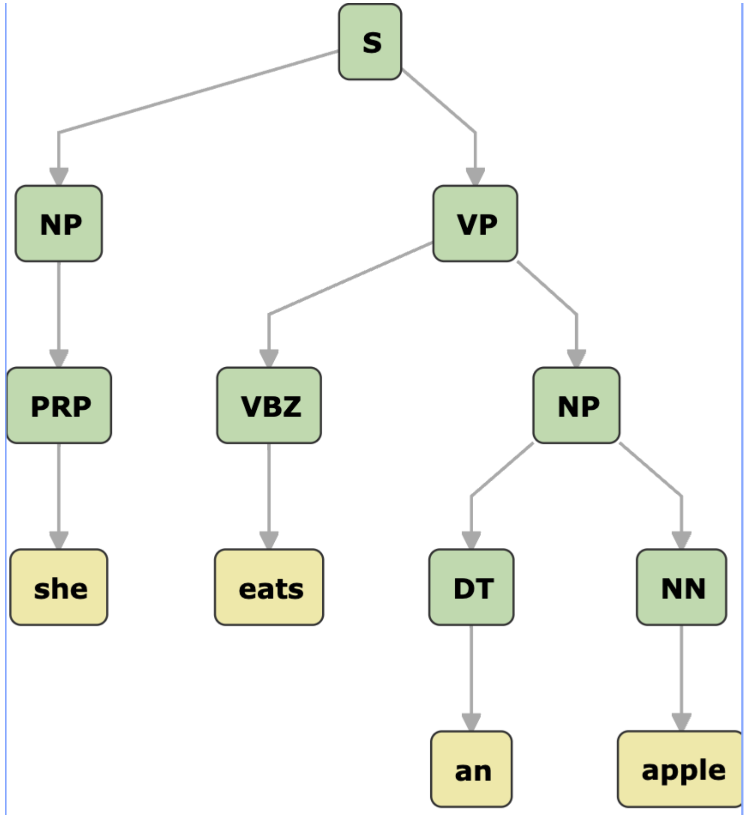


Figure generated by Stanza

Analysis Tools

Lemma, POS, morphological features, and dependency

- Stanza : https://stanfordnlp.github.io/stanza
  - Support 66 languages
- SpaCy : https://spacy.io
  - Support 75+ languages
- UDPipe : https://lindat.mff.cuni.cz/services/udpipe
  - Support 100+ languages

Analysis Tools

Constituency parsing tools

- Stanza
  - Support 10 languages
- Berkeley parser : https://github.com/nikitakit/self-attentive-parser
  - Support 11 languages
- SuPar : https://github.com/yzhangcs/parser
  - Support 11 languages

Linguistic Datasets

- Universal Dependencies : https://universaldependencies.org
  - Features: lemma, POS, morphological features, dependency.
  - Contains 200 treebanks in over 100 languages
- Consituney Treebanks
  - Penn Treebank : https://catalog.ldc.upenn.edu
    - English, Chinese, Korean, and Arabic.
  - SPMRL2013/2014 : https://www.spmrl.org/spmrl2014-sharedtask.html
    - Arabic, Basque, English, French, German, Hebrew, Hungarian, Korean, Polish, and Swedish
  - Asian Language Treebank : https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT
    - English, Japanese, and Myanmar.

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
  - Input features embedding
  - Multiple and mixed encoder
  - Syntax distance
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

Motivations

- Word form ambiguity

1. *We thought a win like this might be close.*

adjective

2. *Wir dachten, dass ein solcher Sieg nah sein könnte.*
3. **Wir dachten, ein Sieg wie dieser könnte schließen.*

verb

Solution : give a POS tag as a hint to disambiguate the word that shares the same form across word types.

- Word order discrepancy

4. *Gefährlich ist die Route aber dennoch .*
   dangerous is the route but still .
5. *However the route is dangerous .*
6. **Dangerous is the route , however .*

4. in German have a verb-second (V2) order.
But English is generally SVO.

Solution : give a syntactic annotation as a hint to guide the model which words to attend and translate first.
e.g., where is root, nsubj, etc.

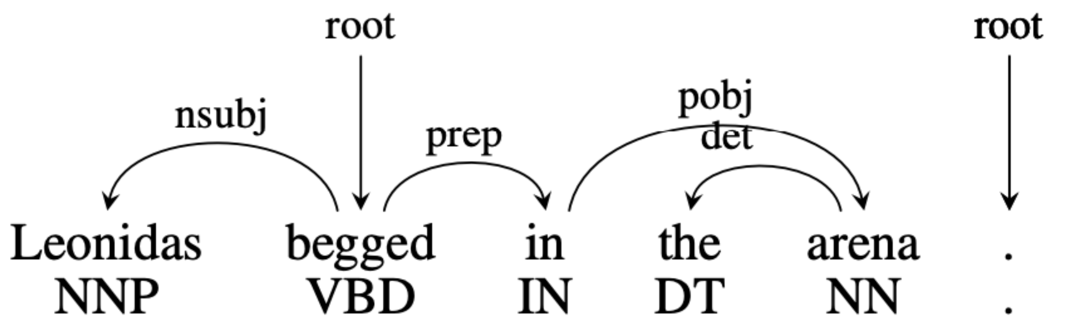
Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016

Other Features

- Lemma
  - Share information between words with the same base form
- Morphological features
  - Like POS tags but with more detail properties
- Word boundary or subword tags
  - Give models clues which symbols to attend to, and when to forget the information

Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016

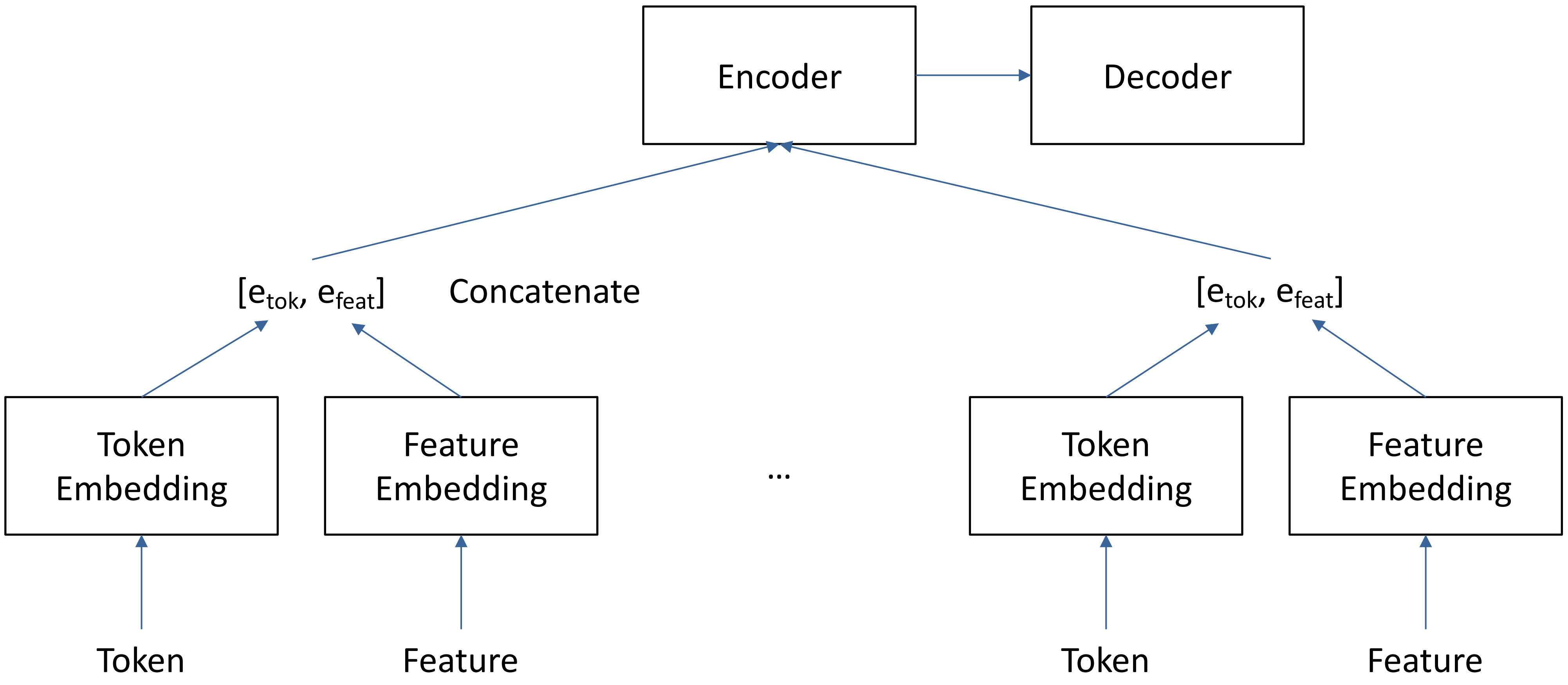
Examples of Features



words	Le:	oni:	das	beg:	ged	in	the	arena	.
lemmas	Leonidas	Leonidas	Leonidas	beg	beg	in	the	arena	.
subword tags	B	I	E	B	E	O	O	O	O
POS	NNP	NNP	NNP	VBD	VBD	IN	DT	NN	.
dep	nsubj	nsubj	nsubj	root	root	prep	det	pobj	root

Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016

How To Integrate Into NMT?



Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016

Features Relevance

- Some features are more relevant to the translation than the others
- Intuition: weighting the features would give better translation

system	German→English ppl ↓ dev	German→English BLEU ↑ test15	German→English BLEU ↑ test16	German→English CHRF3 ↑ test15	German→English CHRF3 ↑ test16	English→German ppl ↓ dev	English→German BLEU ↑ test15	English→German BLEU ↑ test16	English→German CHRF3 ↑ test15	English→German CHRF3 ↑ test16
baseline	47.3	27.9	31.4	54.0	58.0	54.9	23.0	27.8	52.6	56.0
all features	46.2	28.7*	32.9*	54.8	58.5	52.9	23.8*	28.4*	53.9	57.2
lemmas	47.1	28.4	32.3*	54.6	58.7	53.4	23.8*	28.5*	53.7	56.7
subword tags	47.3	27.7	31.5	54.0	58.1	54.7	23.6*	28.1	53.2	56.4
morph. features	47.1	28.2	32.4*	54.3	58.4	-	-	-	-	-
POS tags	46.9	28.1	32.4*	54.1	57.8	53.2	24.0*	28.9*	53.3	56.8
dependency labels	46.9	28.1	31.8*	54.2	58.3	54.0	23.4*	28.0	53.1	56.5

Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016

Features Relevance Weighting

Figure 1: Self relevance of a feature.

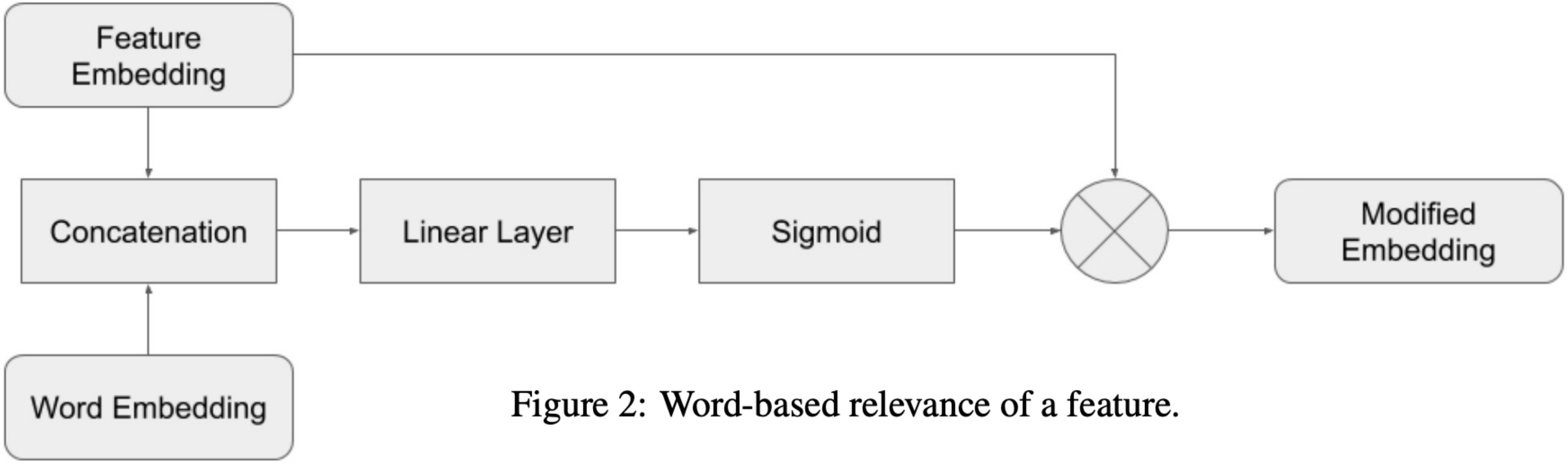


Figure 2: Word-based relevance of a feature.

Abhisek Chakrabarty, et al., Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation, 2020

Features Relevance

- Extremly low-resource languages

		en-bg	en-fi	en-hi	en-id	en-khm	en-ms	en-my	en-vi
Baseline models	Base	4.97	25.59	18.54	27.93	22.88	32.40	13.93	24.99
	Concat	5.56	23.75	20.69	27.99	23.53	32.92	14.92	26.50
	Add	4.66	22.02	15.45	24.78	21.65	30.45	11.86	22.78
	Linear	4.89	24.26	20.65	27.17	23.42	32.64	13.79	25.36
Proposed models	Self-rel	6.10	**26.26**	21.27	**30.41**	24.76	**34.71**	**16.53**	**27.74**
	Word-rel	**6.25**	26.01	**21.63**	26.53	**25.13**	33.20	15.62	27.66

Table 3: BLEU scores of the models for all reference language pairs.

