

Natural Language Processing

Lecture 1: Course Overview and Introduction.

1/24/2020

N L P

COMS W4705
Yassine Benajiba

The 4705 Team

- **Instructor:** Yassine Benajiba <yb2235@columbia.edu>
Office Hours: Mon 2:00pm-3:30pm
(starting next week)
Room 7LW1A
- **Assistants:**
 - Let's look at Courseworks
- IA office hours / recitations start next week.
Time/Location TBA by email.

Lectures & Recitation Sessions

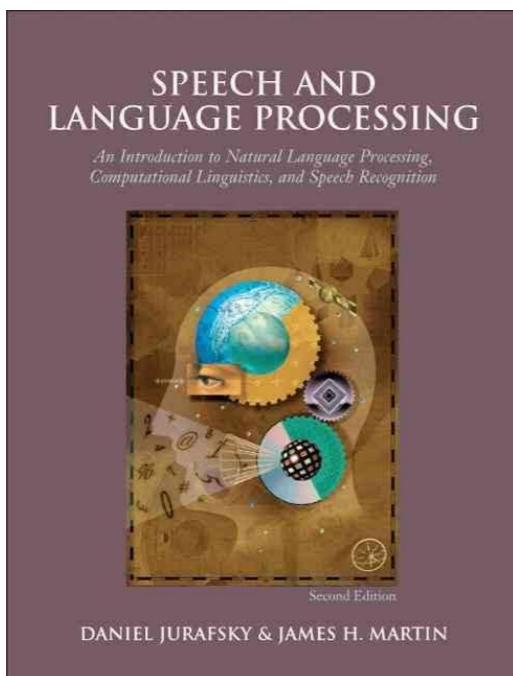
- **Lectures:** Fri 1:10pm-3:40pm
Chandler 402
- **Recitation Sessions:**
 - Optional recitation sessions, led by the IAs
(schedule TBA)

Course Resources

- **Courseworks 2 (a.k.a Canvas):**
 - All course materials: Lecture notes, code, announcements, assignments, reading materials
 - Homework submission, grade book.
- **Piazza** used for Q & A.
Do not email the instructor or IAs with questions about the course content.

Textbook / Reading

- There is **NO official textbook** for this course.
- Recommended textbook (somewhat outdated, we won't follow too closely):



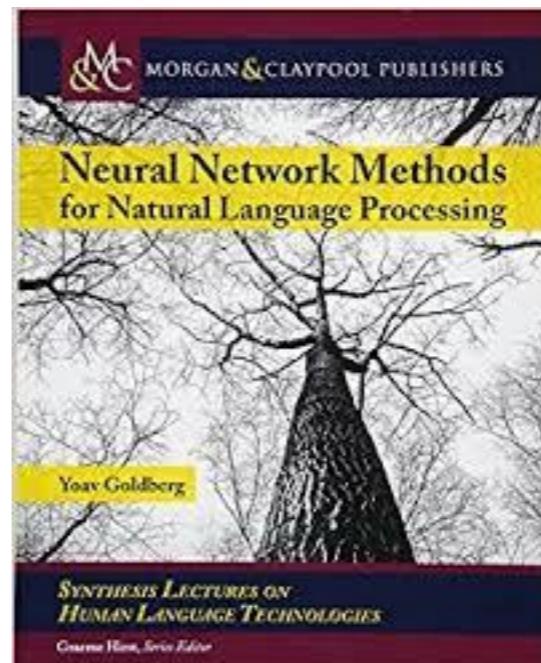
Dan Jurafsky & James Martin
Speech and Language Processing
2nd Ed. Prentice Hall (2009).

- Draft of most 3rd edition chapters:
<https://web.stanford.edu/~jurafsky/slp3/>
- We will also read a number of research papers.

Textbook / Reading

- Recommended textbook (mostly relevant later in the course):

Yoav Goldberg
Neural Network Methods for
Natural Language Processing
Morgan & Claypool. 2017



- Available as an ebook through the CU library
<https://clio.columbia.edu/catalog/13420294>

Prerequisites

- Data Structures (COMS W3134 or COMS W3137)
- Discrete Math (COMS W3202, recommended)
- Some previous or concurrent exposure to AI and machine learning is beneficial, but not required.
- Some experience with basic probability/statistics.
- Some experience with Python is helpful.

