Deep Neural Networks for Natural Language Processing (AI6127)

JUNG-JAE KIM

SUBWORD MODELS

Lecture Plan

- Motivation of subword models
- Purely character-level models
- Byte Pair Encoding
- Hybrid NMT

Below the word: Writing systems

- Most of deep learning NLP works begin with language in its written form – it’s the easily processed, found data
- But human language writing systems aren’t one thing!
- Terminology
  - Grapheme: the smallest unit of a written language whether it carries meaning or corresponds to a single phoneme
    - E.g. English alphabet characters, 안녕 (ㅇ ㅏ ㄴ ㄴ ㅕ ㅇ)
  - Phoneme: the smallest unit of sound
    - E.g. natural / ˈnætʃ ər əl, ˈnætʃ rəl / (IPA; International Phonetic Alphabet)
  - Syllable: a sequence of sounds/phonemes with at least one vowel

Below the word: Writing systems

- Phonemic (maybe digraphs)   jiyawu ngabulu
  - graphemes (written symbols) correspond to phonemes
- Fossilized phonemic   thorough failure
- Syllabic/moraic   ᑐᖑᔪᐊᖓᔪᖅ
  - characters represent syllables and are combined to indicate morphemes
- Ideographic   去年太空船二号坠毁
  - ‘ideogram’ symbols represent elements of language
- Combination of the above   インド洋の島

Wambaya

English

Inuktitut

Chinese

Japanese

Source: CS224n

1. Words in writing systems

- Writing systems vary in how they represent words – or don't

- No word segmentation 美国关岛国际机场及其办公室均接获

- Words (mainly) segmented
  - Clitics? (have form of affixes, but distribution of function words; e.g. it'**s**, we'**ve**)
    - Separated **Je vous ai apporté** des bonbons
    - Joined فقلناها = ها + نا + قال + ف = so+said+we+it
  - Compounds?
    - Separated life insurance company employee
    - Joined Lebensversicherungsgesellschaftsangestellter

Source: CS224n

Morphology: Parts of words

- Traditionally, we have morphemes as smallest **semantic** unit
  - $[[\text{un } [[\text{fortun(e) }]_{\text{ROOT}} \text{ ate}]_{\text{STEM}}]_{\text{ STEM}} \text{ ly}]_{\text{WORD}}$
  - A root/stem is a form which is not further analysable
  - ‘fortunate’ is the stem of ‘unfortunate’

Source: CS224n

Models below the word level

- Need to handle large, open vocabulary
  - Rich morphology: nejneobhospodařovávatelnějšímu Czech
    (“to the worst farmable one”)
  - Transliteration: Christopher ↦ Kryštof
  - Informal spelling:

Source: CS224n

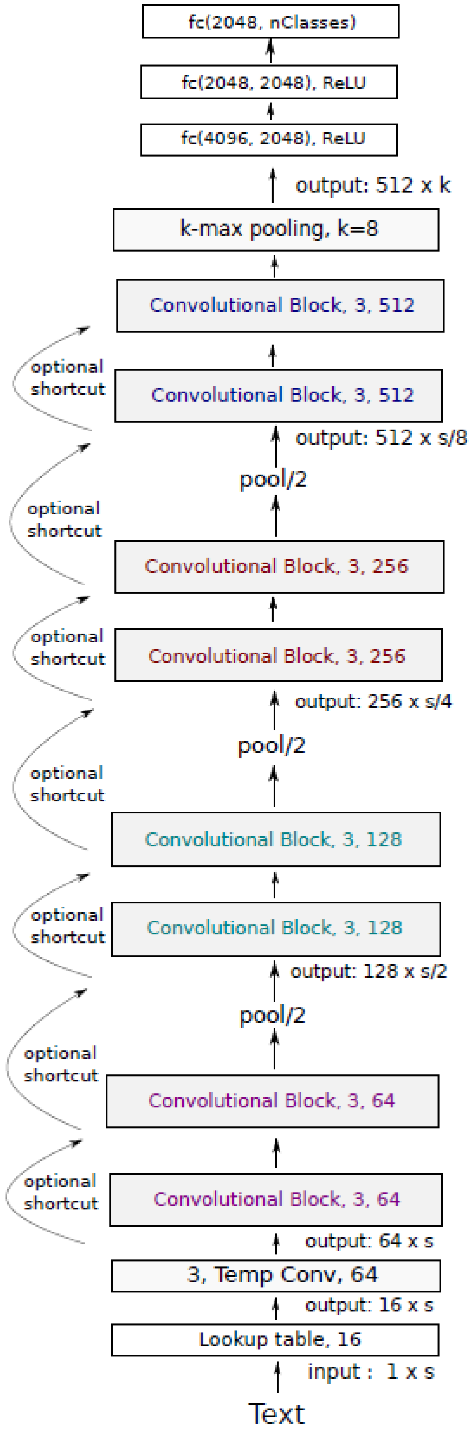
Character-Level Models

- Word embeddings can be composed from character embeddings
  - Generates embeddings for unknown words
  - Similar spellings share similar embeddings
  - Solves OOV problem

Source: CS224n

2. Purely character-level models

- Strong results via a deep convolutional stack
  - Very Deep Convolutional Networks for Text Classification
  - Conneau, Schwenk, Lecun, Barrault. EACL 2017



Source: CS224n

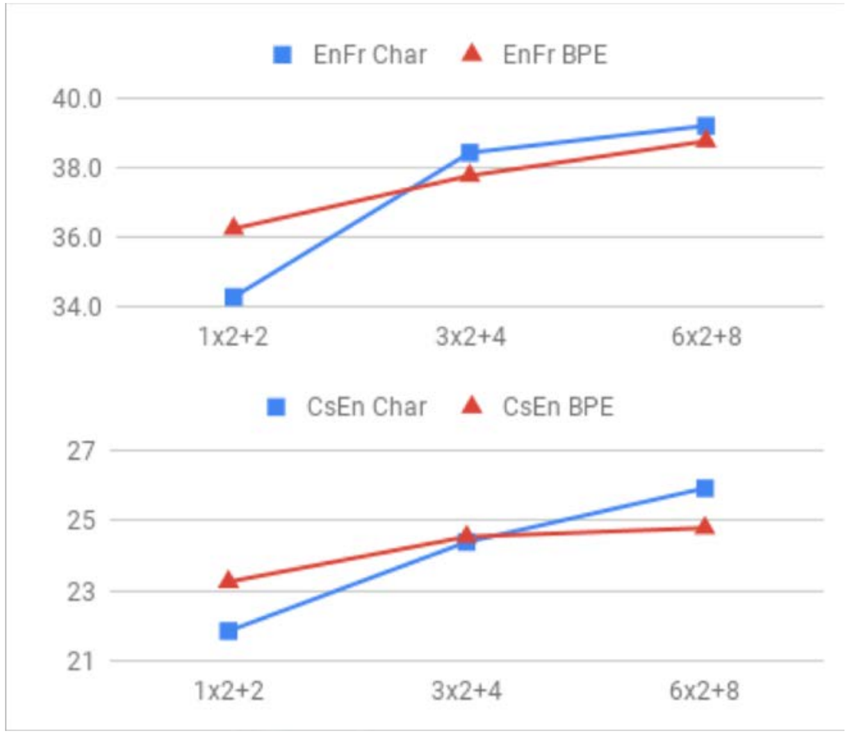
Purely character-level NMT models

- Initially, **unsatisfactory** performance
  - (Vilar et al., 2007; Neubig et al., 2013)
- Subword-level encoder + **Character-level decoder** (w/o segmentation)
  - (Junyoung Chung, Kyunghyun Cho, Yoshua Bengio. arXiv 2016).
- Then **promising** results
  - (Wang Ling, Isabel Trancoso, Chris Dyer, Alan Black, arXiv 2015)
  - (Thang Luong, Christopher Manning, ACL 2016)
  - (Marta R. Costa-Jussà, José A. R. Fonollosa, ACL 2016)

Source: CS224n

Stronger character results with depth in LSTM seq2seq model

- Revisiting Character-Based Neural Machine Translation with Capacity and Compression. 2018. Cherry, Foster, Bapna, Firat, Macherey, Google AI
  - **X-axis: E.g. 1x2+2 indicates 1 BiLSTM encoder layer and 2 LSTM decoder layers**
  - **Y-axis: bleu scores**



Source: CS224n

Hands-on: Character-level recurrent sequence-to-sequence model

- Configuration
- Download and prepare data
- Build LSTM model
- Train the model
- Run inference

https://colab.research.google.com/github/keras-team/keras-io/blob/master/examples/nlp/ipynb/lstm_seq2seq.ipynb

3. Byte Pair Encoding

- To segment word into subword tokens
- Use tokenizer for segmenting text to words
  - Simple space tokenization (e.g. GPT-2, Roberta)
  - Rule-based tokenization (e.g. Moses http://www.statmt.org/moses/?n=Development.GetStarted)

Source: CS224n

Byte Pair Encoding

- Originally a compression algorithm:
  - Most frequent byte pair $\mapsto$ a new byte.