Abhisek Chakrabarty, et al., Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation, 2020

Features in Other NMT Settings

- Features in pretraining stage of BART [Chakrabarty+22]
  - Both content tokens/spans and features are masked.
  - Improvement was observed in an extremely low-resource settings

- Features in multilingual settings [Chakrabarty+23]
  - Language ID and dummy features are important in multilingual setting.
  - Dummy features
    - Each sentence have two representation, w/ dummy or w/ linguistic features

Abhisek Chakrabarty, et al., FeatureBART: Feature Based Sequence-to-Sequence Pre-Training for Low-Resource NMT, 2022
Abhisek Chakrabarty, et al., Low-resource Multilingual Neural Translation Using Linguistic Feature-based Relevance Mechanisms, 2023

Grammar Relations for Augmented Input Features

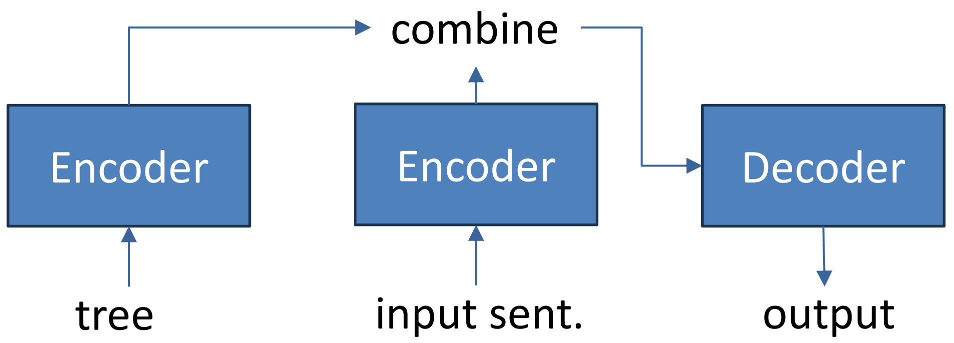
- Grammars (dependency or constituency) are relations between two or multiple words instead grammatical properties for individual words.
- Approaches to integrate the grammars more effectively
  - Multiple encoders
  - Mixed encoder
  - Transformation to syntactic distance

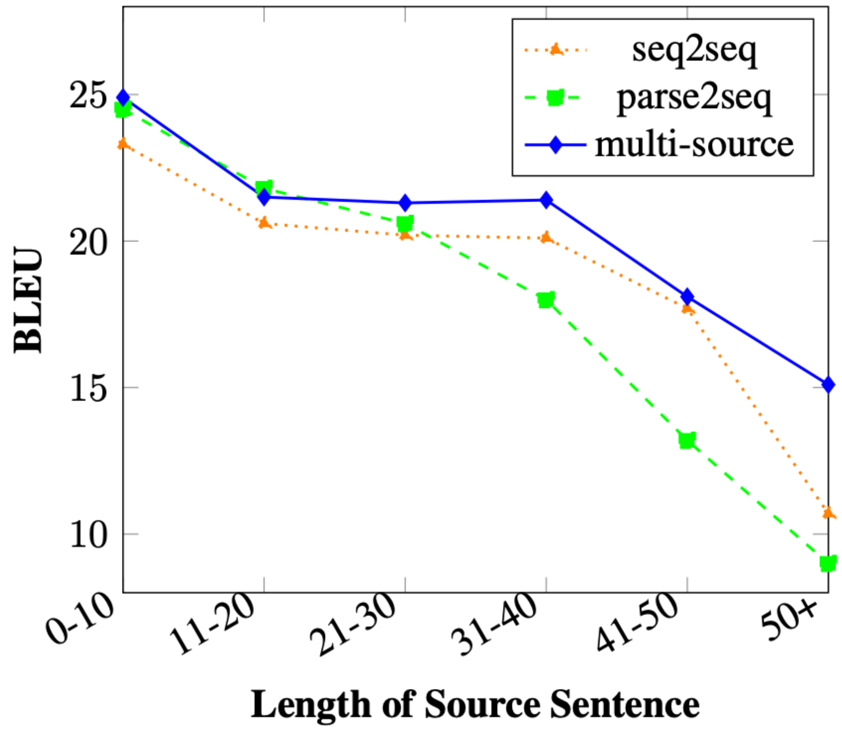
Multiple Encoders

- Linearized tree

	Example Sentence
sequential	history is a great teacher .
lexicalized parse	(ROOT (S (NP (NN history)) (VP (VBZ is) (NP (DT a) (JJ great) (NN teacher))) (. .)))
unlexicalized parse	(ROOT (S (NP (NN)) (VP (VBZ) (NP (DT) (JJ) (NN))) (. .)))
target sentence	die Geschichte ist ein großartiger Lehrmeister .

- Multiple-Encoder NMT
  - Length-agnostic between tree and input
  - RNN NMT with hierarchical attention

