Lecture 17: Machine Translation

Zhizheng Wu

Agenda

- Recap
- Introduction
- Why machine translation is difficult
- Statistical MT
- Neural MT

Probability of next word

class are the best"



$P(\text{best} | \text{Students from my class are the}) = \frac{C(\text{Students from my class are the best})}{C(\text{Students from my class are the})}$

C(Students from my class are the best) is count of the phrase "Students from my

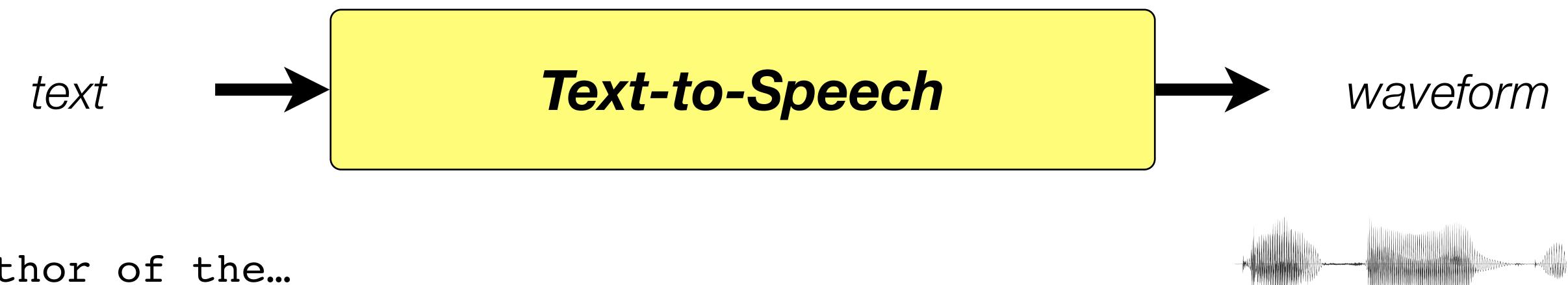
Generalizing bigram to n-gram

From bigram to n-gram

 $P(w_n | w_{1:n-1}) \approx P(w_n | w_{n-N+1:n-1})$

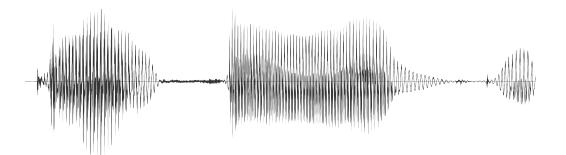
- ► N = 2: bigram
- ► N = 3: trigram
- ► N = 4: 4-gram
- ► N = 5: 5-gram





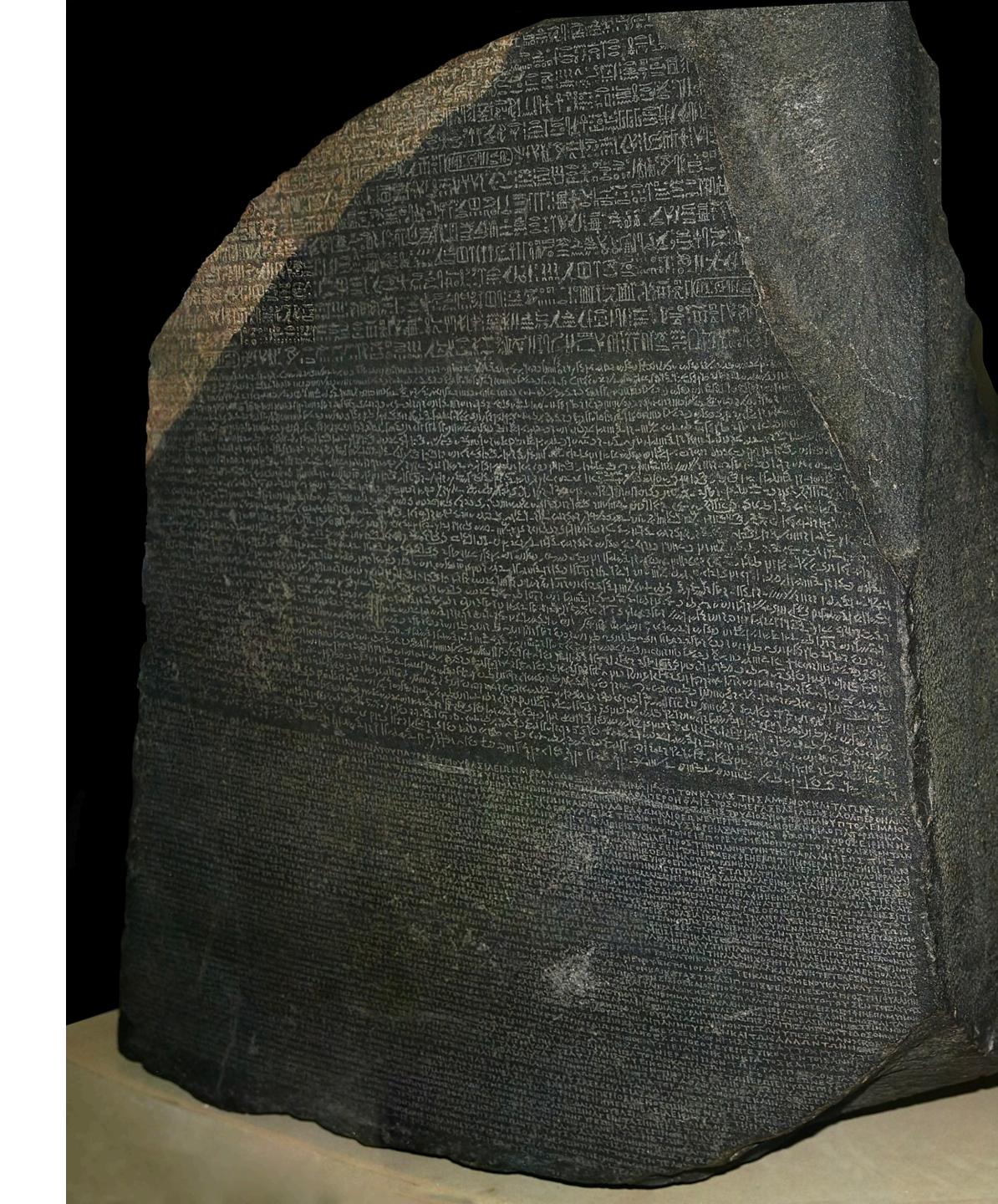
Author of the...







Author of the...



