

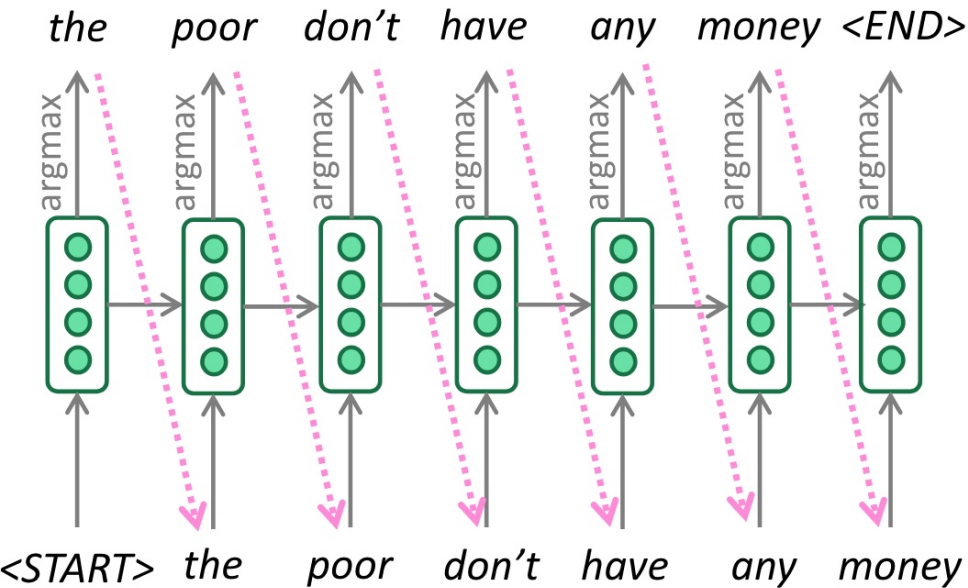
Machine Translation

Jordan Boyd-Graber

University of Maryland

Phrase-Based Models

Adapted from material by Mohit Iyyer, Luke Zettlemoyer, Kalpesh Krishna, Karthik Narasimhan, Greg Durrett, Chris Manning, Dan Jurafsky



Argmax at every time step

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (1)$$

- top- k
- Nucleus / top- p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (1)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (1)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p : Only sample from highest items with at least p probability
- Temperature