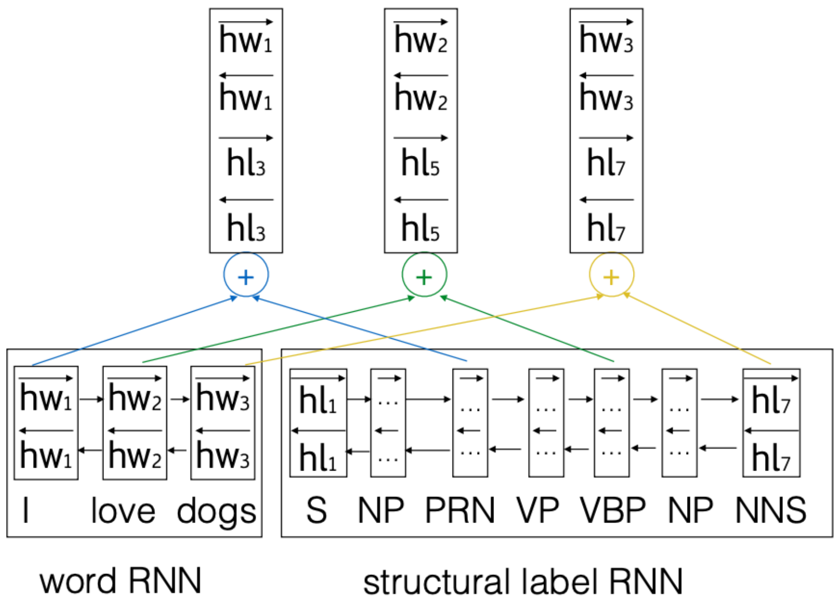




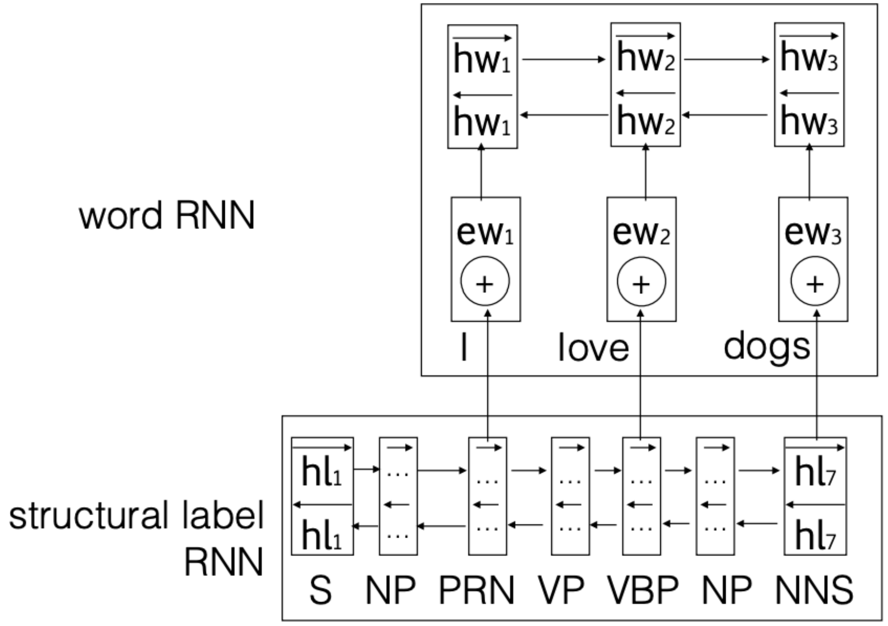
Anna Currey, et al., Multi-Source Syntactic Neural Machine Translation, 2018

Multiple Encoders

- Use only the output vectors of terminal nodes
- len(terminal nodes) == len(sentences)



(a) Parallel RNN encoder

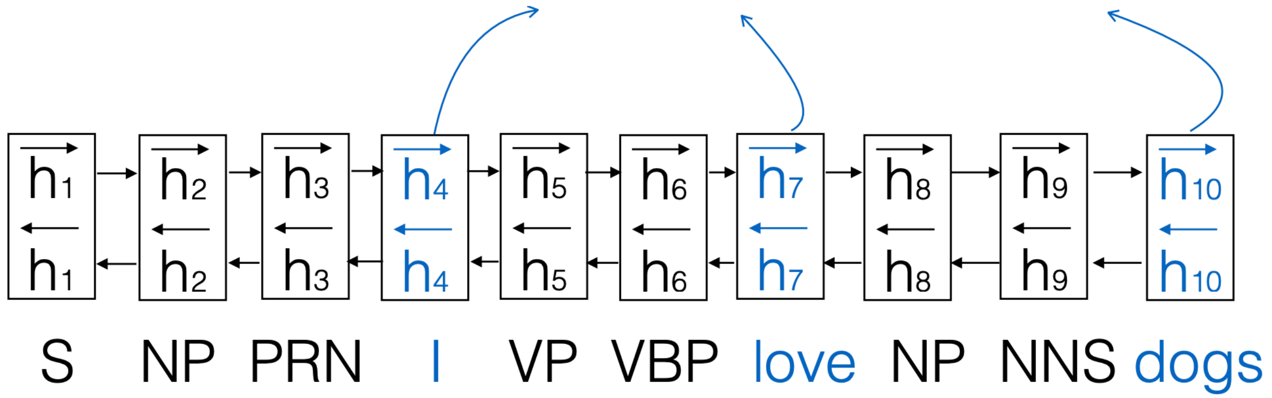


(b) Hierarchical RNN encoder

Junhui Li, et al., Modeling Source Syntax for Neural Machine Translation, 2017

Mixed Encoder

- Encode the linear lexicalized tree
- Take only output vectors of words for decoding



Mixed RNN Encoder

#	System	#Params	Time	MT06	MT02	MT03	MT04	MT05	All
1	cdec	-	-	33.4	34.8	33.0	35.7	32.1	34.2
2	RNNSearch	60.6M	153m	34.0	36.9	33.7	37.0	34.1	35.6
3	Parallel RNN	+1.1M	+9m	34.8†	**37.8‡**	34.2	38.3‡	34.6	36.6‡
4	Hierarchical RNN	+1.2M	+9m	35.2‡	37.2	34.7†	38.7‡	34.7†	36.7‡
5	Mixed RNN	+0	+40m	**35.6‡**	37.7†	**34.9‡**	**38.6‡**	**35.5‡**	**37.0‡**

Junhui Li, et al., Modeling Source Syntax for Neural Machine Translation, 2017

Mixed Encoder

- Data augmentation
  - Combine source parses and sentences
  - Shuffle so that two-sentence pairs are not seen together during traning

(ROOT (S (NP you) (VP have not (VP been (VP elected))) .))	→ no ha sido elegido .
you have not been elected .	→ no ha sido elegido .

Table 2: Example of English→Spanish training data for the mixed encoder system.

EN→*	**base**	**mixed enc.**
LV	26.5	28.1 (+1.6)
LT	23.5	**24.6** (+1.1)
DA	39.5	40.1 (+0.6)

Small scale

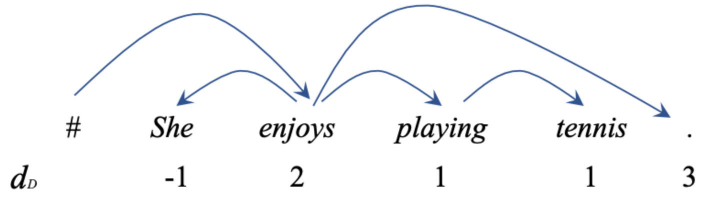
System	**newstest2017**	**newstest2018**
baseline	9.6	8.8
mixed enc.	9.6 (==)	9.3 (+0.5)

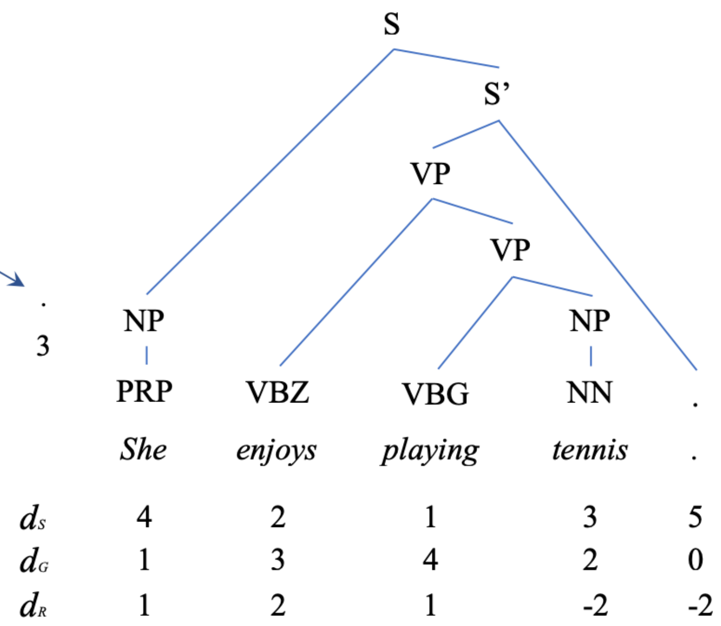
Large scale

Anna Currey, et al., Incorporating Source Syntax into Transformer-Based Neural Machine Translation, 2019