Replace bytes with character ngrams

  - Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. ACL 2016.
    - https://arxiv.org/abs/1508.07909
    - https://github.com/rsennrich/subword-nmt
    - https://github.com/EdinburghNLP/nematus

Source: CS224n

Byte Pair Encoding

A word segmentation algorithm:

- Start with a unigram vocabulary of all (Unicode) characters in data
- Most frequent ngram pairs ↦ a new ngram in the vocabulary

Dictionary

5 l o w
2 l o w e r
6 n e w e s t
3 w i d e s t

Vocabulary

l, o, w, e, r, n, s, t, i, d

Start with all characters in vocab

(Example from Sennrich)

Source: CS224n

Byte Pair Encoding

A word segmentation algorithm:

- Start with a unigram vocabulary of all (Unicode) characters in data
- Most frequent ngram pairs ↦ a new ngram in the vocabulary

Dictionary

5 l o w
2 l o w e r
6 n e w **es** t
3 w i d **es** t

Vocabulary

l, o, w, e, r, n, s, t, i, d, **es**

Add a pair (e, s) with freq 9

(Example from Sennrich)

Source: CS224n

Byte Pair Encoding

A word segmentation algorithm:

- Start with a unigram vocabulary of all (Unicode) characters in data
- Most frequent ngram pairs ↦ a new ngram in the vocabulary

Dictionary

5	l o w
2	l o w e r
6	n e w **est**
3	w i d **est**

Vocabulary

l, o, w, e, r, n, s, t, i, d, es, **est**

Add a pair (es, t) with freq 9

(Example from Sennrich)

Source: CS224n

Byte Pair Encoding

A word segmentation algorithm:

- Start with a unigram vocabulary of all (Unicode) characters in data
- Most frequent ngram pairs ↦ a new ngram in the vocabulary

Dictionary

5	**lo** w
2	**lo** w e r
6	n e w est
3	w i d est

Vocabulary

l, o, w, e, r, n, s, t, i, d, es, est, **lo**

Add a pair (l, o) with freq 7

(Example from Sennrich)

Source: CS224n

Example word segmentation with BPE

- Example: newest -> n e w e s t
- Find the most frequent pair in BPE vocabulary: es
  - n e w es t
- Find the most frequent pair in BPE vocabulary: est
  - n e w est
- Stop if no more pair is found in the vocabulary: n e w est

Dictionary

5	**lo** w
2	**lo** w e r
6	n e w est
3	w i d est

Vocabulary

l 7, o 7, w 16, e 11, r 2, n 6, s 9, t 9,
i 3, d 3, es 9, est 9, lo 7

Source: CS224n

Example of MT using BPE for both source and target languages

- Input sentence in English: “health research institutes”
- Input sentence segmentation by using BPE of English:
  - health research institutes
- Output of MT with decoder based on BPE of German:
  - Gesundheits ##forsch ##ungsin ##stitute
- Post-processing of combining word pieces into word
  - Gesundheitsforschungsinstitute

https://arxiv.org/abs/1508.07909

Source: CS224n

Byte Pair Encoding

- Have a target vocabulary size and stop when you reach it
- Do deterministic longest piece segmentation of words
- Segmentation is only within words identified by some prior tokenizer (commonly Moses tokenizer for MT)
- Automatically decides vocab for system
  - No longer strongly "word" based in conventional way

Byte-level BPE

- The base vocabulary that gets all base characters can be quite big if one allows for all unicode characters
- GPT-2 uses bytes as the base vocabulary (which gives a size of 256)
  - Can tokenize any text in Unicode without needing an unknown token
  - vocabulary size of 50,257, which corresponds to the 256 bytes base tokens, a special end-of-text token and the symbols learned with 50,000 merges.

https://huggingface.co/transformers/master/tokenizer_summary.html

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners. 2019.

