自然语言处理如何落地互联网

（打造你自己的Google Translate?）

李志飞

出门问问 CEO

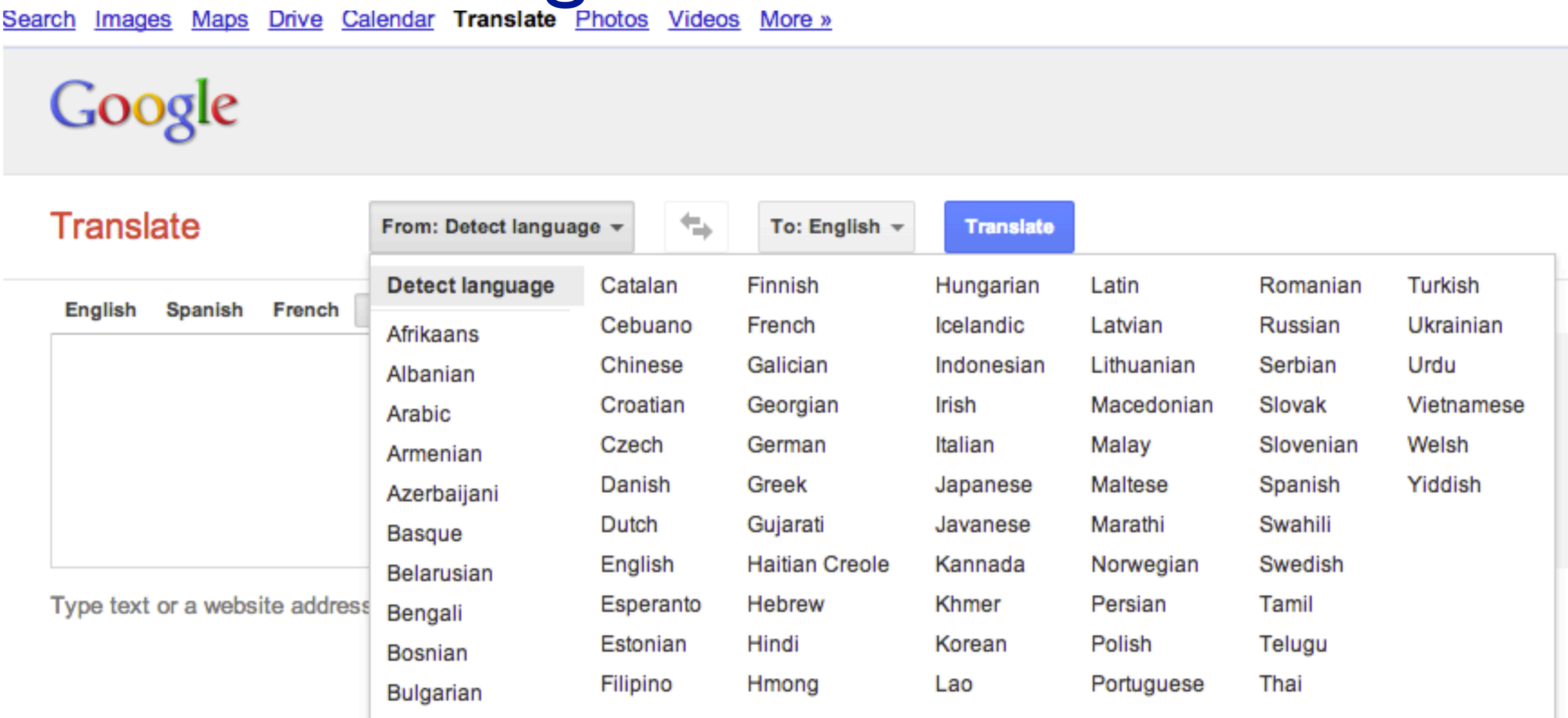
邮件：zfli@mobvoi.com

微博：@李志飞-出门问问

Google Translate

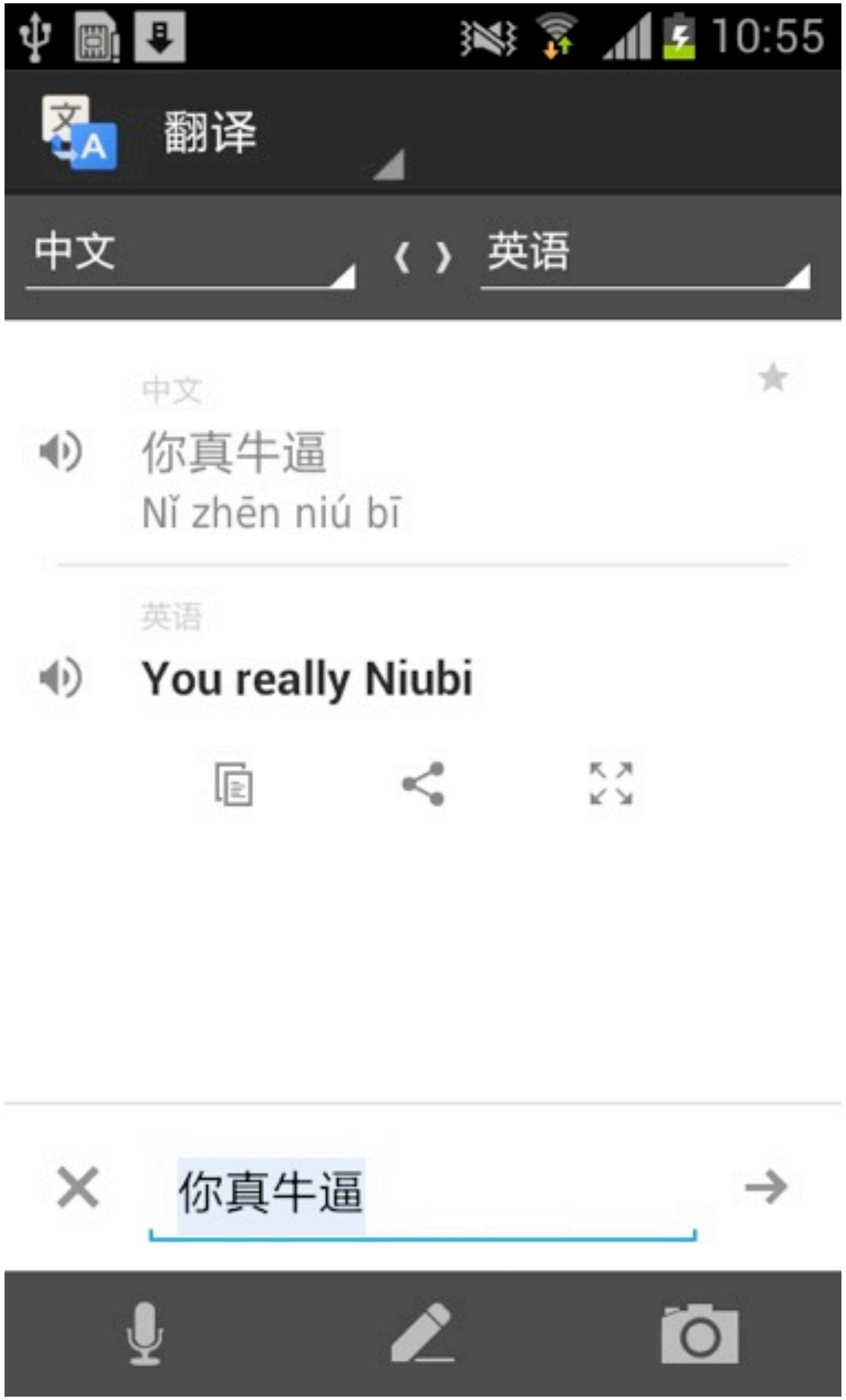
NLP在互联网上最成功的应用？

Google Translate网站



- 支持71种语言
- 流量全世界排在20以内（类似于bing.com）
- 用户数超过3亿，每天的翻译请求10亿级别

Google Translate 移动应用



- Google官方最流行应用之一
- 支持文字，语音，手写，图片等多媒体输入

Outline

- Google Translate
- 机器翻译for dummy
- 机器翻译基础理论和算法
  - 机器学习
  - 数据结构，模型，算法
- 工业界机器翻译系统实战

Training a Translation Model

垫子 上 的 猫

dianzi shang de mao

a cat on the mat

word alignment?