NMT with Syntactic Distance

- Syntax distances
  - Constituency
    - $d_S(w_i)$ is the height of the lowest common ancestor between w_i and w_{i+1}
    - $d_G(w_i)$ is the number of common ancestors between w_i and w_{i+1}
    - $d_R(w_i) = d_G(w_i) – d_G(w_{i-1})$ if $i > 1$ else $d_G(w_i)$
  - Dependency
    - $d_D(w_i) = i – h(i)$
    - $h(i)$ is the index of the head of w_i





- Integration
  - Input features
  - Positional encoding

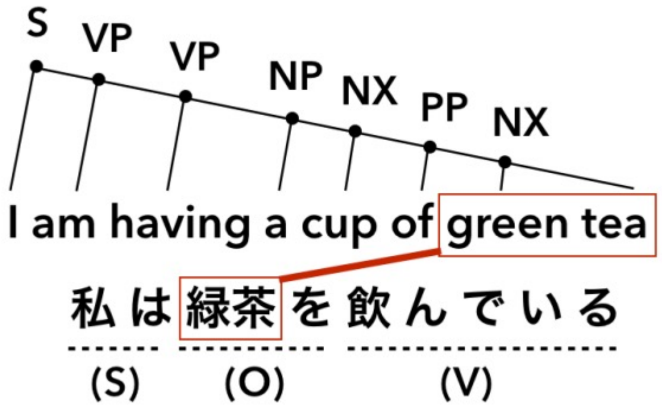
Chunpeng Ma, et al., Improving Neural Machine Translation with Neural Syntactic Distance, 2019

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

Tree Encoder

- Linguistic distance languages
  - Different syntax construction
  - Different lexical units such as words/phrase
  - E.g., if ‘緑茶’ only align with ‘green’ and ‘tea’,
  - Then ‘a cup of’ will align with ‘null’.



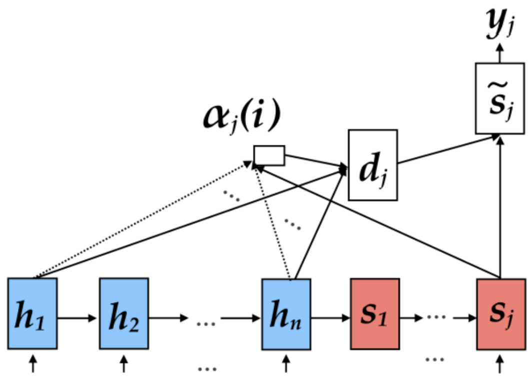


Figure 2: Attentional Encoder-Decoder model.

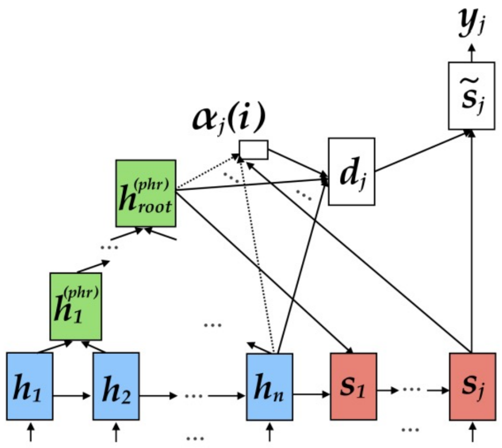


Figure 3: Proposed model: Tree-to-sequence attentional NMT model.

Tree-LSTM

$$h_k^{(phr)} = f_{tree}(h_k^l, h_k^r),$$

$$s_1 = g_{tree}(h_n, h_{root}^{(phr)}),$$

Context vector

$$d_j = \sum_{i=1}^{n} \alpha_j(i) h_i + \sum_{i=n+1}^{2n-1} \alpha_j(i) h_i^{(phr)}.$$

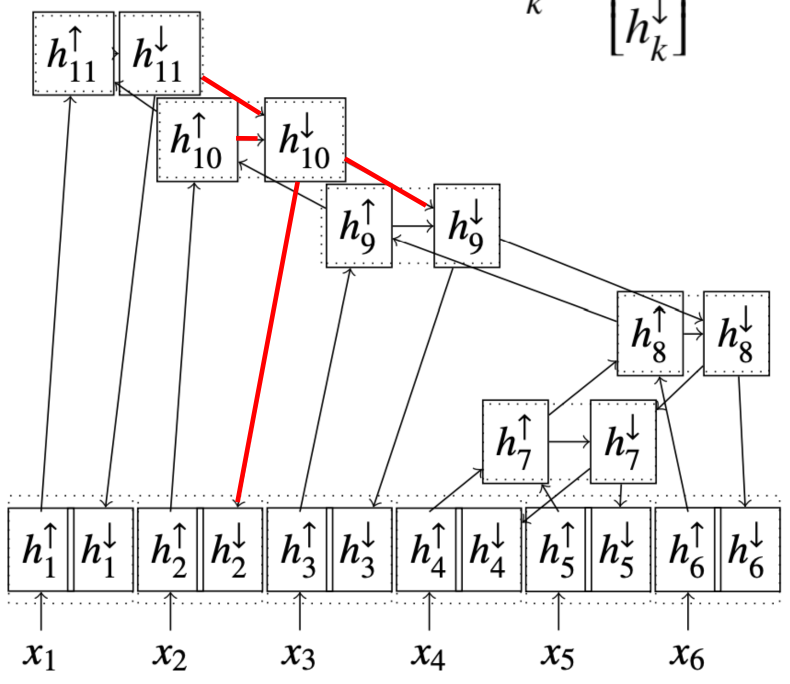
Akiko Eriguchi, et al., Tree-to-Sequence Attentional Neural Machine Translation, 2016