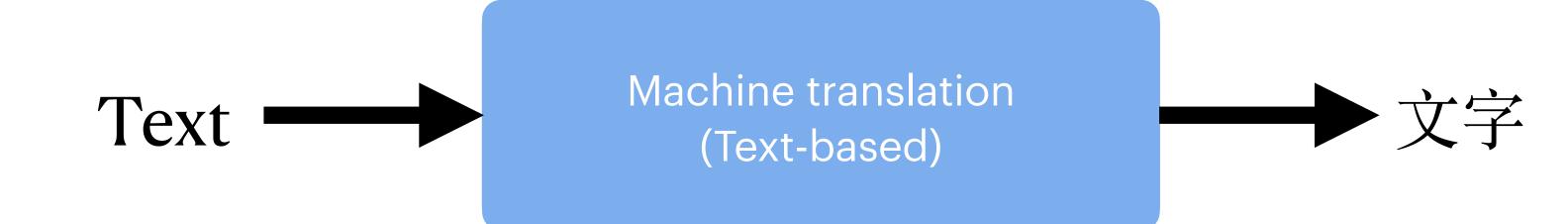
なた命令文『記』をおいて
一般的现象是以10年三大学生会社长长安计会社会社会社会社会社会社会社会社会社会社会社会社会社会社会社会社会社会社会社
and a start a star
三日本には、「「「「「「「「「「「「「「「「「「「」」」」」」」」」」」」」」」」」」
いる意がえるに加加量を引き、気気にないないので、「「「「「「「「「」」」」
「「川田人類王室忍と好意やふいいが近き」「別日言山」を執きにている。「「「川田人類王室忍」」
…。21元川之水川和Y色生化的产品。面Y各是一些名的TILLE和TI都是主要的人们是在这些们会了
」(金豆仁川川田登宅西川三田堂に公室大原会)も認知人の「正室」に認用たい
水川部司が化層で可知器をえる言語が沿向利用にきたる自認なの言語のに置いた」と、「「「「「「「「「」」」、「「」」、「」」、「」」、「」」、「」」、「」」
う語見たるので、「「「「「「「「「「」」」、「「「」」、「「「」」、「「」」、「「」、「」、「
TISTINAL SALAS ALL ALL ALL ALL SALAS SAL
ניבל- עמעשעיבי לעולי גוויבליע געויבי לעלי גוויבליע געויבי לעלי געויבי לעלי געויבי לעולי גוויבי לעולי גוויבי לעלי גוויבי אין איין איין איין איין איין איין אי
アンシーシーシーシーシーシーシーシーシーシーシーシーシーシーシーシーシーシーシー
بد المراجع المراجع وحد معالم المراجع وحد معالم المراجع ومعالم المراجع وحد المراجع ومعالي ومعالي ومعالية ومعالية ومعالية معالية ومعالية ومعالية ومعالية معالية المراجع ومعالية معالية ومعالية ومعالية ومعالية وم
در من
ジャッシュンシンシンシンシンシンシンシンシンシンシンシンシンシンシンシンシンシンシン
ווגראי ביז ביל אלוו וואל ביל אלו וואל ביל אלו ביל אלי איין ביל אלי איין ביו אלי ביז אלי ביאווע ברב איין ביל געראיין ביל געראיין ביל געראיין ביל געראיין ביל געראיין ביל געראיין
ערבווגלוועריו ביו אינג אוויינג להיין אווידער אינגער
いいたいの」、「「「「「」」、「「」」、「」、「」、「」、「」、「」、「」、「」、「」、「
און גיון אוגעוטובוווגלאין גייגאוגעריין אוגעויין געויין געוי אוגעויען געויען געויען געויין געוי
5/5/10/20/20/20/20/20/20/20/20/20/20/20/20/20
المعالية من من المعالية من من المعلية من
(+++, =+================================
- 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1
BALILEYONTSTOYNEDYKA I MARALABONTOS THNBASI LEIANMARATOY MATPOSKYPIOYBASILEIANMERALOAO TOYTOYTHNAIPY OT TOHKATAS THE AMENOY KAI TA OROVSEYSEVSTANTINARHYOD TEPOY TOYTONBION TALANO PA DA NEGANOPOSKYPIOYBASILEIANMERALOAO TOYTONTOHKATAS THE AMENOY KAI TA MERALDASINEYSTANTEA PAKATASIN TONBION TALANO PA DA NEGANOPOSKYPIOYBASILEIAN TOSKYPIOY THAKON TAE THPIOANKAOA DEPON DA IS TOS OMERALDASIL
ALANOBIOYHEANHENDY TO TOYABAE TOY IENA TOYE HEREALAE TOYTOYANG WHANISTOLEAKIMALEHAIOHADIEAAKEHAMHI KHEKONO ZATIHZTOYAHETIOY TOYKAHTA Deoyenida No yieyahiya dooroo dooro Thinto a mangamana analae ya a sa a a a a a a a a a a a a a a a a
O EINKINT FORD PALKANIE PORALMALTIER MALTALE AND THE AND THE ALL ENTRY AND AND THE AND
TO TE MAY TO SOM TALKAITON AND TO THIN ASTRONOMY TO THAT SOME HAVE AN ANTAL THAT IS A CONTRACT OF THE MAY AND THE STATE AND THE SAME AN
E TOMENE IN EINTH LEVY TOY BENERALTON AND AND AND AND AND AND AND AND AND AN
IS FRHEUMAN TETA AT MALE TA NEW AND THE TA NEW AND THE TAN TO BE TALLEN THE SENDERAL THE STATE AT CALL TO THE AT ANALANTAN THE COLLEGE TO TAT TAN THE TAIL TAN TAN THE COLLEGE TO TAT TAN THE COLLEGE TO TAT TAN TAN TAN TAN TAN TAN TAN TAN TAN
FRANKINGTAN ENTRY CONTRACTION CONTRACTOR CONTRACTION OF THE AND A THE PERMIT ON THE AND A THE AN
NOTOR ALL AT A MARKINT HAVENTAL BOY TO IT ALL HAVEN A TELAHAME NAKALA YADALA NA ANA ANA ANA ANA ANA ANA ANA ANA A
REALKATELXENEK NCA AAH TONA NOXY PATAL TAL TOMATA TANNOTAMAN XOPHTHIA SELLAY TAXIHMANABAIH HERAKHANIMLAMENDY EXTA IOTABAIE TEKANIOISTKATA AY TANEHOA WAIXIONAITHNTE NOAINKA TAKPATOLEIAE KAITOY LEHAY THIAS EBELS MAN TAXIHMATA MAD EDYKOAI FONKAIKATA SITABAIE TEKANIOIST TEKANI SOY ATA
ΤΟΠΑΥΤΙΚΑΊΤΗΙΕΑΥΤΟΥ ΑΤΙΛΕΊΑΙΠΑΝ ΤΑΣΕΚΟΛΑΤΕΝΚΑΊ ΤΛΙΚΟΝΤΑΤΚΑΣΟΥΥ ΤΟΥΠΑΤΙΟΙΚΑΙΤΗΝΥΛΙΛΑΝΟ. ΤΑ ΓΑΙΚΑΤΑΙΕΛΑΔΙΚΗ ΕΛΗΤΑΕΠΑΛΑ ΡΕΝΟΗΕΙΕΝΟΗ ΤΟΠΑΥΤΙΚΑΊΤΗΙΕΑΥΤΟΥ ΑΣΙΛΕΊΑΙΠΑΝ ΤΑΣΕΚΟΛΑΤΕΝΚΑΙ ΕΝΚΟΝΤΑΙΚΑΟΝΚΑΙ ΜΑΝΠΑΛΤΕΙΝΚΟΗΠΡΙΙΤΟΣΤΗΛΕΛΕΙΟΗ ΙΑΥΤΑΙΚΑΙΤΑΙΚΑΛΑΤΗΠΑΛΑΤΗΠΑΛΑΤΗΠΑΛΑΤΗΝΟΕΕΙΕΝΟΗ ΤΗΣΙΕΓΕΙΣΟΦΕΙΛΦΜΕΝΑΕΙΣΤΟΒΑΣΙΔΙΚΟΝ ΕΛΕΛΟΥΟΓΔΙΑΥΣ ΤΑΥ ΣΟΝΤΑΣΙΣΙ ΤΟΥ ΤΕΚΛΙΑΛΟΥΛΟΥΠΑ ΠΟΙΟΥΤΗΛΙΤΗΛΙΑΥΤΗ ΤΑΝΑΙΚΗΤΑΙΤΑΙΑΛΑΤΗΠΑΛΑΙΚΗΕΙΑΙΟΙΕΙ ΕΝΟΕΕΙΕΝΟΗ
TABLE ANON THIARY PAITALTEAN THE VEHING A ABAANIEAT OR ALT OR ALT ALL AND THE PAITALE ALT ALL AND THE ANTAL THE ANTAL AND ANTAL THE ANTAL AND ANTAL THE ANTAL AND ANTAL AN
TATE I TO TECHNER TO TECHNER TO THE ALTER AND THE TAKANA TO THE ALTER TO THE ALTER A
THEBAILLEIAIDIANENSYLNIAY TAIKAITOITTERNOLLEILTONAILATAADEREN AANTYKHEGOSENTOILIBEYD TANKATATHNA A MIEPAN MAN TAY MANATOTATAN TOLANA TAY ANTAN TAY
H NESTENWATEHIETAINTOARMANY TEVENAMINAN TOTTHIAN TOTTHIAN TOTALING THE TALE TAIL TO DEST TOVIEROYALOOYIAY TAILONAL THE THOMALE TAIL THE
Η πρέξειναλεφήτεται τολεμάλογ τονεπολητικά που τημαχή πονεπολητικά του μαργολογιού 0000 του τολαργολογιέλον το μεργολογιά τον τελικα το του ται του
TERE JOHI TAMONIZOME NA THINA FALH4EI THI SALI A EIA I ENIOEI NAIGENAIGENAI START OTAR I ARIA TANOY KATA TON FOEIFI ME NONBAIA EIONS YAAK THIA MA TIEITIN TW BATIAEAI TOYE NI AANHINGIHIJAN TATI HIM TEANAY APANKAI THIKATA KUENEI THITTIAKAAATOY TOYMEZOPHE KIII TATE ME GAMA TOYEAII AE GAMA TOYEAII AE TAI O ROMAI ASKAM TIEITIN TW BATIAEAI TAYEAI I AANHINGIHIJAN TATI HIM TEANAY APANKAI THIKATA KUENEI THITTIAKAAATOY TOYMEZOPHE KIII TATE ME GAMA TOYEAII AE TAI O ROMAI ASKAM
FYATONIE POLIKA TAMHMAKAII YN TEAEIN ENAY TOI I OY ZIAL KAIZABHAAL FAITAAAA TA HOMIZONENARAAAANIA TAI CAANACAANA TA TETIN TA L TEFIN MENAL TAADA TA HOMIZONENARAAANIA TAI CAANACAANA TA TETIN TA L TEFIN MENAL TATA TA TETIN TALA TA TA TETIN TALA TA
KANYAN ANO TRENOVMINIAL TOYOCHIMEN INEN TEENAILVALE TEKANKAPALO TANE IN TANYA UTALIAN INANYA UTALIANI TANYA TANA TILANYA TANYA TANA TILANYA TANYA TANY
TEPEOYA IDEY TOILTEIEFOILKAIE TXANII KAIEN AHNIKOIZFRAMMALIN KAILTAILENE KALTAITANTEARATAMKAISEY TEPA