Word Alignment

垫子 上 的 猫

a cat on the mat

我 看见 猫

I saw a cat

我 有 猫 和 狗

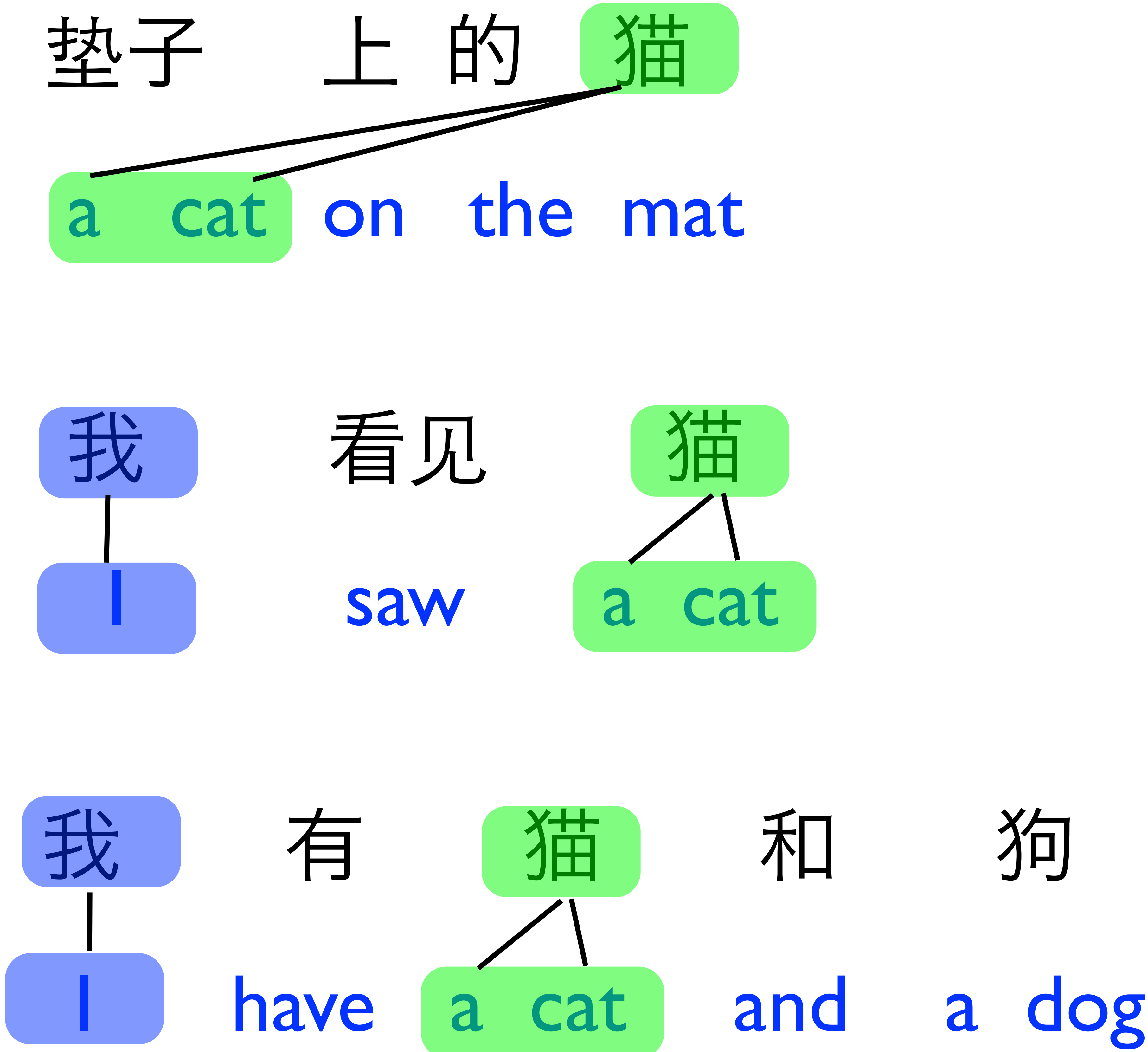
I have a cat and a dog

Word Alignment

数据冗余

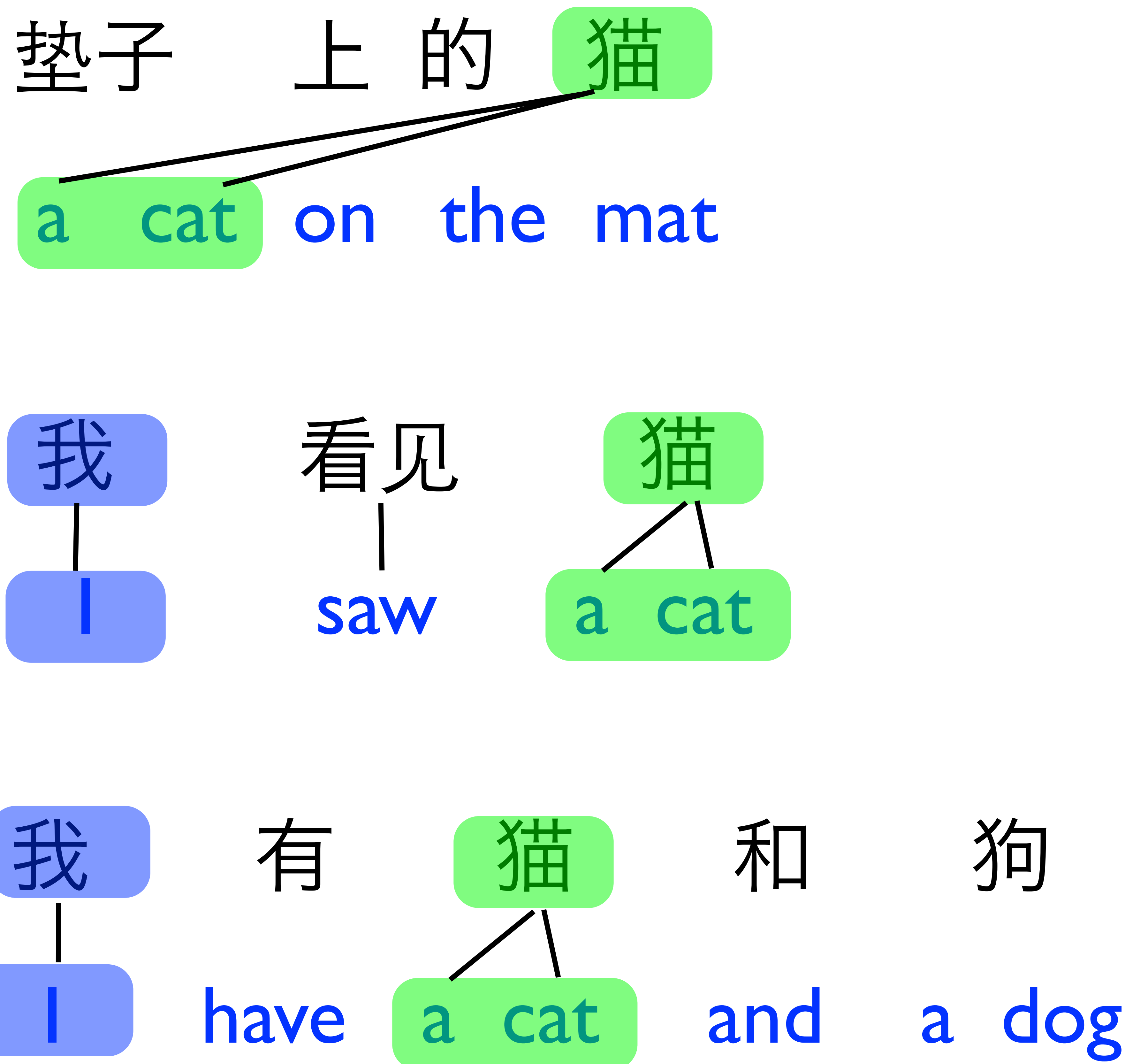
Word Alignment

数据冗余

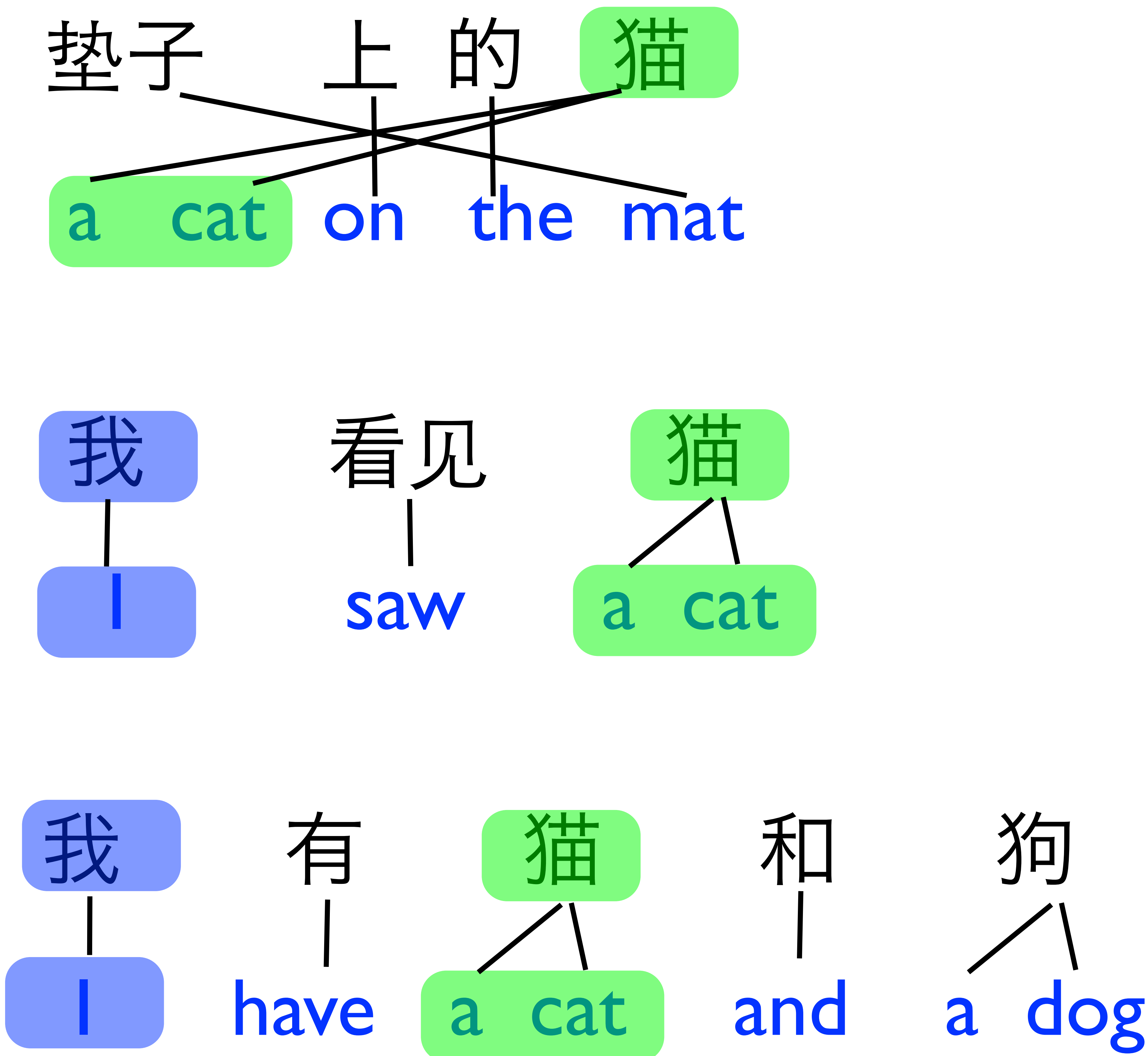


Word Alignment

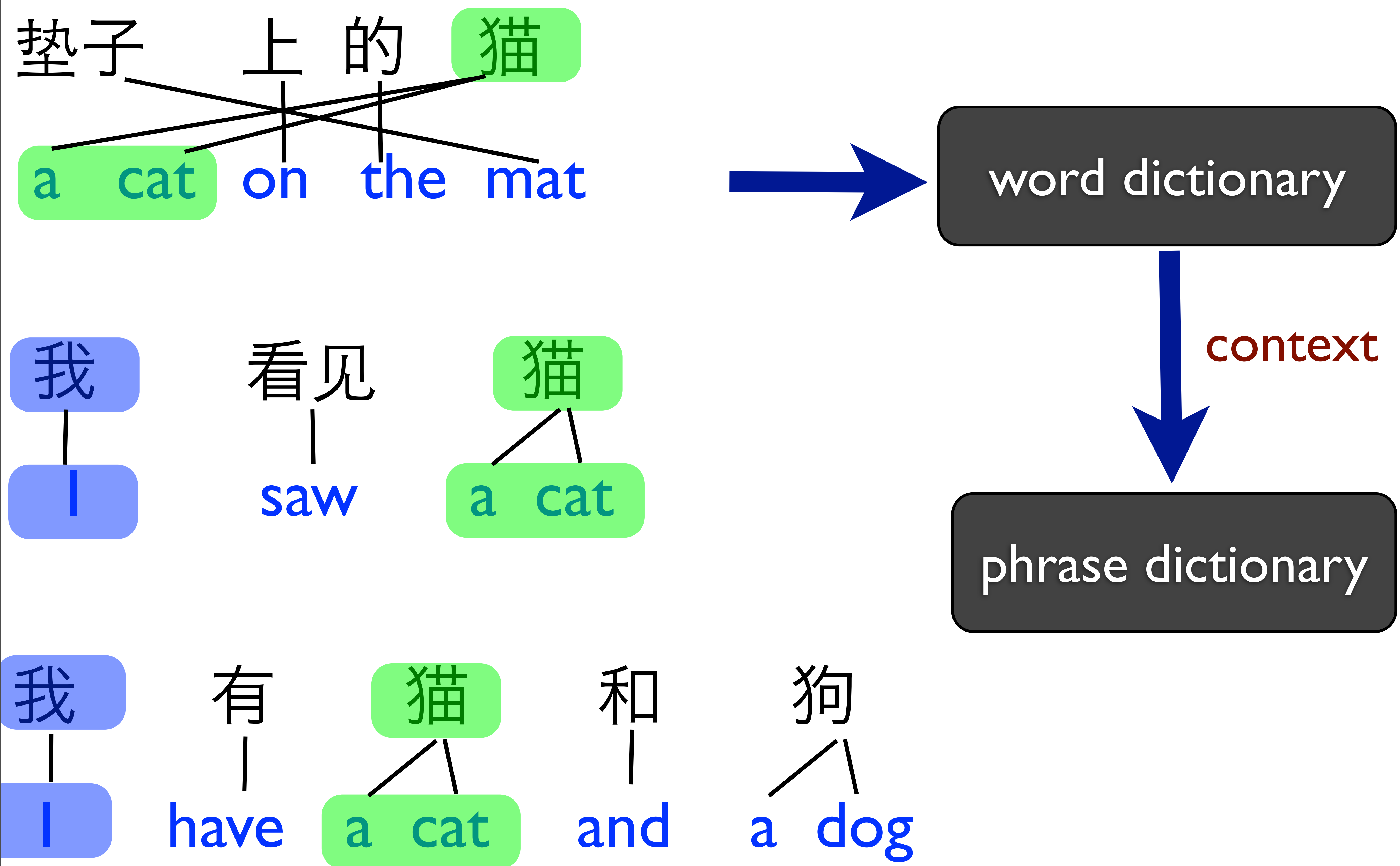
数据排除



Word Alignment



Word Alignment



Phrase Extraction

垫子 上 的 猫

dianzi shang de mao

a cat on the mat

$$X \rightarrow \langle \text{ dianzi shang , the mat } \rangle$$
$$X \rightarrow \langle \text{ mao , a cat } \rangle$$

Phrase Extraction

垫子 上 的 猫

dianzi shang de X_0

X_0 on the mat

$$X \rightarrow \langle \text{ dianzi shang , the mat } \rangle$$

$$X \rightarrow \langle \text{ mao , a cat } \rangle$$

$$X \rightarrow \langle \text{ dianzi shang de } X_0 \text{ , } X_0 \text{ on the mat } \rangle$$

Phrase Extraction

垫子 上 的 猫

X_0 de mao

a cat on X_0

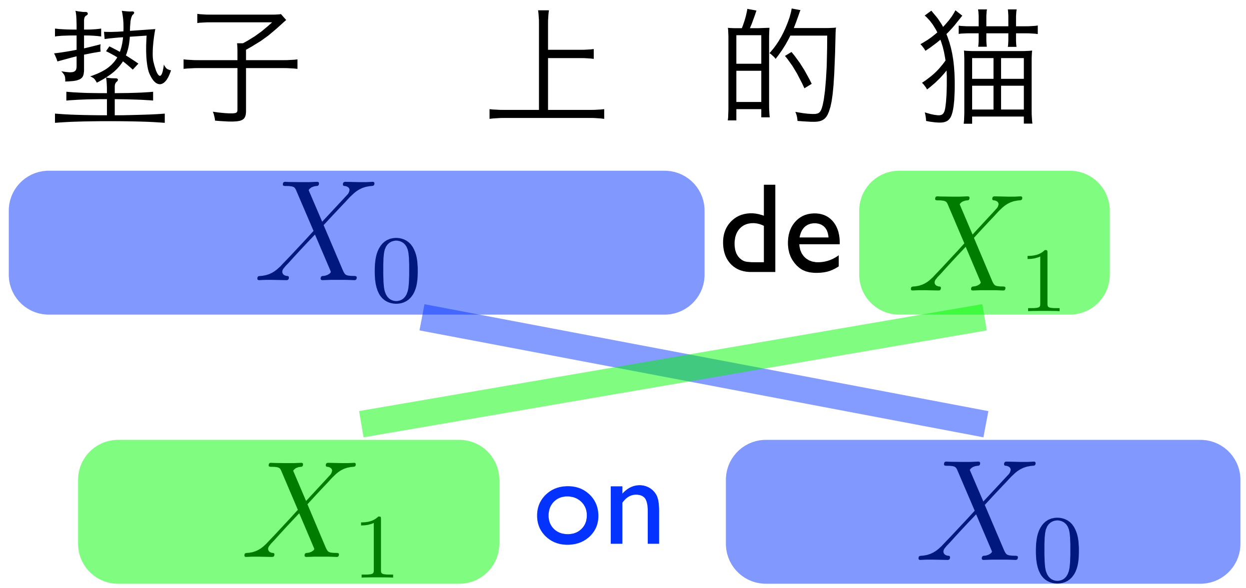
$$X \rightarrow \langle \text{ dianzi shang , the mat } \rangle$$

$$X \rightarrow \langle \text{ mao , a cat } \rangle$$

$$X \rightarrow \langle \text{ dianzi shang de } X_0 \text{ , } X_0 \text{ on the mat } \rangle$$

$$X \rightarrow \langle X_0 \text{ de mao , a cat on } X_0 \rangle$$

Phrase Extraction



$X \rightarrow \langle$ dianzi shang , the mat $\rangle$

$X \rightarrow \langle$ mao , a cat $\rangle$

$X \rightarrow \langle$ dianzi shang de X_0 , X_0 on the mat $\rangle$

$X \rightarrow \langle$ X_0 de mao , a cat on X_0 $\rangle$

$X \rightarrow \langle$ X_0 de X_1 , X_1 on X_0 $\rangle$

Decoding a Test Sentence

垫子 上 的 狗

dianzi shang de gou

the dog on the mat

$$X \rightarrow \langle \text{ dianzi shang }, \text{ the mat } \rangle$$
$$X \rightarrow \langle \text{ gou }, \text{ the dog } \rangle$$
$$X \rightarrow \langle X_0 \text{ de } X_1 , X_1 \text{ on } X_0 \rangle$$
$$S \rightarrow \langle X_0 , X_0 \rangle$$

$$\mathbf{S} \rightarrow \langle \mathbf{X_0}, \mathbf{X_0} \rangle$$

$$\mathbf{X} \rightarrow \langle \mathbf{X_0} \textbf{ de } \mathbf{X_1}, \mathbf{X_1} \text{ on } \mathbf{X_0} \rangle$$

$$\mathbf{X} \rightarrow \langle \textbf{dianzi shang}, \textbf{the mat} \rangle \qquad \mathbf{X} \rightarrow \langle \textbf{gou}, \textbf{the dog} \rangle$$

dianzi shang de gou

Translation is easy?