Bidirectional Tree Encoder

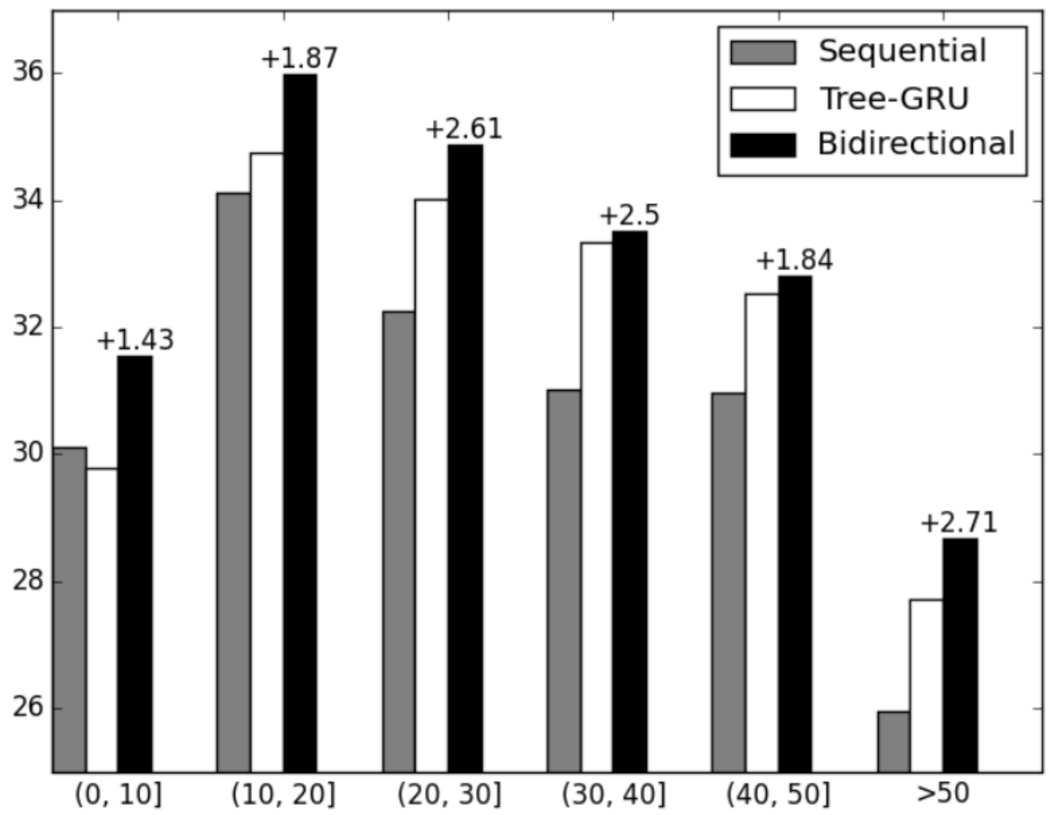
- The node representation is based on its nodes only
- And contains no information from the higher nodes

Tree-GRU

$$h_k^{\updownarrow} = \begin{bmatrix} h_k^{\uparrow} \\ h_k^{\downarrow} \end{bmatrix}.$$



(b) Bidirectional Tree Encoder



Huadong Chen, et al., Improved Neural Machine Translation with a Syntax-Aware Encoder and Decoder, 2017

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

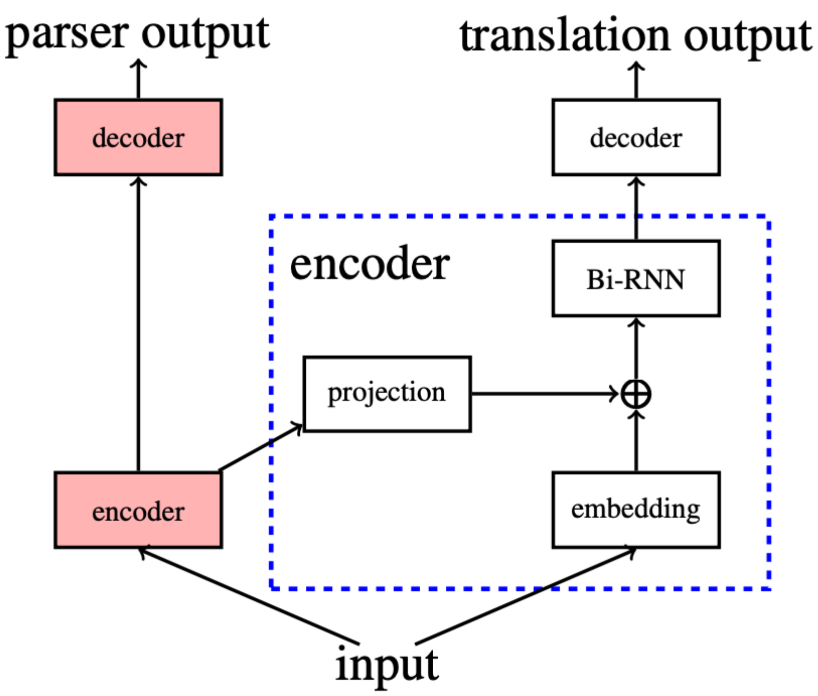
Syntax-aware representation

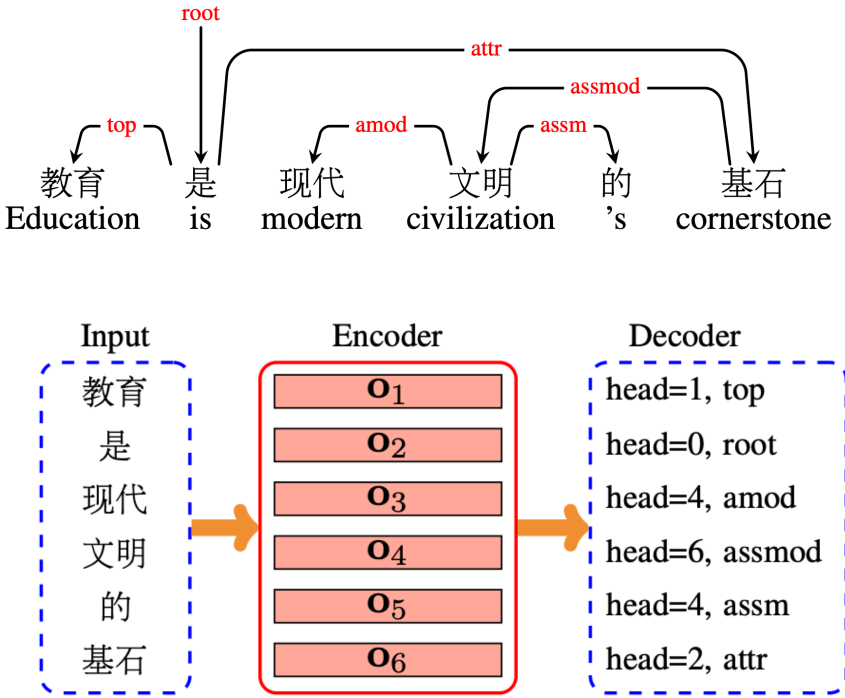
- Generated linguistic features could be irrelevant or errors.
- Linguistic features are always generated alongside the translation.