Grading

- Midterm 20%
- Final 30%
- 5 Homework assignments, each contains an analytical and a programming part, 10% each
- Regrade requests should be submitted within 3 days!

Homework

- Homework uploaded through Courseworks. Do not email!
 - Analytical part: Must be a plain txt or pdf documents (give LaTeX a shot).
 - Programming part: We will use Python 3.

Homework Late Policy

- Written homework and programming problems may be submitted up to four days late for a 20 point penalty.
- No homework will be accepted more than four days after the deadline.
- Other extensions will only be granted in exceptional circumstances.

Academic Honesty

- Submit your own answers and code.
- Review academic honesty policy on the syllabus (Courseworks).
- When in doubt, ask.
- When in trouble, ask for help (and early).

NLP in the Movies



I am fluent in over
six million forms
of communication

Open the pod bay
doors HAL!



I'm sorry Dave, I'm
afraid I can't do
that!

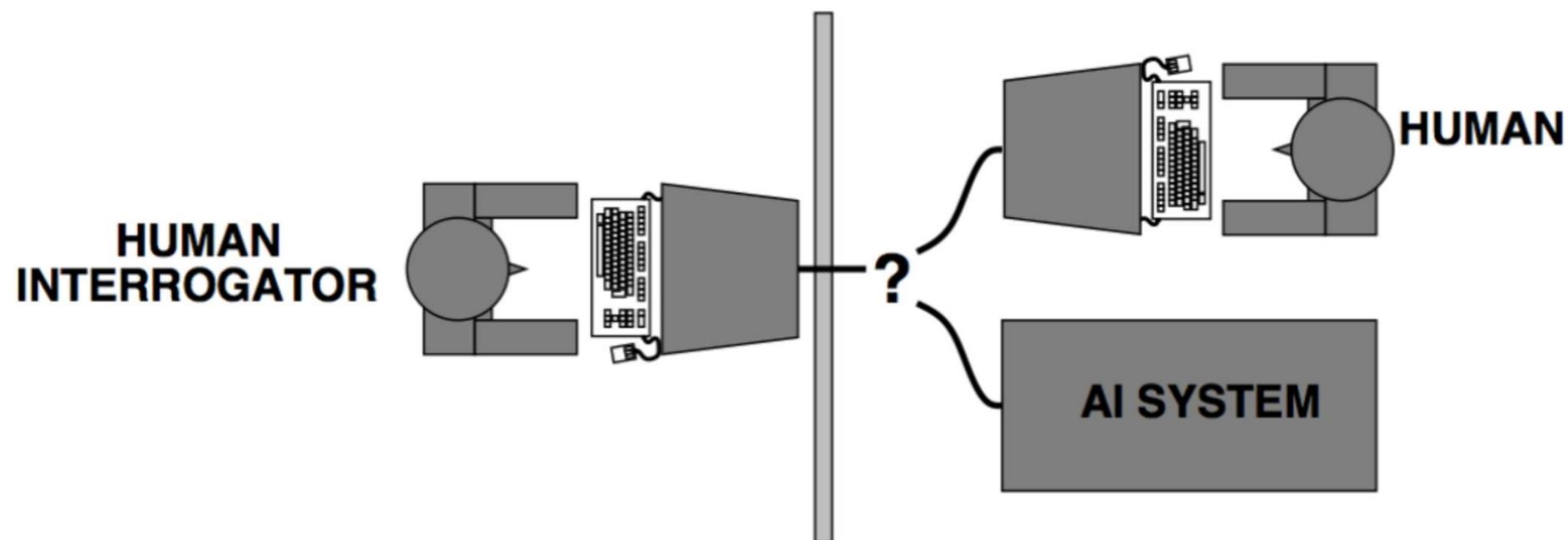
Natural Language Processing

- Important and active research area within AI.
- Timely: Most of our activities online are text based (web-pages, email, social media, blogs, news, product descriptions and reviews, medical reports, course content, ...)
- NLP leverages more and more available training data and modern Machine Learning techniques.
- Communicating with computers is the “holy grail” of AI.