Source: CS224n

Wordpiece/Sentencepiece model

- Google NMT (GNMT) uses a variant of this
  - V1: wordpiece model
  - V2: sentencepiece model
- Difference to BPE
  - Rather than char *n*-gram count, uses a greedy approximation to maximizing language model log likelihood to choose the pieces
  - Add *n*-gram that maximally reduces perplexity

Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. In *ICASSP* (pp. 5149–5152).

Source: CS224n

Wordpiece/Sentencepiece model

- Wordpiece model tokenizes inside words (like BPE)
- Sentencepiece model works from raw text
  - Treats raw text just as a sequence of Unicode characters
  - Whitespace is handled specially, replaced with e.g. '_'
  - You can reverse things at end by joining pieces

  - https://github.com/google/sentencepiece
  - https://arxiv.org/pdf/1804.10959.pdf

Source: CS224n

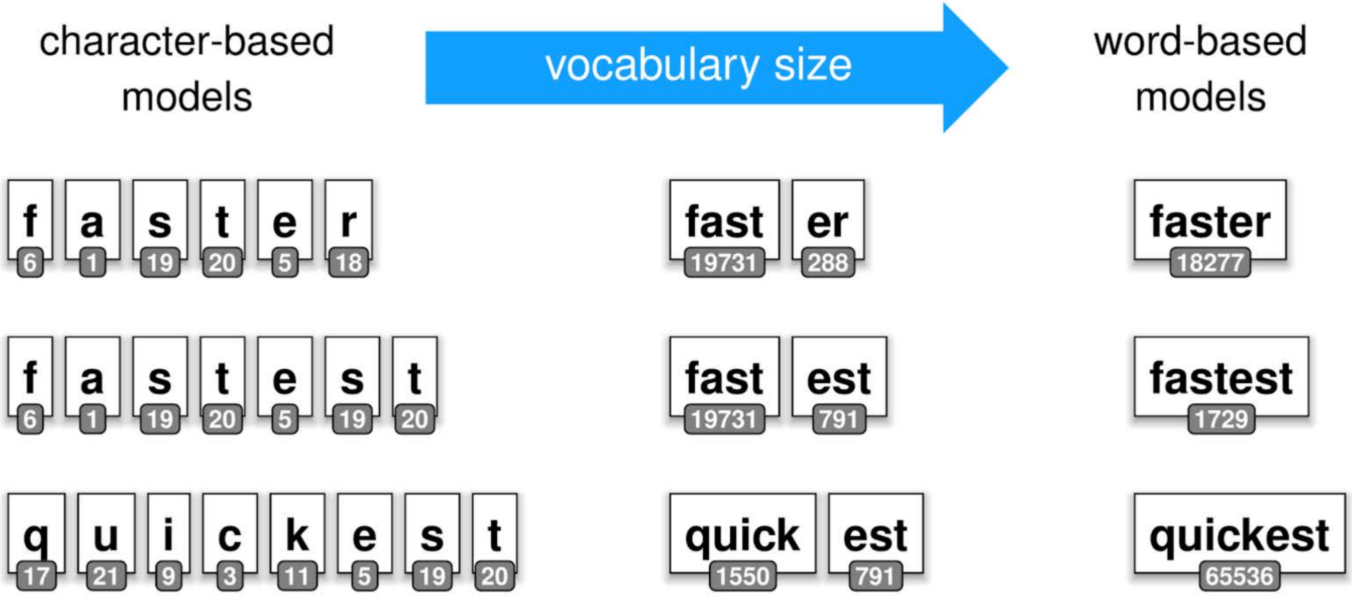
Wordpiece/Sentencepiece model

- BERT uses a variant of the wordpiece model
  - (Relatively) common words are in the vocabulary:
    - at, fairfax, 1910s
  - Other words are built from wordpieces:
    - hypatia = h ##yp ##ati ##a
- If you're using BERT in an otherwise word based model, you must deal with this

Source: CS224n

Vocabulary size trade-off

- Subword model reduces vocab size to train a machine learning model
- On the other side, it increases input sequence's length
  - Issue on model with non-linear complexity over the input sequence's length



https://colab.research.google.com/github/huggingface/transformers/blob/master/notebooks/01-training-tokenizers.ipynb

4. Hybrid NMT

- A best-of-both-worlds architecture:
  - Translate mostly at the word level
  - Only go to the character level when needed (rare words)

