

Machine Translation

การแปลภาษาด้วยเครื่อง

Machine Translation

- Machine Translation เป็นการแปลโดยอาศัยคลังข้อมูลคู่ตัวอย่างการแปลจำนวนมหาศาล
- MT เป็นหนึ่งในโจทย์ที่ยากที่สุดของ AI และ NLP

☰ Google แปลภาษา



汶 ข้อความ

เอกสาร

ตรวจภาษา

อังกฤษ

ไทย



ไทย

อังกฤษ

ญี่ปุ่น



Once when I was six I saw a magnificent picture in a book about the jungle, called True Stories.



เมื่อฉันอายุหกขวบฉันเห็นภาพที่สวยงาม
ในหนังสือเกี่ยวกับป่าที่เรียกว่าเรื่องจริง



Meùx cħan xāyu ħk khwb cħan hĕn phāph thi
riwyngām ni hñaqngħiż kċiċċ kapt pā thi reiħy k wā
rebiex ng cring



96/5000





Carlos Greg Diuk

April 10 at 7:54 AM ·

•••

Las dos noticias relevantes del día : llegaron los fondos del FMI, y la foto del agujero negro por el que se van a ir.

The two relevant news of the day: IMF funds arrived, and the photo of hole black for which they are going to go.

· Rate this translation



kyoungtaek46
South Korea

●●●●● Reviewed 3 weeks ago □ via mobile

너무 더울땐 이곳에서 하루를

Google Translation

하루정도 보내기 무리없이 좋습니다. 구경할것도 많고 음식점과 디저트도 다양합니다. 너무 더울땐 하루 정도는 씨암 파라곤에서 보내곤 하는데 아이들도 좋아해서 갈때마다 들리게 되네요

X

"If it's too hot,"

●●●●● Mar 31, 2019 kyoungtaek46, South Korea

It's good to spend a day or so. There are many things to see, restaurants and desserts. When I'm too hot, I spend a day at Siam Paragon, and I like children so I can hear it every time I go.

This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC.

Translated by Google™

- speech to speech translation

- augmented-reality translation

การณ์การใช้ MT

- User-initiated on-demand real-time
คุณภาพต่ำ (Bing หรือ Google)
- Author-initiated คุณภาพสูง

AI-assisted Translation

1 Les rivières isolées s'écoulent vers la mer,

|

The isolated rivers flow to the sea,

2 Les rivières isolées soupirent attendez moi, atte

ส่วนประกอบในการสร้างเครื่องแปล

1. ตัวอย่างการแปลของคู่ภาษาที่ต้องการ
(parallel corpus)
2. คอมพิวเตอร์แรง ๆ ในการเทรนโมเดล

Parallel corpora

corpus	doc's	sent's	en tokens	th tokens
OpenSubtitles v2018	4353	3.5M	28.4M	7.8M
OpenSubtitles v2016	3656	2.9M	23.4M	6.5M
OpenSubtitles v2013	1474	1.0M	8.5M	2.2M
Tanzil v1	15	93.5k	2.8M	3.4M
GNOME v1	1201	0.5M	2.3M	3.5M
Tatoeba v2	1	0.2k	3.6M	1.5k
KDE4 v2	567	92.0k	0.5M	0.2M
Ubuntu v14.10	253	46.6k	0.4M	0.2M
OpenSubtitles v2012	37	34.4k	0.3M	0.1M
OpenSubtitles v2011	20	16.5k	0.1M	57.6k
<i>total</i>	11577	8.1M	70.5M	24.0M

We going to get rich , Pa ?

เราจะรวยหรือจ้า พ่อ

Who knows ?

ใจจะรู้

Patrick !

แพทริค

Wait a minute .

เดี๋ยวก่อน

Look at the filth on your boots .

Clean 'em .

ดูผุ่นที่รองเท้าสิ เช็คชะ

ແຜນພໍມນາເສຣະຫຼຸກົງຈແລະສັງຄມແຫ່ງໜາຕີ ຂັບທີ ๑

4.6.3 Develop an efficient database to help monitor and evaluate.

1) Create a comprehensive database.

Information networks of government agencies at the policy level should be developed to monitor and evaluate significant issues, changes and conditions that have affected the country's progress.

๔.๖.๓ ພໍມນາຮບບຸການຂໍ້ມູນໃຫ້ເຊື່ອມໂຍງເປັນ ເຄື່ອງຂ່າຍໃນທຸກຮະດັບ ສໍາຮັບກາຣຕິດຕາມປະເມີນ ຜລທີ່ມີປະສິທິກາພ

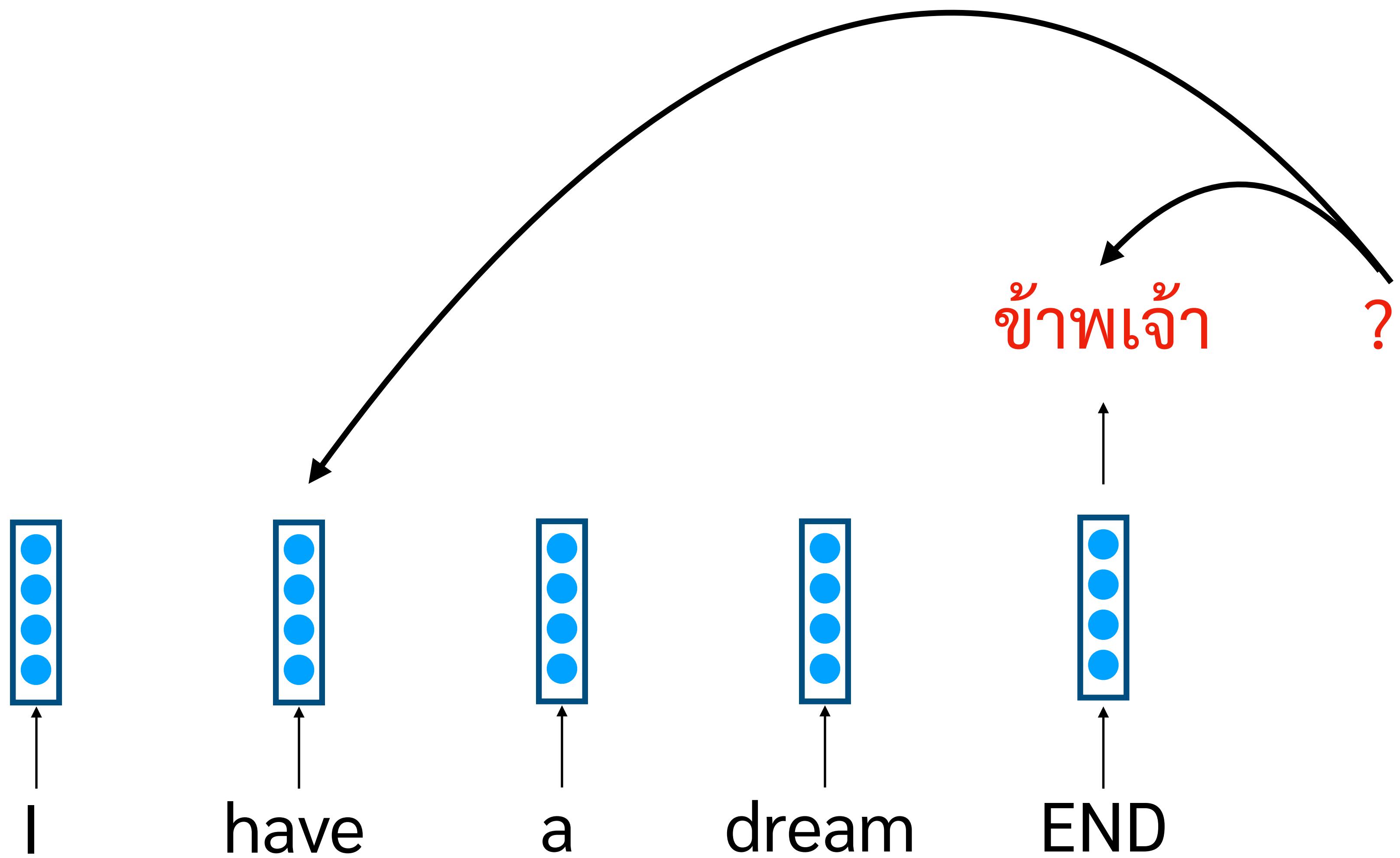
(๑) ພໍມນາຮບບຸການຂໍ້ມູນຮະດັບກາພຽມ ໂດຍ ພໍມນາຮບບົການຂໍ້ມູນຂ່າຍຂໍ້ມູນຂ່າວສາຮອງໜ່ວຍງານ ຮະດັບນໂຍບາຍ ໃນກາຣຕິດຕາມຜລກາຣດຳເນີນງານ ຕາມປະເດືອນກາຣພໍມນາສໍາຄັລູ ກາຣເປົ່າຍິນແປລັງ ຂອງສຖານກາຣນໍແລະເງື່ອນໄຂຕ່າງໆ ທີ່ມີຜລກະທບ ຕ່ອກກາຣພໍມນາປະເທສ ໂດຍປະຍຸກຕື່ໃໝ່ເທໂນໂລຢີ ສາຮສນເທສໃນກາຣເພີມປະສິທິກາພແລະ ປະສິທິຜລຂອງບຸການຂໍ້ມູນທີ່ມີອຸ່ງເປັນຈຳນວນມາກ

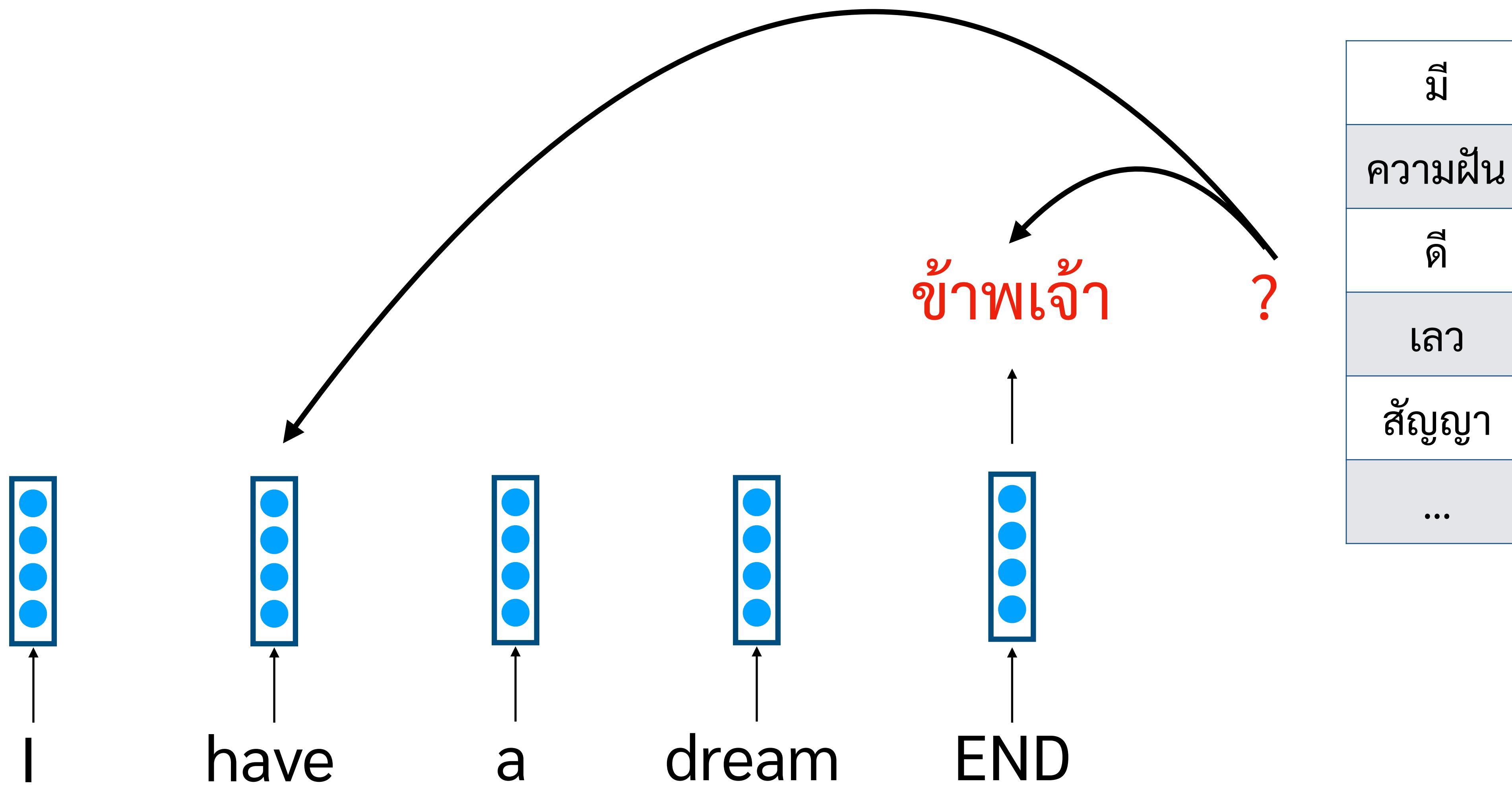
ส่วนประกอบในการสร้างเครื่องแปล

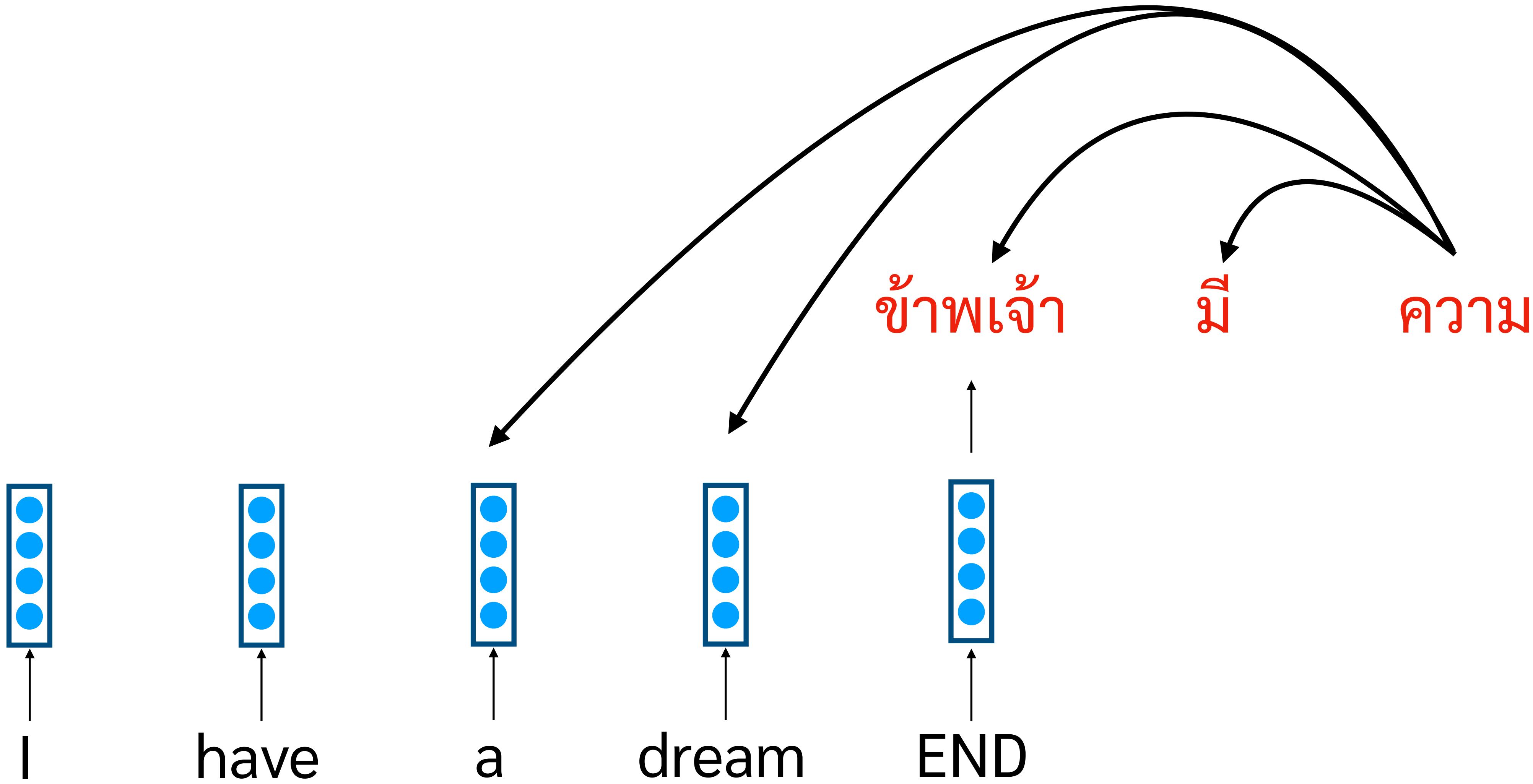
1. ตัวอย่างการแปลของคู่ภาษาที่ต้องการ
(parallel corpus)
2. คอมพิวเตอร์แรง ๆ ในการเทรนโมเดล

Encoder-Decoder Model

- เรียนรู้วิธีเก็บความหมายที่ sensitive ต่อบริบท
- เรียนรู้วิธีการเลือกคำ สร้างประโยคให้สวยงาม







I

have



a



dream



END

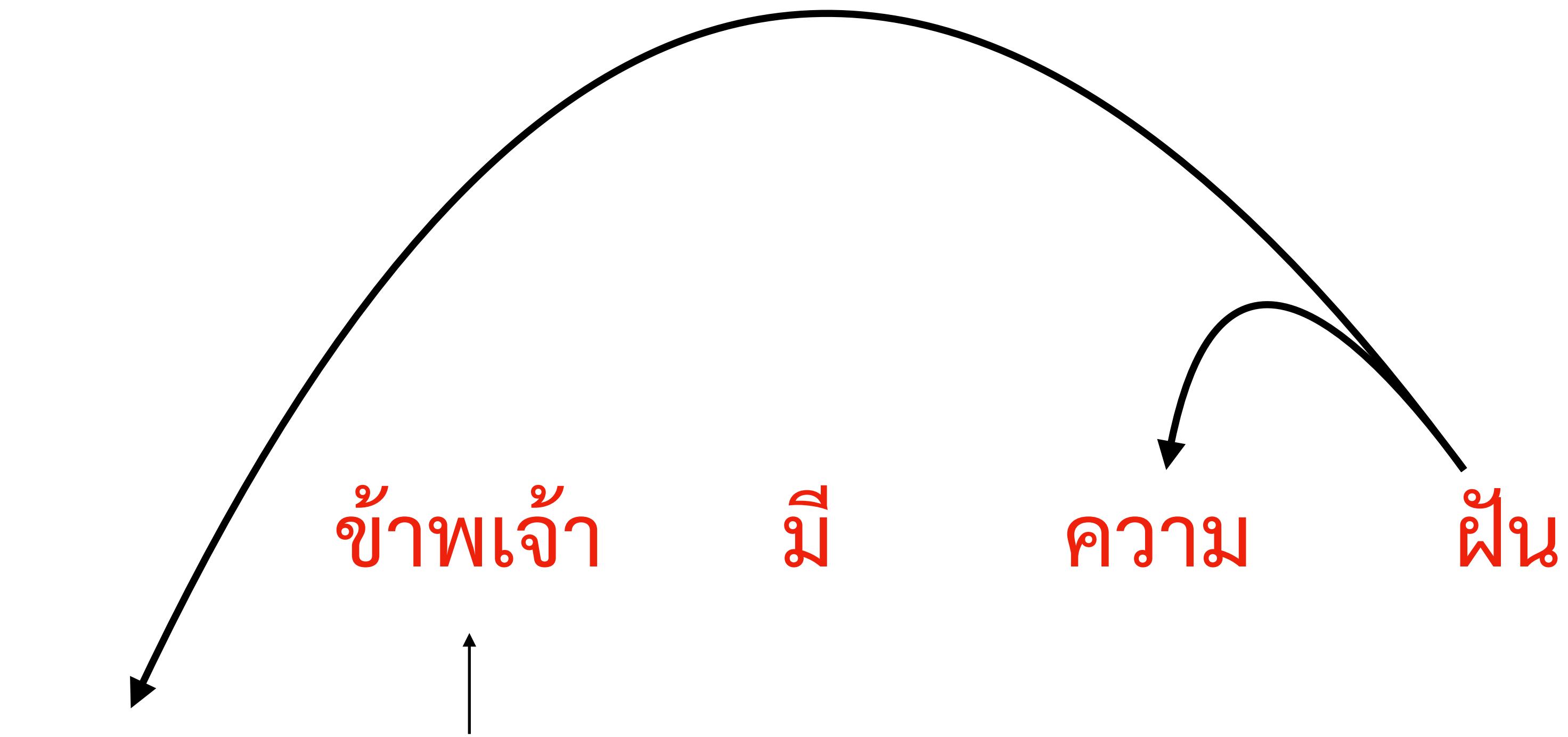


ข้าพเจ้า

เป็น

ความ

ผืน



สรุป

- Machine Translation เป็นโจทย์ของ AI และ NLP ที่
ยกที่สุดก็ว่าได้ และจัดว่าเป็นตัวบ่งชี้ความก้าวหน้า
ของศาสตร์นี้
- MT เรียนการแปลจากตัวอย่างการแปลจำนวนมาก
โดยอาศัยโมเดล Machine Learning ที่ความ слับซับ
ซ้อนสูง

Parallel Corpus

for Machine Translation

ภาษาต้นฉบับ (Source language)

ภาษาปลายทาง (Target language)

A geographer is too important to go wandering about. เขาสำคัญเกินกว่าที่จะมาเดินเล่นได้

He never leaves his study.