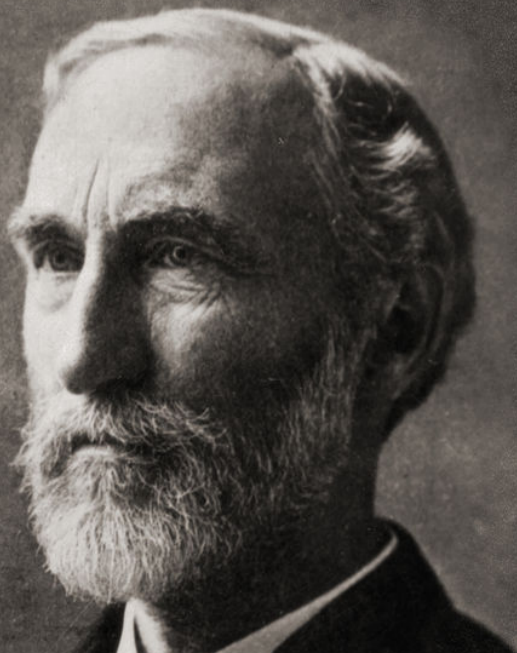
Sampling Methods

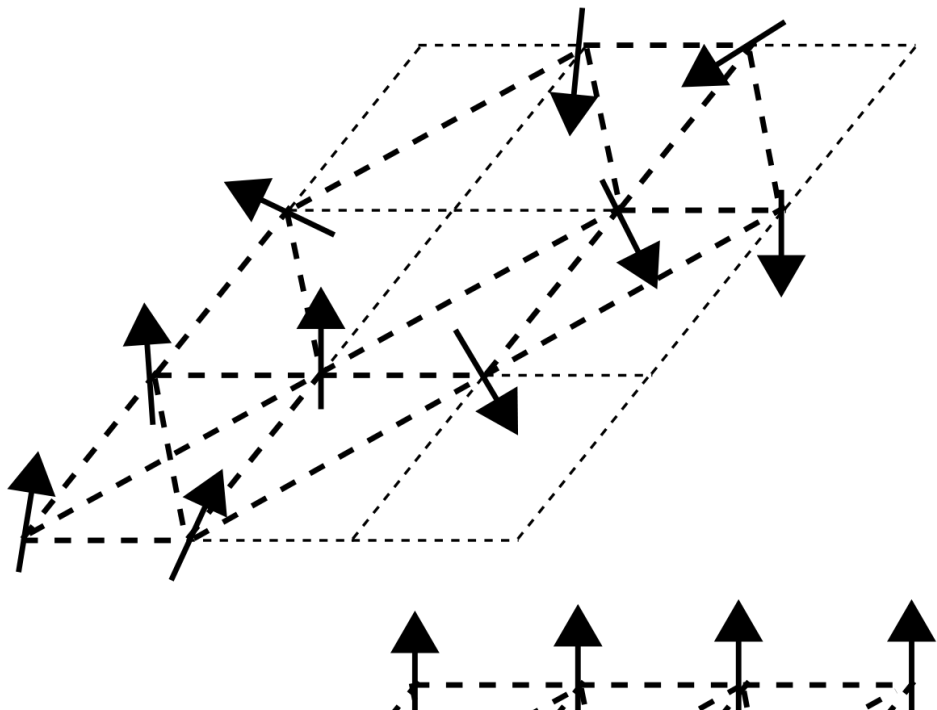
Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (1)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p : Only sample from highest items with at least p probability
- Temperature

$$p(w) = \frac{\exp\left\{\frac{\beta \cdot \vec{f}(w)}{T}\right\}}{\sum_{w'} \exp\left\{\frac{\beta \cdot \vec{f}(w')}{T}\right\}} \quad (2)$$





Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (3)$$

- top- k
- Nucleus / top- p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (3)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (3)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p : Only sample from highest items with at least p probability
- Temperature

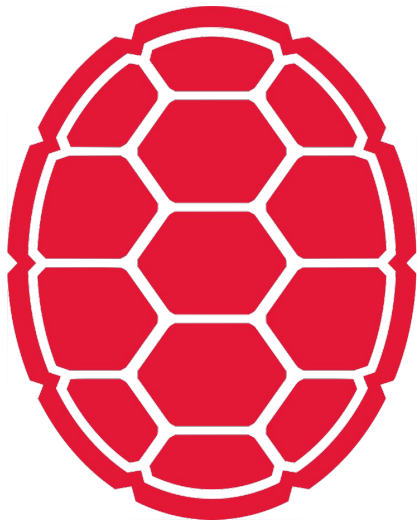
Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\{\beta \cdot \vec{f}(w)\}}{\sum_{w'} \exp\{\beta \cdot \vec{f}(w')\}} \quad (3)$$

- top- k : Only sample from k items with highest probability
- Nucleus / top- p : Only sample from highest items with at least p probability
- Temperature

$$p(w) = \frac{\exp\left\{\frac{\beta \cdot \vec{f}(w)}{T}\right\}}{\sum_{w'} \exp\left\{\frac{\beta \cdot \vec{f}(w')}{T}\right\}} \quad (4)$$



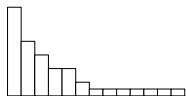
Top- k

Nucleus

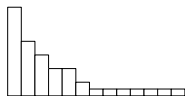
Temperature ($T=0.1$)

Temperature ($T=2$)

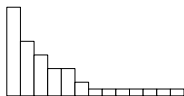
Top-k



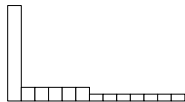
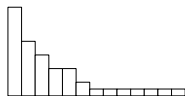
Nucleus



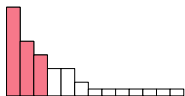
Temperature ($T=0.1$)



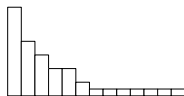
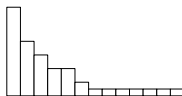
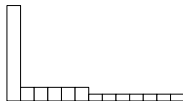
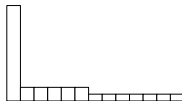
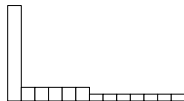
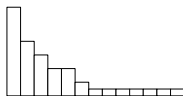
Temperature ($T=2$)