Translation Ambiguity

垫子 上 的 猫

dianzi shang de mao

a cat on the mat

$$X \rightarrow \langle X_0 \text{ de } X_1, X_1 \text{ on } X_0 \rangle$$

---

zhongguo de shoudu

capital of China

$$X \rightarrow \langle X_0 \text{ de } X_1, X_1 \text{ of } X_0 \rangle$$

---

wo de mao

my cat

$$X \rightarrow \langle X_0 \text{ de } X_1, X_0\ X_1 \rangle$$

---

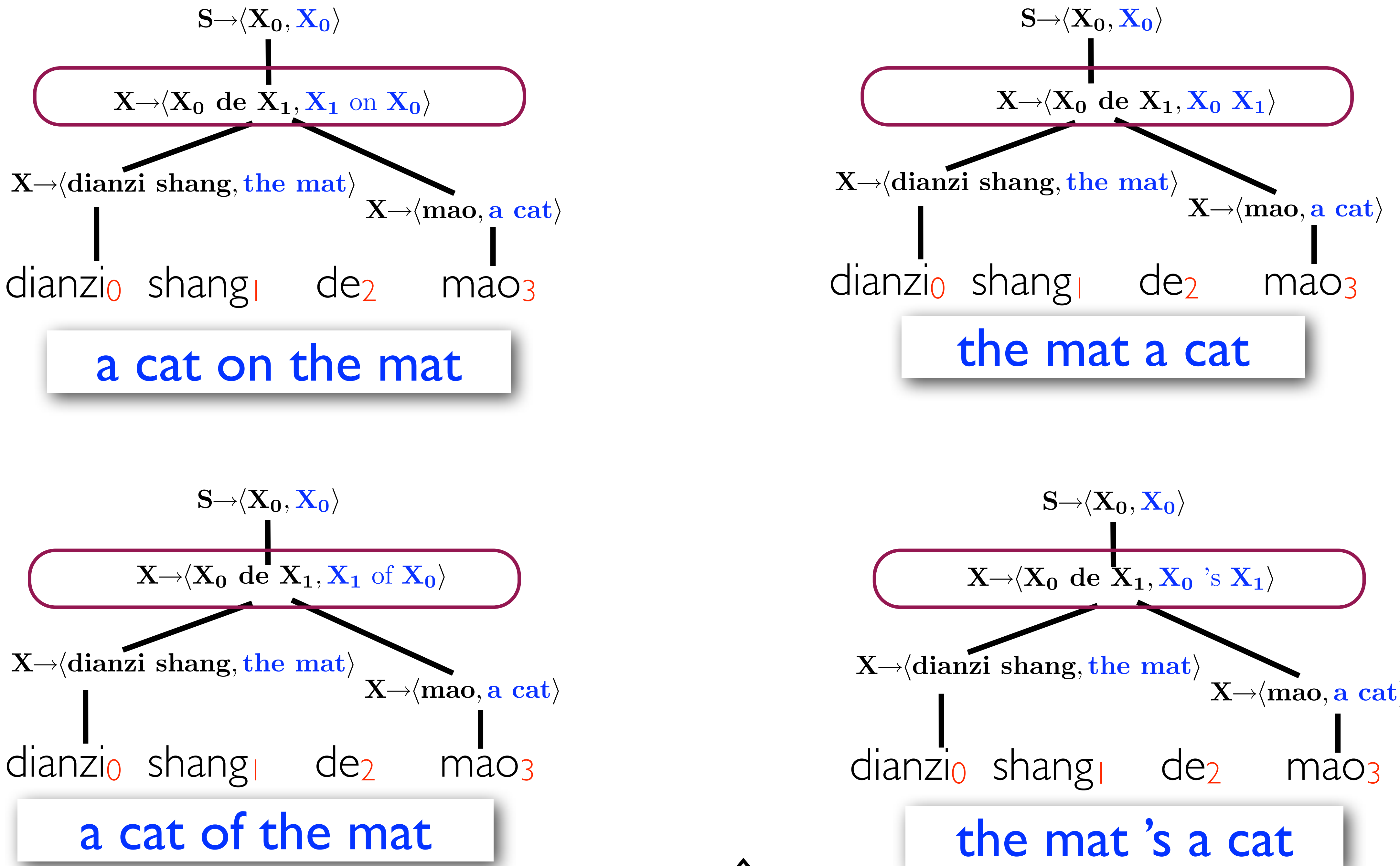
zhifei de mao

zhifei 's cat

$$X \rightarrow \langle X_0 \text{ de } X_1, X_0 \text{ 's } X_1 \rangle$$

Decoder
(e.g. Joshua)

dianzi shang de mao



Decoder
(e.g. Joshua)

dianzi shang de mao

Language Model

a cat on the mat

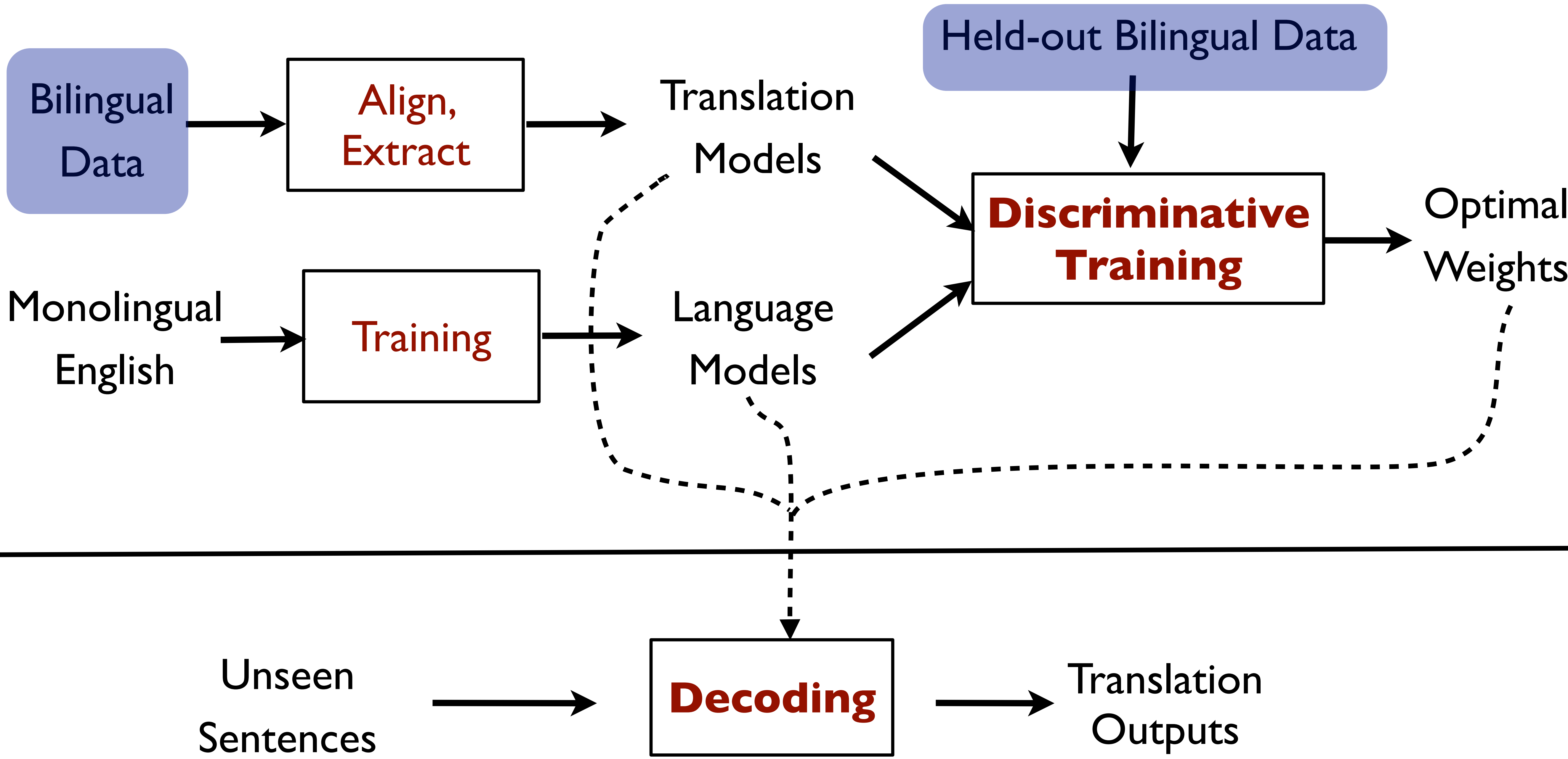
the mat a cat

a cat of the mat

the mat 's a cat

在没看到中文原文情况下，能看出哪个英文句子更靠谱吗？

Statistical Machine Translation Pipeline



Numbers in Real World

- 训练句子对
  - 几千万（一个语言对）
- Phrase Dictionary
  - 亿级条目（一个语言对）
- 语言模型
  - 亿级ngrams（一个语言对）

Outline

- Google Translate
- 机器翻译for dummy
- 机器翻译基础理论和算法
  - 机器学习
  - 数据结构，模型，算法
- 工业界机器翻译系统实战

机器学习：分类器

【机器学习实战】机器学习是人工智能研究领域中一个极其重要的研究方向，在现今的大数据时代背景下，捕获数据并从中萃取有价值的信息或模式，成为...http://t.cn/zHNXceF。想看更多"机器学习"的资讯，猛戳→http://t.cn/zjNCS5w

？

微博

体育新闻

政治新闻

军事新闻

- 分类（Classification）
  - 输入：特征
  - 输出：类别
  - Naive Bayes，最大熵，SVM，神经网络等

Structured Prediction(SP): 结构化预测

- 词性标注是一个典型的SP问题

I	like	machine-learning
名词	动词	名词

I: 名词

like: 介词，动词

machine-learning: 名词，动词

Structured Prediction as Classification

- SP可以看成是**特殊**的分类问题
  - 类别的个数随着输入的长度而指数级增长

类别	I	like	machine-learning
1	名词	动词	名词
2	名词	动词	动词
3	名词	介词	名词
4	名词	介词	动词

Structured Prediction as Classification

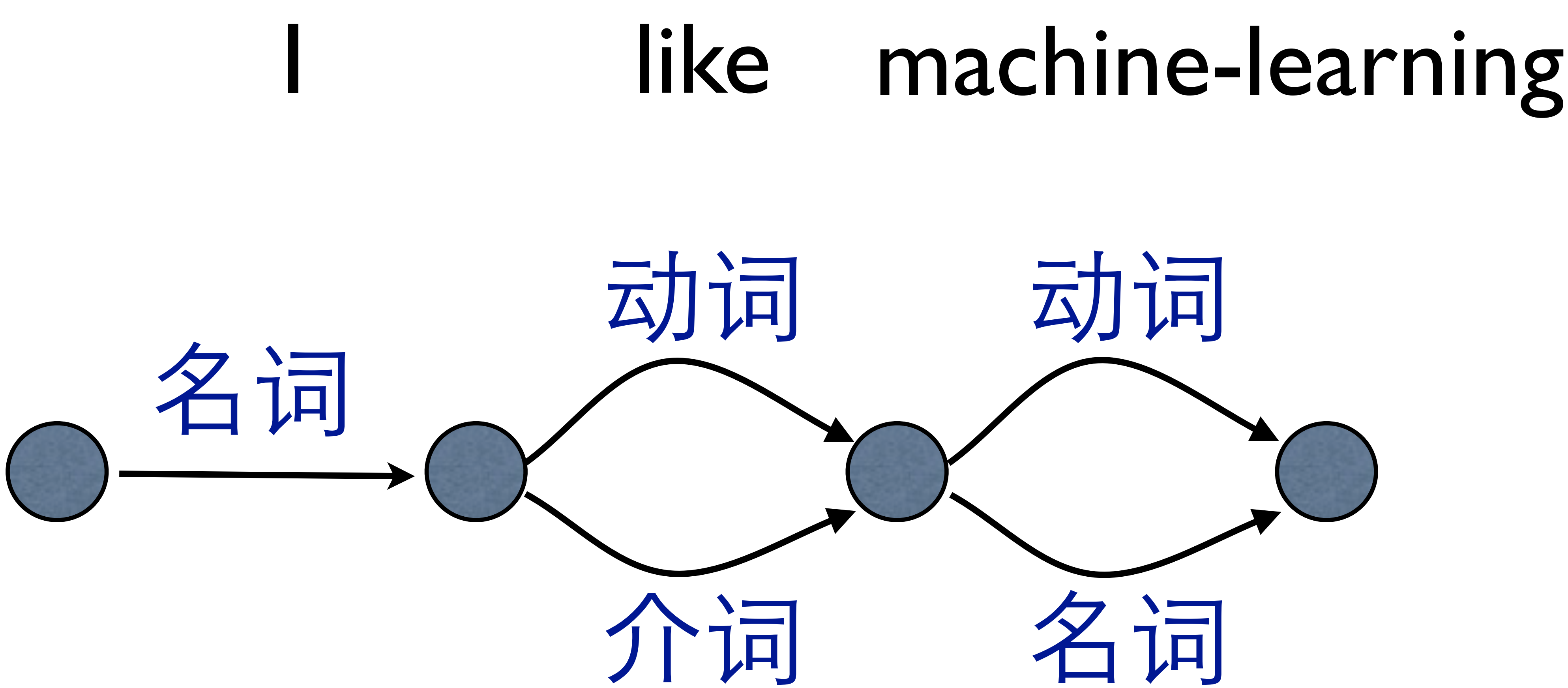
- SP可以看成是**特殊**的分类问题
  - 类别的个数随着输入的长度而指数级增长
  - 类别内部有联系

I like machine-learning

名词 动词 名词

Structured Prediction as Classification

- SP可以看成是**特殊**的分类问题
    - 类别的个数随着输入的长度而指数级增长
    - 类别内部有联系
    - 类别之间有联系



Structured Prediction as Classification

- SP可以看成是**特殊**的分类问题
  - 类别的个数随着输入的长度而指数级增长
  - 类别内部有联系
  - 类别之间有联系

这些特殊性使得SP的难度增大,

尤其是在算法上!