เขาจะไม่ออกไปนอกที่ทำงานของเขา

But he receives the explorers there.

แต่เขาจะต้อนรับนักสำรวจ

He questions them, and he writes down what they remember

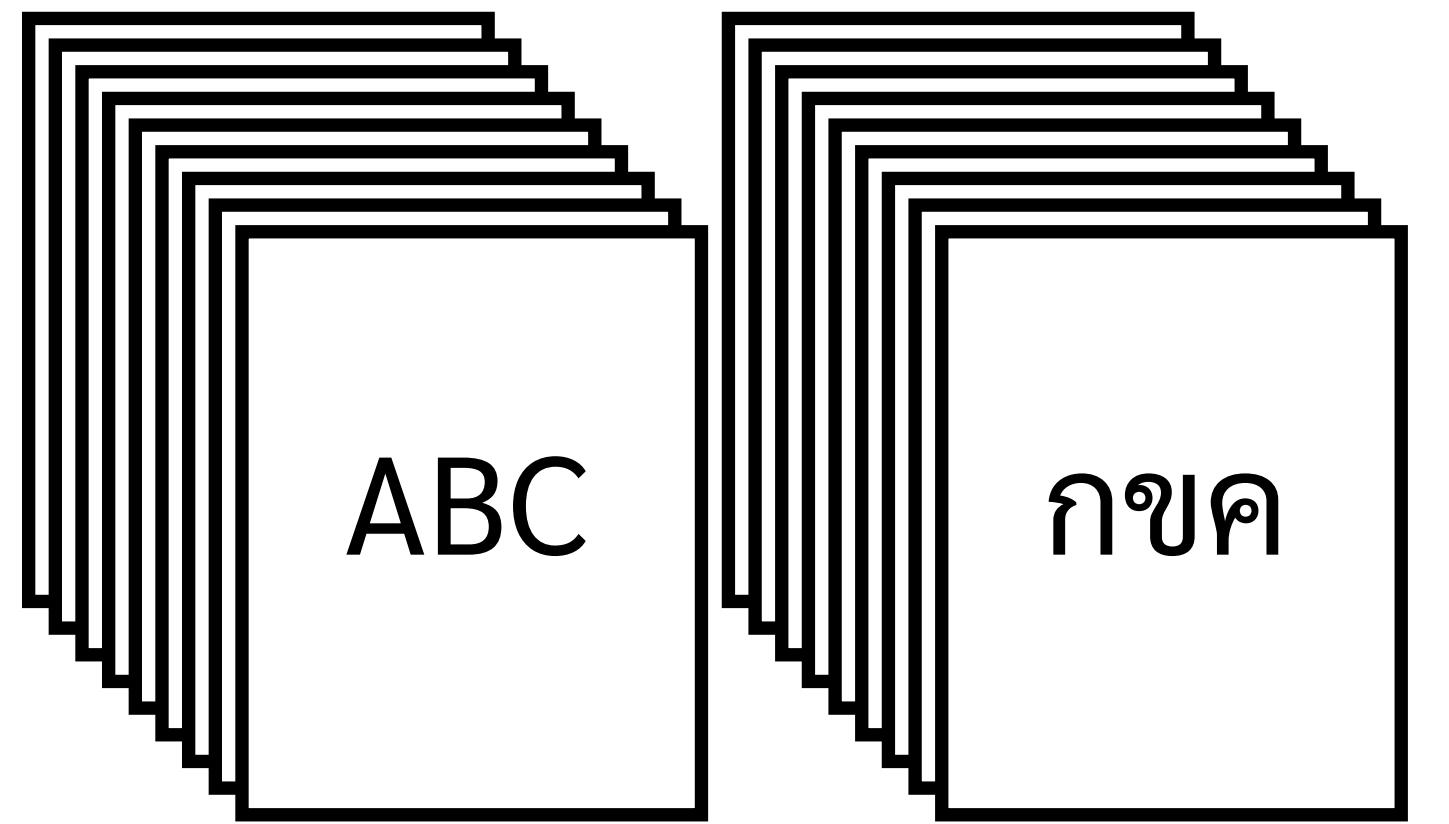
เขากتابบันทึกความทรงจำของนักสำรวจ

...

...

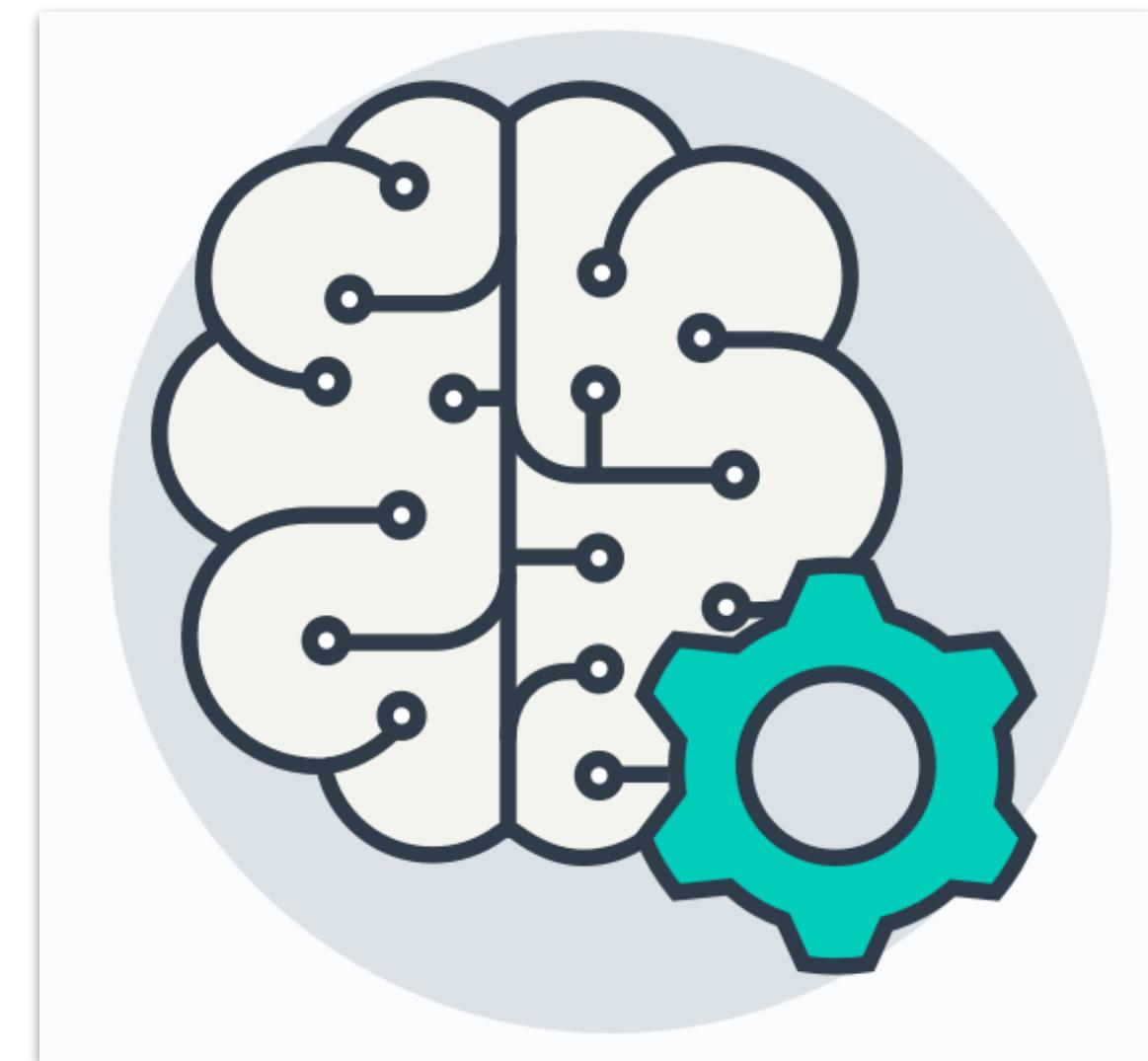
...

...

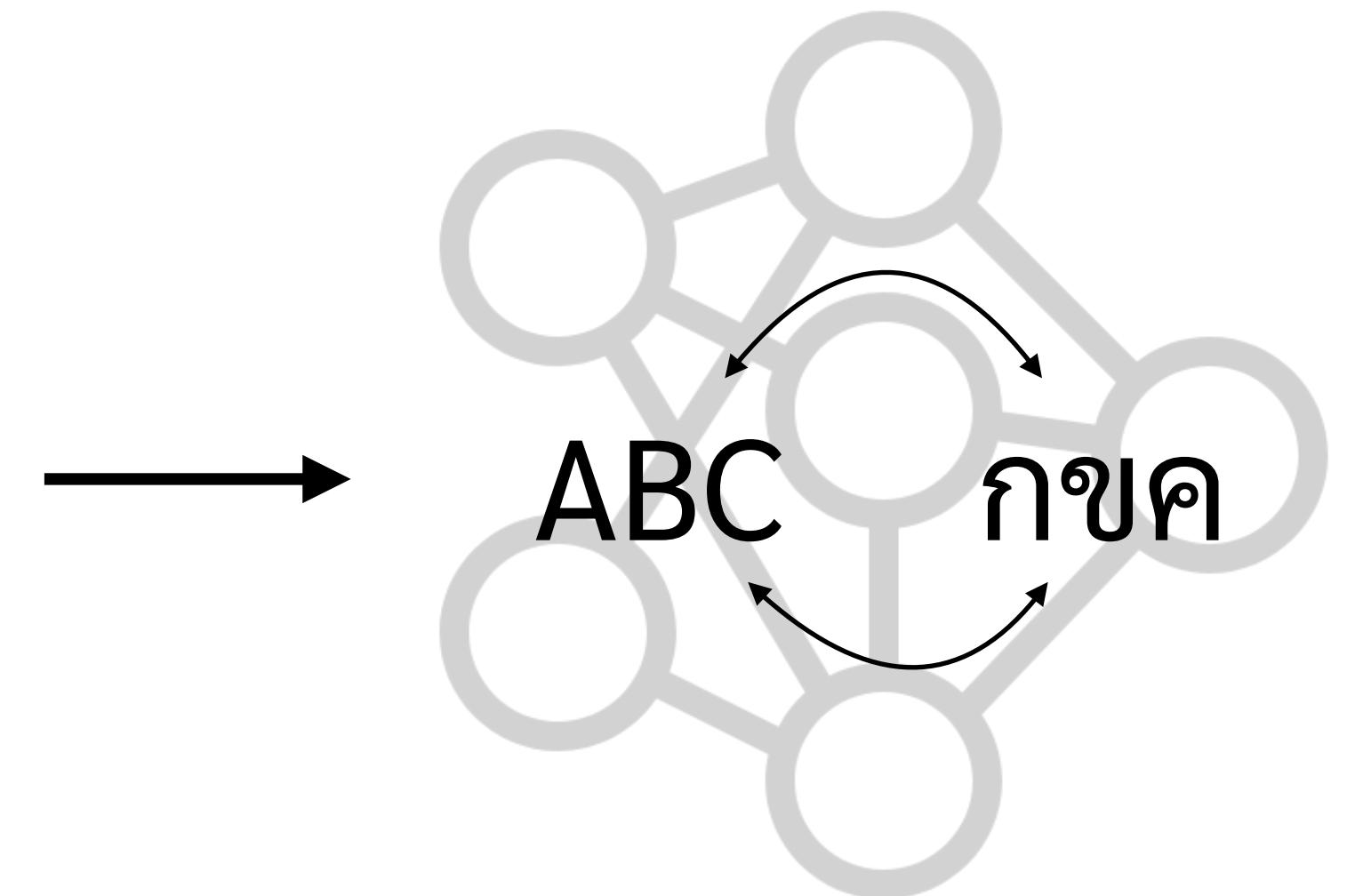


Parallel Corpus

+



Machine Learning
Algorithm



Machine Translation

Limited Domain MT is very effective.

- domain: science vs sports vs law
- modality: spoken vs written

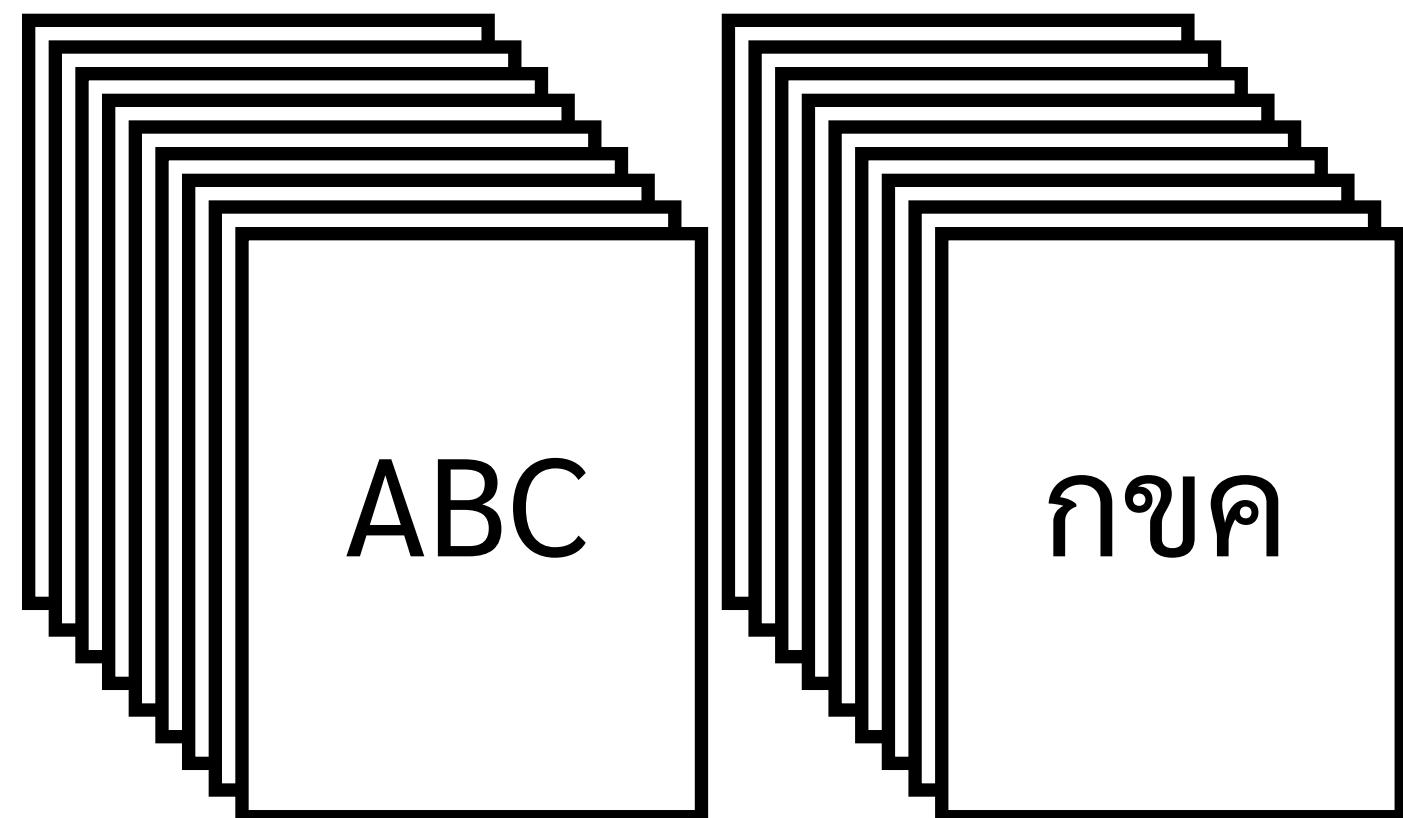
Parallel corpus มาจากไหน

1. จ้างนักแปลมาแปล text ที่เตรียมไว้
2. เอกสารและประกาศของหน่วยงานต่าง ๆ
3. ชุดมาจากอินเตอร์เน็ต