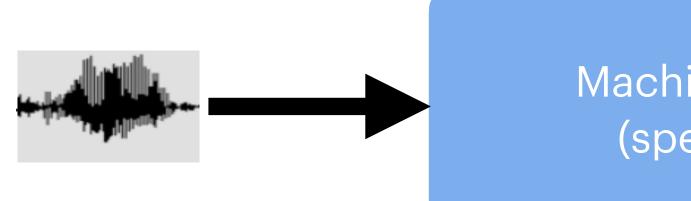


The Rosetta Stone

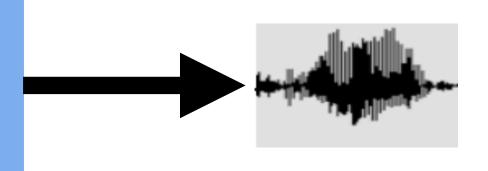
- The same text in three languages
- This is an instance of parallel text
 - The Greek inscription allowed scholars to decipher the hieroglyphs

Machine Translation

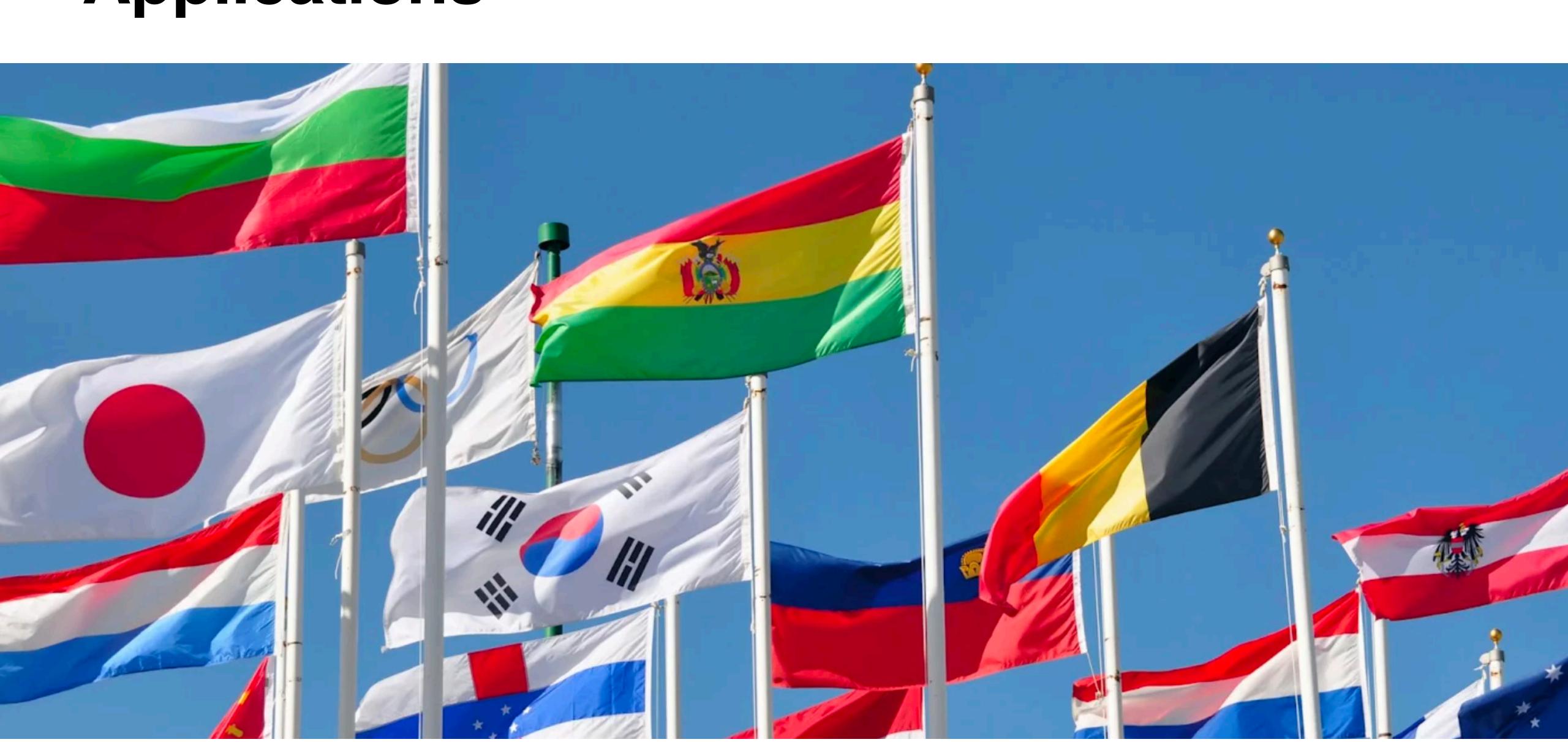


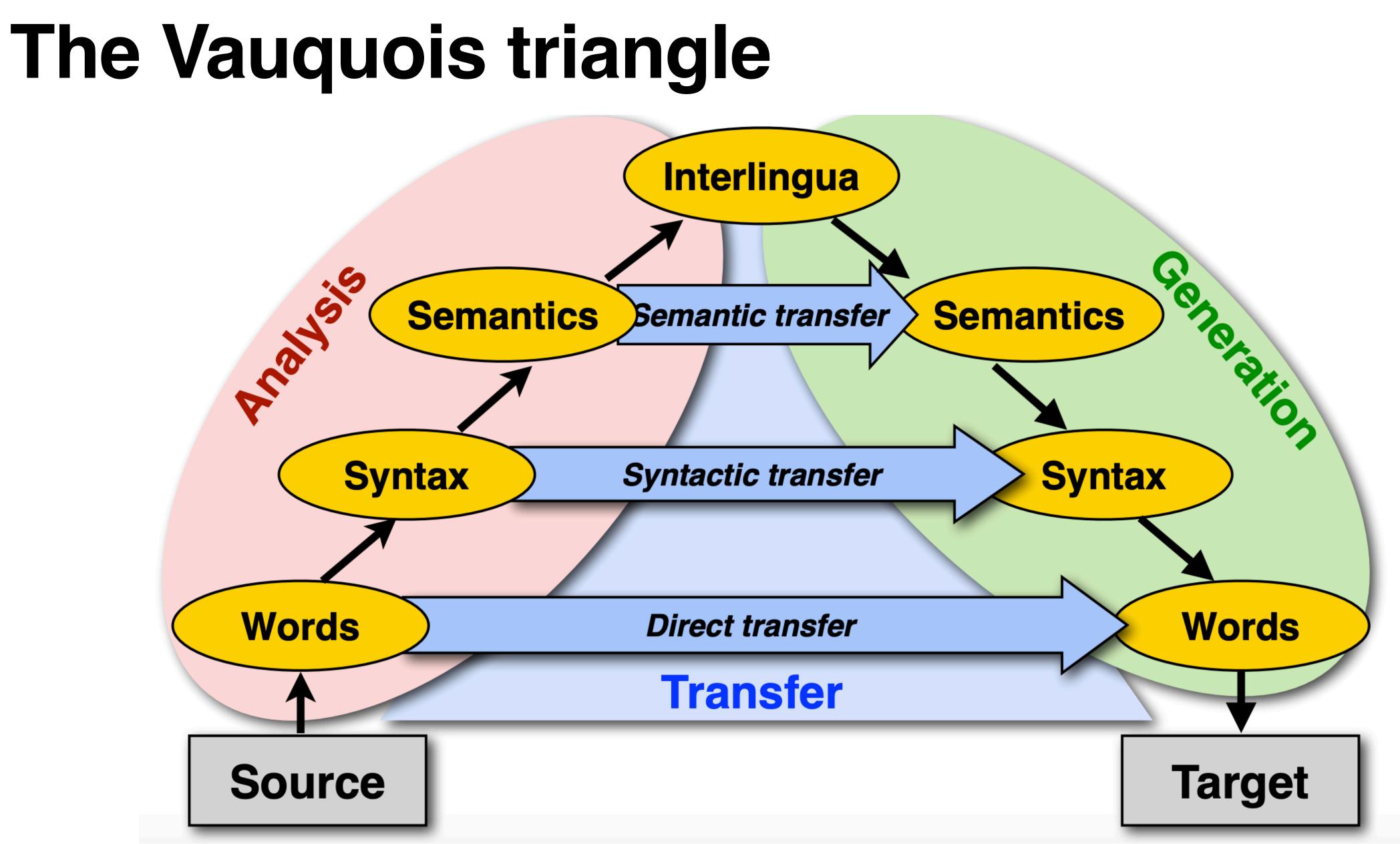


Machine translation (speech-based)



Applications





https://www.youtube.com/watch?v=cgI_Zb1VMBE

Why machine translation is difficult?

- Lexical divergences
- Syntactic divergences
- Semantic divergences
- Word-to-Word correspondence divergences

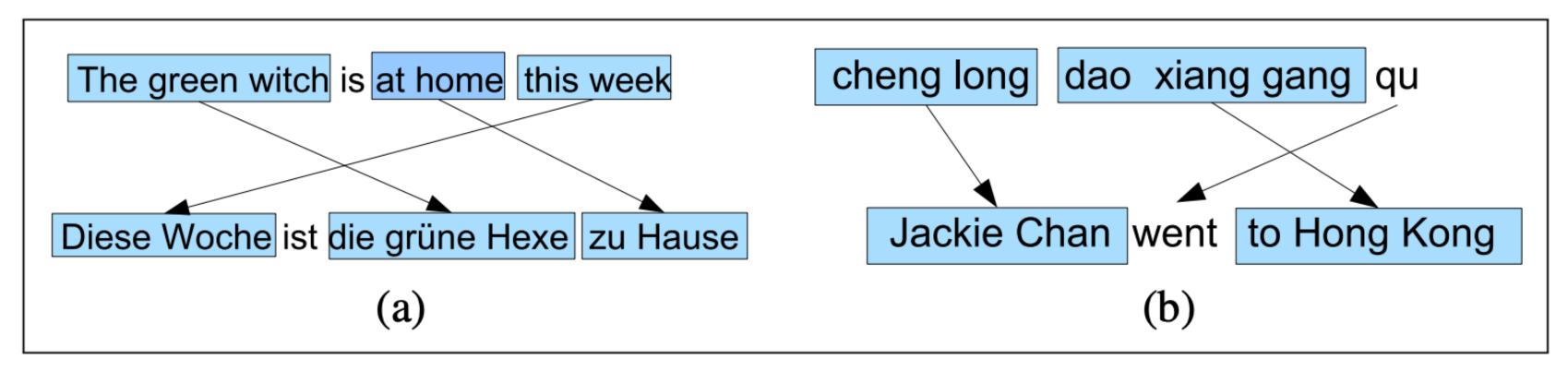
Lexical divergences

- ▶ English: brother vs Chinese: gege, didi (哥哥、弟弟)
- English: sister vs Chinese: jie jie, meimei (姐姐、妹妹)
- ► English: Play vs Chinese: 打、弹、...
- - English-German:
 - (river) bank Ufer
 - (financial) bank Bank

The different senses of homonymous words generally have different translations:

Syntactic divergences

Word order



- Head-marking vs. dependent-marking
 - Dependent-marking (English): the man's house
 - Head-marking (Hungarian): the man house-his

an's house house-his

Semantic divergences

- English has a progressive aspect
 - 'Peter swims' vs. 'Peter is swimming'
- German can only express this with an adverb
 - 'Peter schwimmt' vs. 'Peter schwimmt <u>gerade</u>' ('swims currently')



Word-to-word correspondences

One to-one:

One-to-many: (and reordering)

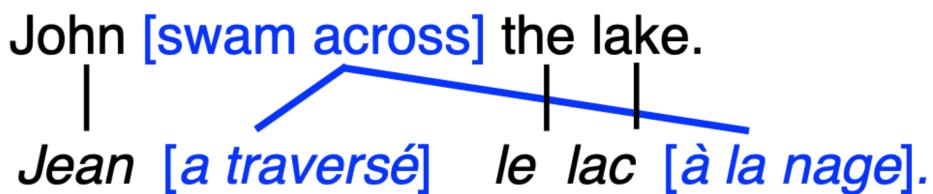
Many-to-one: (and elision)

Many-to-many:

John loves Mary. Jean aime Marie.



- John is a [computer scientist].
- Jean est informaticien.



Statistical MT

Given input in the source language, S,... e.g. a Chinese sentence... 主席: 各位議員, 早晨。

... return the best translation in the target language, T^* e.g in English:

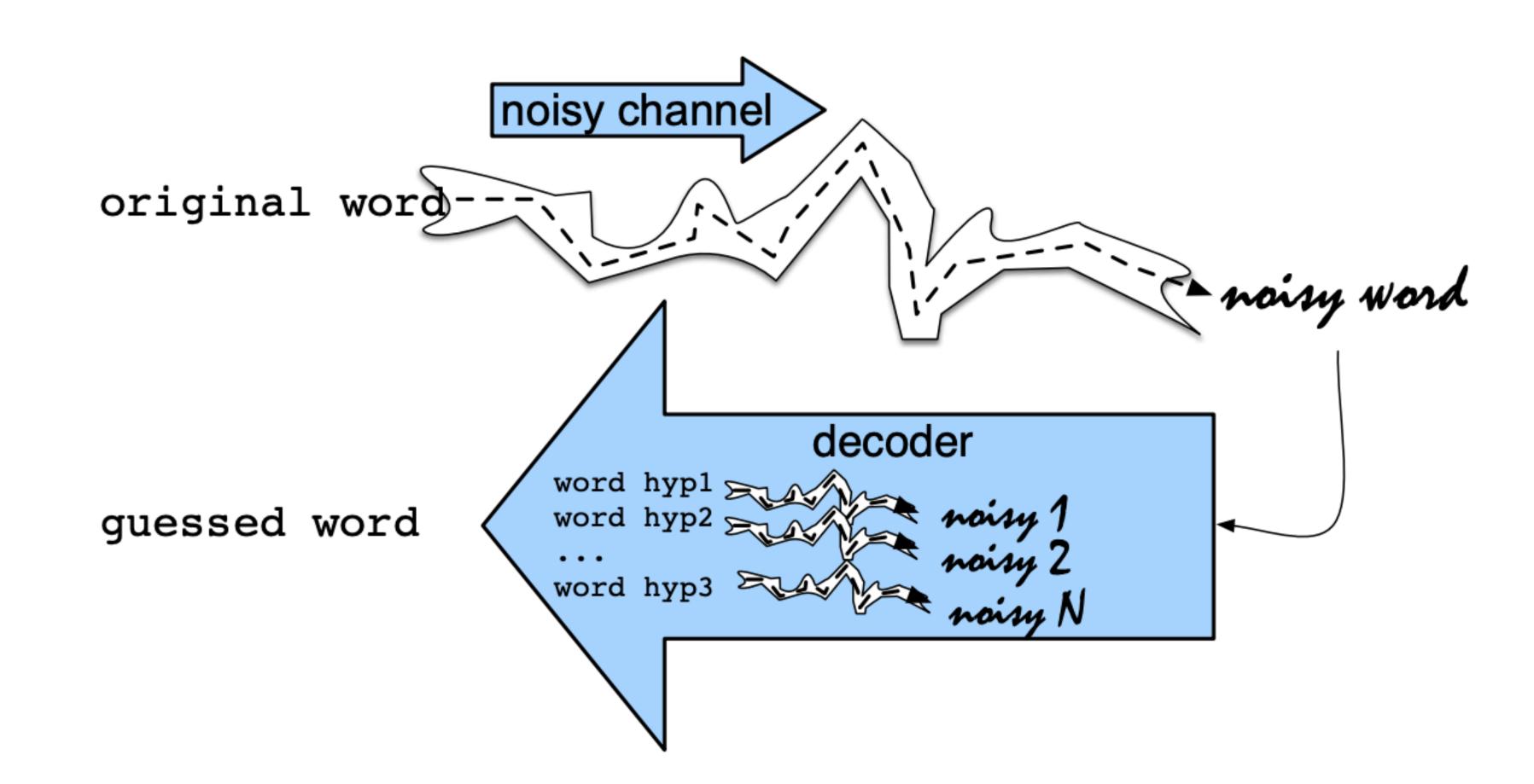
President: Good morning, Honourable Members.