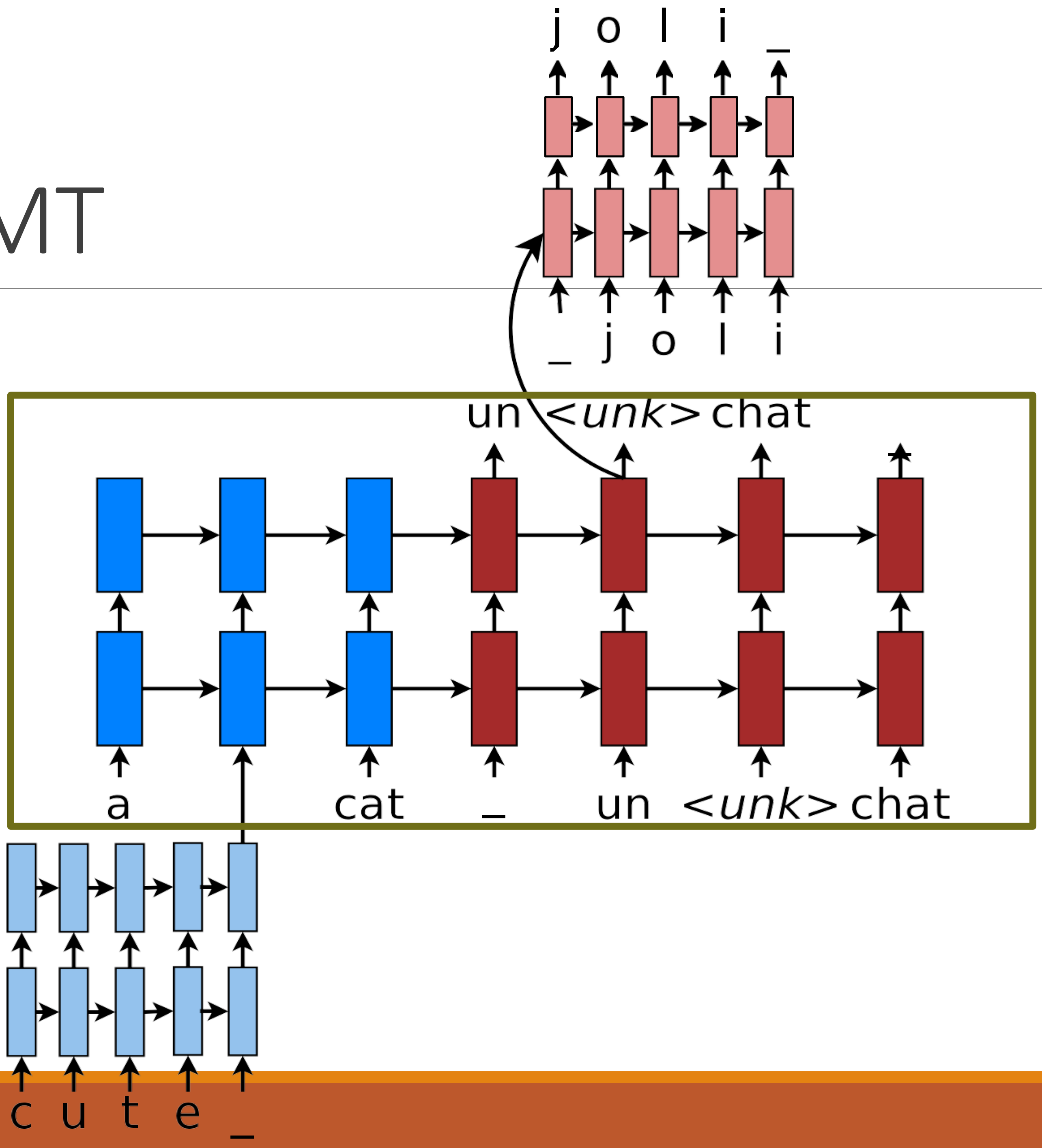
- More than 2 BLEU improvement over a copy mechanism (exact word string from source to target sentence) to try to fill in unknown words

Thang Luong and Chris Manning. ***Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models****. ACL 2016.*