1. จ้างนักแปล

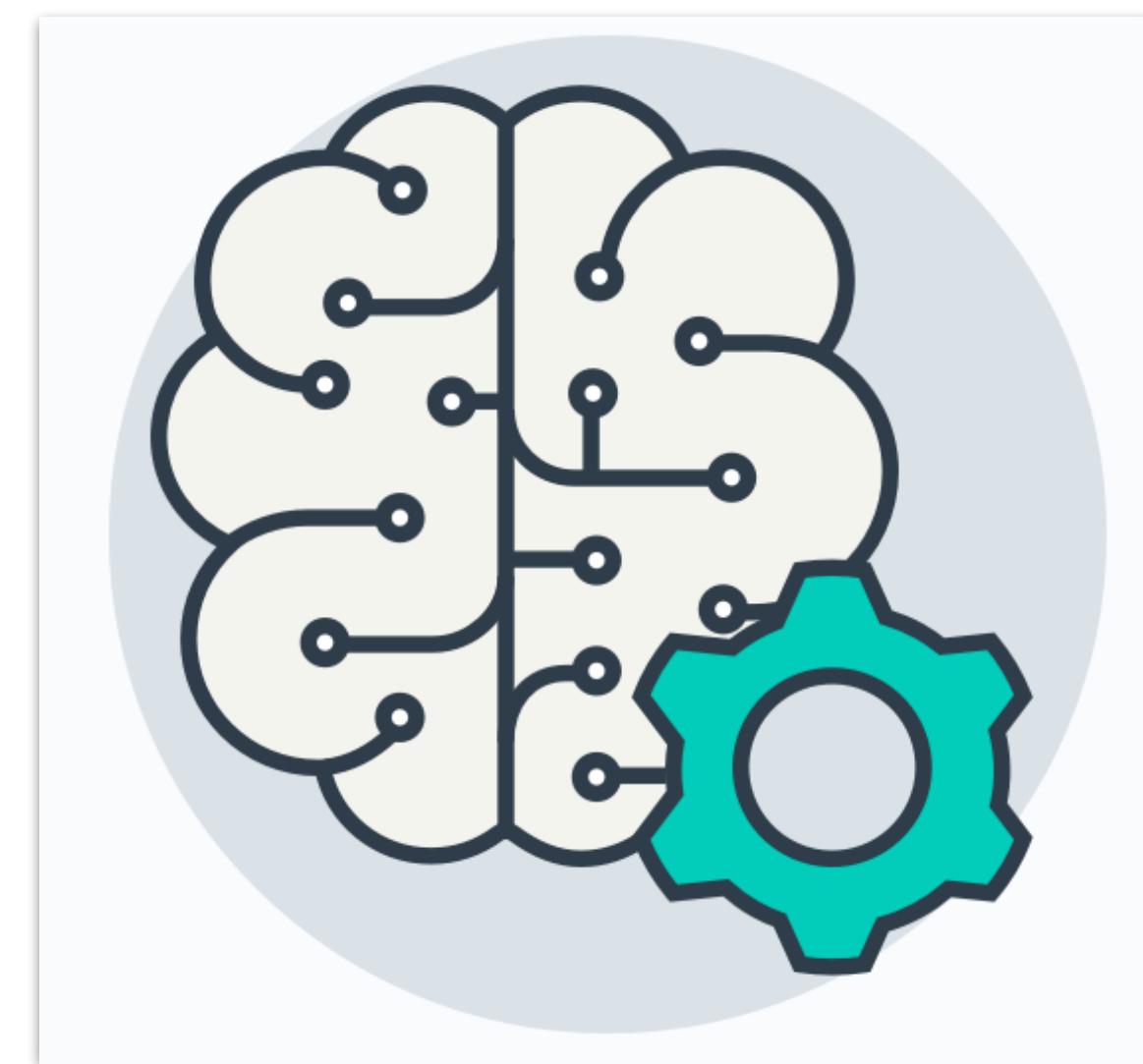
- ข้อดี
 - คุณภาพการแปลสูง
 - แพงมาก (2-10+ บาท/ประโยค)
 - บางครู่ภาษาหายาก

ต้องเจาะจงคุณภาษา ก่อน

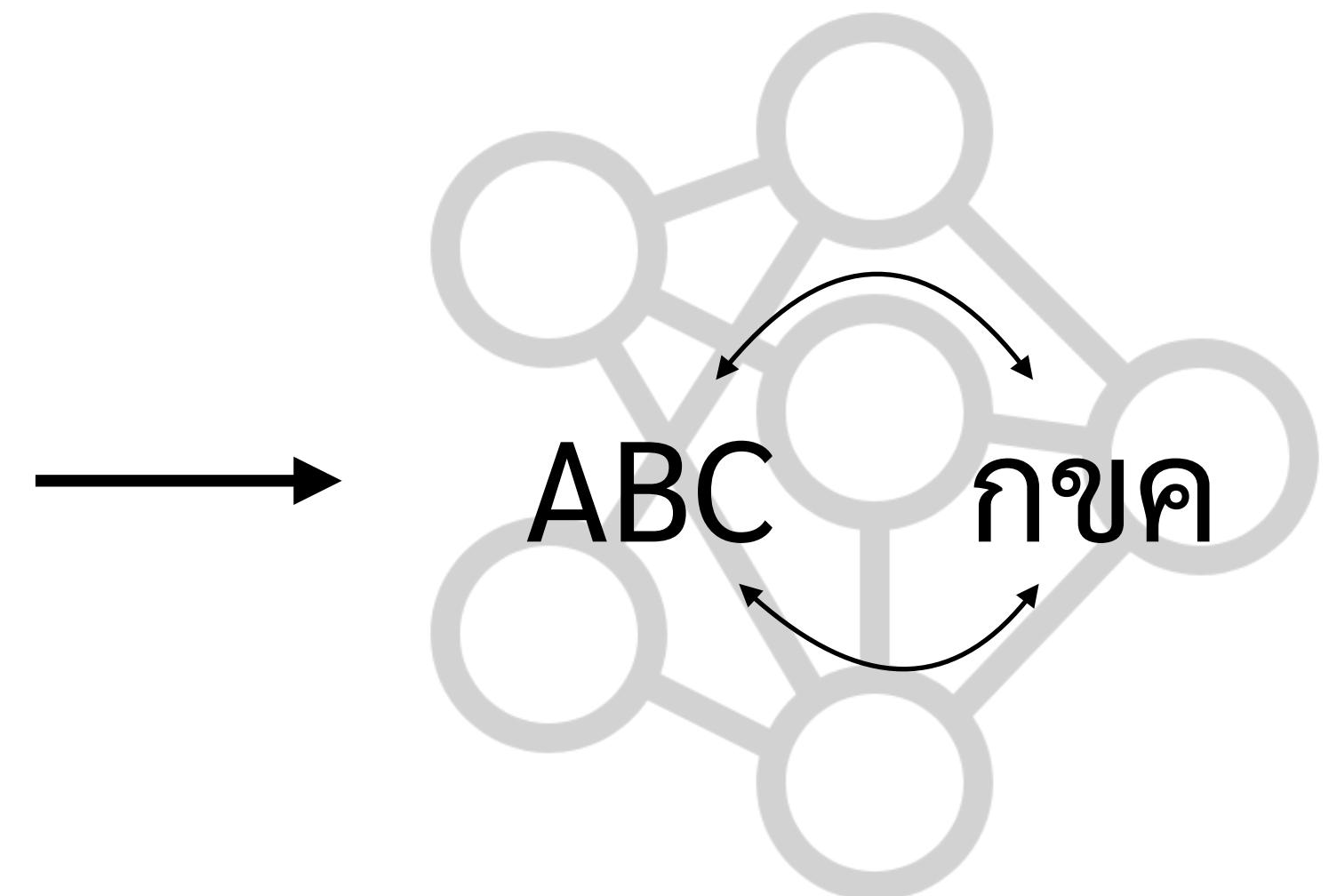


Parallel Corpus

+



Machine Learning
Algorithm



Machine Translation

Parallel corpus มาจากไหน

1. จ้างนักแปลมาแปล text ที่เตรียมไว้
2. เอกสารและประกาศของหน่วยงานต่าง ๆ
3. ชุดมาจากอินเตอร์เน็ต

Parallel corpora ที่เด่น ๆ

ชื่อ corpus	คู่ภาษา	Domain	ขนาด
OpenSubtitles	60+ ภาษา	หนังและ series	1 - 400M
Europarl	20+ ภาษา	รายงานการประชุม	0.4 - 2M
United Nations	6 ภาษา	เอกสารสภากาชาด	11M (lines)

2. เอกสารและประกาศของหน่วยงานต่าง ๆ

- ข้อดี
 - พรี เยอะ และคุณภาพการแปลสูงมาก
 - ส่วนใหญ่แปลโดยคต่อโดยค
- ข้อเสีย
 - domain อาจจะไม่ตรงกับที่อยากรenameใช้

The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

He has long lived in the woods.
He likes the isolation and solitude of his house.
It's small, but cozy.
The next village is miles away.
He only goes there once a week.

It just after dusk.
The hot sun finally set.
The forest was still abuzz in chatter.
Voices of birds and insects fill the air.
A comforting sound.
But the shadow was larger than those animanls
Only little creatures live here, not this.
It seemed almost as large as a man.
But why that?
Nobody comes ever here.
So the man's eyes keep looking.

As the minutes passed, nothing happens.
But then, cast against the bright moonlit, it returns.

- Der Mann schaut aus dem Fenster.
- Er sieht einen Schatten in den Bäumen.
- Was war das?
- Er war alamiert und wach.
- Er hat schon lange im Wald gelebt.
- Er genießt die Einsamkeit des Hauses
- Es ist klein.
- Aber es ist gemütlich.
- Das nächste Dorf ist meilenweit entfernt.
- Er geht dorthin nur einmal in Monat.
- Es ist nach der Untergang der heißen Sonne.
- Der Wald is voller Geschwätz.
- Stimmen von Vögeln und Insekten dringen herüber.
- Aber der Schatten war größer als diese Tiere.
- Nur Kleingetier lebt hier.
- Nicht soetwas Großes.
- Es erschien fast so groß wie ein Mensch.
- Aber warum, wenn hier niemand jemals herkommt?
- Der Mann schaut.
- Sein Augen aus dem Fenster gerichtet.
- Minuten vergehen, aber nichts passiert.
- Dann plötzlich kehrt er im Mondschein zurück.

Parallel corpus มาจากไหน

1. จ้างนักแปลมาแปล text ที่เตรียมไว้
2. เอกสารและประกาศของหน่วยงานต่าง ๆ
3. ชุดมาจากอินเตอร์เน็ต

Typical Pipeline

- หาเว็บที่น่าจะมีสองภาษาขนานกัน
 - document alignment หา web page ที่น่าจะขนานกัน
 - sentence alignment หาคู่ประโยคที่น่าจะขนานกัน
- เลือยกับเว็บอื่น ๆ (web crawling) ที่มีลิงค์จากหน้าเว็บนั้น

<https://www.central.co.th/en/dyson-cyclone-vacuum-cleaner-v10-fluffy-cds18033144>

<https://www.central.co.th/th/dyson-cyclone-vacuum-cleaner-v10-fluffy-cds18033144>

STIEBEL ELTRON Water Heater

DYSON Cord-free Cyclone Vacuum Cleaner

- One main tool : QR Soft roller cleaner (Fluffy) head
- Four additional tools : QR mini motorhead, QR crevice tool, QR combination tool and QR mini soft dusting brush
- 2 Tier RadialTM Cyclones increasing airflow and capture fine dust without loss of suction
- ...

เครื่องดูดฝุ่นไซโคลนแบบไร้สายรุ่น V10 Fluffy SV12FLUFFY จากแบรนด์ DYSON ที่สุดของประสิทธิภาพในการทำความสะอาด ขจัดคราบฝุ่นได้สะอาดหมดจด ไม่เปลือยแรง

- เครื่องดูดฝุ่นไซโคลนแบบไร้สาย
- หัวดูดหลัก 1 หัว: หัวดูดแบบลูกกลิ้งนุ่ม (Fluffy)
- อุปกรณ์เสริม 4 ชิ้น: หัวดูดมอเตอร์ขนาดเล็ก, หัวดูดปากแcap หัวดูด 2-in-1, และแปรงปัดฝุ่นขนาดนุ่ม
- ระบบไซโคลน 2 ชั้น: ไม่เสียแรงดูด เพิ่มประสิทธิภาพการดูด
- ...

<https://www.chula.ac.th/en/impact/6081/>

About research: Department of Food Technology, Faculty of Science, Chulalongkorn University has various research studies regarding the development of carbohydrate and protein based edible films. Those films can be used as food packaging, where they can readily be consumed together with the foods, and are 100% biodegradable. These films provide unique characteristics, including a high gloss and transparent appearance, good barrier properties, and can incorporate desirable compounds, such as antioxidants and antimicrobial agents, in order to increase the shelf life of the food products.

<https://www.chula.ac.th/impact/3903/>

เกี่ยวกับโครงการ: ภาควิชาเทคโนโลยีทางอาหาร คณะวิทยาศาสตร์ จุฬาฯ มีงานวิจัยเกี่ยวกับการพัฒนาฟิล์มบริโภคได้จากโพลิเมอร์ชีวภาพในกลุ่มคาร์บอไฮเดรตและโปรตีน เพื่อประยุกต์เป็นบรรจุภัณฑ์อาหารที่สามารถบริโภคได้พร้อมอาหาร และสามารถย่อยสลายได้ตามธรรมชาติ จึงเป็นมิตรต่อสิ่งแวดล้อม นอกจากนี้ยังมีการพัฒนาให้มีคุณสมบัติโดดเด่น เช่น พิล์มบริโภคได้บางชนิดมีความใส่ไกล์เคียงกับพิล์มจากพลาสติกที่นิยมใช้ทั่วไป บางชนิดสามารถป้องกันการซึมผ่านของก๊าซต่างๆ ได้ดี หรือสามารถผสมสารอื่นๆ เช่น สารต้านออกซิเดชัน สารต้านจุลินทรีย์ ไว้ในฟิล์มดังกล่าว เพื่อช่วยยืดอายุการเก็บรักษาอาหารได้ ซึ่งผลงานเชิงนวัตกรรมเหล่านี้มีข้อมูลพร้อมเผยแพร่ต่อผู้ประกอบการและประชาชนทั่วไปที่สนใจ

<https://www.chula.ac.th/en/impact/6081/>

About research:

Department of Food Technology, Faculty of Science, Chulalongkorn University has various research studies regarding the development of carbohydrate and protein based edible films.

Those films can be used as food packaging, where they can readily be consumed together with the foods, and are 100% biodegradable.

These films provide unique characteristics, including a high gloss and transparent appearance, good barrier properties, and can incorporate desirable compounds, such as antioxidants and antimicrobial agents, in order to increase the shelf life of the food products.

<https://www.chula.ac.th/impact/3903/>

เกี่ยวกับโครงการ:

ภาควิชาเทคโนโลยีทางอาหาร คณะวิทยาศาสตร์ จุฬาฯ มีงานวิจัยเกี่ยวกับการพัฒนาฟิล์มบริโภคได้จากพอลิเมอร์ชีวภาพในกลุ่ม карт์โรบอไฮเดรตและโปรตีน เพื่อประยุกต์เป็นบรรจุภัณฑ์อาหารที่สามารถบริโภคได้พร้อมอาหาร และสามารถย่อยสลายได้ตามธรรมชาติ จึงเป็นมิตรต่อสิ่งแวดล้อม

นอกจากนี้ยังมีการพัฒนาให้มีคุณสมบัติโดดเด่น เช่น ฟิล์มบริโภคได้บางชนิดมีความใส่ใจล้ำคุณภาพมาก สามารถป้องกันการซึมผ่านของก๊าซต่างๆ ได้ดี หรือสามารถสมรสารอื่นๆ เช่น สารต้านออกซิเดชัน สารต้านจุลินทรีย์ไว้ในฟิล์มดังกล่าว เพื่อช่วยยืดอายุการเก็บรักษาอาหารได้

บางชนิดสามารถป้องกันการซึมผ่านของก๊าซต่างๆ ได้ดี หรือสามารถสมรสารอื่นๆ เช่น สารต้านออกซิเดชัน สารต้านจุลินทรีย์ไว้ในฟิล์มดังกล่าว เพื่อช่วยยืดอายุการเก็บรักษาอาหารได้

ซึ่งผลงานเชิงนวัตกรรมเหล่านี้มีข้อมูลพร้อมเผยแพร่ต่อผู้ประกอบการและประชาชนทั่วไปที่สนใจ

Scraped Corpus

ชื่อ corpus	คุณภาษา	Domain	ขนาด
News Commentary	12 ภาษา	ข่าว	1,000 - 400,000
ParaCrawl	8 ภาษา	Website	1 - 73M

3. ขุดมาจากอินเตอร์เน็ต

- ข้อดี
 - ฟรี เยอะ
 - ข้อเสีย
 - ลิขสิทธิ์อาจเป็นปัญหาได้
 - Document + Sentence alignment ทำได้ยาก
 - คุณภาพค่อนข้างหลากหลาย ควบคุมได้ยาก

Parallel corpus มาจากไหน

1. จ้างนักแปลมาแปล text ที่เตรียมไว้
2. เอกสารและประกาศของหน่วยงานต่าง ๆ
3. ชุดมาจากอินเตอร์เน็ต

Statistical Machine Translation

"Маленький принц" -
очень популярная
книга, переведенная на
множество языков.

"The Little Prince" is a very
popular book that was
translated into many
languages

When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

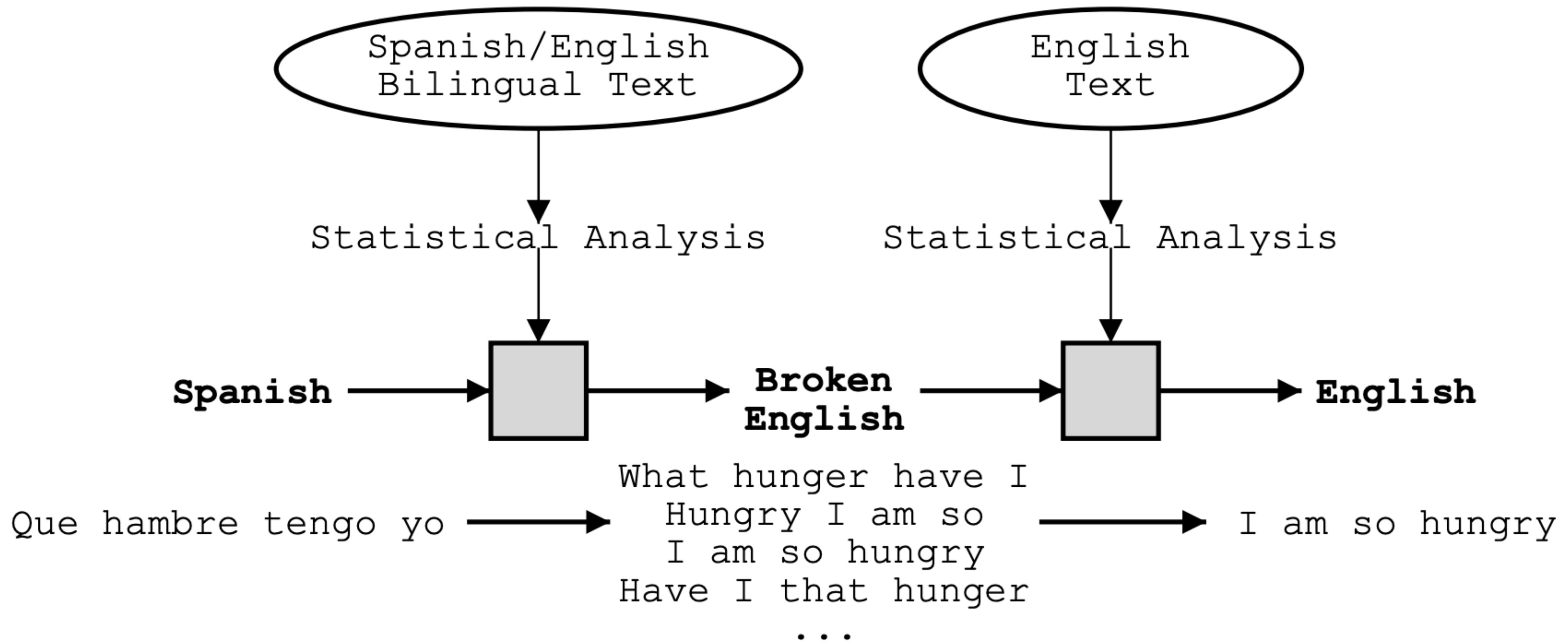
เวลาผมเห็นบทความภาษาอังกฤษ เชิญ ผู้พูดกับตัวเองว่า "ที่จริงมันเป็นภาษาอังกฤษ เพียงแต่ถูกแปลงให้เป็นรหัสที่ใช้สัญลักษณ์แปลกๆ ข้าพเจ้าขอดำเนินการถอดรหัส ณ บัดนี้"