We can formalize this as $T^* = \operatorname{argmax}_T P(T \mid S)$

https://courses.grainger.illinois.edu/cs447/sp2023/Slides/Lecture16.pdf



Noise channel model

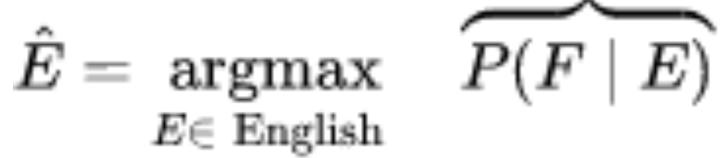


https://web.stanford.edu/~jurafsky/slp3/B.pdf

Noise Channel model

Suppose we want to translate a foreign language to English, we can model P(E|F)

by Bayes law, we have the equivalent equation

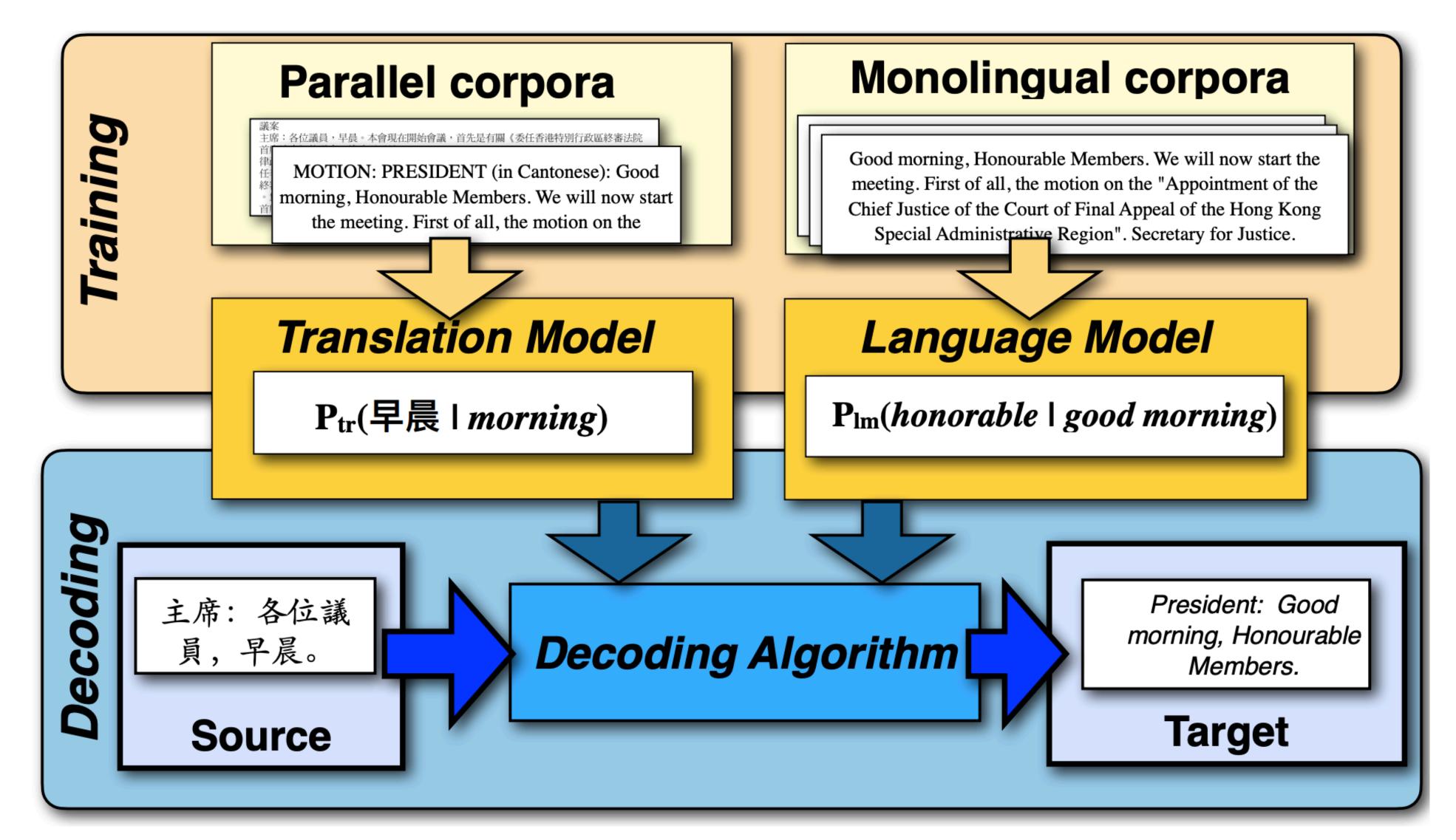


Intuitively, a good model of English, and a good English-to-foreign translator