- Implicit integration
  - Syntax-aware representation injection
  - Multi-task learning

Syntax-Aware Representation Injection

- First stage: pretrained linguistic model
- Second stage: inject the linguistic encoder output into the input embedding



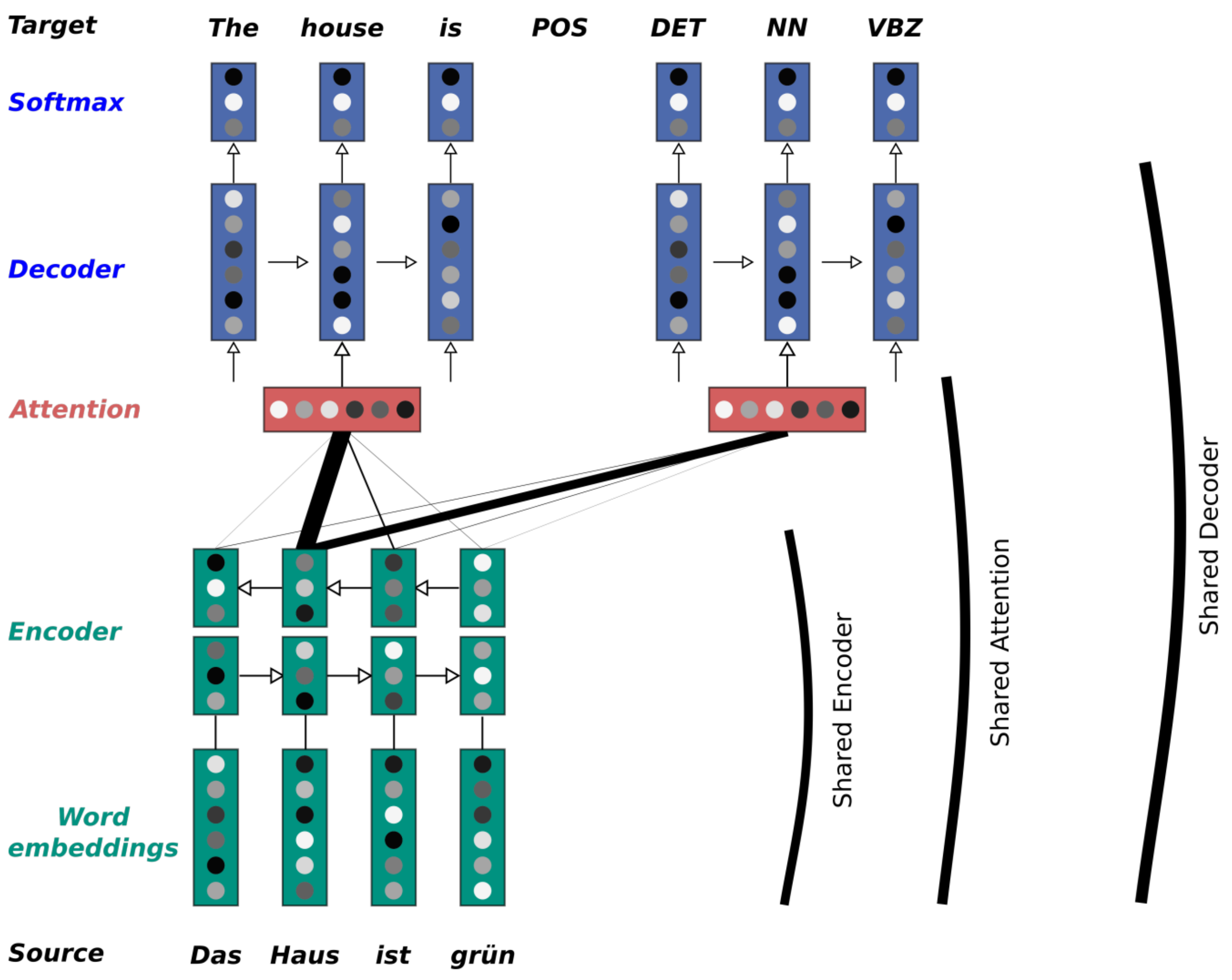


Meishan Zhang, et al., Syntax-Enhanced Neural Machine Translation with Syntax-Aware Word Representations, 2019

Multi-Task Learning

- Training a model to perform two tasks jointly
  - Translation
  - Linguistic prediction

- Task adaptation
  - Train on both tasks
  - Finetune on the main task



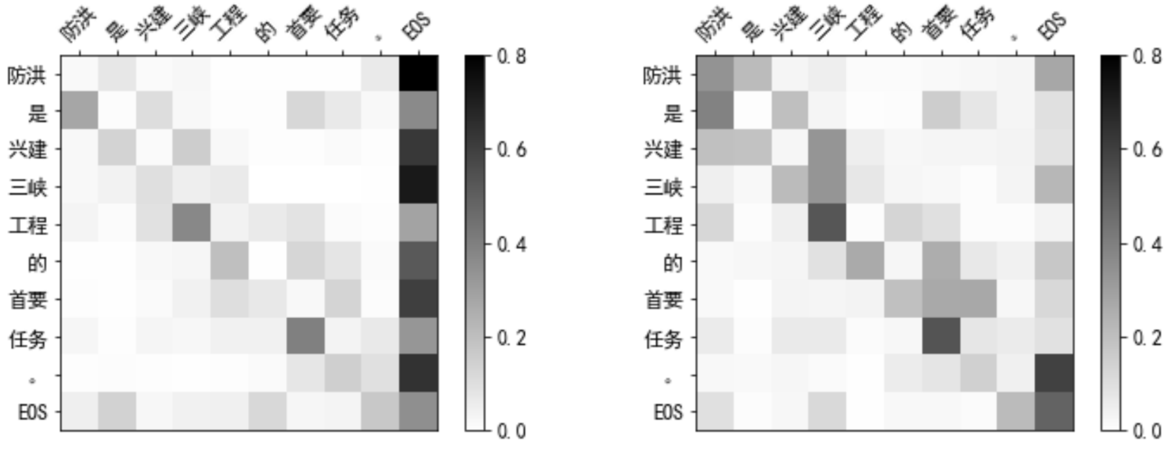
Jan Niehues, et al., Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning, 2017

Agenda

- Linguistic features, tools and datasets
- Augmented input feature NMT
- Tree encoder
- Syntax-aware representation
- Syntax-aware self-attention