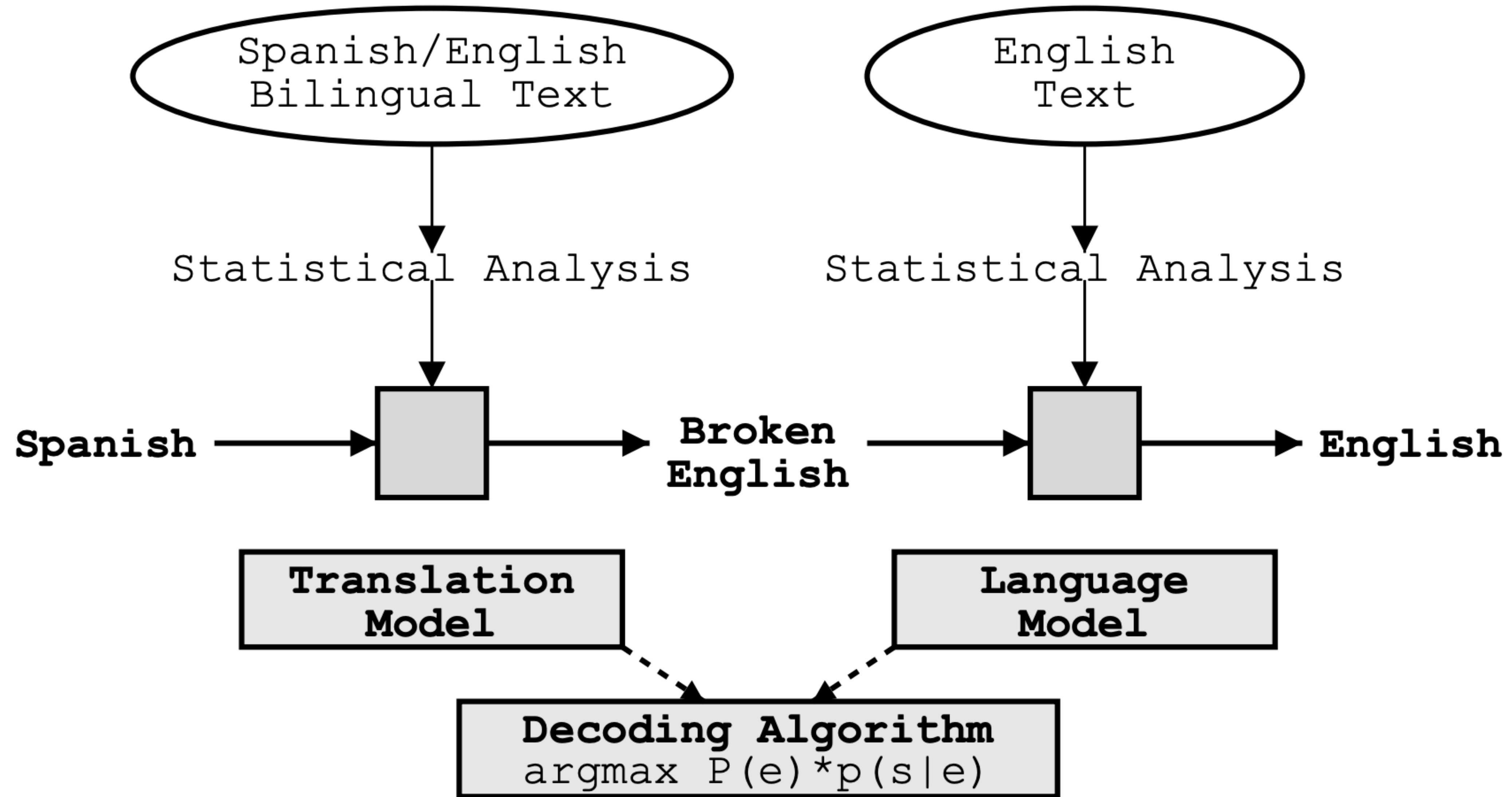
-Warren Weaver



Statistical Machine Translation

- ไอเดียหลัก: สร้าง probabilistic model จากข้อมูลการแปลง
- $\operatorname{argmax} P(Y|X) = \operatorname{argmax} P(X|Y) P(Y)$





Translation Model

Y	# Y Haus	Y	P(Y Haus)
house	8000	house	0.8
building	1600	building	0.16
home	200	home	0.02
household	150	household	0.015
shell	50	shell	0.05

Language Model

- Bigram Language Model

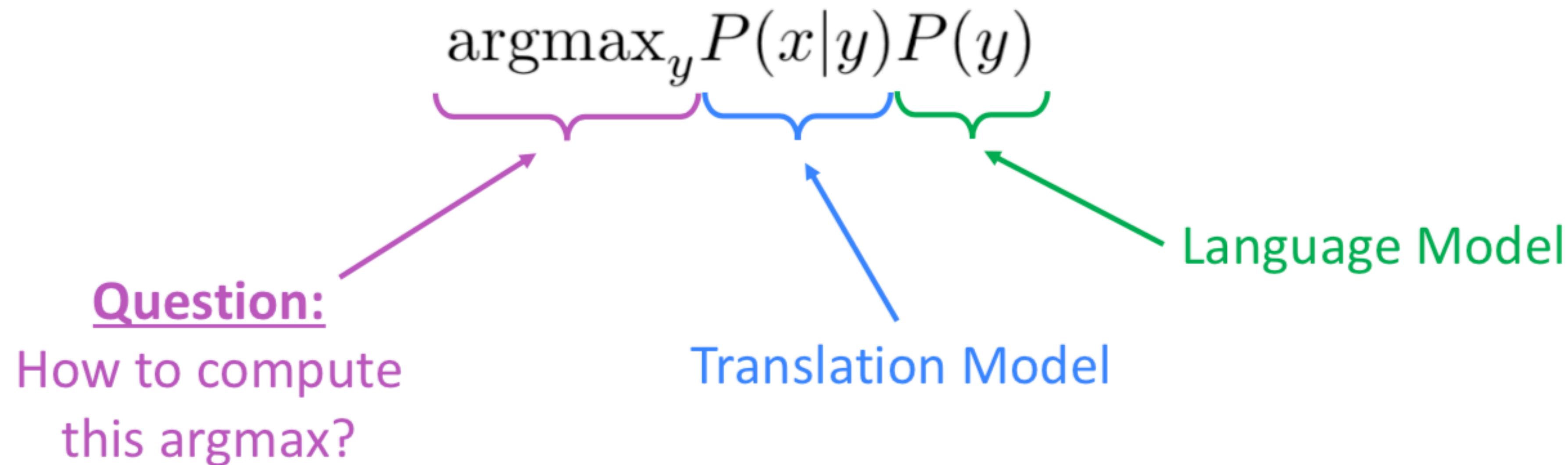
$$P(w_1, w_2, w_3, \dots, w_n) = P(w_1|\text{START}) P(w_2|w_1) P(w_3|w_2) \dots P(w_n|w_{n-1})$$

- Bangkok has many **high** buildings vs
Bangkok has many **tall** buildings

วิธีแปลแบบง่ายสุด

- Bangkok hat viele hohe Gebäude.
- Bangkok has many tall buildings.

Decoding for SMT



- $y = \text{sentence}$ ต้องลองคำนวนค่าแนวโน้มจากประโยคที่เป็นไปได้ในภาษาอังกฤษ

Word Alignment

Statistical Machine Translation - Part II

ปัญหา

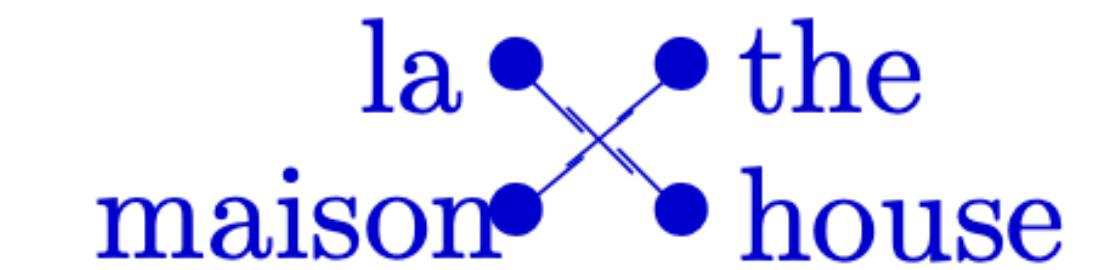
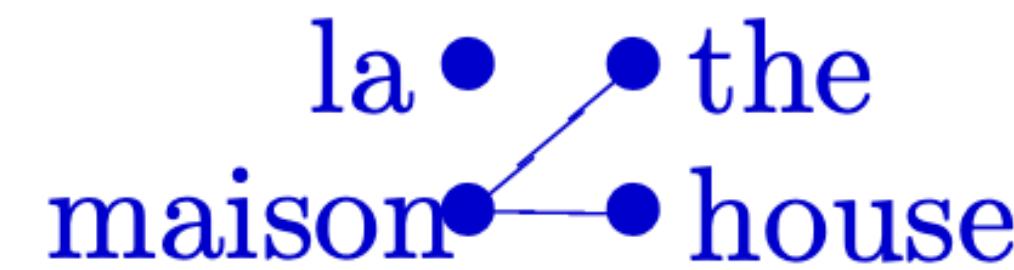
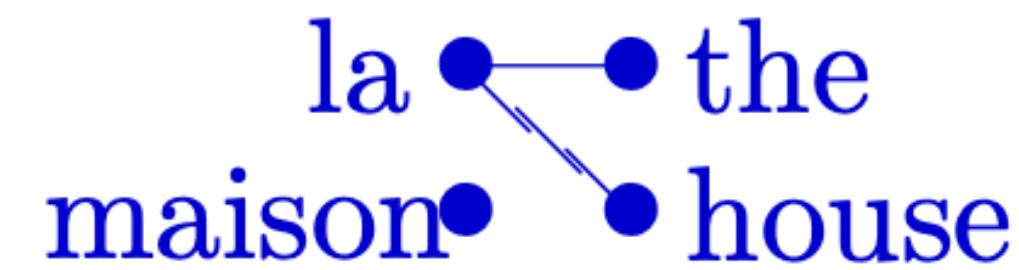
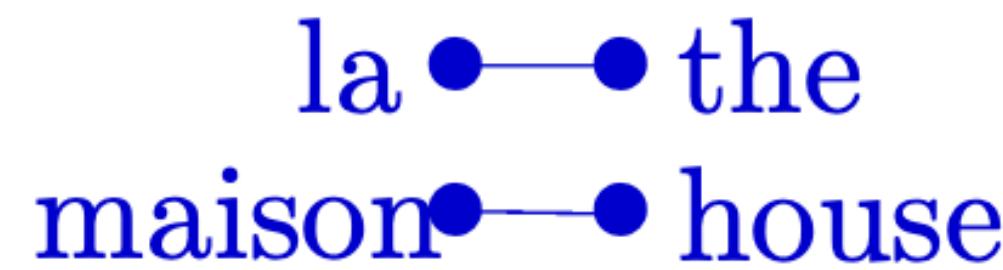
- Parallel corpus คือ ชุดประโยคคู่ขنان แต่ไม่ได้บอกว่าคำไหนแปลเป็นคำไหน
- ถ้าเรารู้คำไหน align กับคำไหน ก็สามารถไปแล้ว

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7$$
$$p(\text{the}|\text{maison}) = 0.1$$

$$p(\text{house}|\text{la}) = 0.05$$
$$p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**