许多在分类上特别简单的算法（如解码）在SP上变得很复杂

Structured Prediction 问题

任务	输入	类别
中文分词	句子	词序列
词性标注	句子	词性序列
语法解析	句子	语法树
机器翻译	英文句子	中文句子
语音识别	声音	句子
手写识别	笔话	句子
光学识别	图片	句子

Outline

- Google Translate
- 机器翻译for dummy
- 机器翻译基础理论和算法
  - 机器学习
  - 数据结构，模型，算法
- 工业界机器翻译系统实战

$S \rightarrow \langle X_0, X_0 \rangle$

$X \rightarrow \langle X_0 \text{ de } X_1, X_1 \text{ on } X_0 \rangle$

$X \rightarrow \langle \text{dianzi shang}, \text{the mat} \rangle$ $X \rightarrow \langle \text{mao}, \text{a cat} \rangle$

dianzi$_0$ shang$_1$ de$_2$ mao$_3$

a cat on the mat

$S \rightarrow \langle X_0, X_0 \rangle$

$X \rightarrow \langle X_0 \text{ de } X_1, X_0\ X_1 \rangle$

$X \rightarrow \langle \text{dianzi shang}, \text{the mat} \rangle$ $X \rightarrow \langle \text{mao}, \text{a cat} \rangle$

dianzi$_0$ shang$_1$ de$_2$ mao$_3$

the mat a cat

$S \rightarrow \langle X_0, X_0 \rangle$

$X \rightarrow \langle X_0 \text{ de } X_1, X_1 \text{ of } X_0 \rangle$

$X \rightarrow \langle \text{dianzi shang}, \text{the mat} \rangle$ $X \rightarrow \langle \text{mao}, \text{a cat} \rangle$

dianzi$_0$ shang$_1$ de$_2$ mao$_3$

a cat of the mat

$S \rightarrow \langle X_0, X_0 \rangle$

$X \rightarrow \langle X_0 \text{ de } X_1, X_0 \text{ 's } X_1 \rangle$

$X \rightarrow \langle \text{dianzi shang}, \text{the mat} \rangle$ $X \rightarrow \langle \text{mao}, \text{a cat} \rangle$

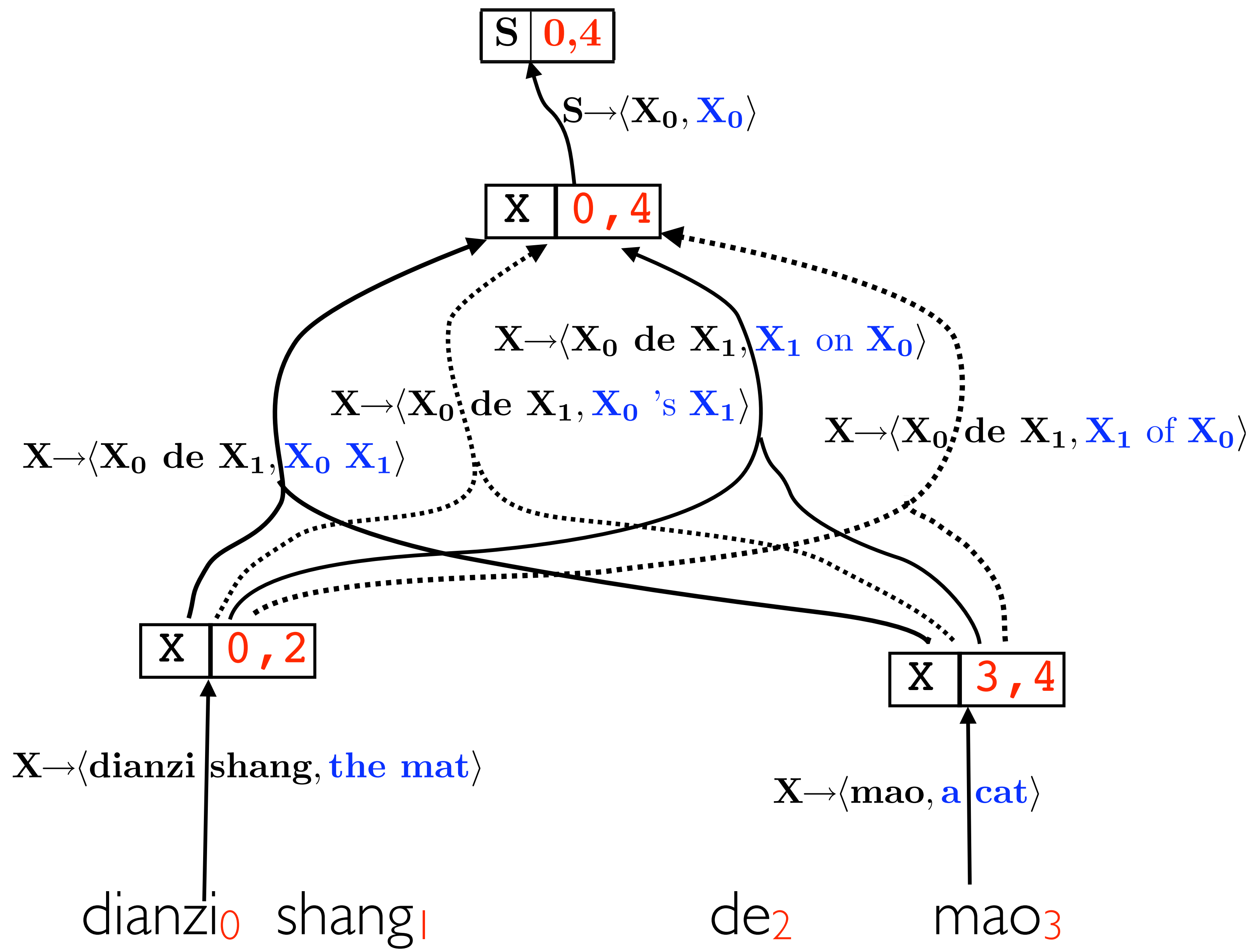
dianzi$_0$ shang$_1$ de$_2$ mao$_3$

the mat 's a cat

Decoder
(e.g. Joshua)

dianzi shang de mao

hypergraph

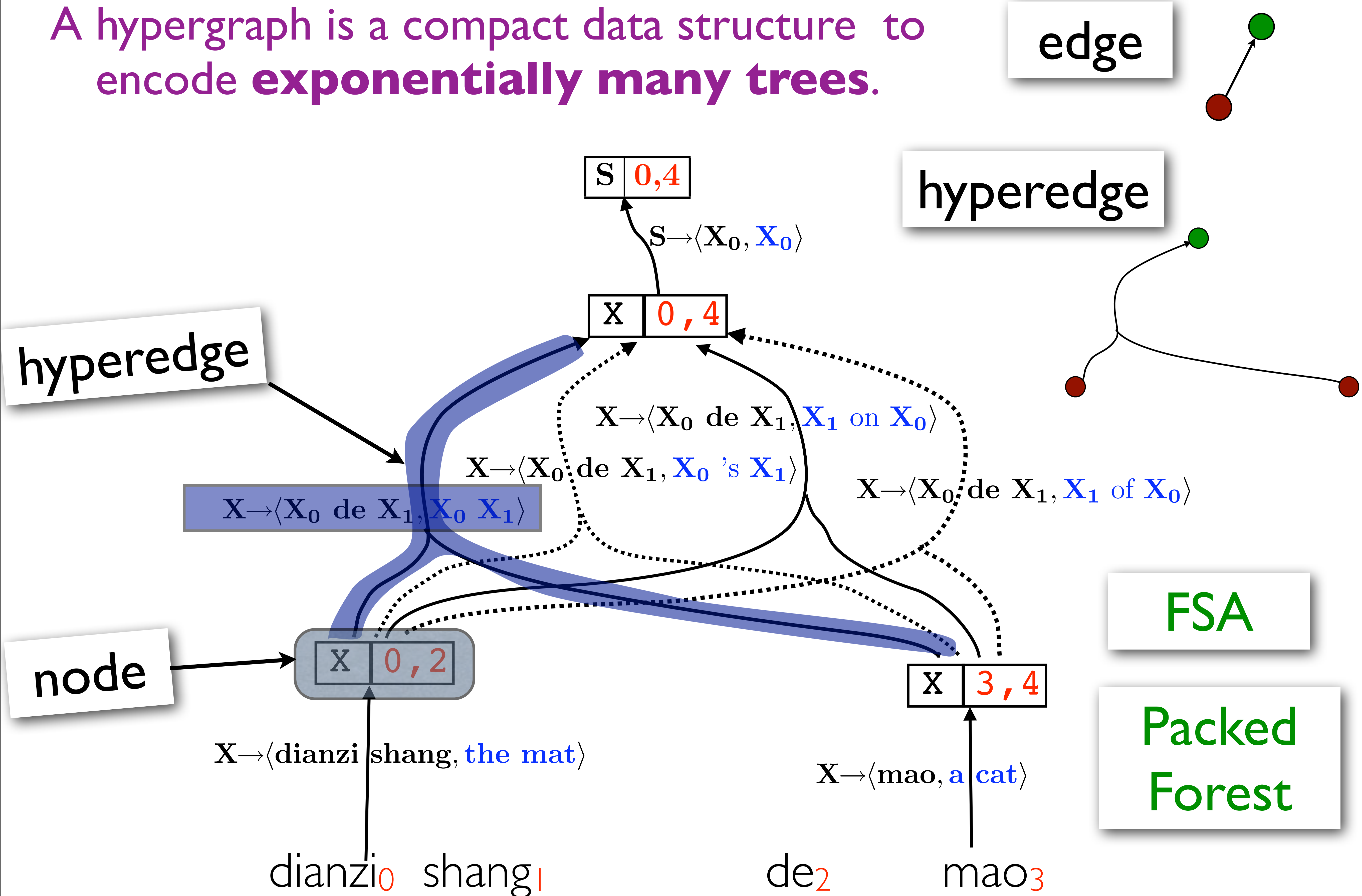


Decoder

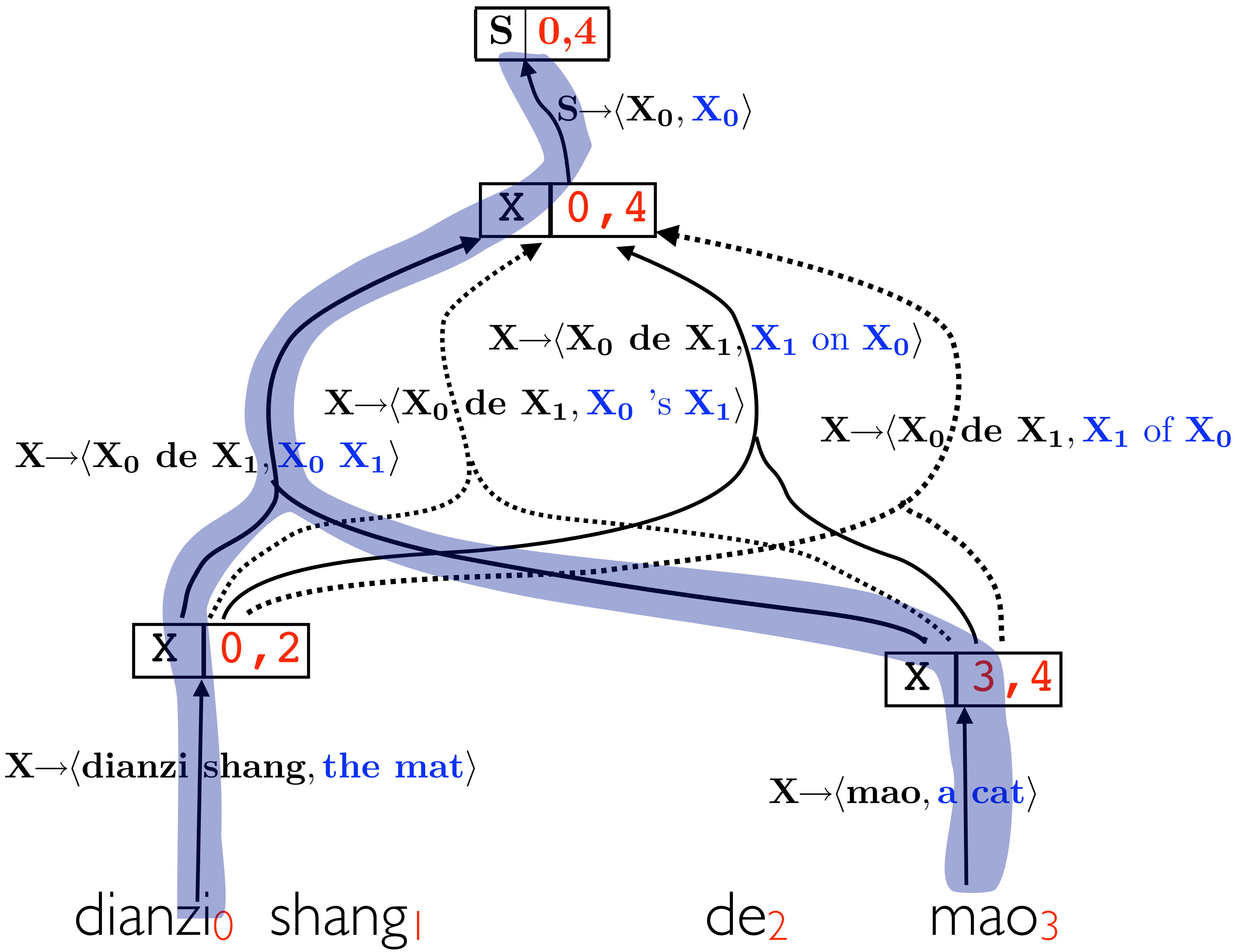
(e.g. Joshua)