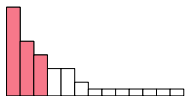
Top- k



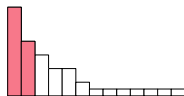
Nucleus

Temperature ($T=0.1$)Temperature ($T=2$)

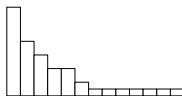
Top-k



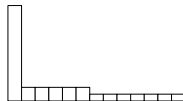
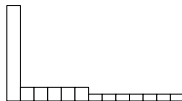
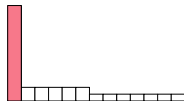
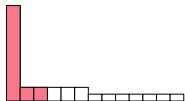
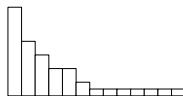
Nucleus



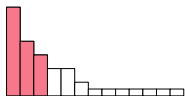
Temperature ($T=0.1$)



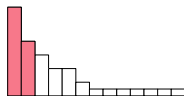
Temperature ($T=2$)



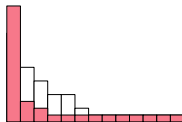
Top-k



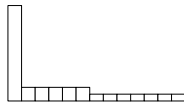
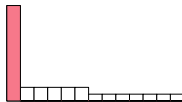
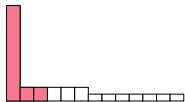
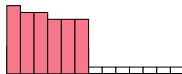
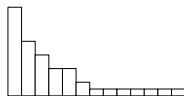
Nucleus



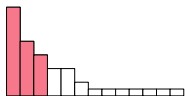
Temperature ($T=0.1$)



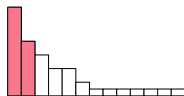
Temperature ($T=2$)



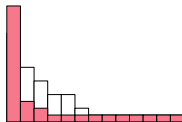
Top-k



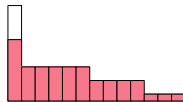
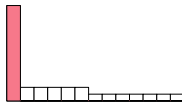
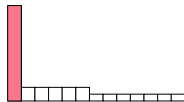
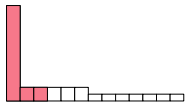
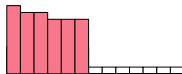
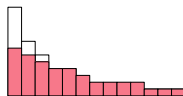
Nucleus



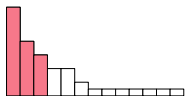
Temperature ($T=0.1$)



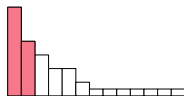
Temperature ($T=2$)



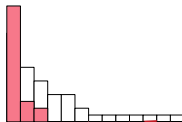
Top-k



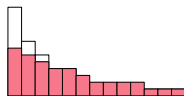
Nucleus



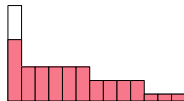
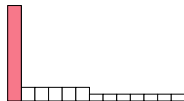
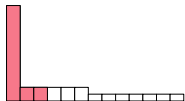
Temperature ($T=0.1$)



Temperature ($T=2$)



Probabilities too small to see!



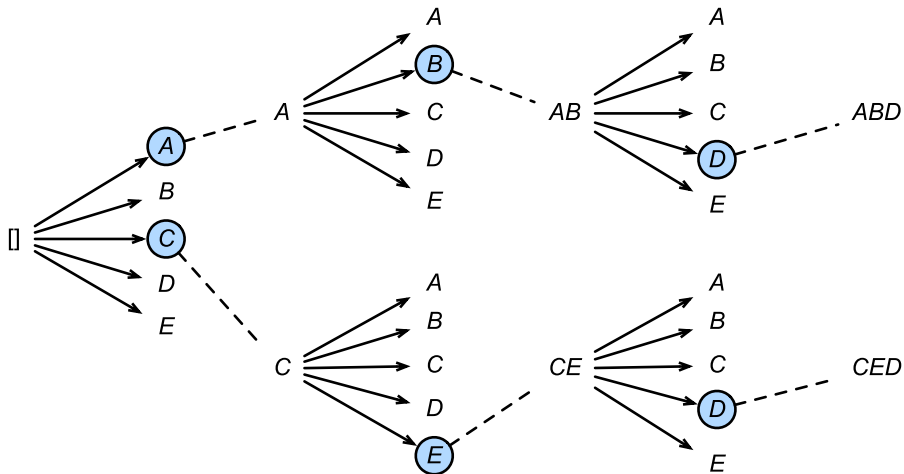
What do you do with samples?