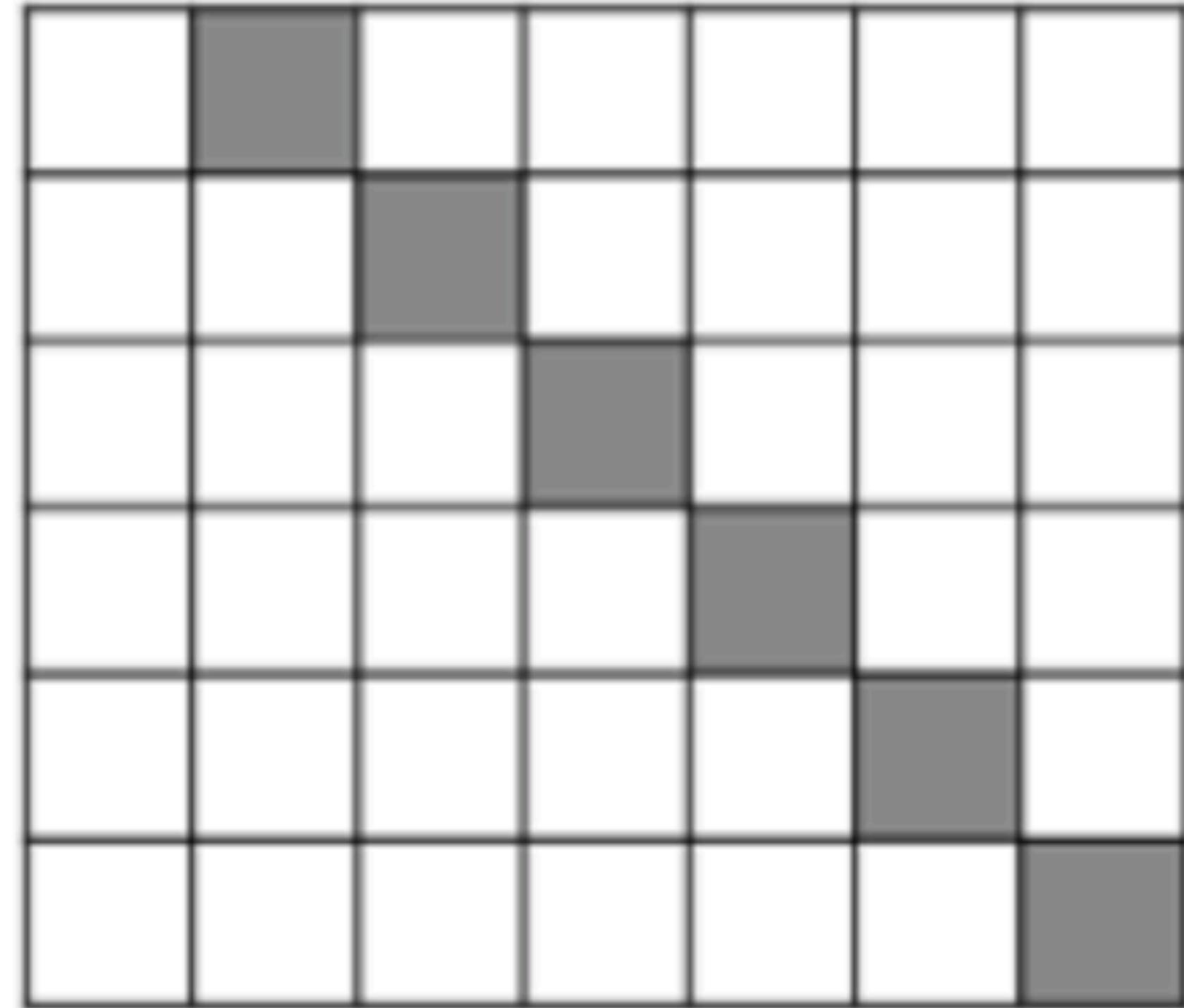


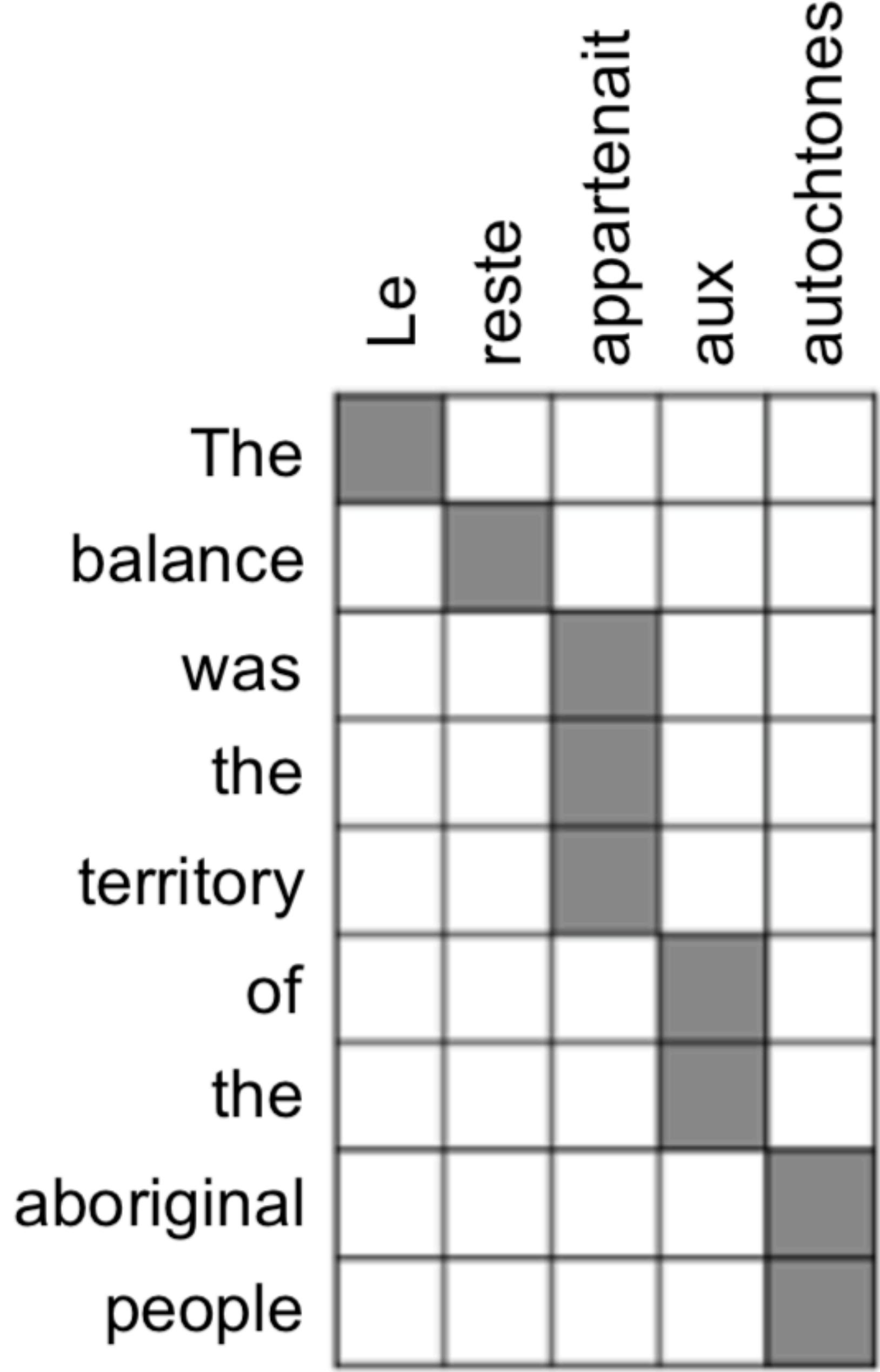
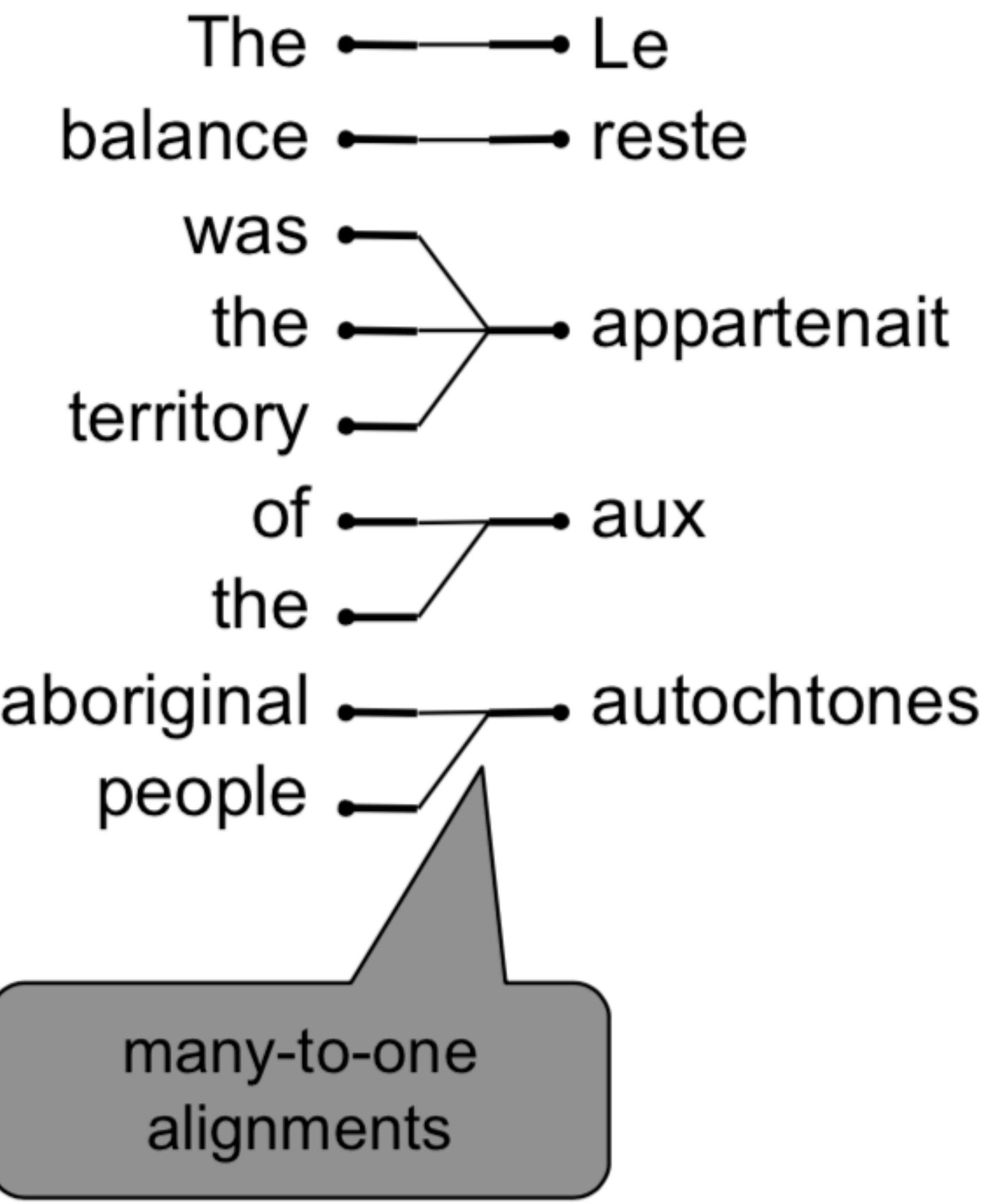
Japan Japon
 shaken secoué
 by par
 two deux
 new nouveaux
 quakes séismes

Le “spurious”
 word

Japan
 shaken
 by
 two
 new
 quakes

“spurious” word





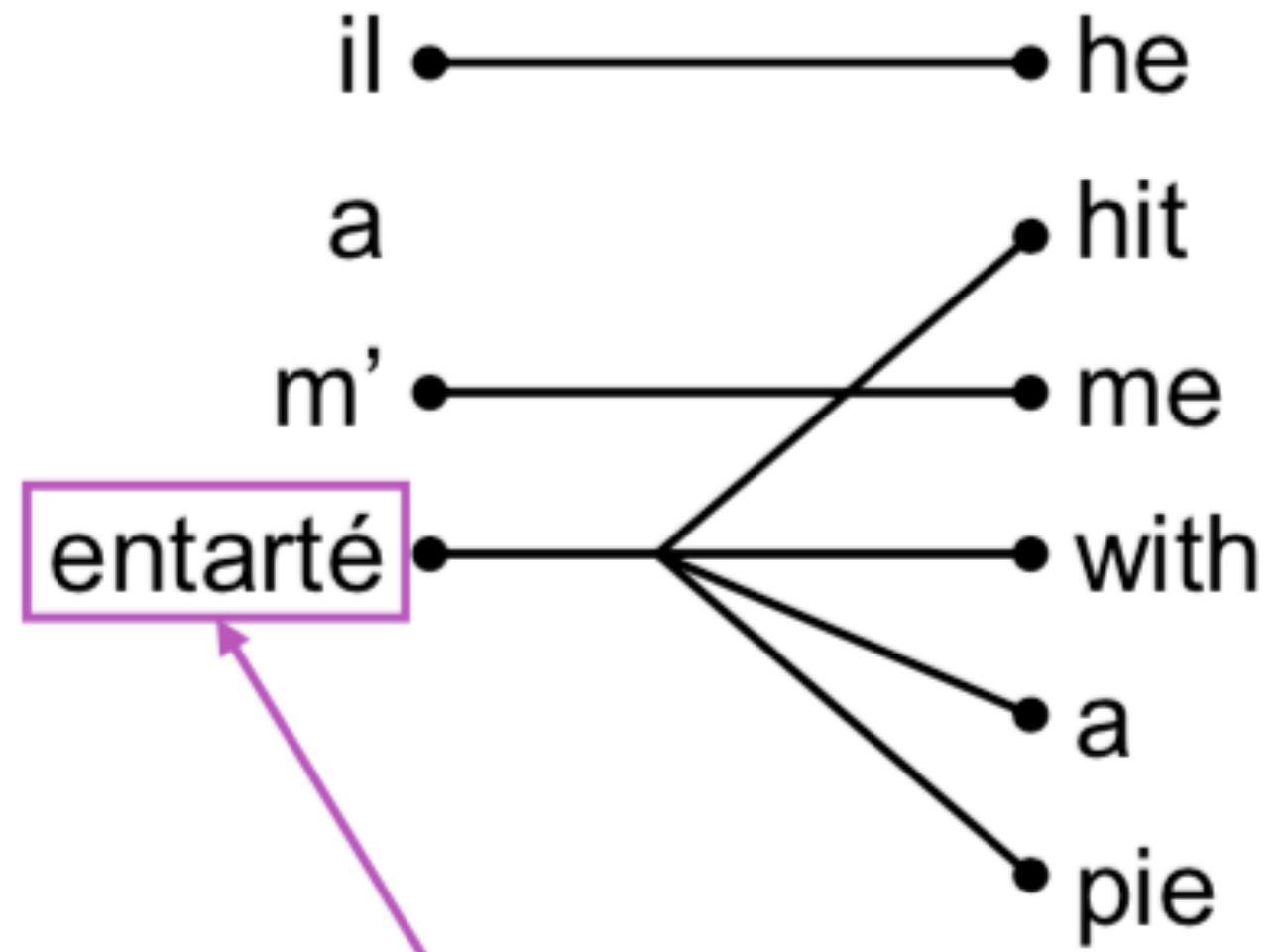
We call this a
fertile word

And
the programme
program a
has été
been mis
implemented en
application

one-to-many
alignment

And
the
program
has
been
implemented

Le	programme
a	éété
mis	en
en	application
application	



This word has no single-word equivalent in English

	he	hit	me	with	a	pie
il						
a						
m'						
entarté						



IBM Model 1	lexical translation
IBM Model 2	adds absolute reordering model
IBM Model 3	adds fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

Phrase-based Machine Translation

Statistical Machine Translation - Part III

Lexical Translation Table

Y	# Y Haus	Y	P(Y Haus)
house	8000	house	0.8
building	1600	building	0.16
home	200	home	0.02
household	150	household	0.015
shell	50	shell	0.05

The Les
poor paupr̄es
don't sont
have démunis
any
money

many-to-many
alignment

Les pauvres sont démunis

The poor don't have any money



The grid illustrates the many-to-many alignment between the two sentences. The first column ('The') has one shaded cell. The second column ('poor') has two shaded cells in the first two rows. The third column ('don't') has one shaded cell in the third row. The fourth column ('have') has two shaded cells in the first two rows. The fifth column ('any') has one shaded cell in the fifth row. The sixth column ('money') has one shaded cell in the fifth row.

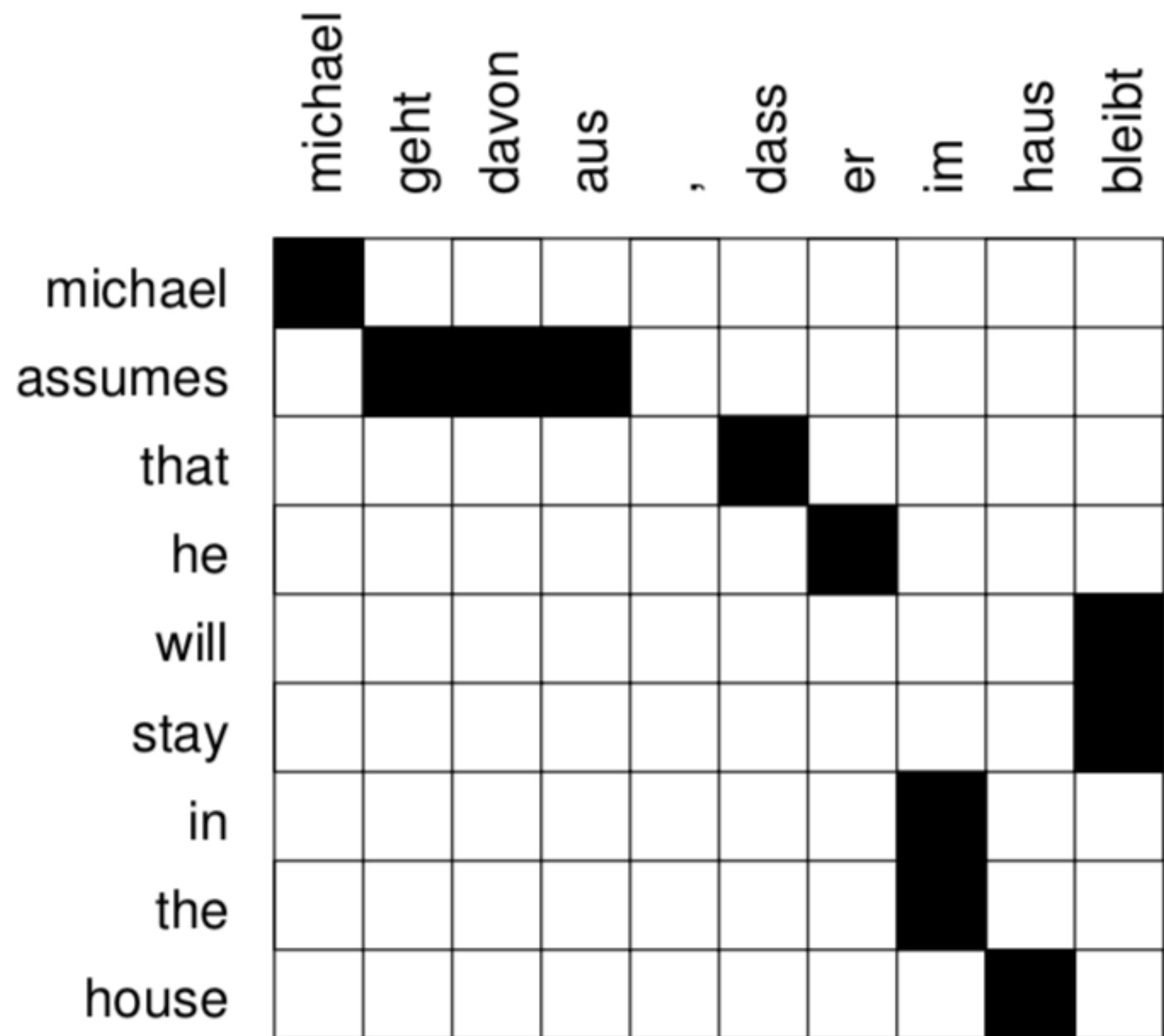
phrase
alignment

แปลเป็นก่อนๆ

- บางส่วนของประโยชน์ควรจะถูกแปลทั้งก่อนพร้อมกัน
- เปลี่ยนจาก lexical translation table เป็น phrase translation table

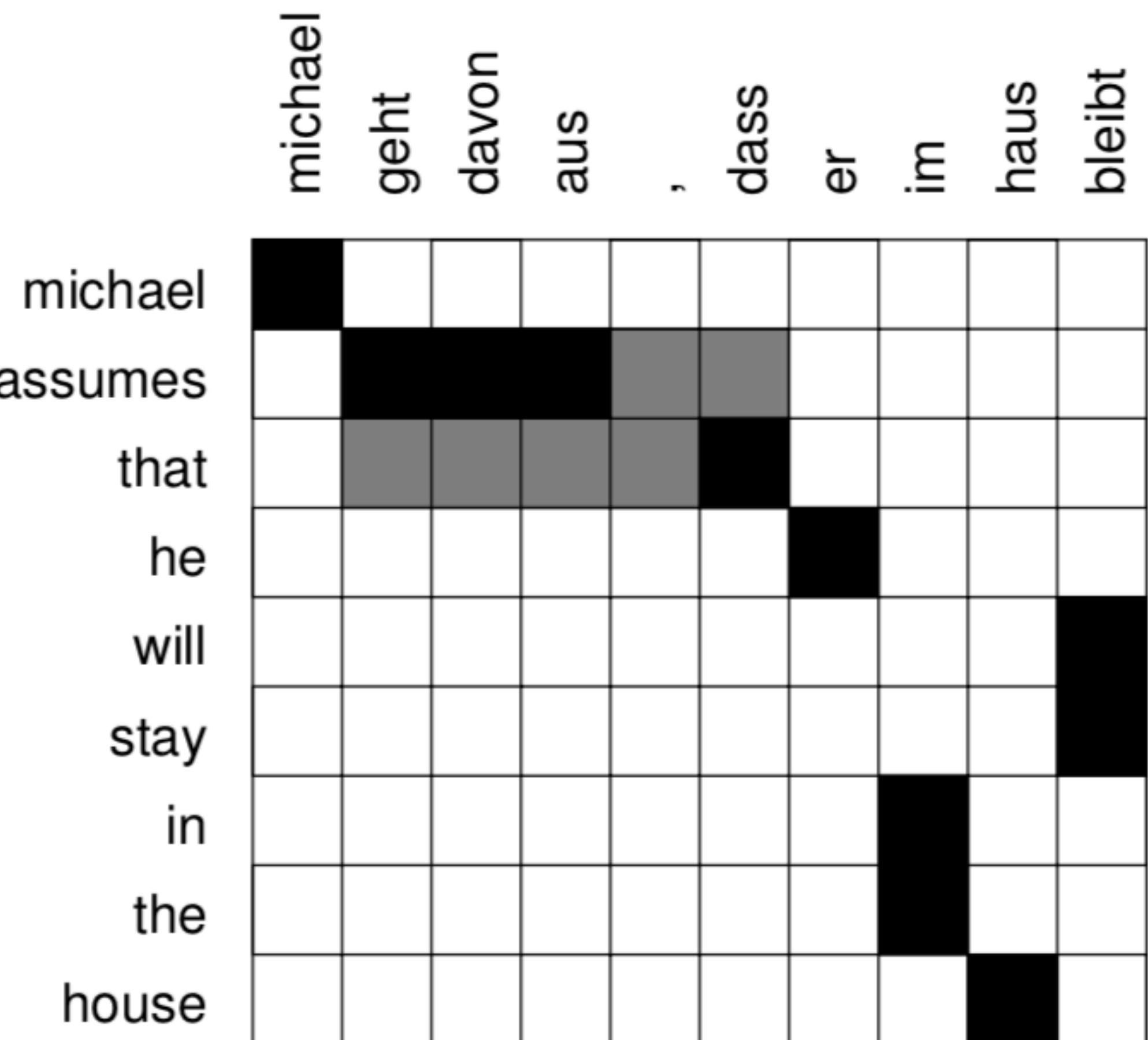
ສັກດ Phrase alignment

- phrase ປະກອບດ້ວຍຄຳທີ່ອູ່ຕິດ
ກຳນທງສອງພາສາ
- ຄ້າຄຳທີ່ອູ່ໃນ phrase ມີ word
alignment ຈະຕ້ອງເອາໄປທຸກຄຳໄປ
ຮວມໃນ phrase



Phrase 'မျိုး' Constituent

- assumes that he
geht davon aus, dass



Dear Dad, How are you doing?

Lieber Papa, wie geht es dir?

(Polish) (Tape) How are you doing?

(Polnisch) (Band) Wie geht es dir?

Comrade Major, how are you doing?

Genosse Major, wie geht's?

Morning, how are you doing?

Guten Morgen, wie geht's?

HAP: How are you doing?

(Eckhart) Wie geht es Ihnen?

How are you doing? I haven't seen you in ages!

Wie geht es Ihnen? Ich habe Sie seit Ewigkeiten nicht gesehen!

How are you doing? I haven't seen you in ages!

Wie geht es dir? Ich habe dich seit Ewigkeiten nicht gesehen!

(Tape) How are you doing?

(Band) Wie geht es dir?

How are you doing? Thank you.

[Laut] Wie geht es dir?

Phrase Translation Table

- ดึง phrases ออกมาให้หมดจาก word alignment (ซ้ำกันบ)
- $P(y \text{ phrase} | x \text{ phrase}) = \frac{\# y \text{ phrase } \longleftrightarrow x \text{ phrase}}{\# x \text{ phrase}}$
- ปัญหา?