dianzi shang de mao

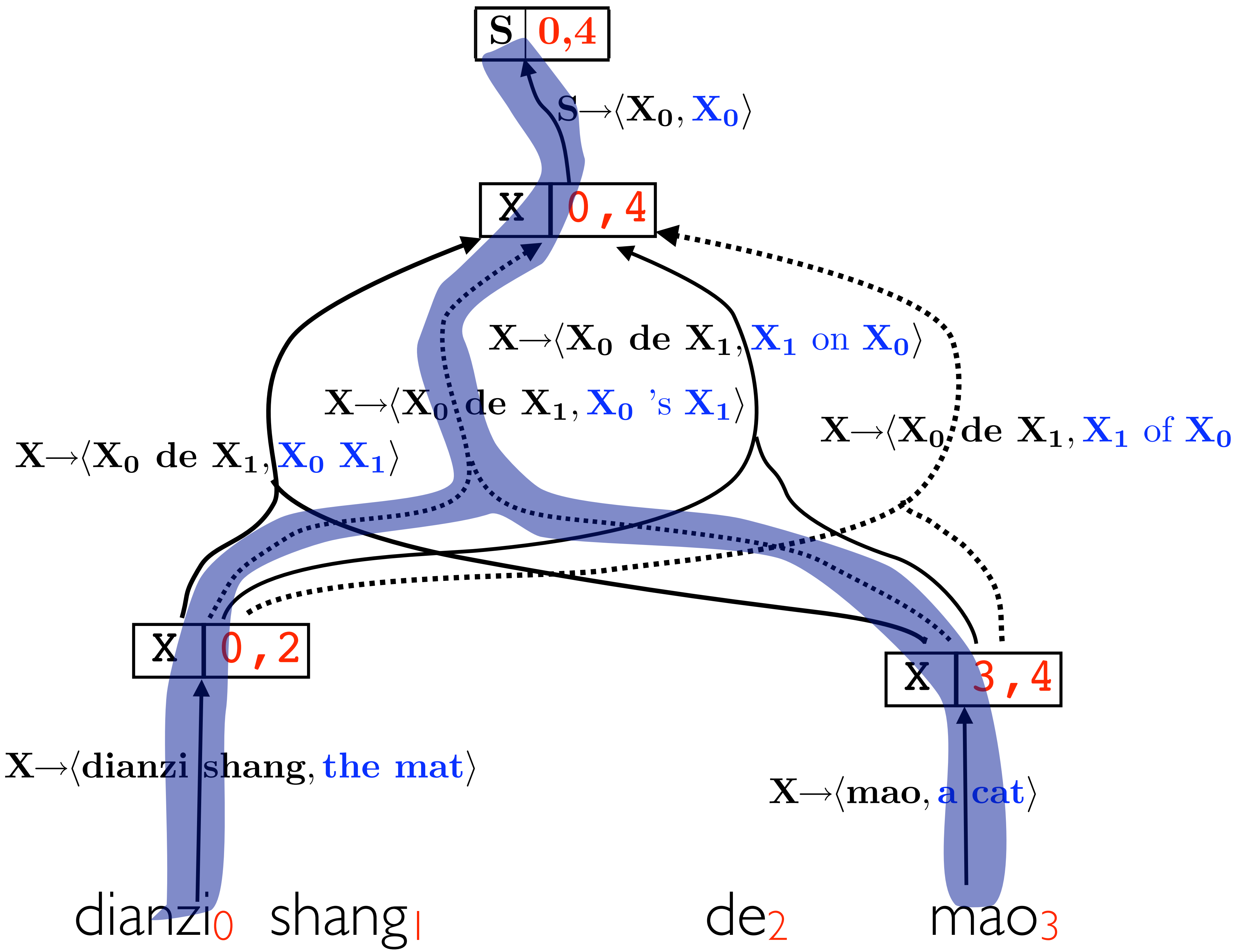
A hypergraph is a compact data structure to encode **exponentially many trees**.



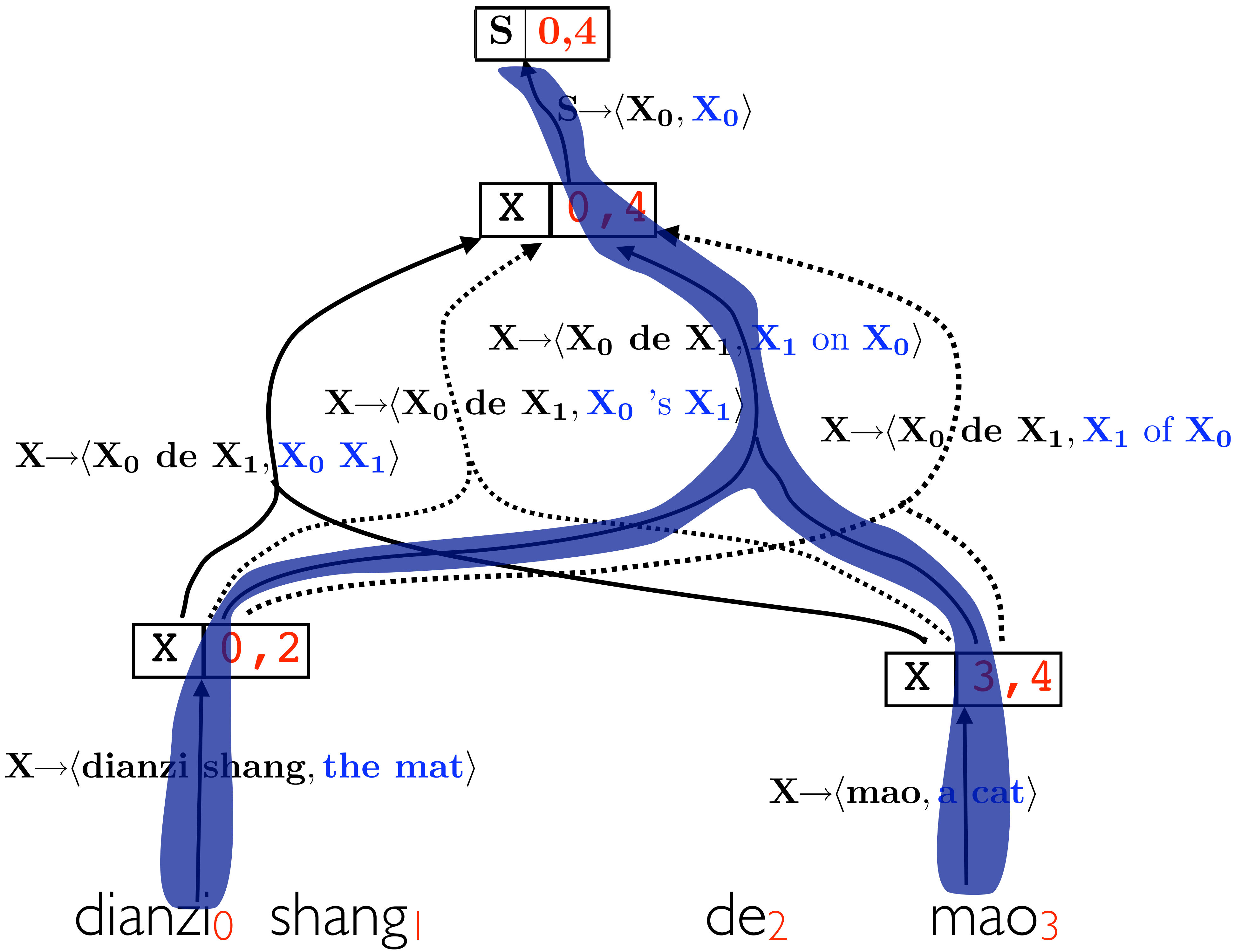
A hypergraph is a compact data structure to encode **exponentially many trees**.



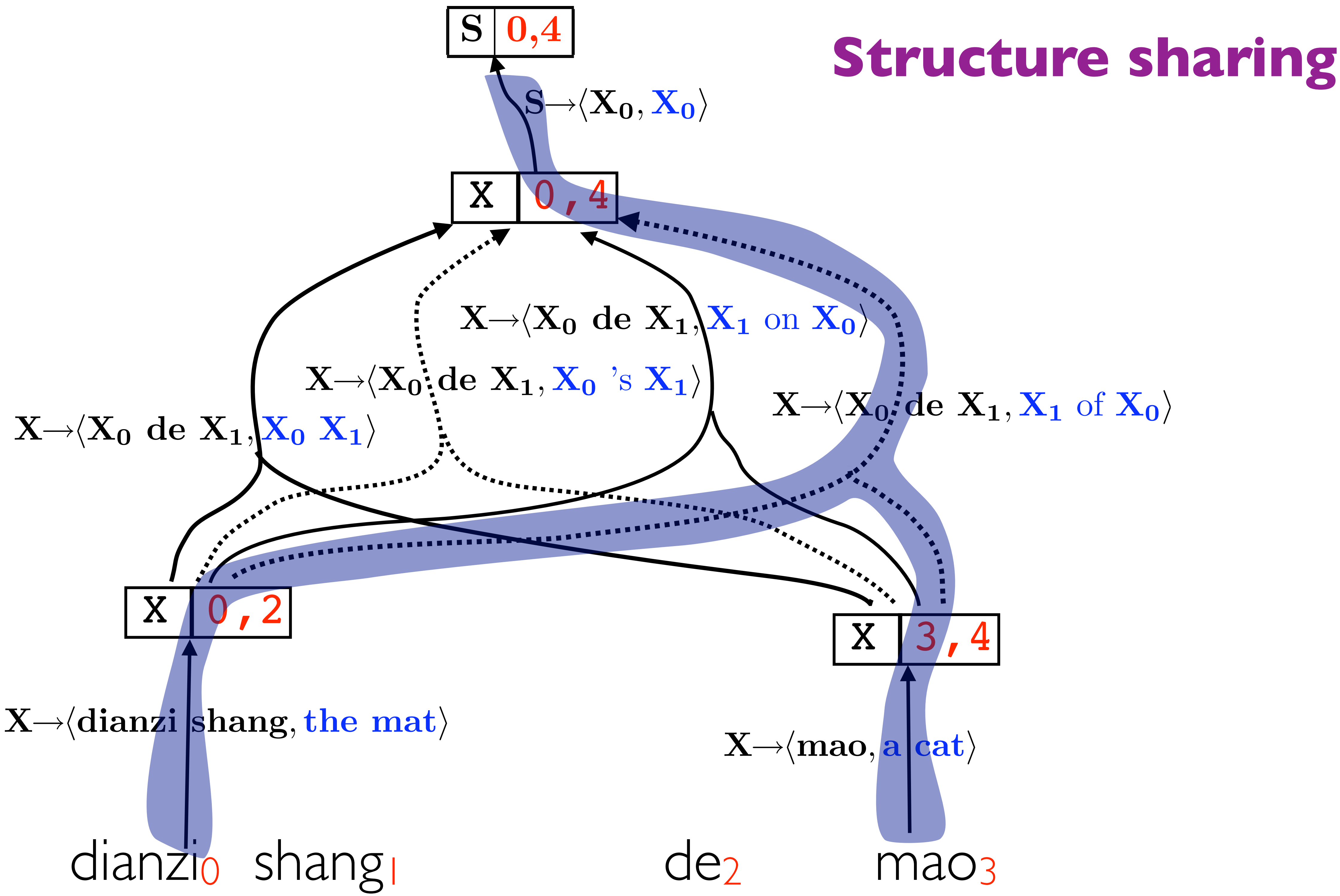
A hypergraph is a compact data structure to encode **exponentially many trees**.



A hypergraph is a compact data structure to encode **exponentially many trees**.