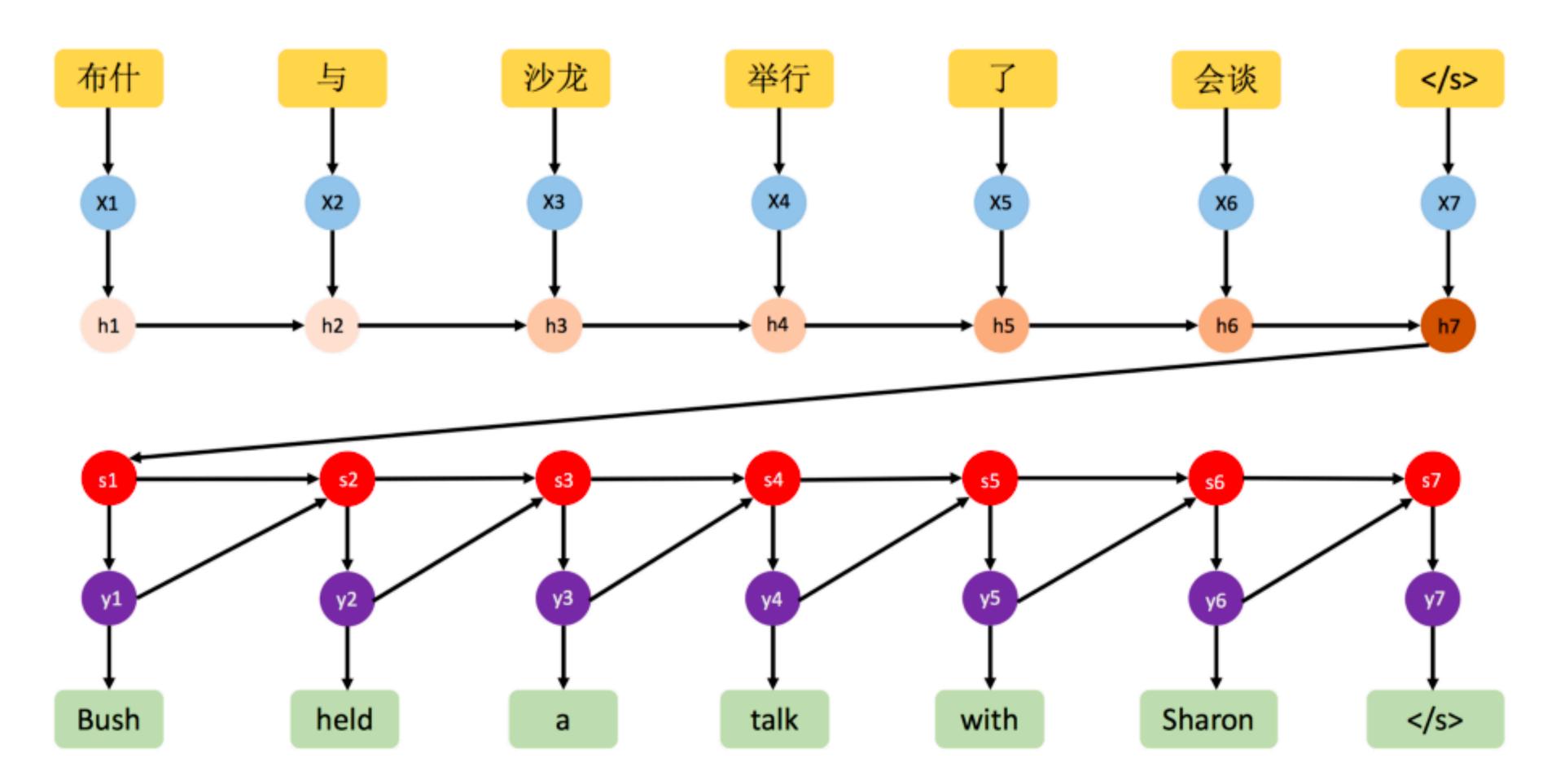


translation model language model P(E)

Training and decoding

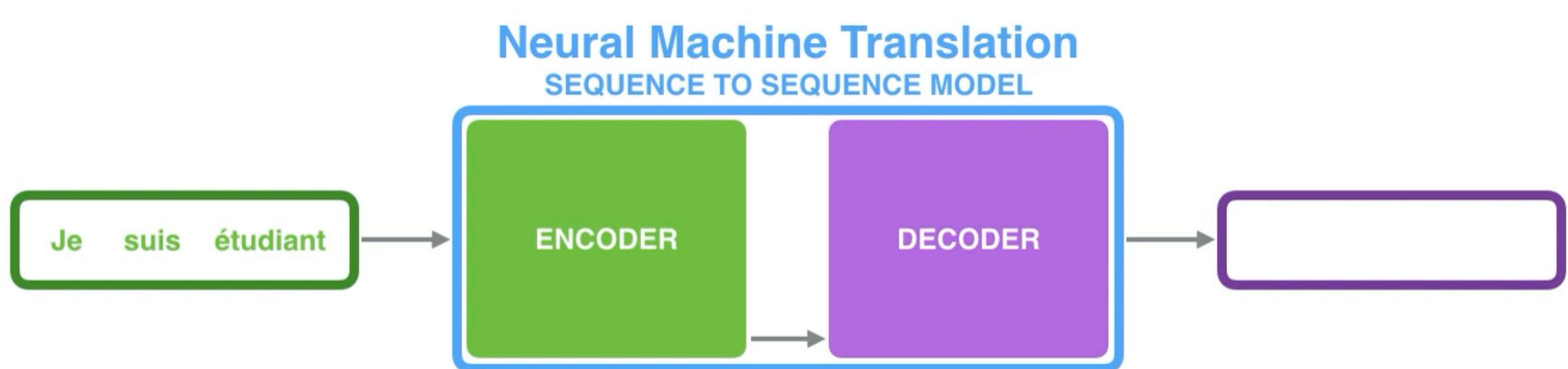


Neural MT



(Sutskever et al., 2014)

Encoder-decoder model

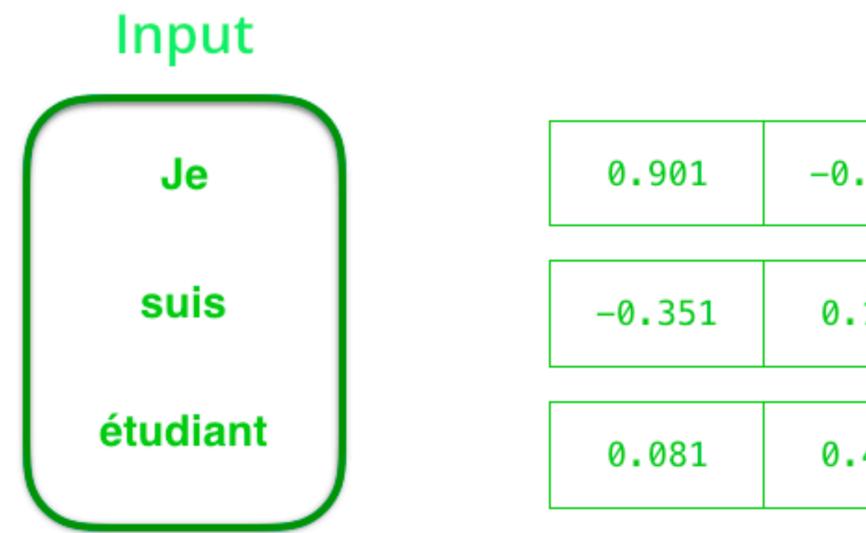


https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

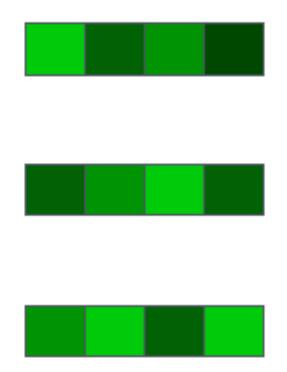


Input to the encoder

Embedding



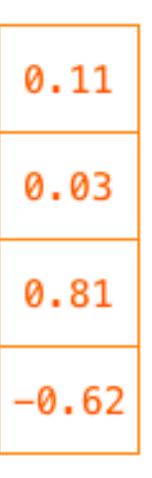
-0.194	-0.822
0.435	-0.200
-0.400	0.480
	0.435