Lexical vs Phrase Translation Table

Y	$P(Y \text{Haus})$	Y	$P(Y \text{geht davon aus, dass})$
house	0.8	assumes that	0.90
building	0.16	assumes	0.05
home	0.02	is based on the fact that	0.03
household	0.015	are regarded as	0.02
shell	0.05		

Log-linear model

$$p(e, a|f) = \exp(\lambda_\phi \sum_{i=1}^I \log \phi(\bar{f}_i | \bar{e}_i) +$$

$$\lambda_d \sum_{i=1}^I \log d(a_i - b_{i-1} - 1) +$$

$$\lambda_{LM} \sum_{i=1}^{|\mathbf{e}|} \log p_{LM}(e_i | e_1 \dots e_{i-1}))$$

Other features

- จำนวนคำ
- จำนวน phrase
- Bi-directional alignment
- แหล่งความรู้อื่น ๆ

Statistical MT

- ระบบ SMT ใหญ่เบ็มมาก เพราะต้องเก็บตารางไว้เยอะแยะ
- ต้องพยายามหา features มาเสริมเยอะๆ ในแต่ละคู่ภาษา

Beam Search Decoding

Decoding

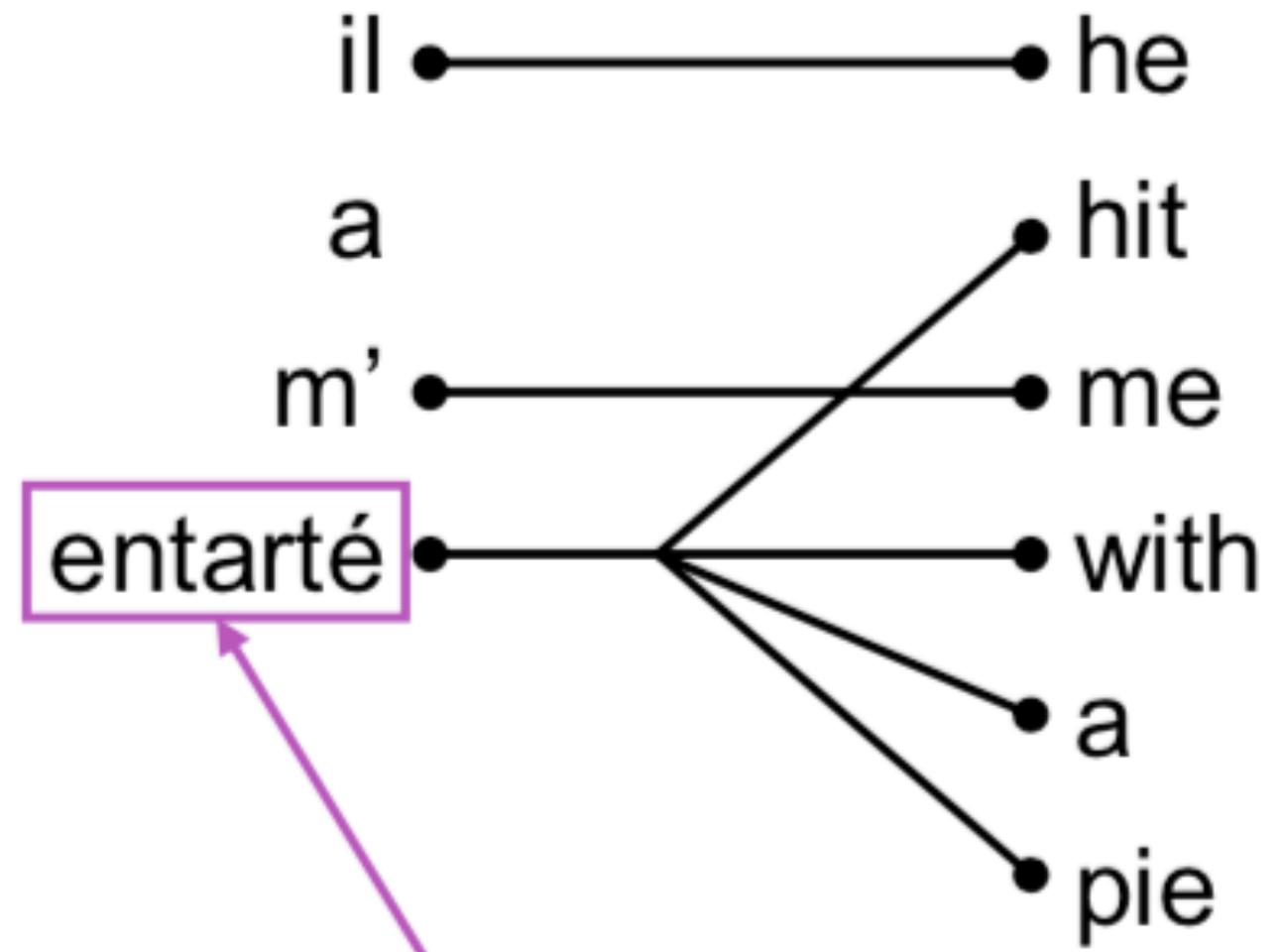
$\text{score}(\text{คำแปล}, \text{ต้นฉบับ}) = \text{adequacyScore}(\text{คำแปล}, \text{ต้นฉบับ}) + \text{fluencyScore}(\text{คำแปล})$

เงื่อนไขของ Scoring Function

- ต้องคิดทีละคำจากซ้ายไปขวาได้
- $\text{Score}(\text{"He hit"}, \text{"Il m'a entarté"}) =$
 $\text{Score}(\text{"He"} | \text{"Il m'a entarté"}) + s(+\text{hit})$
- $\text{Score}(\text{"He hit me"}, \text{"Il m'a entarté"}) =$
 $\text{Score}(\text{"He hit"} | \text{"Il m'a entarté"}) + s(+\text{me})$

Search หาคำแปลที่ดี(ที่สุด?)

- Exhaustive search ค้นหาแบบหมดจด
- Greedy search ค้นหาแบบละโมบ
- Beam search ค้นหาแบบลำแสง



This word has no single-word equivalent in English

	he	hit	me	with	a	pie
il						
a						
m'						
entarté						



Exhaustive Search

Il m'a entarté

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

Exhaustive Search

Il m'a entarté

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

Exhaustive Search (Viterbi)

- เป็นไปไม่ได้ เพราะภาษา มีความเป็นอนันต์
- ถ้าอยากรองทุกประโยคที่มีความยาว k คำ และ vocab size = V เราจะต้องลองทั้งหมด V^k ประโยค
- $30,000^{10} =$ เยอะเกินสมองมนุษย์จะเข้าใจ

Greedy Search

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

Il m'a entarté

he
it
hit
struck
with
on
a
one
pie
tart
END

he
it
hit
struck
with
on
a
one
pie
tart
END

Greedy Search

- เร็วดี แต่...
- ถ้าผิดตอนต้นๆ มันจะส่งผลไปถึงที่เหลือทั้งหมด

Beam Search

- แต่ละ step เก็บ hypothesis เอ้าไว้ k ตัว
- แต่ละ hypothesis เอามาขยายเพิ่มอีกคำ

Beam search decoding: example

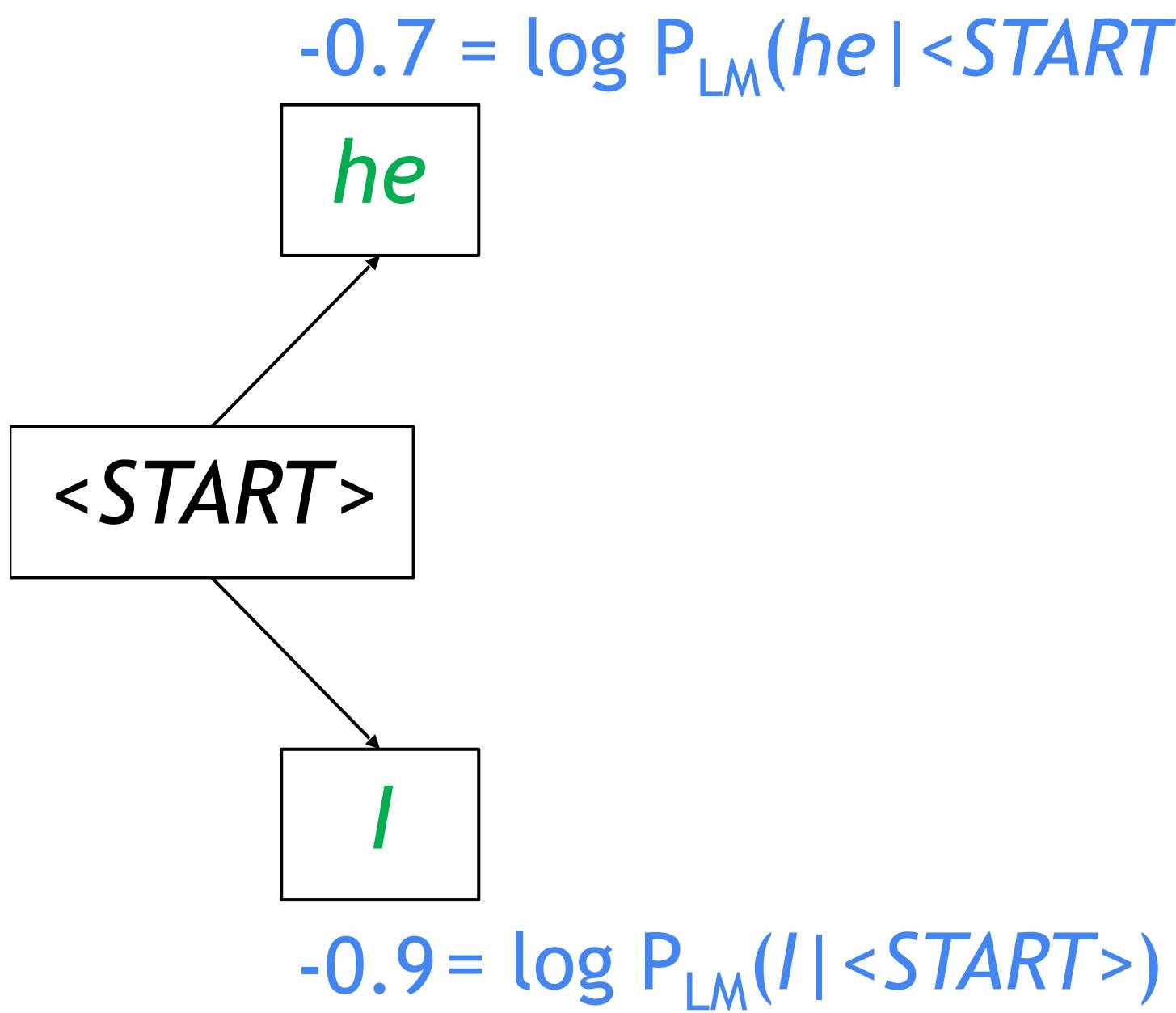
Beam size = k = 2. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$

<START>

Calculate prob
dist of next word

Beam search decoding: example

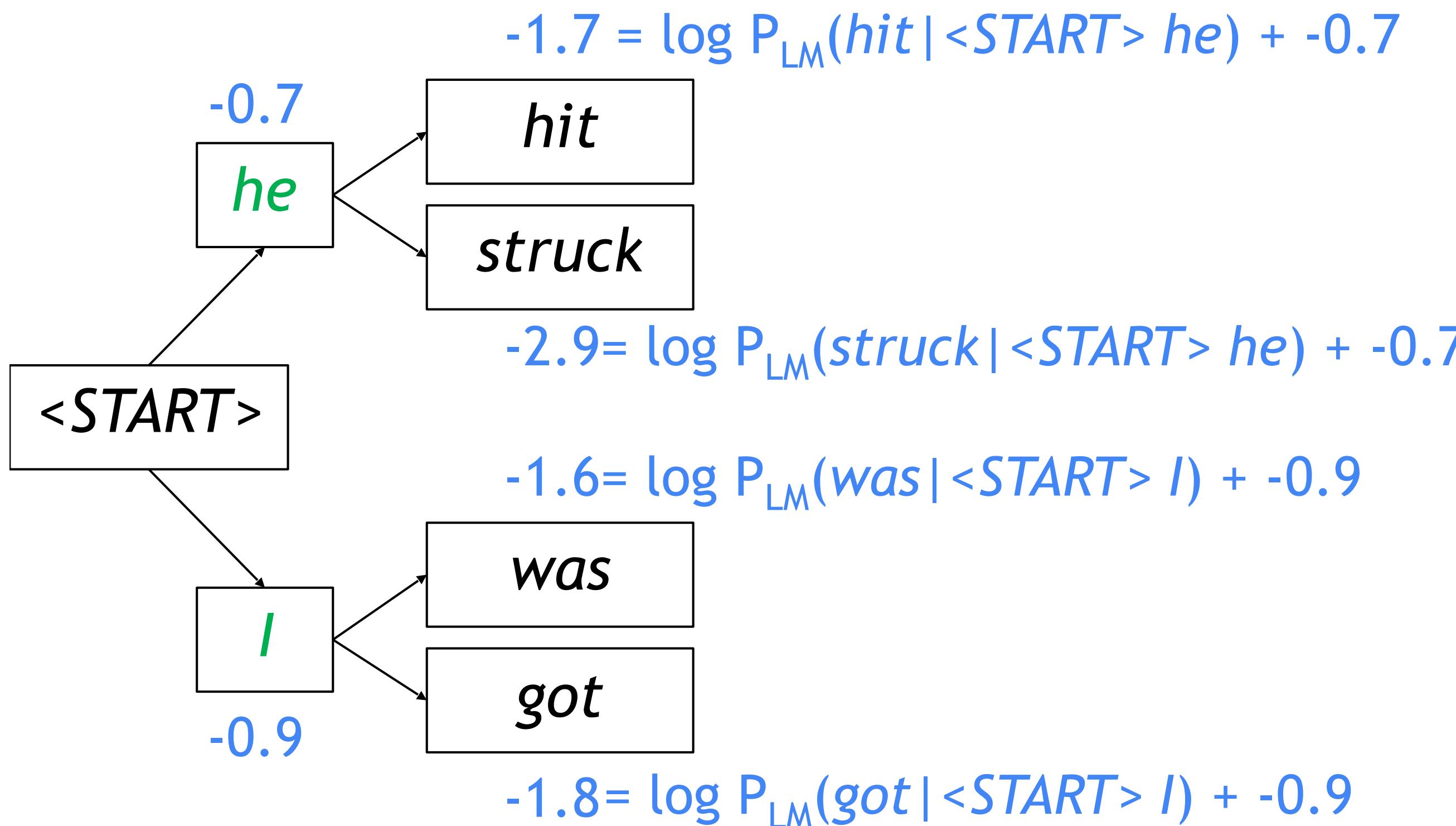
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Take top k words
and compute scores

Beam search decoding: example

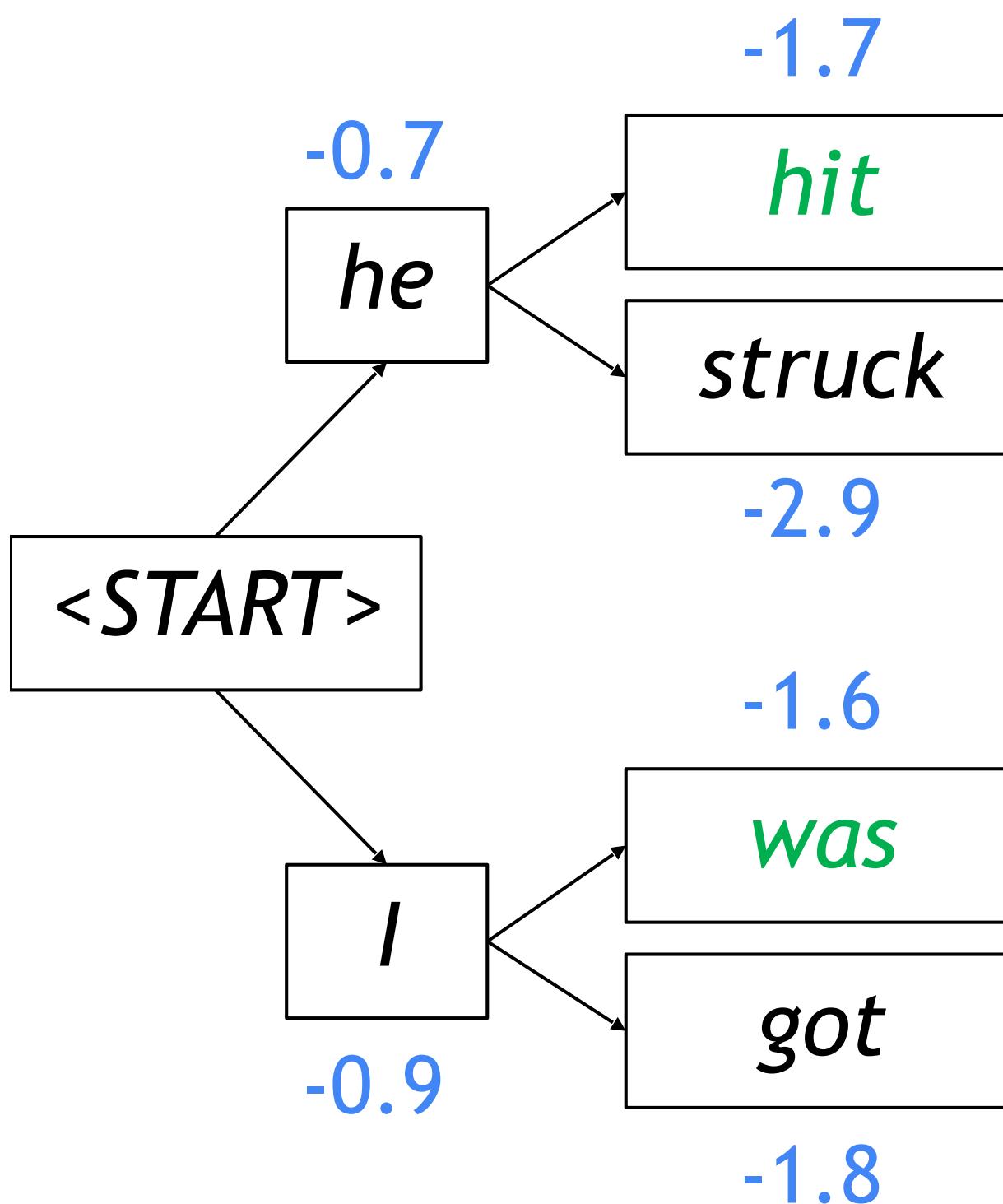
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