Source: CS224n

Hybrid NMT

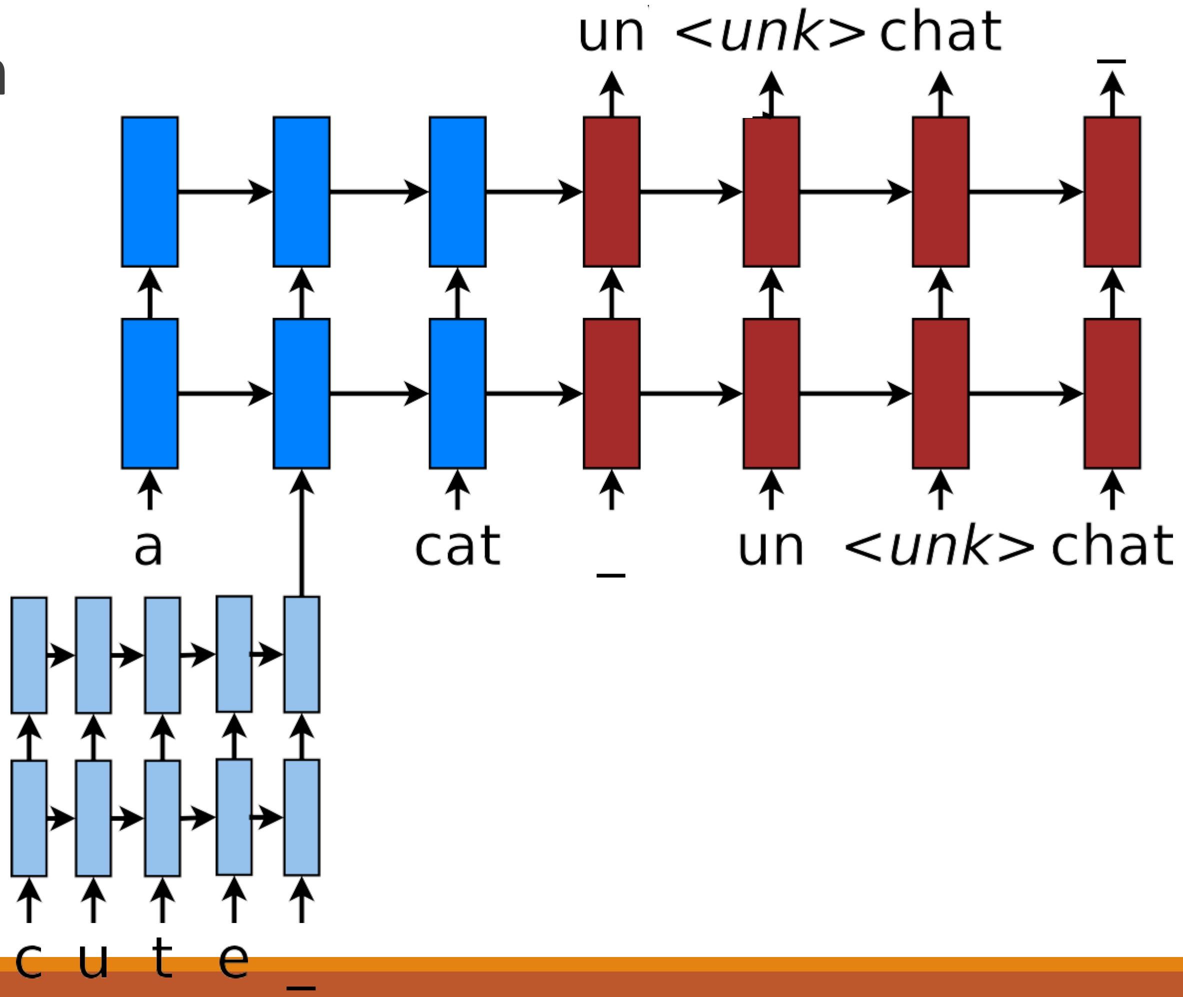
Word-level
(4 layers)