- Getting out of being stuck in a garden path
- Getting diverse outputs
- Combining multiple models together
- Rescoring by a non-probability metric

Time step 1
Candidates

Time step 2
Candidates

Time step 3
Candidates



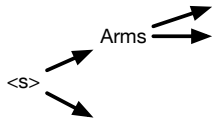
From Zhang et al. (Dive into Deep Learning)

<S>

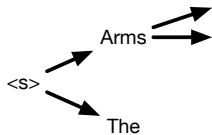
Beam Search Decoding for: Die Arme haben kein Geld



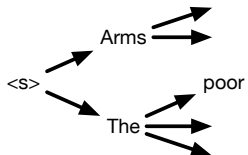
Beam Search Decoding for: Die Arme haben kein Geld



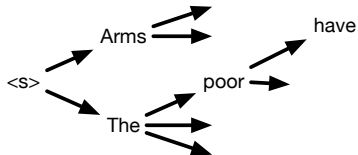
Beam Search Decoding for: Die Arme haben kein Geld



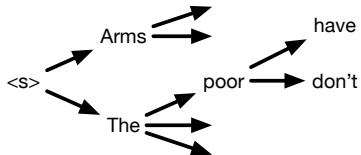
Beam Search Decoding for: Die Arme haben kein Geld



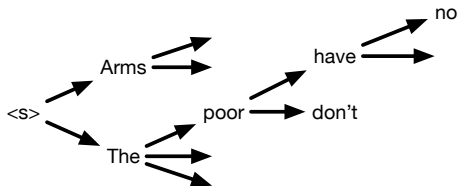
Beam Search Decoding for: Die Arme haben kein Geld



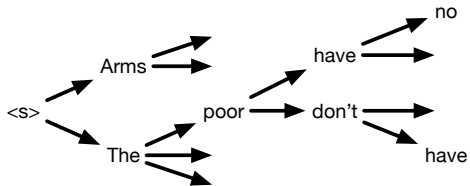
Beam Search Decoding for: Die Arme haben kein Geld



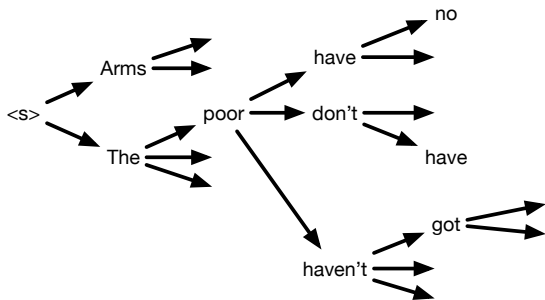
Beam Search Decoding for: Die Arme haben kein Geld



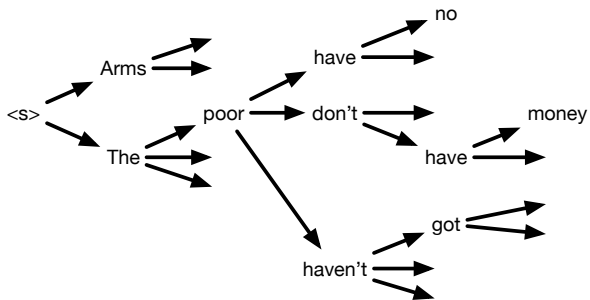
Beam Search Decoding for: Die Arme haben kein Geld



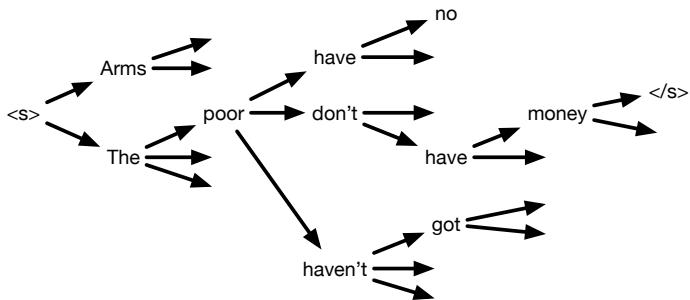
Beam Search Decoding for: Die Arme haben kein Geld