For each of the k hypotheses, find top k next words and calculate scores

Beam search decoding: example

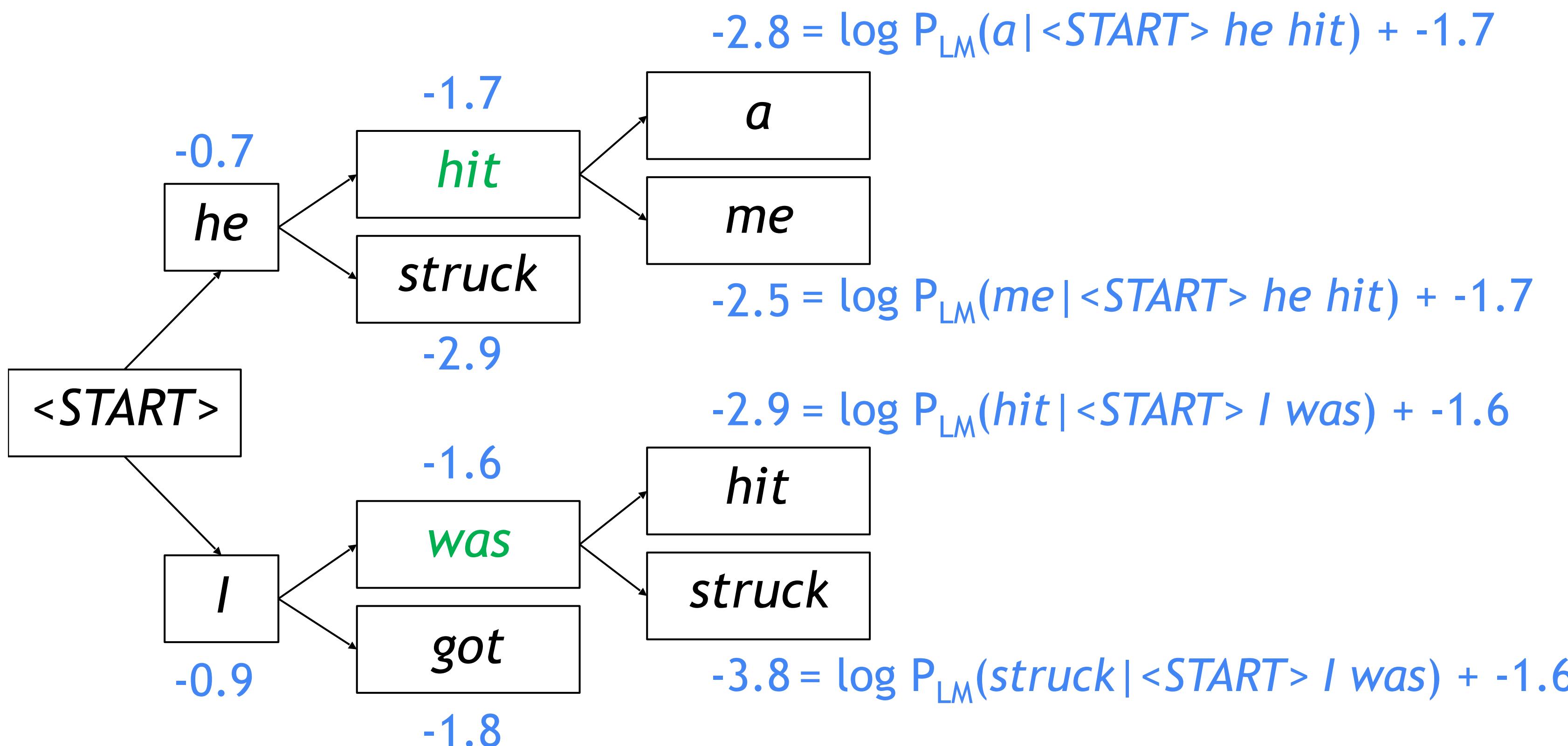
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Of these k^2 hypotheses,
just keep k with highest scores

Beam search decoding: example

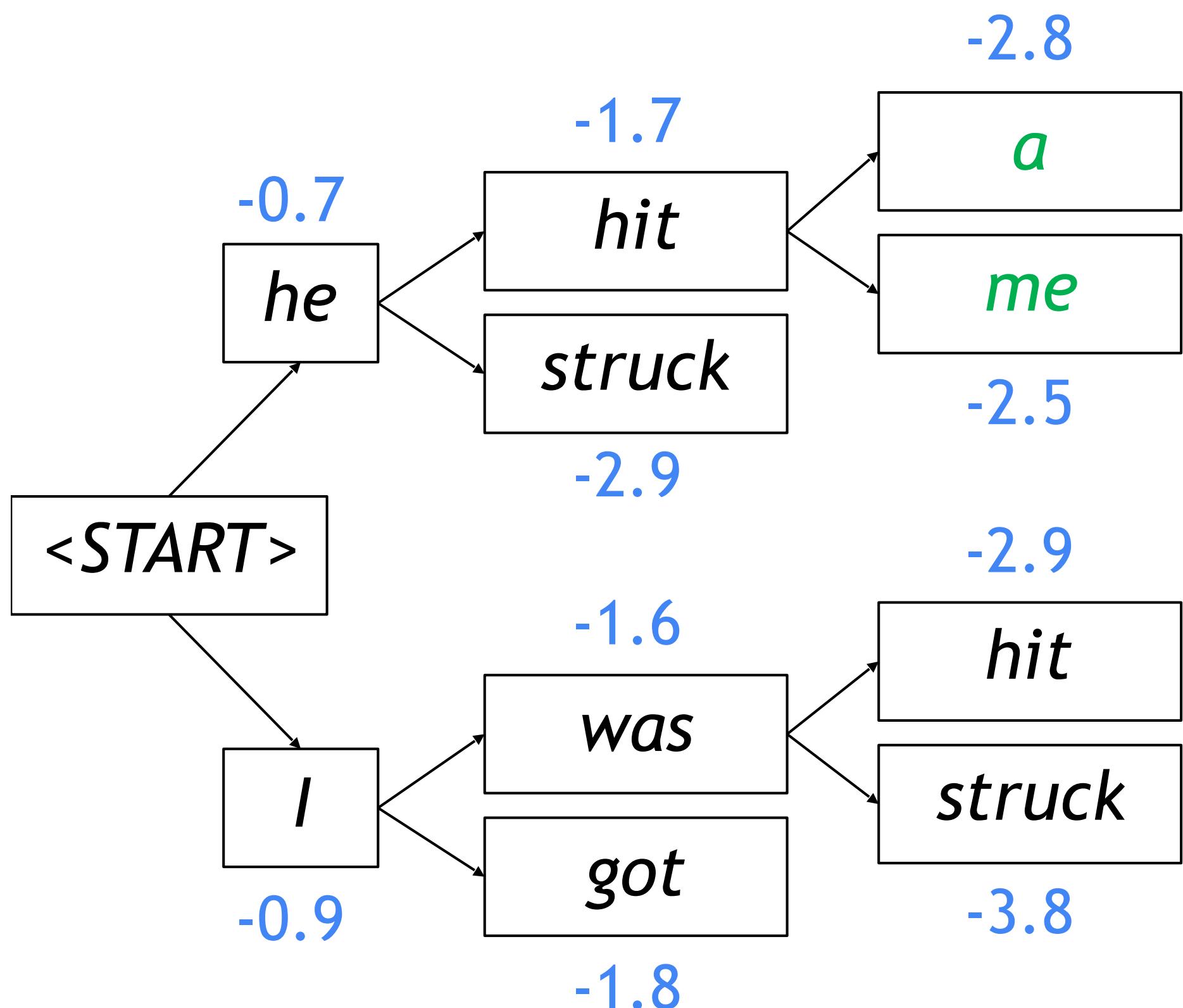
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



For each of the k hypotheses, find
top k next words and calculate scores

Beam search decoding: example

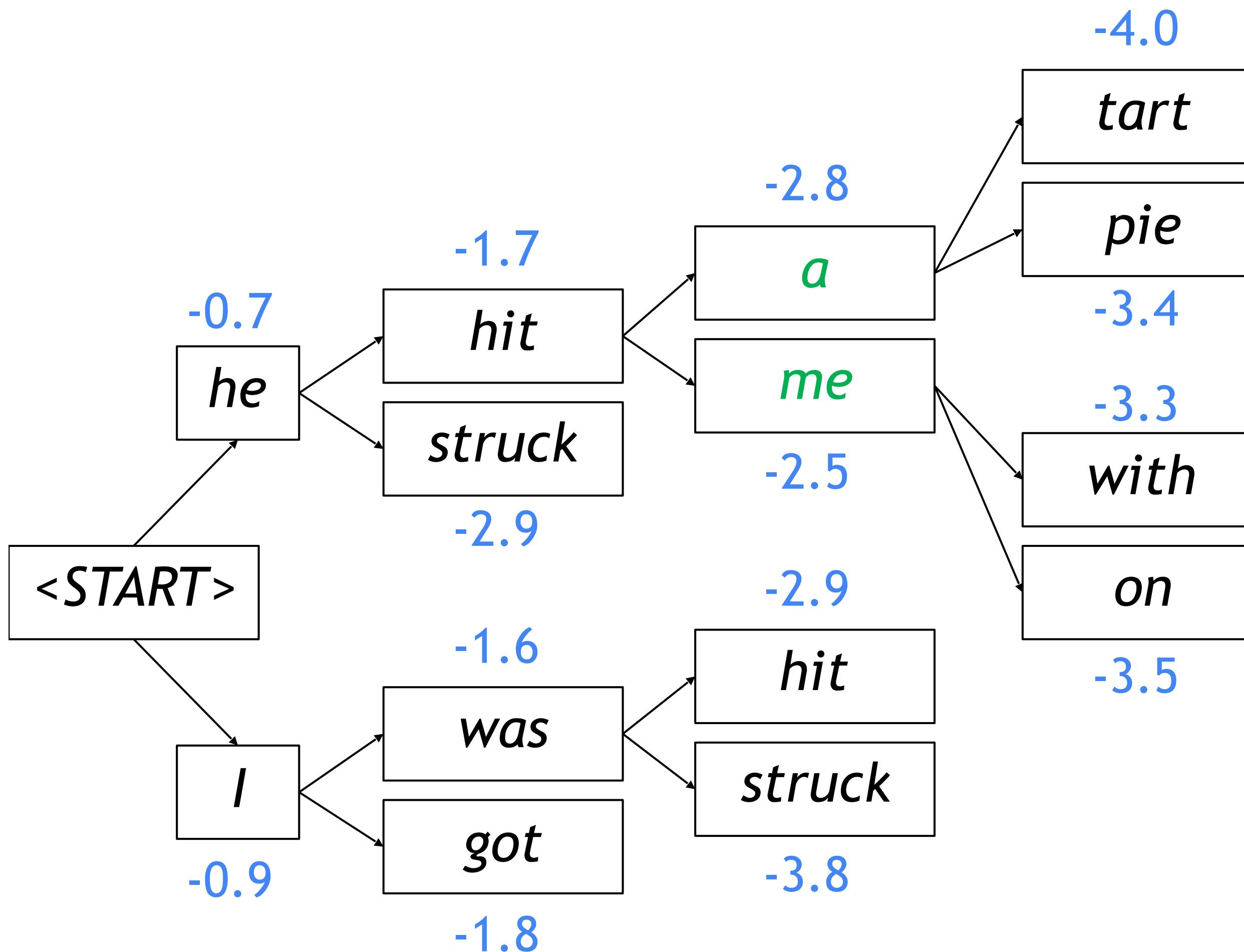
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Of these k^2 hypotheses,
just keep k with highest scores

Beam search decoding: example

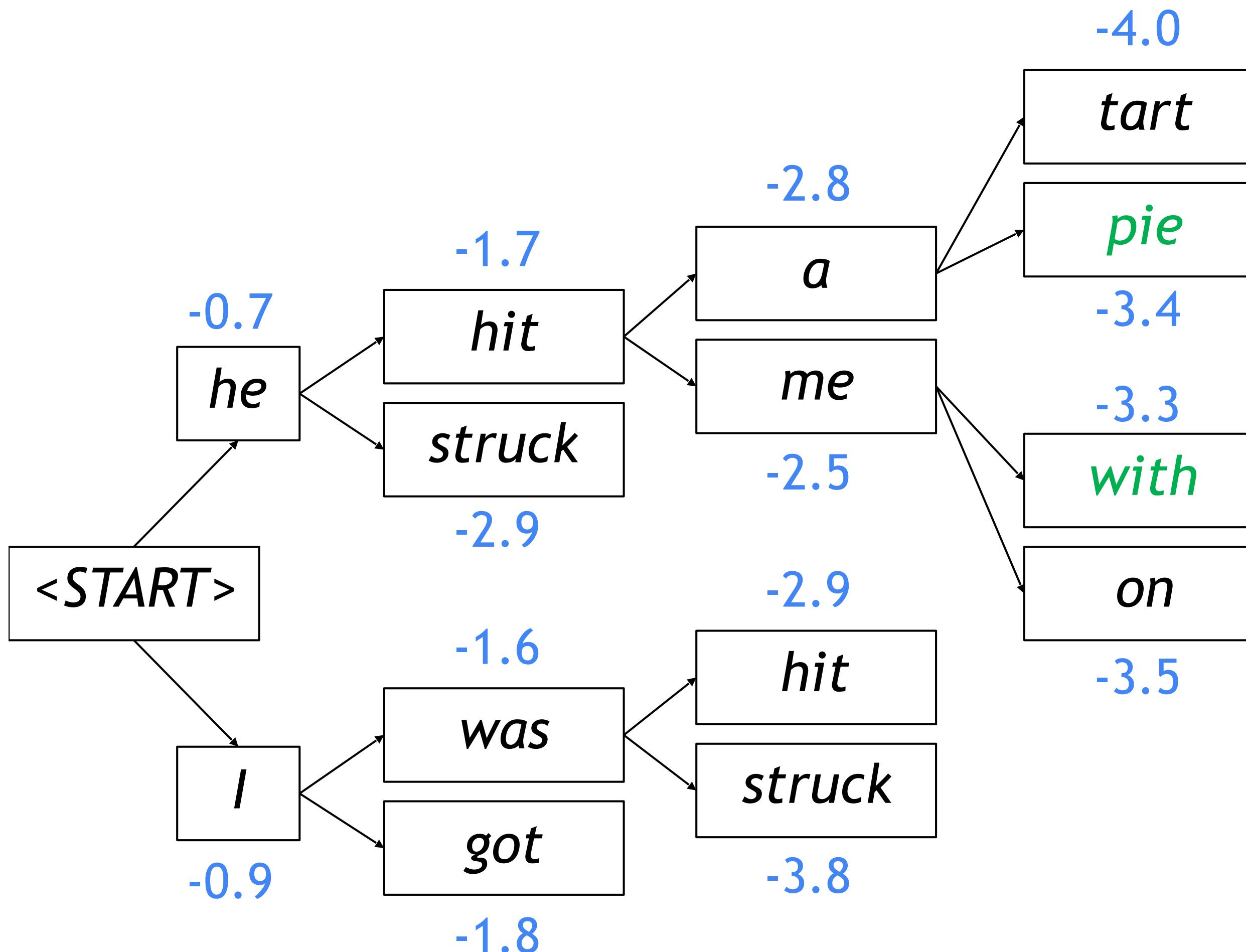
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



For each of the k hypotheses, find top k next words and calculate scores

Beam search decoding: example

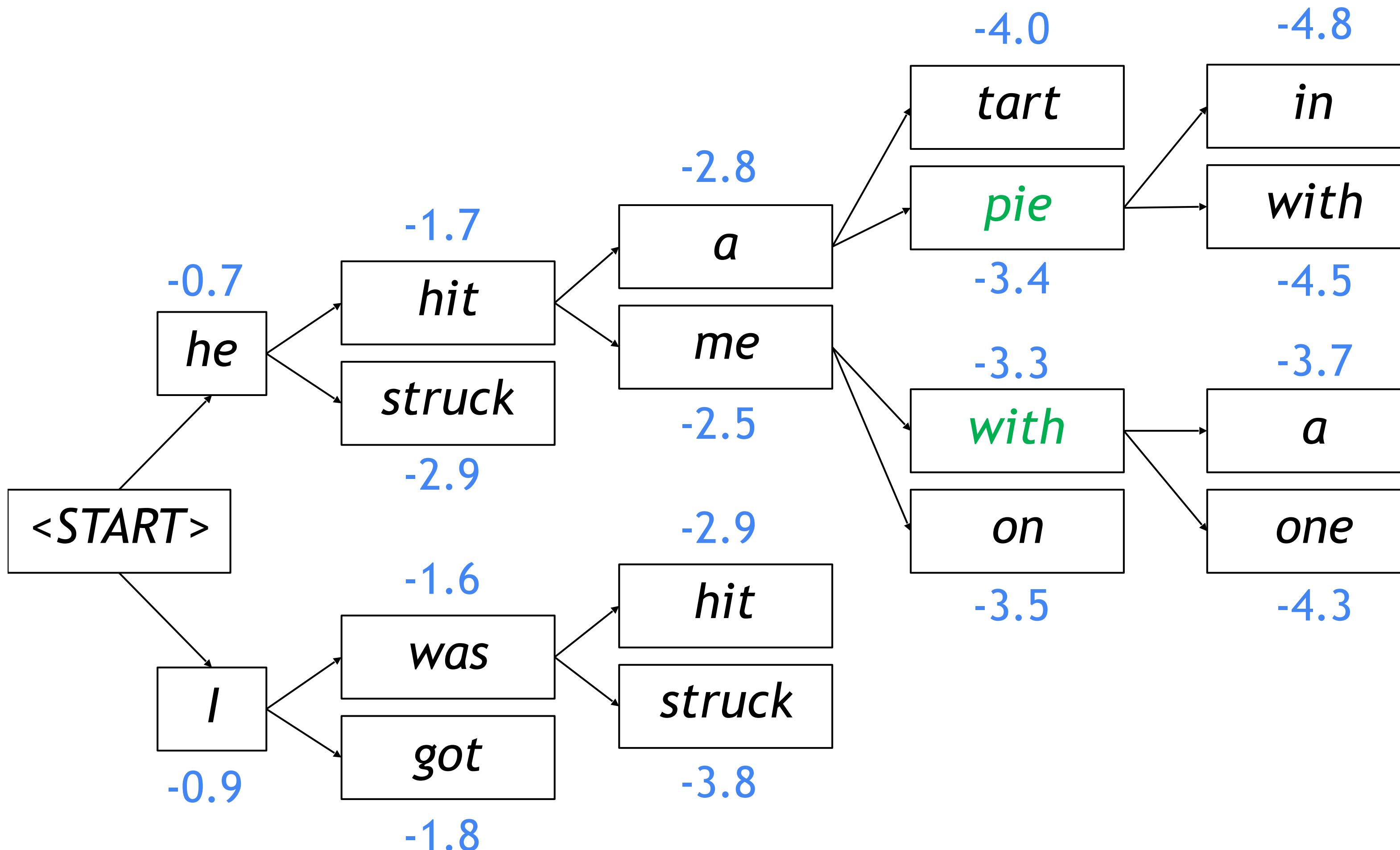
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Of these k^2 hypotheses,
just keep k with highest scores