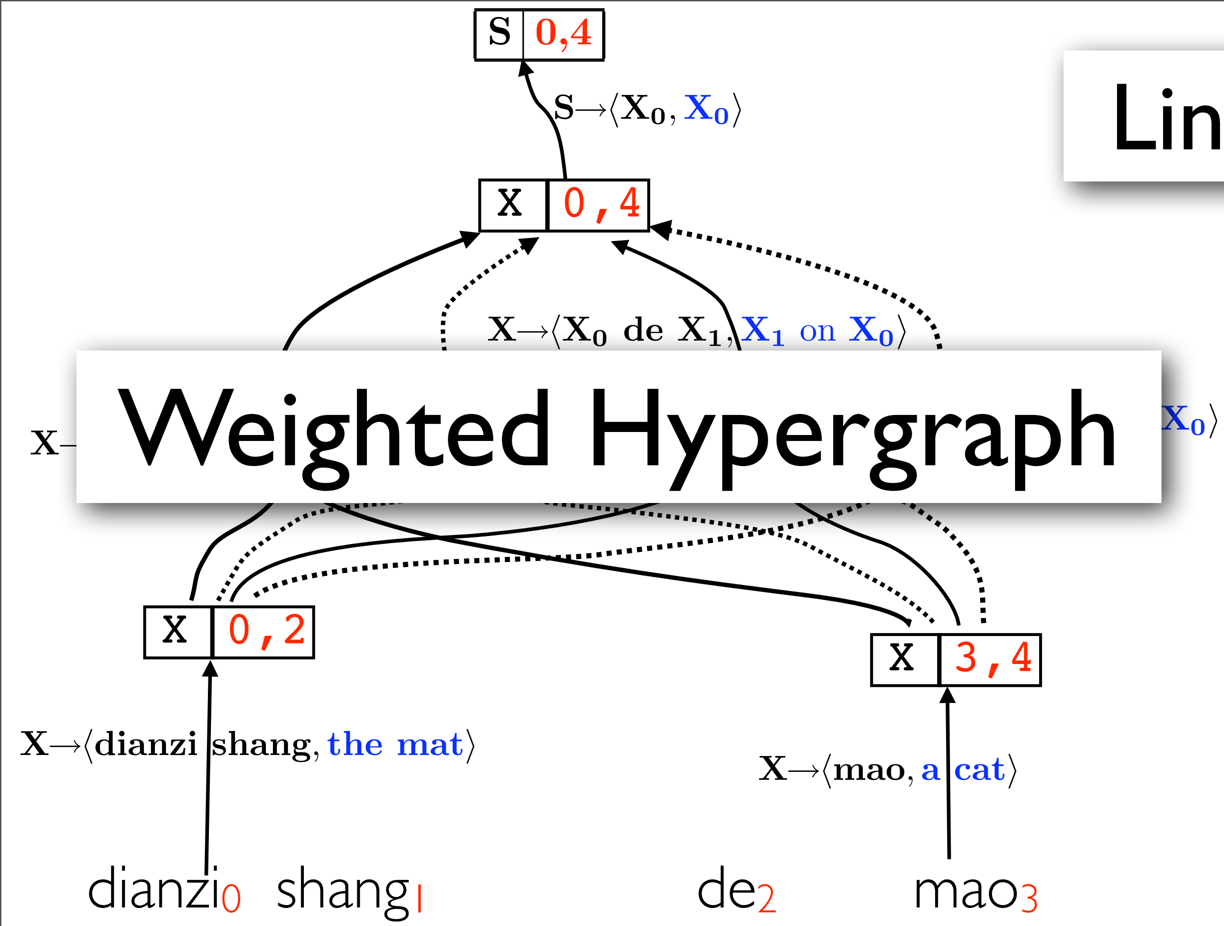
A hypergraph is a compact data structure to encode **exponentially many trees.**



Why Hypergraphs?

- General compact data structure
  - special cases include
    - finite state machine (e.g., lattice)
    - and/or graph
    - packed forest
  - can be used for speech, parsing, tree-based MT systems, and many more

Weighted Hypergraph



Linear model:

$$p(d \mid x) = \theta \cdot \Phi(d, x)$$

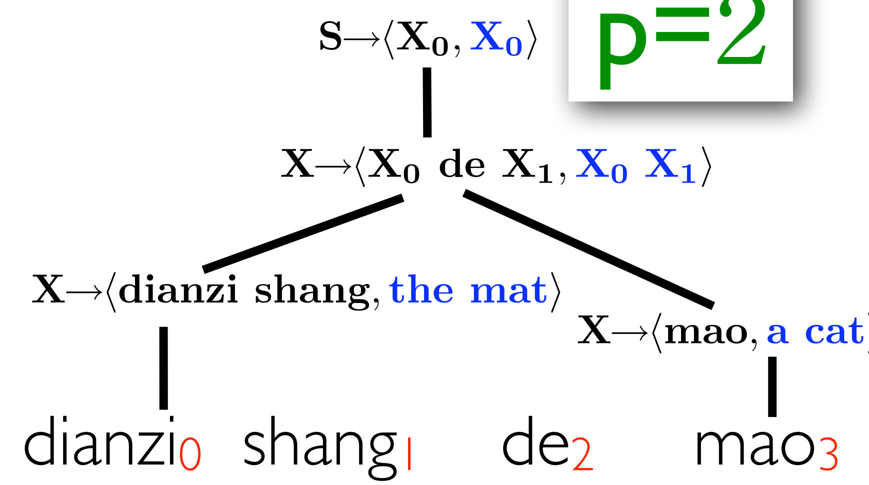
derivation

foreign input

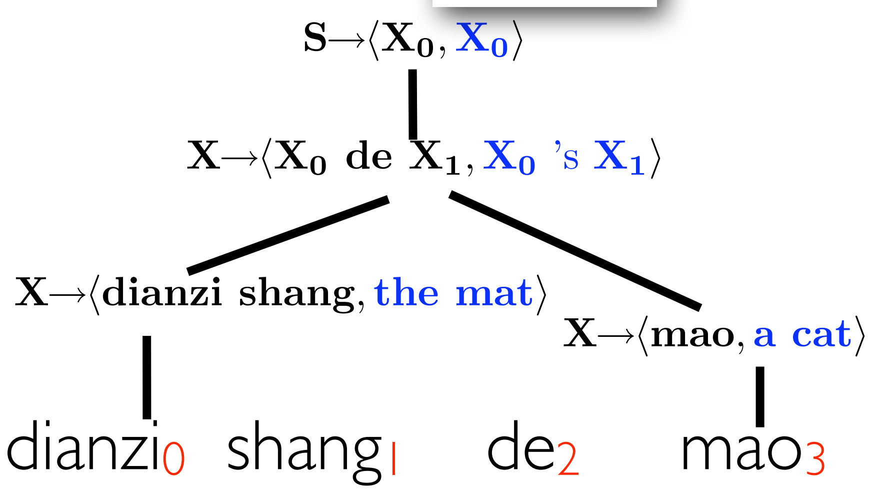
weights

features

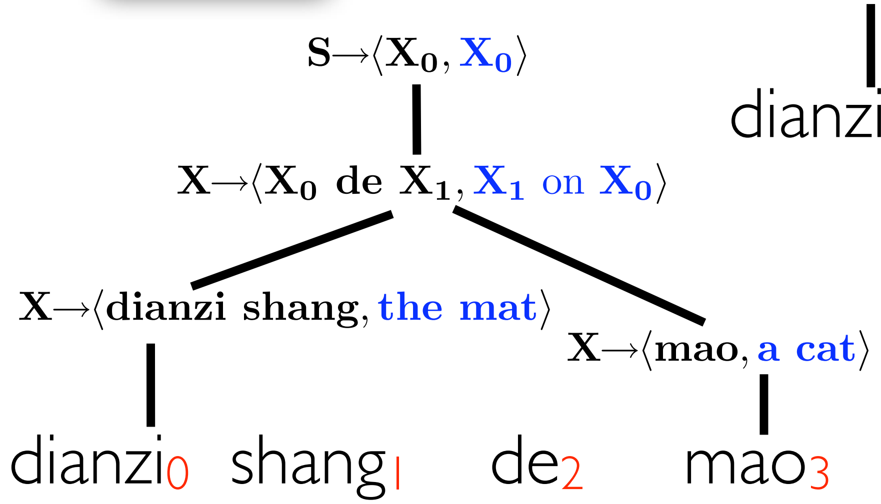
p=2



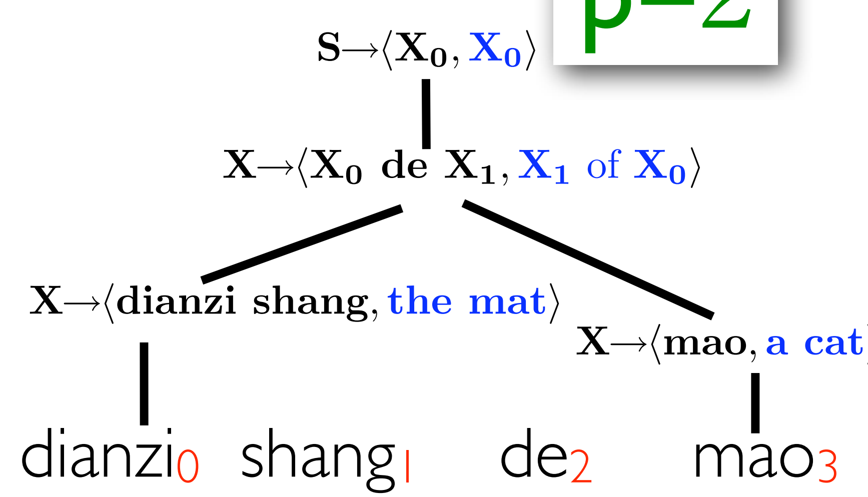
p=3



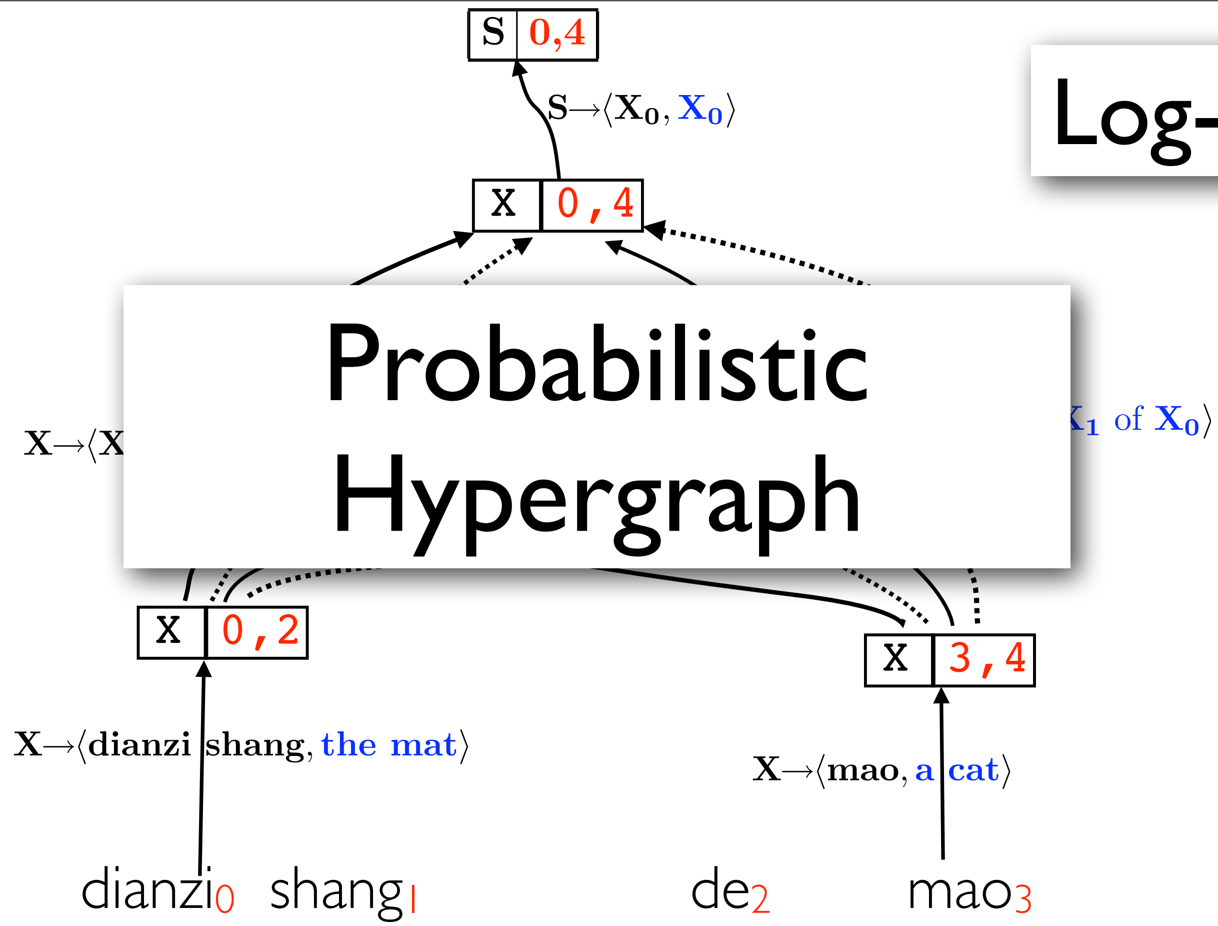
p=1



p=2



Probabilistic Hypergraph

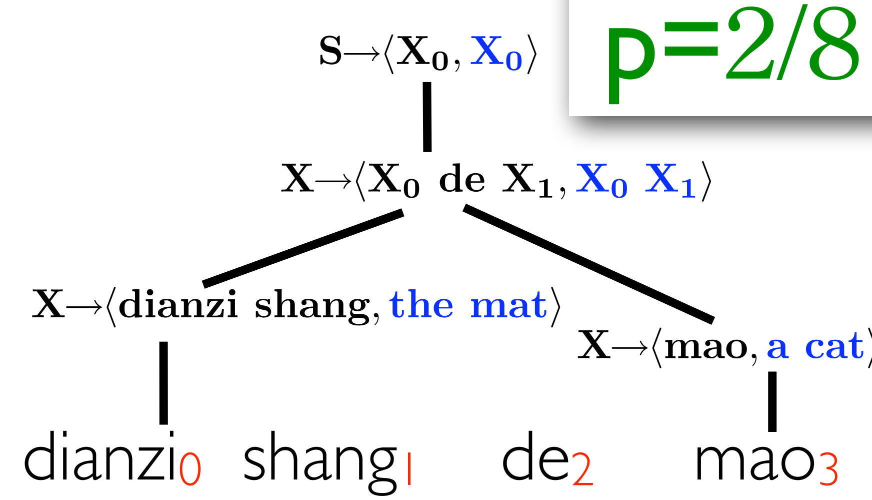


Log-linear model:

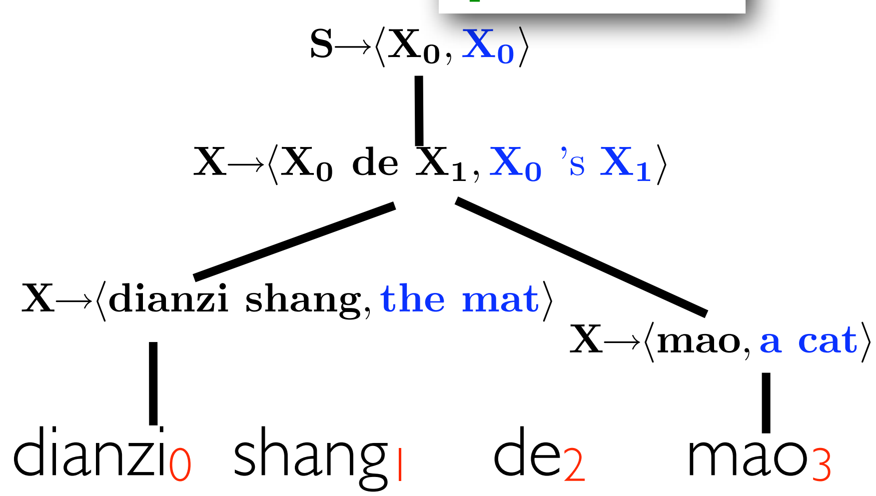
$$p(d \mid x) = \frac{e^{\theta \cdot \Phi(d,x)}}{Z(x)}$$