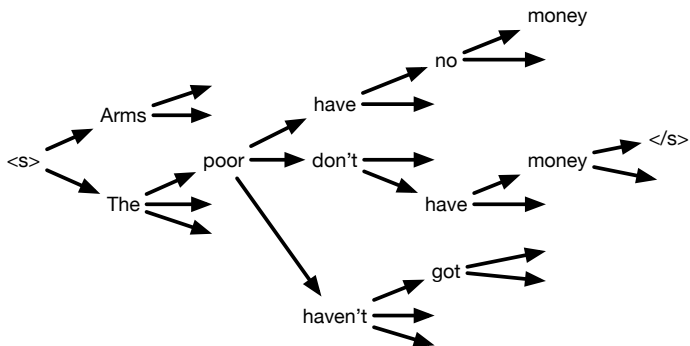
Beam Search Decoding for: Die Arme haben kein Geld



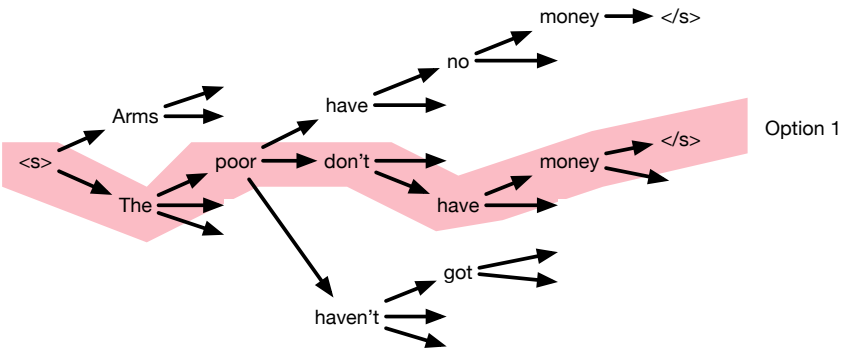
Beam Search Decoding for: Die Arme haben kein Geld



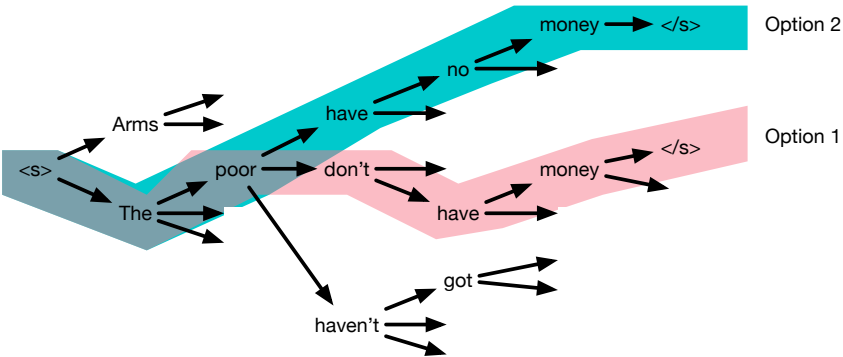
Beam Search Decoding for: Die Arme haben kein Geld



Beam Search Decoding for: Die Arme haben kein Geld



Beam Search Decoding for: Die Arme haben kein Geld



Beam Search Decoding for: Die Arme haben kein Geld

Using multiple sources

- Generate from multiple models
- Generate from multiple directions
- Generate from multiple data
- Generate from multiple temperatures

How to pick?

- Show to a user

How to pick?

- Show to a user
- Take highest probability

How to pick?

Can Neural Machine Translation be Improved with User Feedback?