Beam search decoding: example

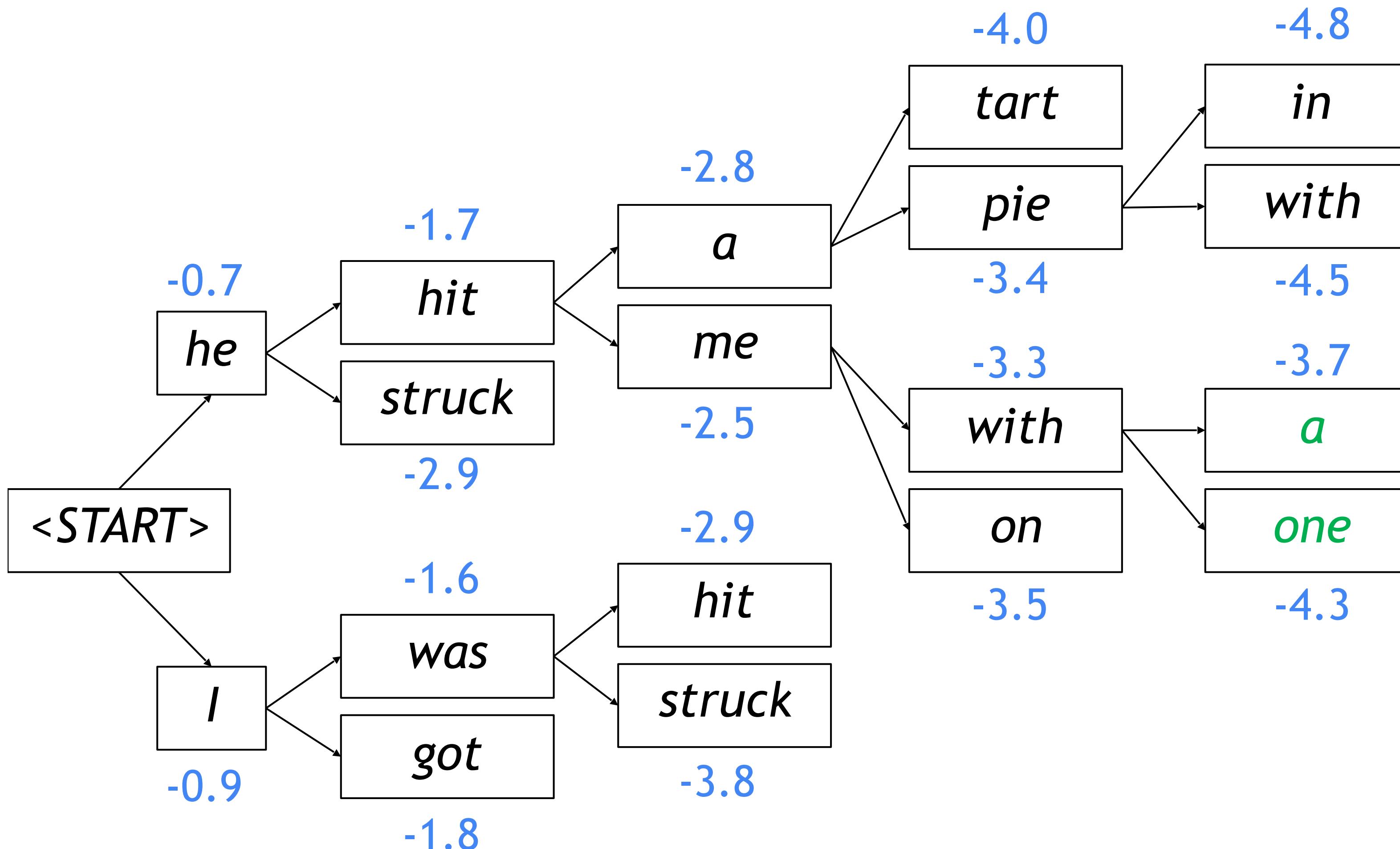
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



For each of the k hypotheses, find top k next words and calculate scores

Beam search decoding: example

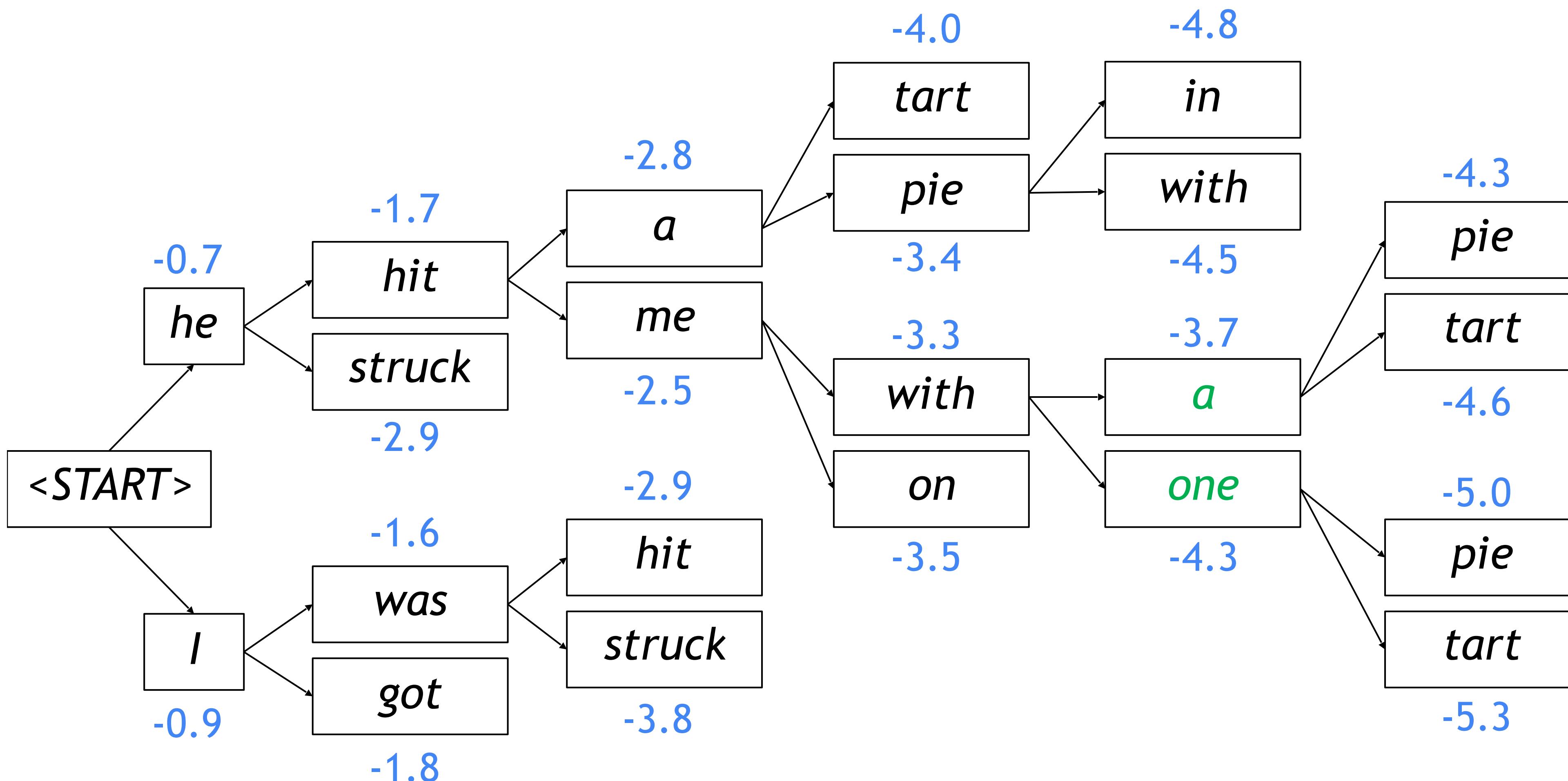
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Of these k^2 hypotheses,
just keep k with highest scores

Beam search decoding: example

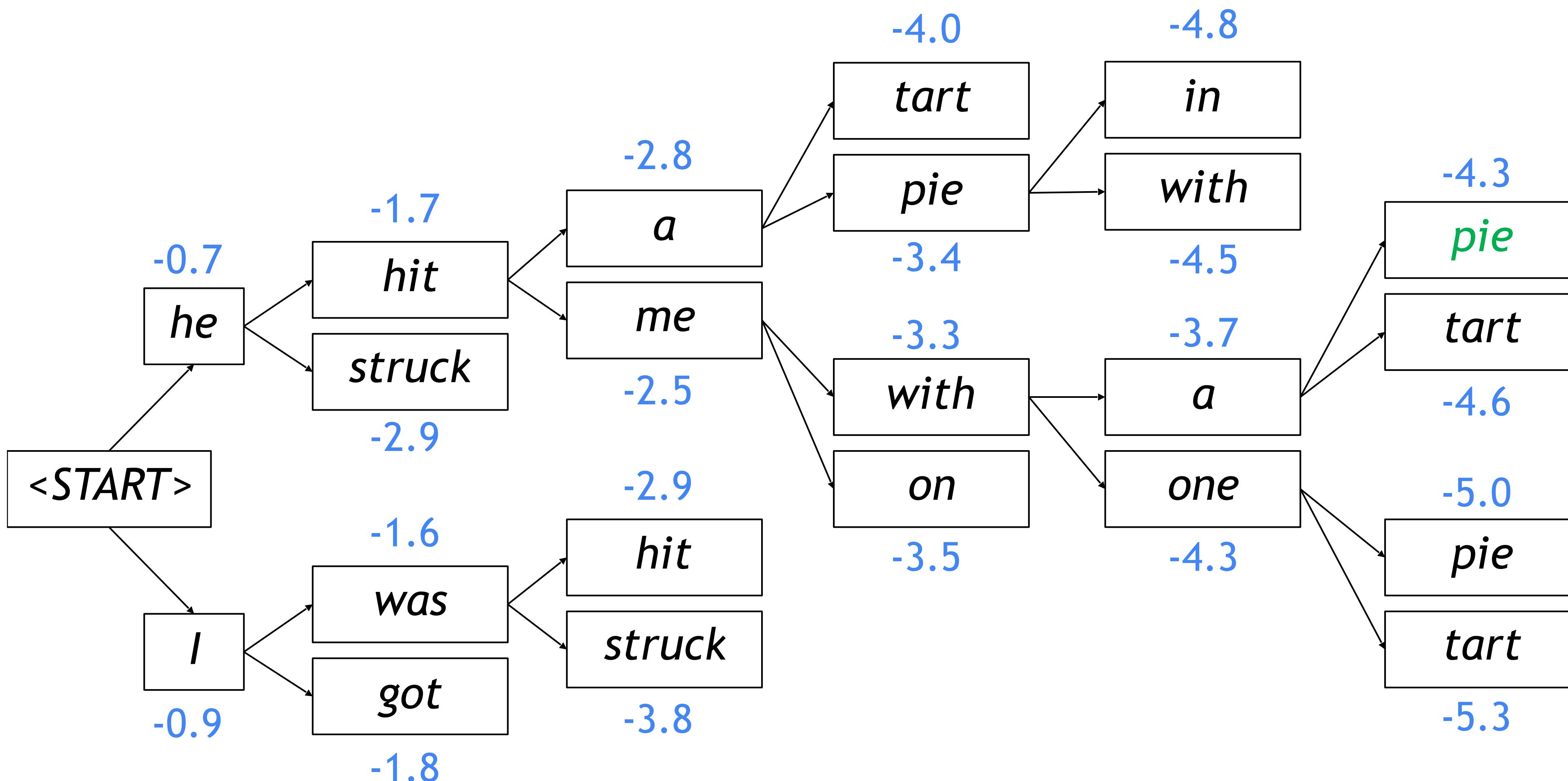
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



For each of the k hypotheses, find
top k next words and calculate scores

Beam search decoding: example

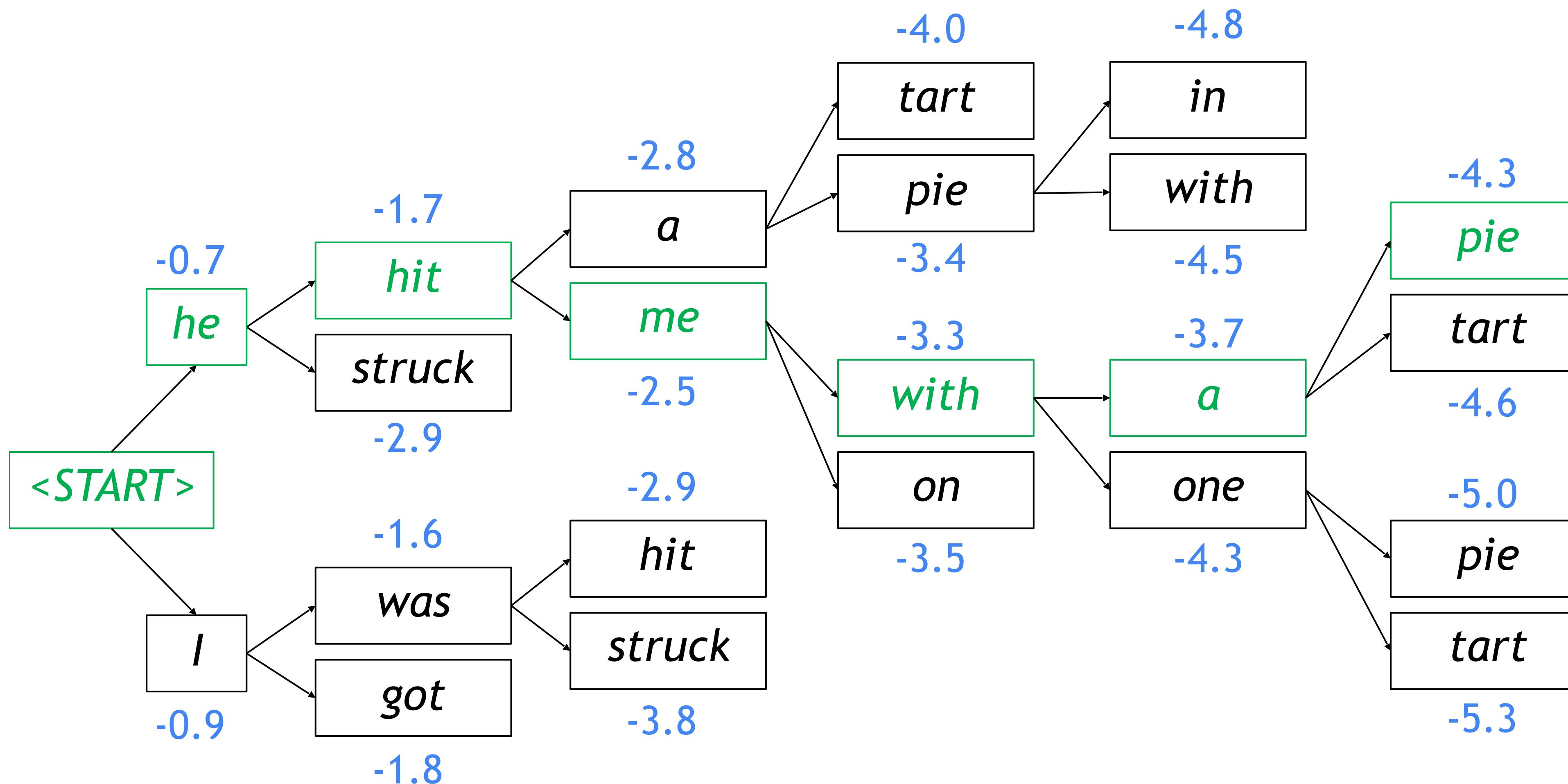
Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



This is the top-scoring hypothesis!

Beam search decoding: example

Beam size = $k = 2$. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Beam Search ຈະຍັງໄຟ

- expand ໄປເຮືອຍໆ ຈນກວ່າຈະເຈອ <END> ຄືວ່າ
ສມບູຮັນ
- Hypothesis ໄහນຍັງໄມ່ເສີຣຈັກ search ຕ່ອໄປຈນກວ່າຈະ
ຄື່ງຄວາມຍາວສູງສຸດ
- ທີ່ຢູ່ຕອນໄດ້ຄໍາແປລທີ່ສມບູຮັນຈຳນວນທີ່ຕ້ອງການ

สรุป

- Beam search decoding เป็นวิธีการนำ translation model และ scoring model อื่นๆ ไปใช้ในการแปลงโดยคจริง ๆ
- ไม่ได้ผลที่ดีที่สุด แต่ว่าเร็วและได้ผลดีแบบยอมรับได้

Evaluation for MT

MT vs Sequence Tagging

ลุง	ตู่	ต่อว่า	ผู้สื่อข่าว	ที่	ตึกไทยคู่ฟ้า	เมื่อ	เช้า
B-PER	I-PER	O	O	O	B-PLACE	B-TIME	I-TIME

การแปลมีมี ground truth ແທ້າ

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israel government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport's security.

Israel is responsible for safety work at this airport.

Israel presides over the security of the airport.

Israel took charge of the airport security.

The safety of this airport is taken charge of by Israel.

This airport's security is the responsibility of the Israeli security officials.

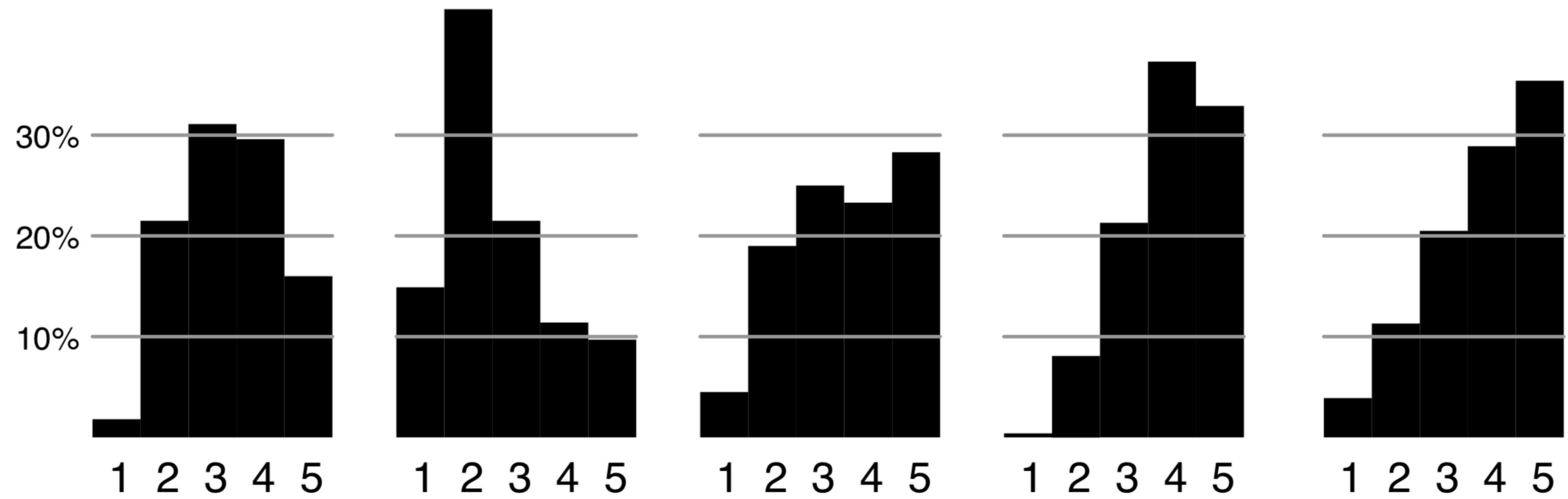
(a typical example from the 2001 NIST evaluation set)

Adequacy and Fluency

- Adequacy:
 - คำแปลสื่อความหมายเดียวกับประโยค input มี
 - สาระขาด เกิน หรือบิดเบือนมี
- Fluency:
 - พังดูเหมือนภาษาที่เจ้าของภาษาพูดรีบเล่า
 - ผิดไวยากรณ์มี ใช้คำผิดไม่เหมาะสมกับความหมายหรือเปล่า

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible



(from WMT 2006 evaluation)

การวัดความพ้องกัน

- Kappa coefficient

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

- $p(A)$: proportion of times that the evaluators agree
- $p(E)$: proportion of time that they would agree by chance
(5-point scale → $p(E) = \frac{1}{5}$)

- Example: Inter-evaluator agreement in WMT 2007 evaluation campaign

Evaluation type	$P(A)$	$P(E)$	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226

จัดลำดับคุณภาพของการแปล

- ระหว่างสองอันนี้ อันไหนดีกว่า หรือดีเท่ากัน
- ผลออกมา consistent กว่า

Evaluation type	$P(A)$	$P(E)$	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226
Sentence ranking	.582	.333	.373

วัดคุณภาพอัตโนมัติ

- ทำไมถึงจำเป็นต้องมีมาตรการวัดคุณภาพโดยอัตโนมัติ

BLEU Score

SYSTEM A: **Israeli officials** responsibility of **airport** safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: **airport security** **Israeli officials are responsible**
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

$$\text{BLEU} = \min \left(1, \frac{\text{output-length}}{\text{reference-length}} \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$