Z=2+1+3+2=8

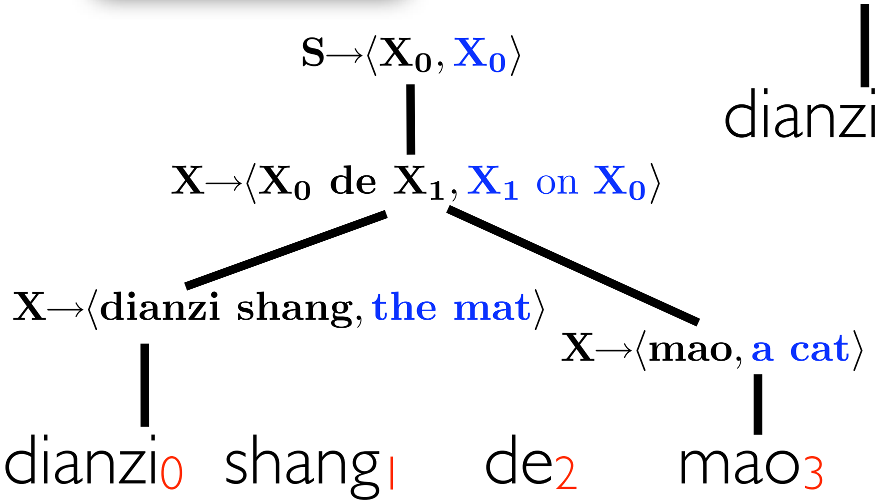
p=2/8



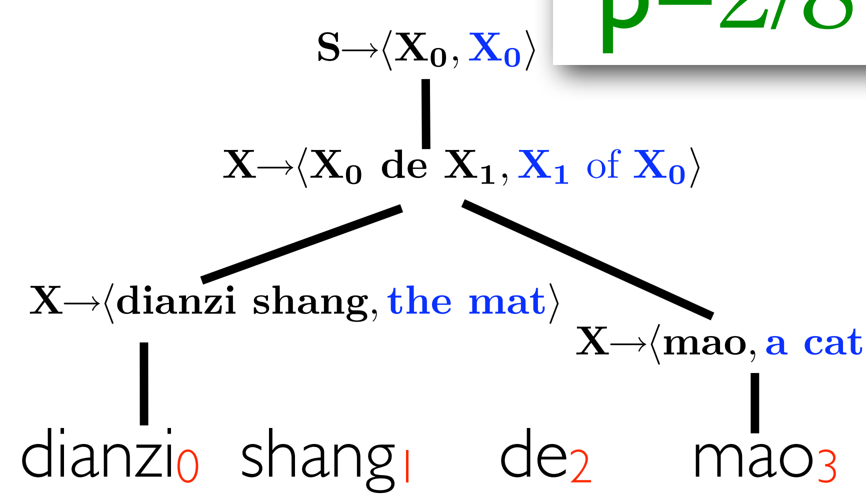
p=3/8



p=1/8

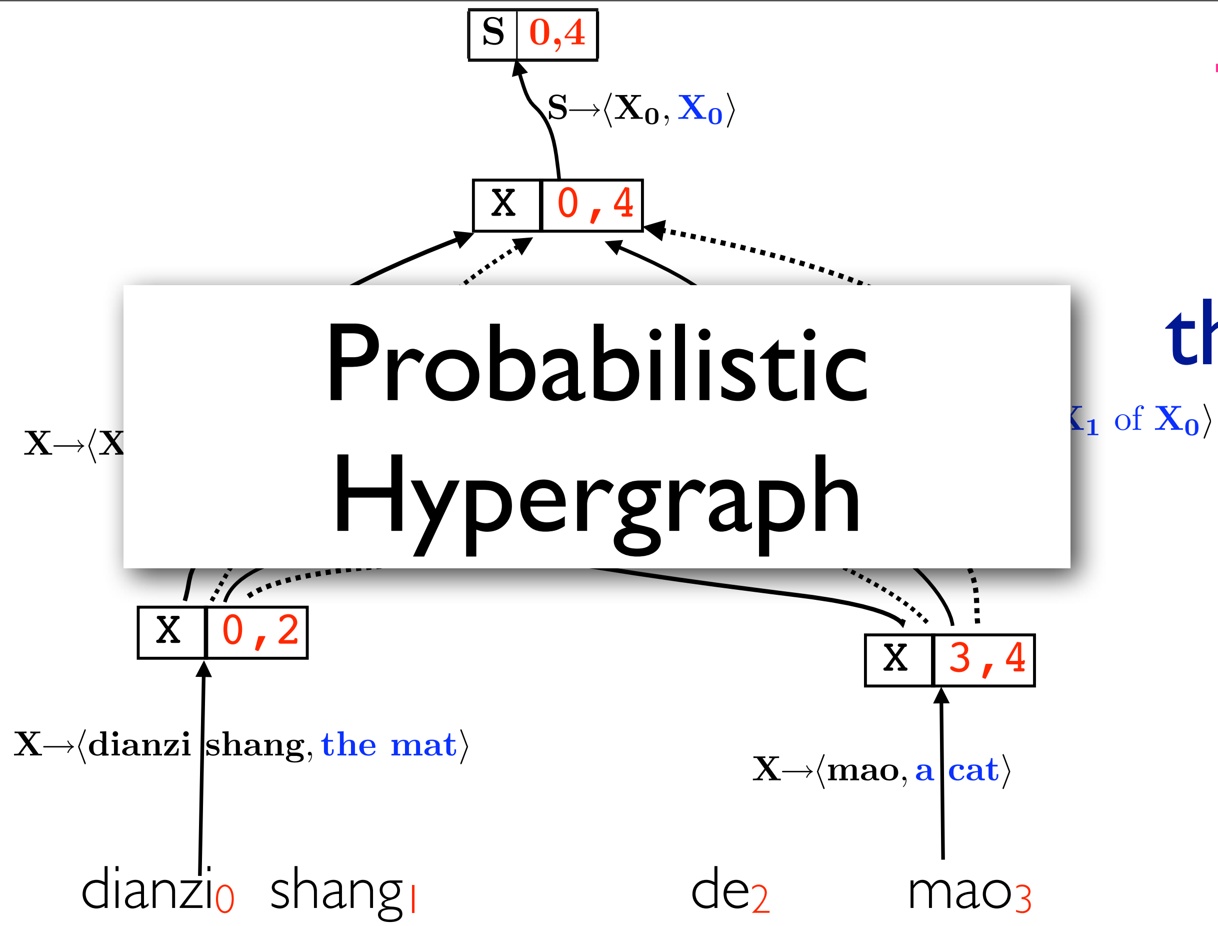


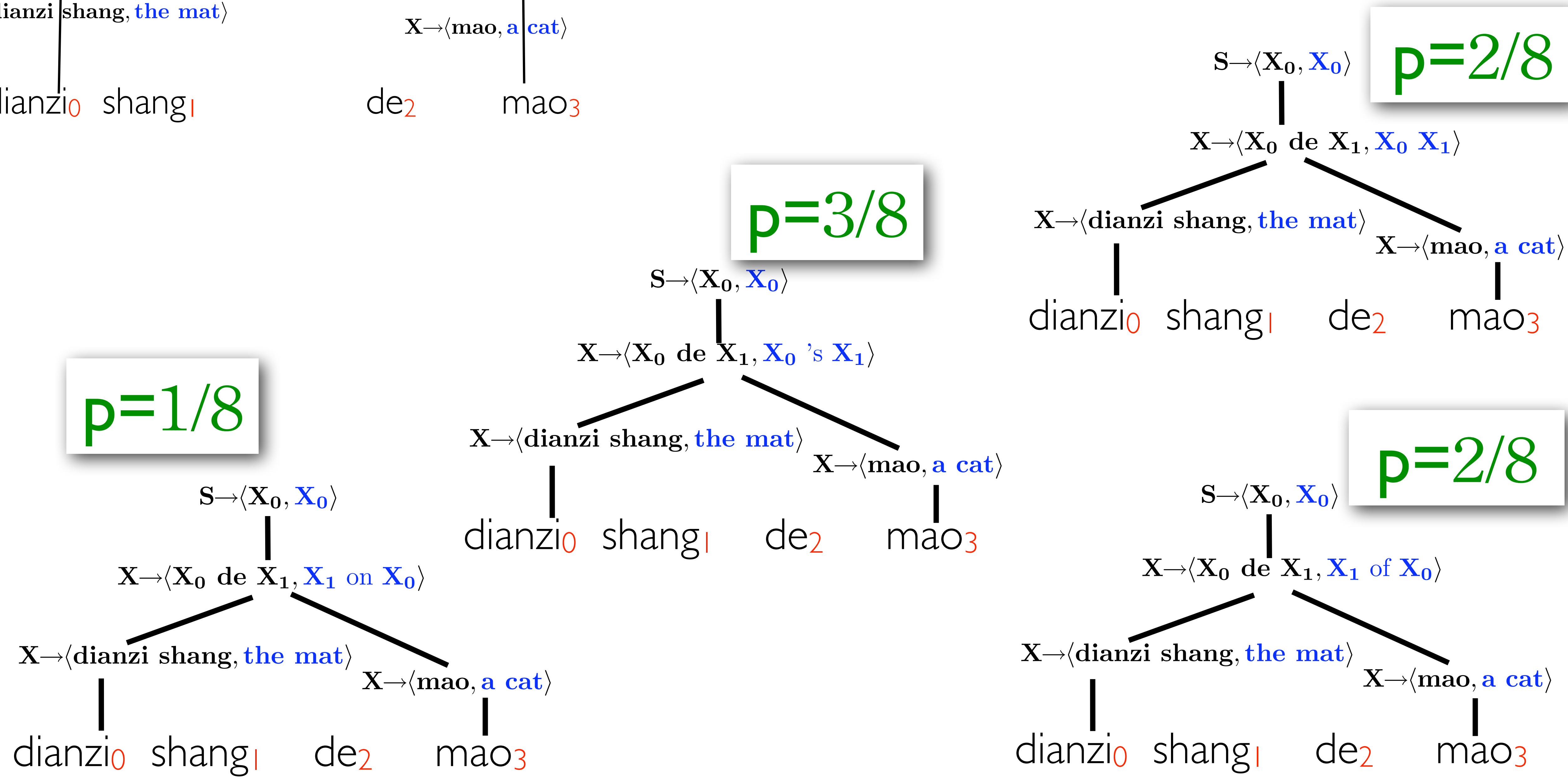
p=2/8

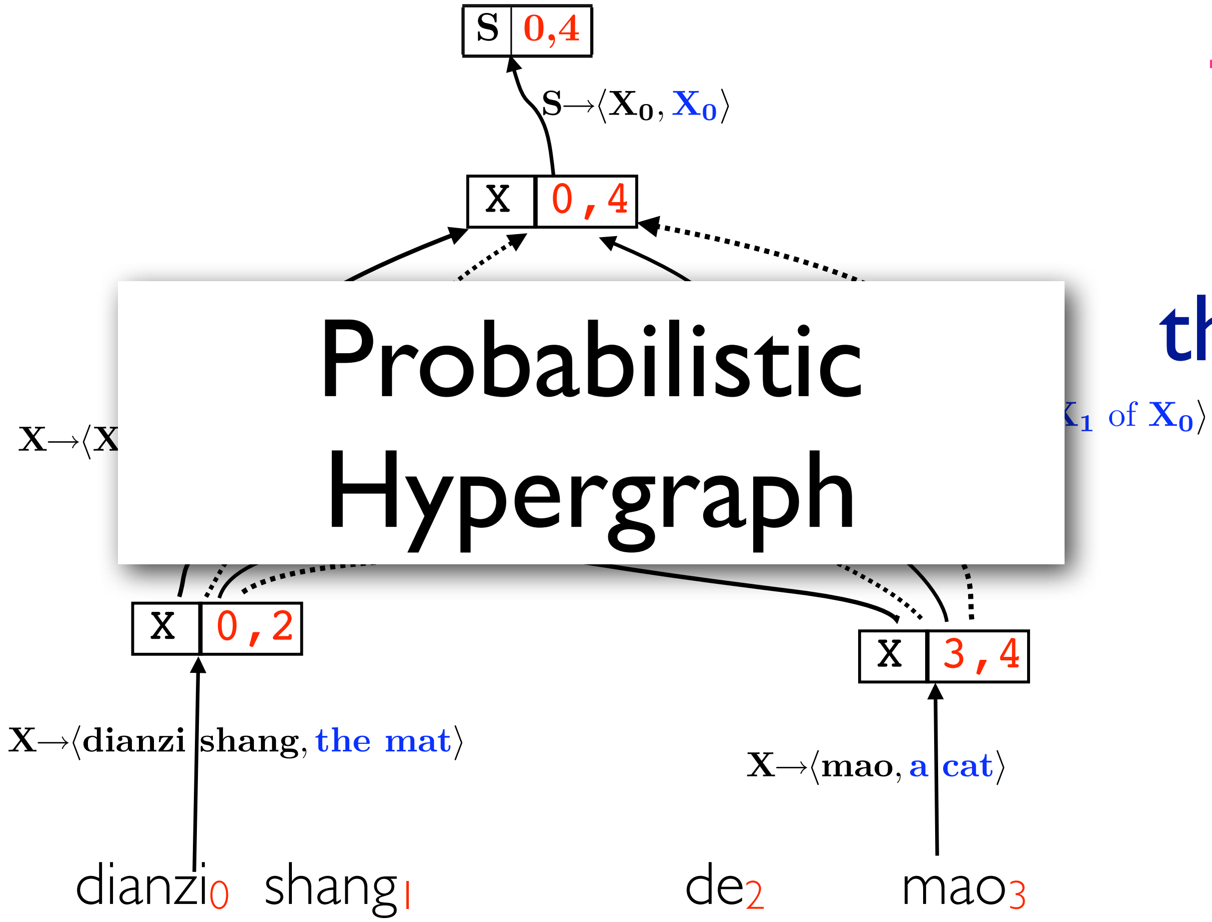


The hypergraph defines a probability distribution over **trees**!

the distribution is parameterized by Θ







The hypergraph defines a probability distribution over **trees**!

the distribution is parameterized by Θ

training (e.g., mert)	**decoding** (e.g., mbr)
atomic inference operations (e.g., finding one-best, k-best or expectation, inference can be *exact* or *approximate*)	

Which translation do we present to a user? Decoding

How do we set the parameters Θ? Training

What atomic operations do we need to perform? Atomic Inference

Why are the problems difficult?

- brute-force will be too slow as there are exponentially many trees, so require sophisticated dynamic programs
- sometimes intractable, require approximations

Inference, Training and Decoding on Hypergraphs

- Atomic Inference Algorithms
  - finding one-best derivations

Graph	Topological	Best-first		
		no heuristic	with heuristic	with hierarchy
FSA	Viterbi	Dijkstra	A^*	HA^*
Hypergraph	CYK	Knuth	Klein and Manning	Generalized A^*

  - finding k-best derivations
  - computing expectations (e.g., of features)

- Training
  - Perceptron
  - Conditional random field (CRF)
  - Minimum error rate training (MERT)
  - Minimum risk
  - MIRA

- Decoding
  - Viterbi
  - Maximum a posterior (MAP)
  - Minimum Bayes risk (MBR)

原理和算法的更多细节

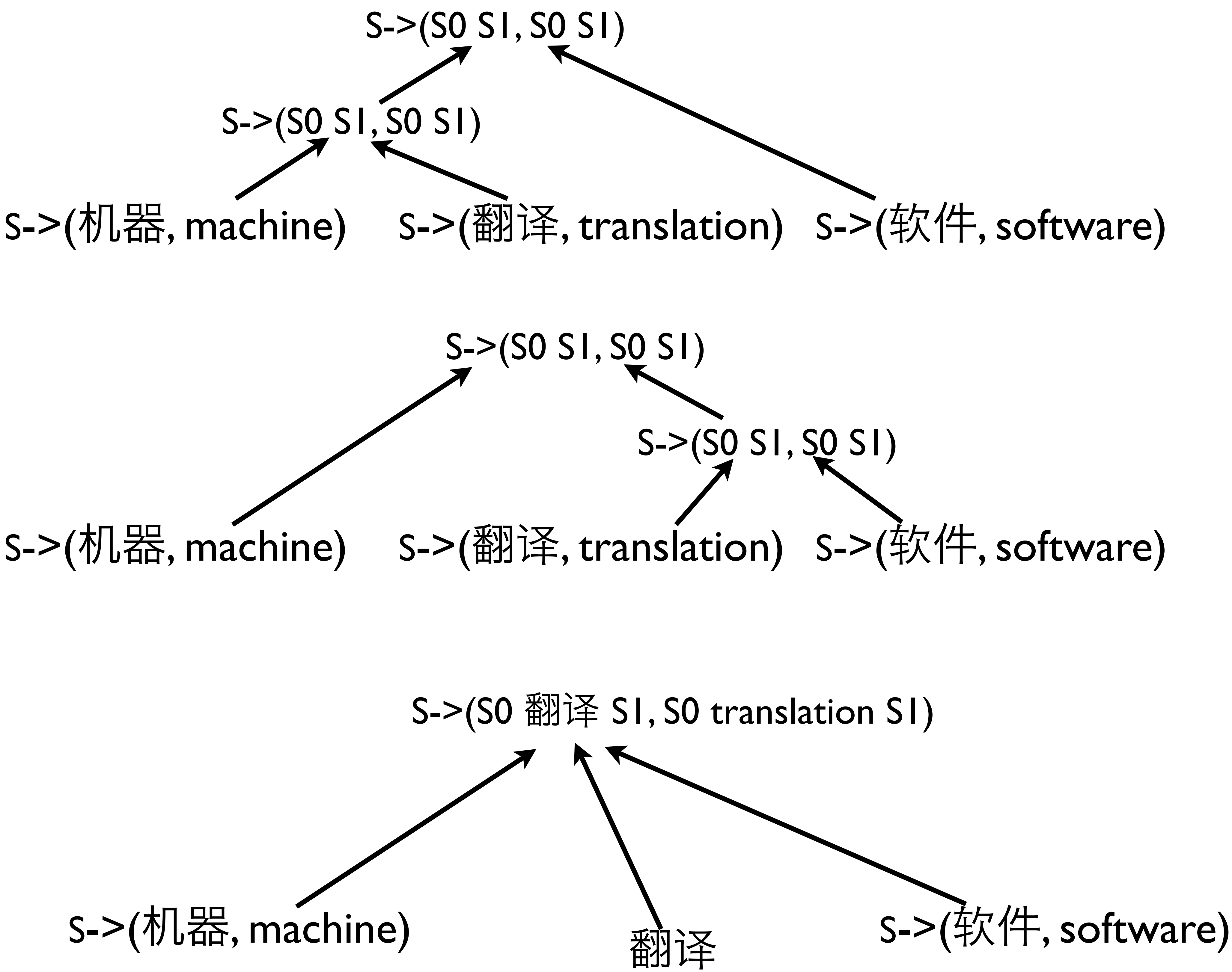
CCF互联网大数据与机器学习讲习班

Structured Prediction在自然语言处理中的应用

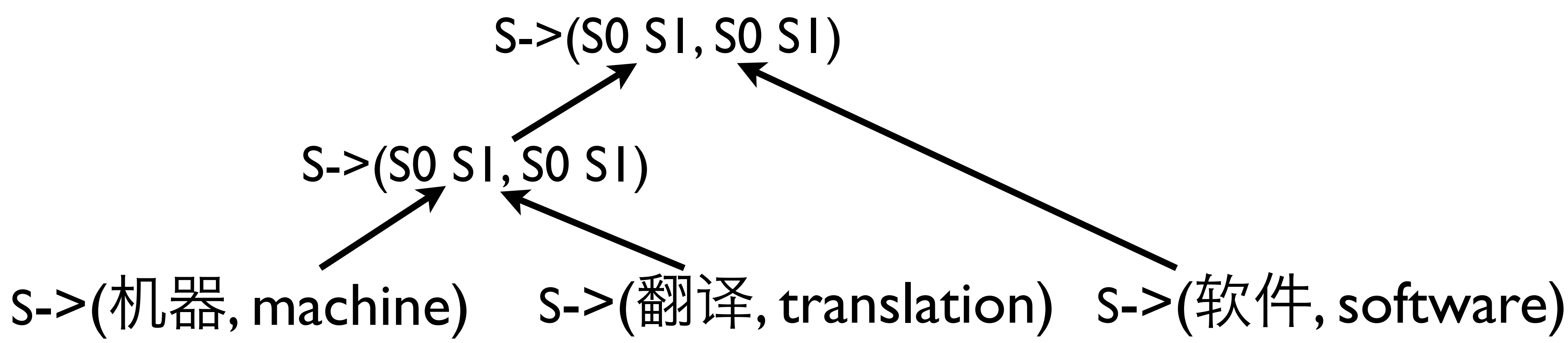
http://mobvoi-resource.oss.aliyuncs.com/ccf2013_noannimation_.pdf

为什么机器翻译算法很复杂？

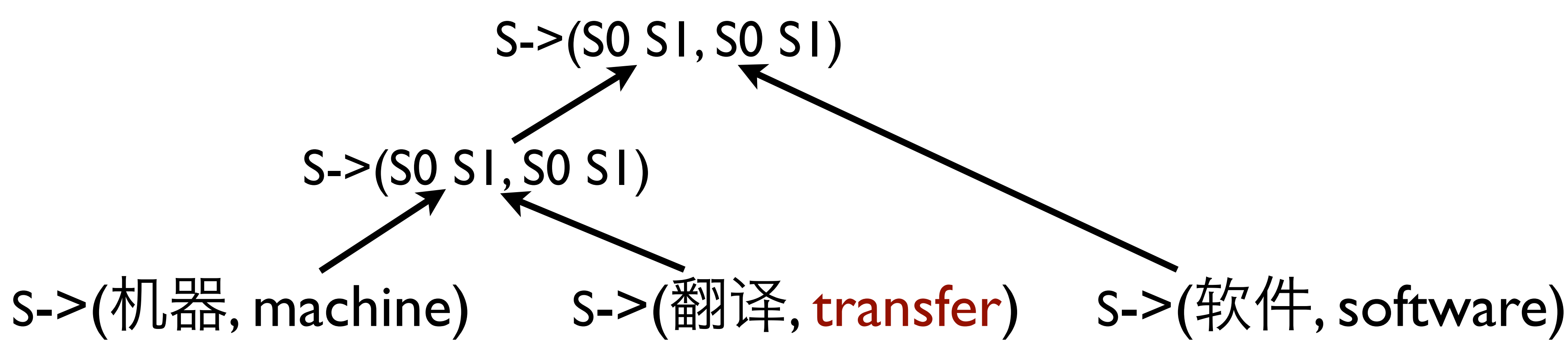
解码器的复杂性：分割的歧义



解码器的复杂性：翻译的歧义



machine translation software



machine transfer software

解码器的复杂性：排序的歧义

S->(S0 翻译 S1, S0 translation S1)

S->(机器, machine)　　翻译　　S->(软件, software)

machine translation software

S->(S0 翻译 S1, S1 translation S0)

S->(机器, machine)　　翻译　　S->(软件, software)

software translation machine

解码器的复杂性

- 给定一个句子，解码过程种要考虑各种歧义
  - 分割的歧义
  - 翻译的歧义　　　所有的歧义都可压缩在超图里！
  - 排序的歧义
- 每一种歧义都会导致组合爆炸
- 穷举不可能，所以需要非常复杂的动态规划

Outline

- Google Translate
- 机器翻译for dummy
- 机器翻译基础理论和算法
  - 机器学习
  - 数据结构，模型，算法
- 工业界机器翻译系统实战

算法，数据，工具

- 一个成功的工业界翻译系统包含

核心算法

数据

支撑工具

工具的重要性

- 一切都应该工具化，自动化
- 好架构和工具会大大加速迭代
  （谷歌翻译系统可以在一天之内重新训练所有语言，训练结果直接以Email的形式发给训练者）

为什么是Google?

- IBM Research是许多NLP核心算法的开创者

- Microsoft Research拥有豪华的NLP科研团队

- 但Google第一个把翻译做成大规模互联网产品，
  为什么？

为什么是Google?

- 为何Google第一个把翻译做成大规模互联网产品？
  - 团队基因：科学家＋工程师
  - 整个谷歌大环境：实用至上
  - 大数据：中英系统用几千万对句子
  - 云架构：GFS, Map-reduce, Big-table
- 很多类似的故事正在上演

语音识别	图像识别	句法解析
深度学习	知识图谱	对话搜索

打造你自己的Google Translate?

后端系统：10人

翻译模型（1）

数据处理（2）

语言模型（1）

解码器（1）

工具和架构（3）

区分训练（1）

NLP基础模块（1）

产品（16人）

推广运营（2）

产品经理（2）

前端开发（2）

后端（10）

创业公司的捷径？

开源软件

整套： Moses Joshua CDec

NLP工具： Stanford NLP Berkeley Parser

语言模型： SRILM

云计算： Hadoop

机器学习： CRF++ libSVM

Outline

- Google Translate
- 机器翻译for dummy
- 机器翻译基础理论和算法
  - 机器学习
  - 数据结构，模型，算法
- 工业界机器翻译系统实战

把机器翻译换成NLP!!

- 微信公号：出门问问

问

听得懂，查得到
一个能理解你的生活助手

微信扫一扫二维码 or
微信搜索“出门问问”

打造Google Now的中文版

- 招聘： www.mobvoi.com

打造中国的Google

Thank you!

XieXie!

谢谢!