Context vector

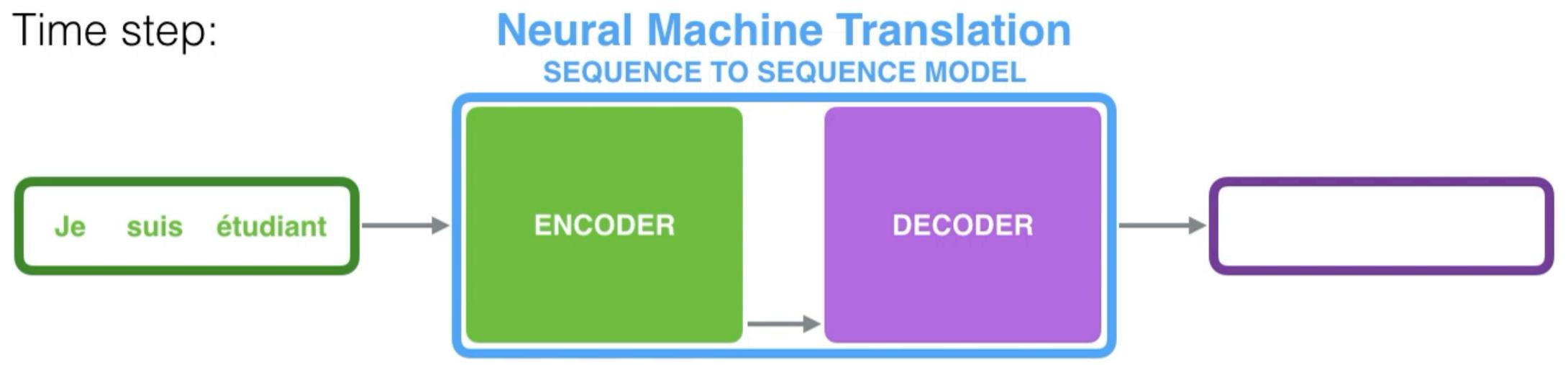
A continuous vector from the encoder





0.11
0.03
0.81
-0.62





Let's pay attention

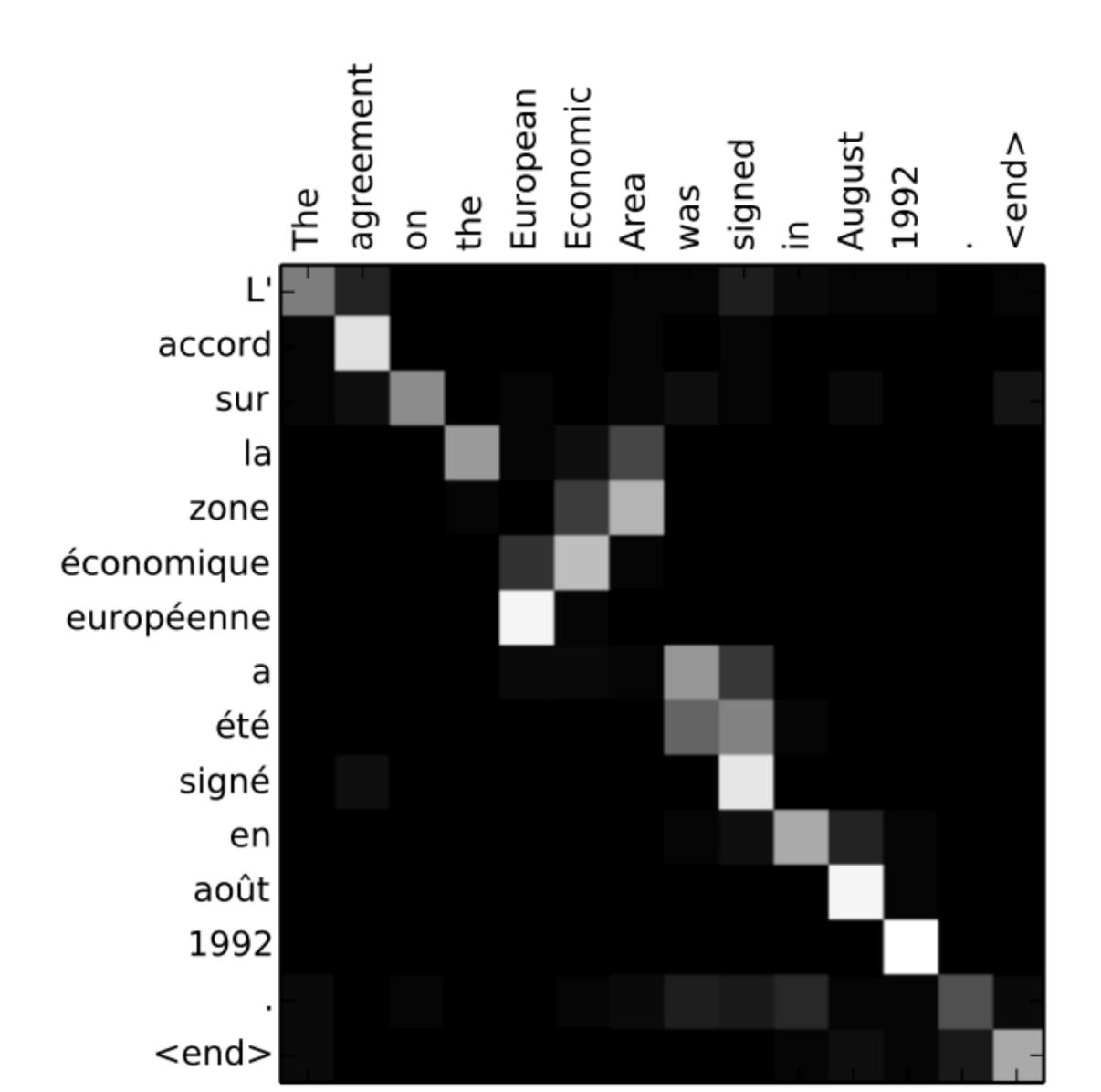
Time step: 7

Neural Machine Translation SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