Julia Kreutzer^{1*} and Shahram Khadivi³ and Evgeny Matusov³ and Stefan Riezler^{1,2}

¹Computational Linguistics & ²IWR, Heidelberg University, Germany

{kreutzer, riezler}@cl.uni-heidelberg.de

³eBay Inc., Aachen, Germany

{skhadivi, ematusov}@ebay.com

- Show to a user
- Take highest probability
- Rerank



RankGen — Improving Text Generation with Large Ranking Models *(EMNLP 2022)*



Kalpesh Krishna



Yapei Chang



John Wieting



Mohit Iyyer

UMass Amherst
Manning College of Information
& Computer Sciences



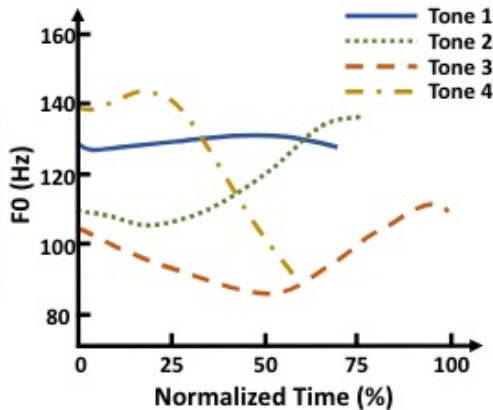
Decoding Method	GPT2-md		GPT2-XL	
	PG19	wiki	PG19	wiki
Nucleus ($p = 0.9$)	73.0	74.6	74.4	75.0
Eta (Hewitt et al., 2022)	76.9	71.2	76.9	74.8
<i>Contrastive methods</i>				
search (Su et al., 2022)	5.3	21.2	54.0	43.2
decode (Li et al., 2022)	65.2	83.2	73.2	84.9
RANKGEN-all-XL (ours)				
rerank full ancestral	79.0	84.9	79.0	86.4
beam search nucleus	76.2	88.9	77.0	89.4

Tone 1: 叔 shū (uncle)

Tone 2: 熟 shú (cooked, familiar)

Tone 3: 鼠 shǔ (mouse, Muroidea)

Tone 4: 树 shù (tree)



Tones in Chinese (for “shu”, not “ma” like I said)

Original Lyrics
(Inconsistent Tone)



sì zài yǎn qián
似 在 眼 前
appear where eye front

As if before my eyes

Inter-syllable pitch alignment score: 0.5

Misheard Lyrics
(Consistent Tone)



sǐ zài yǎn qián
死 在 眼 前
death where eye front

Die before my eyes

Inter-syllable pitch alignment score: 0.75

Misheard lyrics when the tones are wrong

REST: intervals of silence that usually align with word segmentations or punctuation

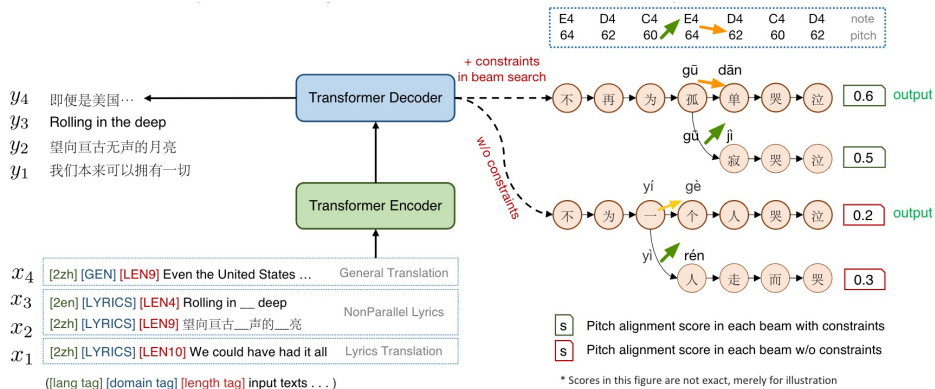
wǒ yǐ wàng jì wǒ céng huó guò diū le gǎn guān bēi shāng cǐ kè wǒ zhàn zài jīng tāo hài
我已 忘记 我曾 活过 丢了 感官 悲伤 此刻 我 站 在 惊涛骇
I have forgotten (that) I've lived. (I've) lost (my) sense (my) sorrow. Right now I'm standing above the terrifying

làng dà hǎi zhī shàng wǒ yǎng tóu wàng wàng xiàng gèn gǔ wú shēng de yuè liang hēi àn
浪 大海之上 我仰头望 望向亘古无声的月亮 黑暗
stormy sea (above). I look up look up to the eternal silence moon. Dark

One character (syllable) aligns
with a group of multiple notes

One character (syllable) aligns
with a single note

Aligning music to translated lyrics



Decoding song translations with tones in decoder