Source: CS224n

2-stage Decoding

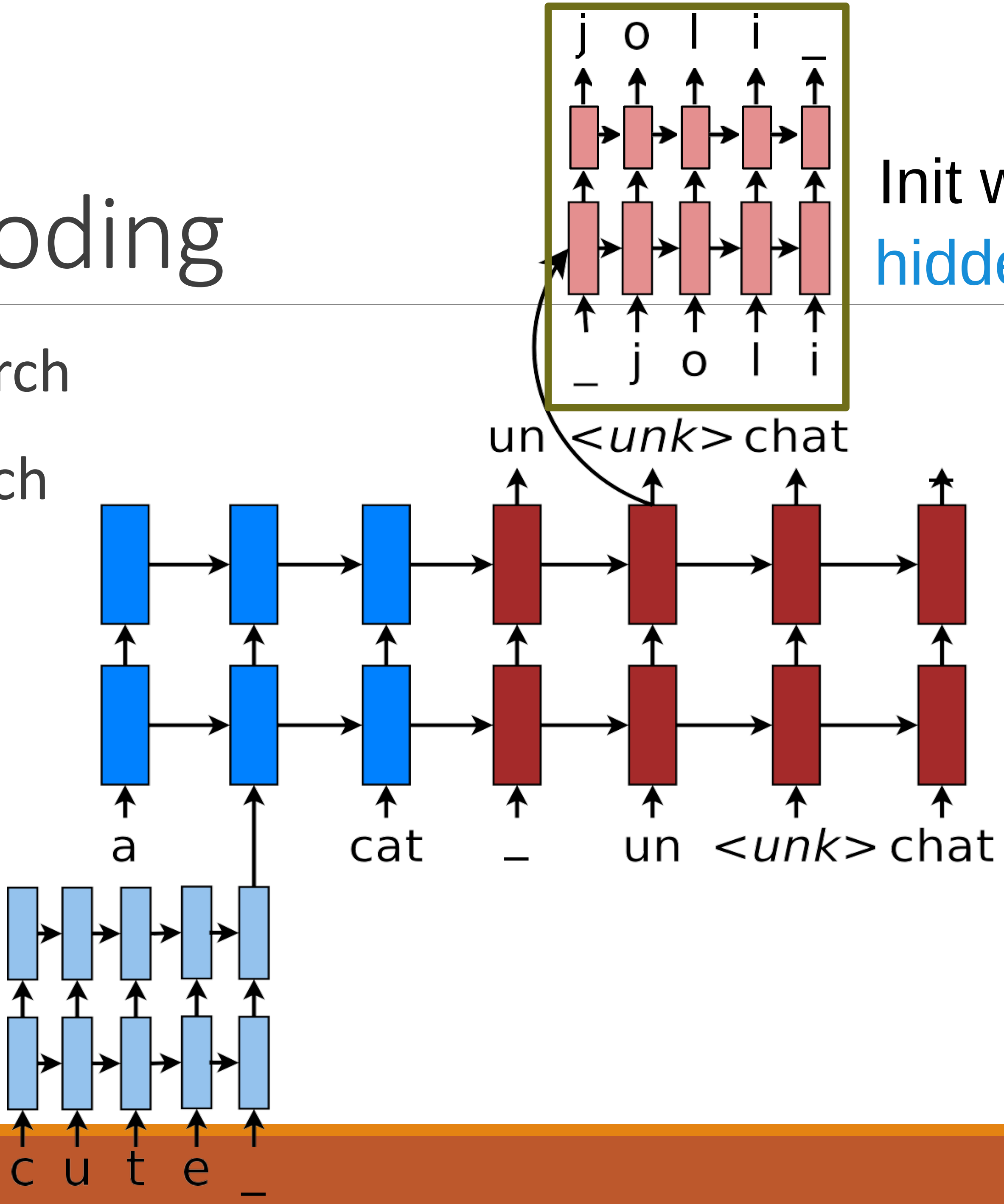
- Word-level beam search



Source: CS224n

2-stage Decoding

- Word-level beam search
- Char-level beam search for <unk>



Init with word hidden states.

Source: CS224n

English-Czech Results

- Train on WMT'15 data (12M sentence pairs)
  - newstest2015

Systems	BLEU
Winning WMT'15 (Bojar & Tamchyna, 2015)	18.8
Word-level NMT (Jean et al., 2015)	18.3
Hybrid NMT (Luong & Manning, 2016)*	**20.7**

30x additional data
3 systems (2 MT, 1 post-editing)

Large vocab
+ copy mechanism

Source: CS224n

Sample English-Czech translations

source	The author **Stephen Jay Gould** died 20 years after **diagnosis** .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

Char-based: wrong name translation

Source: CS224n

Sample English-Czech translations

source	The author *Stephen Jay Gould* died 20 years after *diagnosis* .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

Word-based: incorrect alignment

Source: CS224n

Sample English-Czech translations

source	The author ***Stephen Jay Gould*** died 20 years after ***diagnosis*** .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

Char-based & hybrid: correct translation of diagnóze

Source: CS224n

Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její $\boxed{\textbf{11-year-old}}$ dcera **Shani ,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

Word-based: identity copy fails

Source: CS224n

Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani** **,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

Hybrid: correct, 11-year-old – jedenáctiletá

Wrong: Shani Bartová

Source: CS224n