l am a

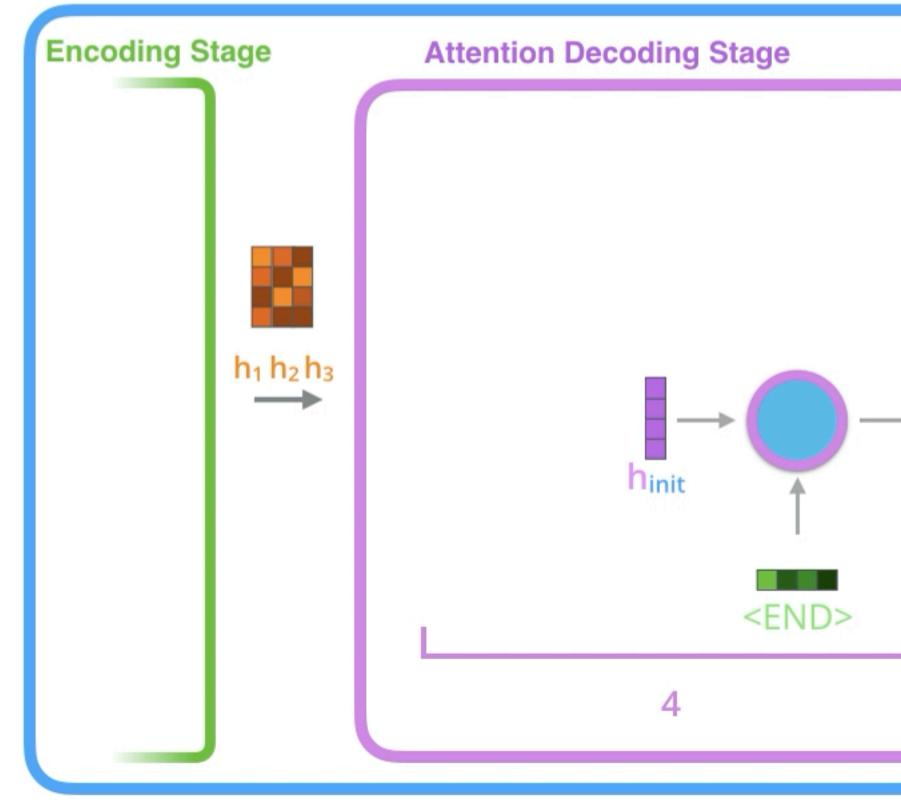
Attention visualization





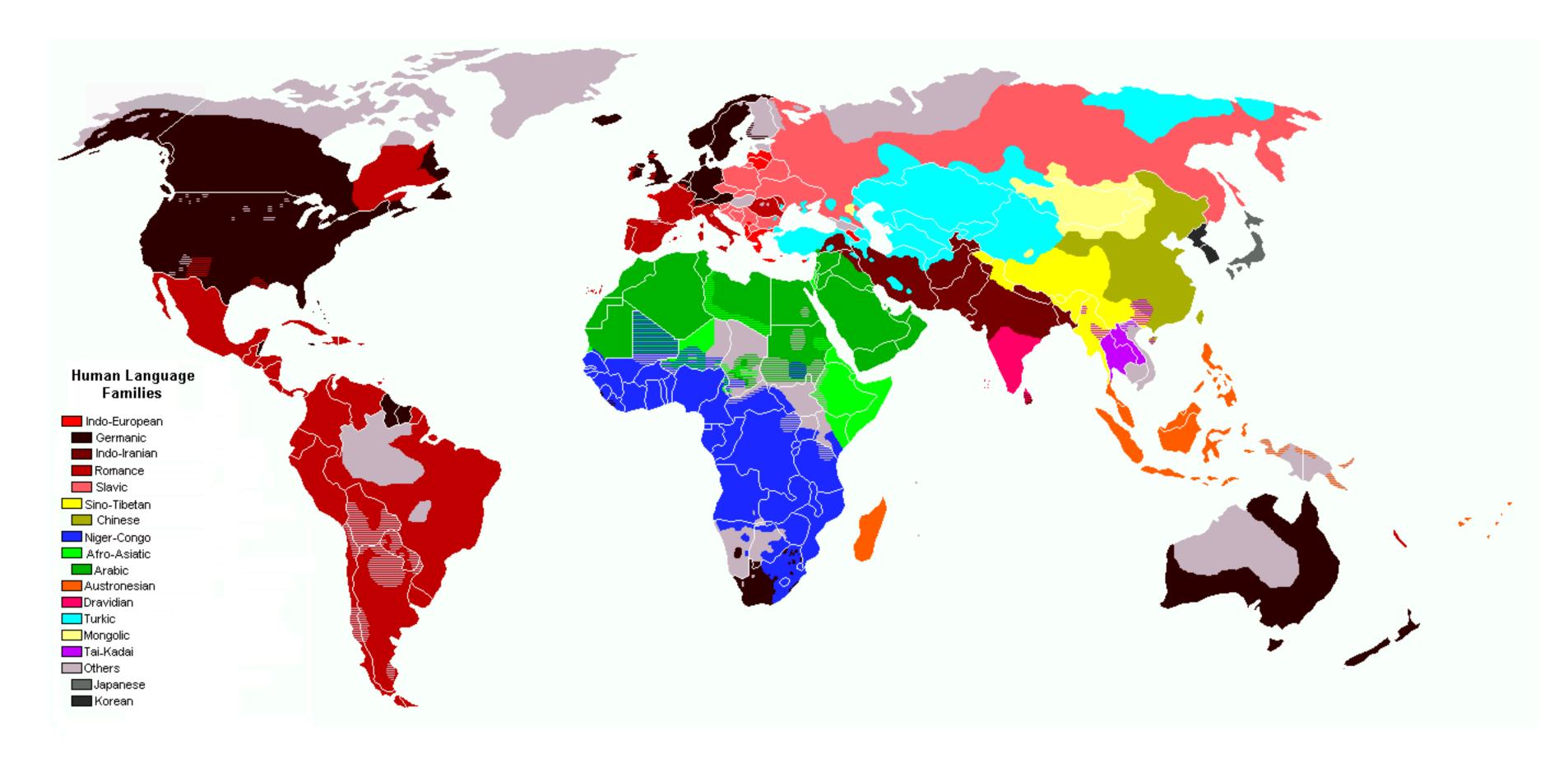
Seq2Seq with attention

Neural Machine Translation SEQUENCE TO SEQUENCE MODEL WITH ATTENTION

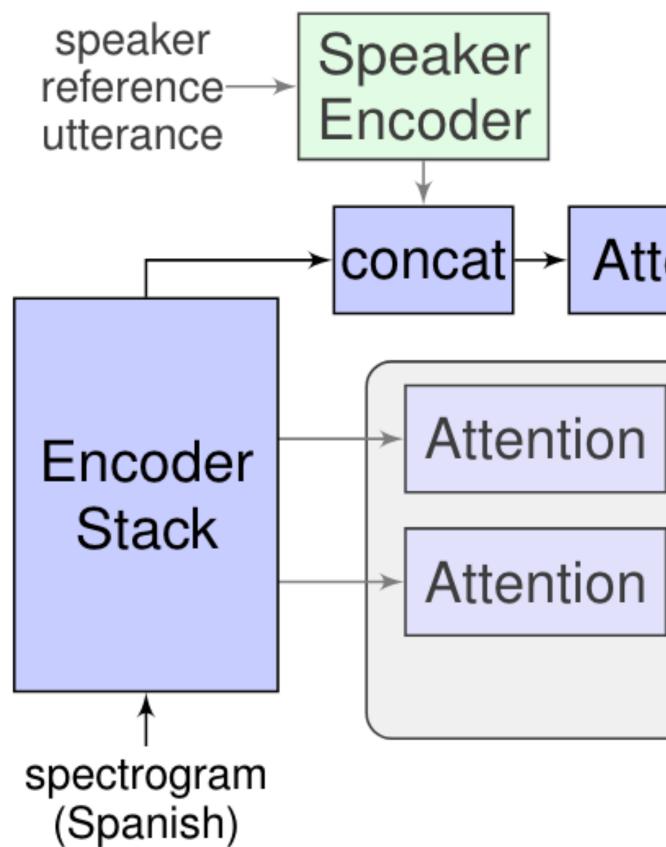


Speech-to-speech translation

Some languages don't have a written form

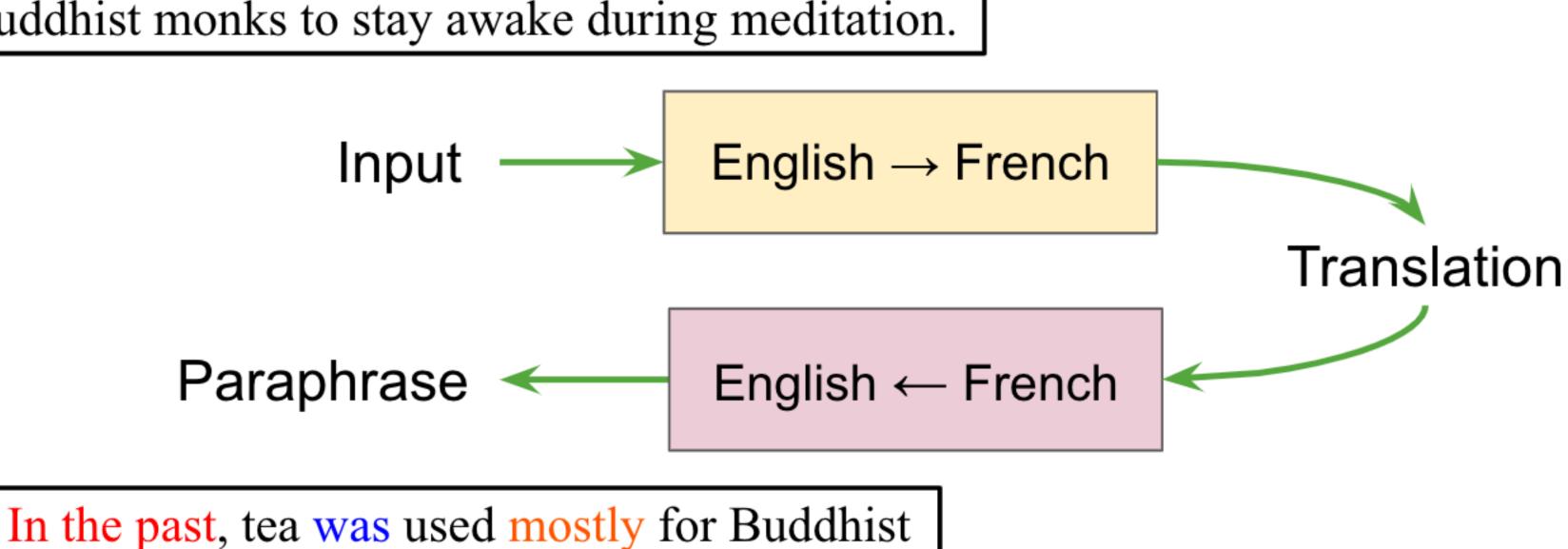


Speech-to-speech translation speaker Speaker waveform Vocoder reference→ (English) Encoder utterance Spectrogram spectrogram Attention concat⊦ (English) Decoder phonemes Attention Decoder Encoder (English) Stack phonemes Attention Decoder (Spanish) Auxiliary tasks



Data augmentation: back translation

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.



monks to stay awake during the meditation.

Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.



Summary

- Lexical, syntactic, semantic divergences make MT difficult
- Statistical machine translation and noise channel model
- End-to-end neural machine translation
- Speech-to-speech translation for languages without written form

Reading

- Chapter 10: Machine translation
 - https://web.stanford.edu/~jurafsky/slp3/13.pdf