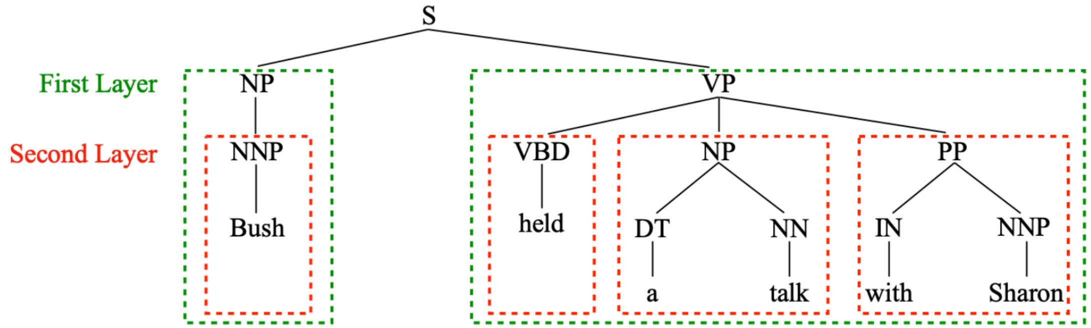
Multi-Granularity Self-Attention

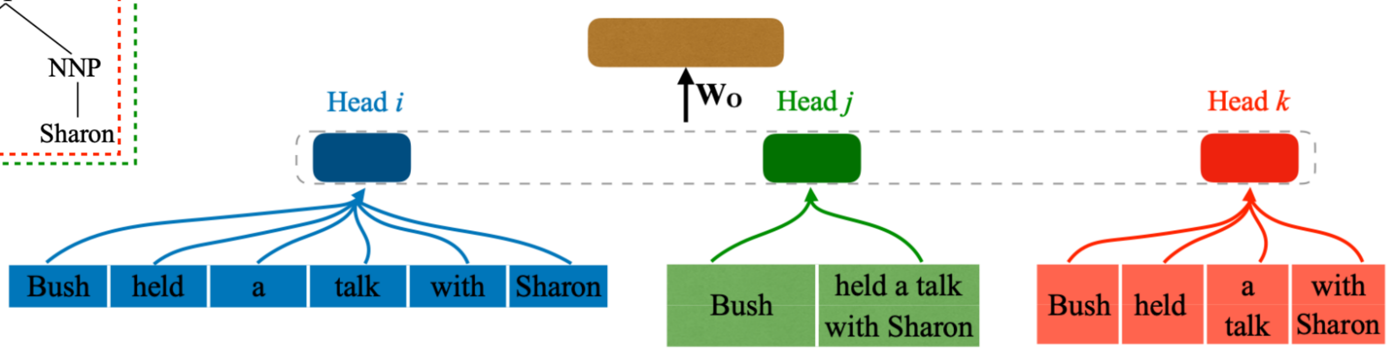
- 3 steps
  - Phrase partition
    - N-gram or constituents
  - Phrase composition
    - CNN, RNN, or selft-attention network
  - Phrase interaction
    - LSTM or tree-like LSTM



(a) Vanilla Multi-Head Self-Attention

(b) Multi-Granularity Self-Attention

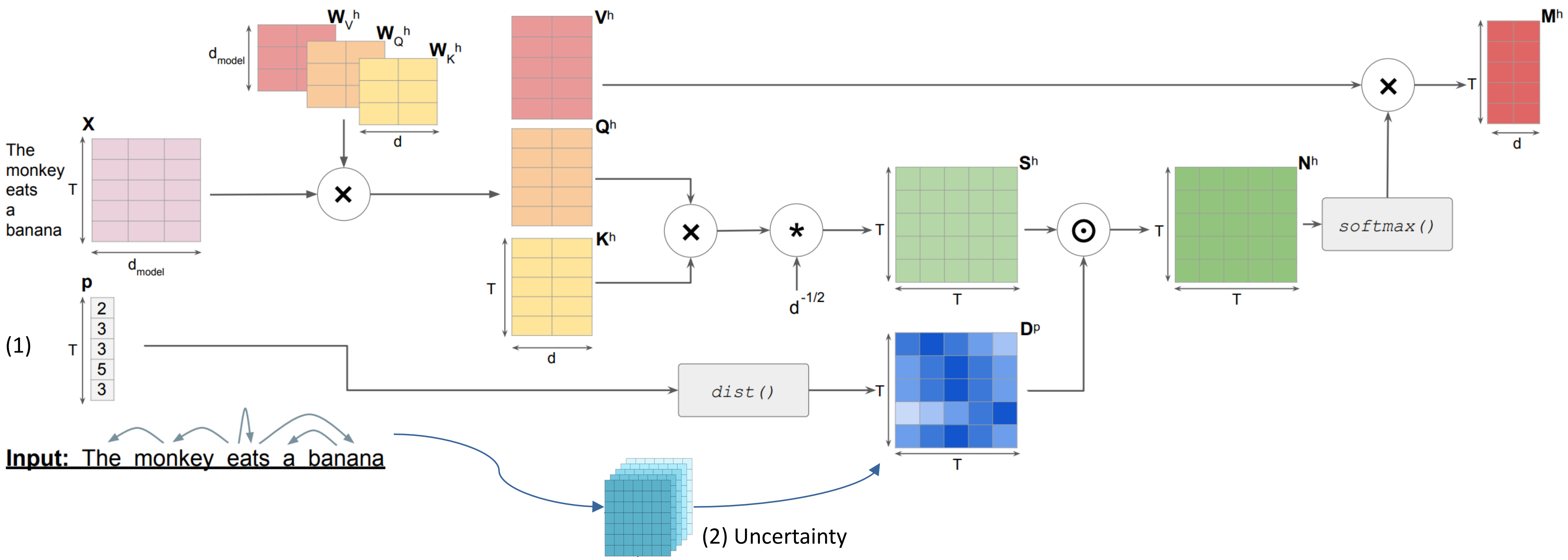




Jie Hao, et al., Multi-Granularity Self-Attention for Neural Machine Translation, 2019

Dependency-Aware Self-Attention

- Guide the attention weights using the dependency distance
- Great improvement for long sentences



(1) Emanuele Bugliarello, et al., Enhancing Machine Translation with Dependency-Aware Self-Attention, 2020
(2) Dongqi Pu, et al., Passing Parser Uncertainty to the Transformer: Labeled Dependency Distributions for NMT, 2020

Summary

	Augmented Input	Tree Encoder	Representation	Self Attention
Morphological features	✓		✓	
Grammar relations	✓	✓	✓	✓

- Linguistic features are usually useful for low-resource setting.
- Grammar relation tend to beneficial long sentences.
- However, the quality of the linguistic features is crucial.
  - This may hurt the translation performance
  - Soft/implicit integration such as data augment or representation learning, would reduce the constraint.

References

- Abhisek Chakrabarty, et al., FeatureBART: Feature Based Sequence-to-Sequence Pre-Training for Low-Resource NMT, 2022
- Abhisek Chakrabarty, et al., Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation, 2020
- Abhisek Chakrabarty, et al., Low-resource Multilingual Neural Translation Using Linguistic Feature-based Relevance Mechanisms, 2023
- Akiko Eriguchi, et al., Tree-to-Sequence Attentional Neural Machine Translation, 2016
- Anna Currey, et al., Incorporating Source Syntax into Transformer-Based Neural Machine Translation, 2019
- Anna Currey, et al., Multi-Source Syntactic Neural Machine Translation, 2018
- Bei Li, et al., Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation, 2020
- Chunpeng Ma, et al., Improving Neural Machine Translation with Neural Syntactic Distance, 2019
- Dongqi Pu, et al., Passing Parser Uncertainty to the Transformer: Labeled Dependency Distributions for NMT, 2020
- Emanuele Bugliarello, et al., Enhancing Machine Translation with Dependency-Aware Self-Attention, 2020
- Huadong Chen, et al., Improved Neural Machine Translation with a Syntax-Aware Encoder and Decoder, 2017
- Jan Niehues, et al., Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning, 2017
- Jie Hao, et al., Multi-Granularity Self-Attention for Neural Machine Translation, 2019
- Jindřich Libovický, et al., Attention Strategies for Multi-Source Sequence-to-Sequence Learning, 2017
- Junhui Li, et al., Modeling Source Syntax for Neural Machine Translation, 2017
- Meishan Zhang, et al., Syntax-Enhanced Neural Machine Translation with Syntax-Aware Word Representations, 2019
- Rico Sennrich et al., Linguistic Input Features Improve Neural Machine Translation, 2016