Turing Test

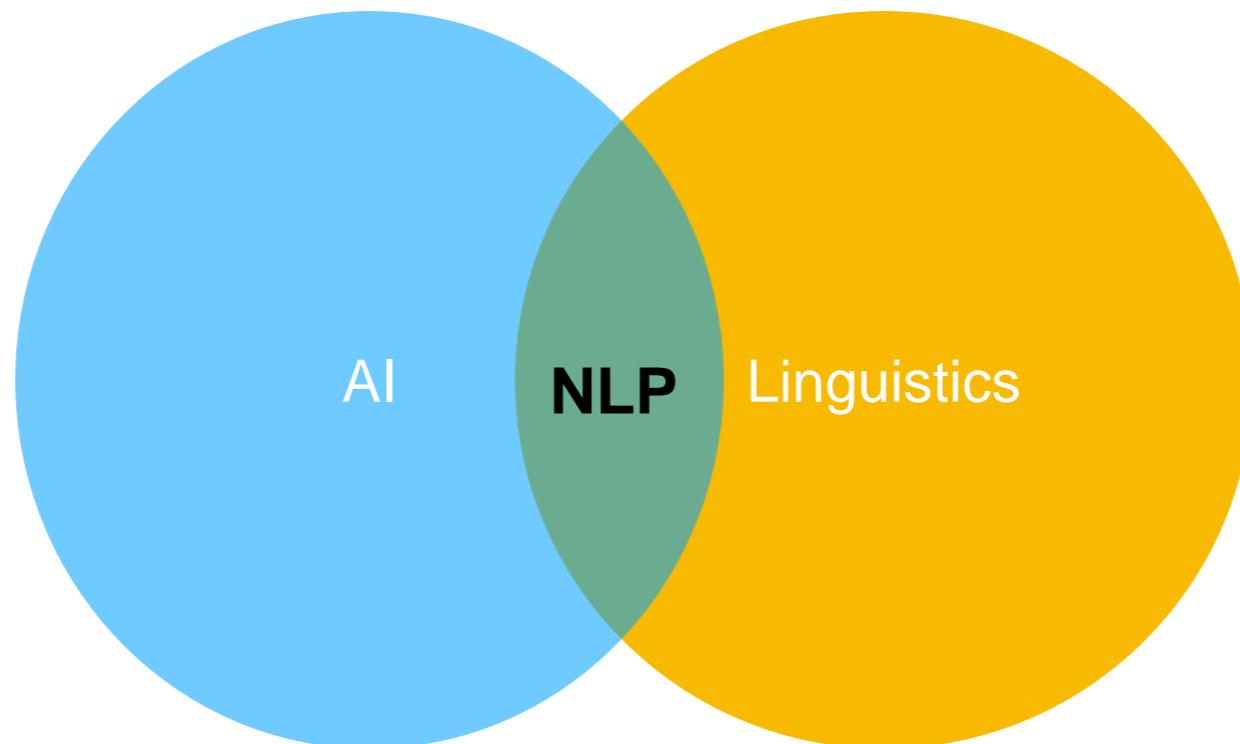
(Alan Turing, 1950)

- A computer passes the test of intelligence if it can fool a human interrogator into believing it is human.



- What skills are needed to build such a system?
 - **Language processing**, knowledge representation, reasoning, learning.

Natural Language Processing



“Every time I fire a linguist, my performance goes up” (Fred Jelinek)

Natural Language Processing vs. Computational Linguistics

- **NLP:** Build systems that can understand and generate natural language. Focus on applications.
- **Computational Linguistics:** Study human language using computational approaches.
- Many overlapping techniques.

Applications: Information Retrieval



Applications: Text Classification

- Spam filtering.
- Detecting topics / genre.
- Sentiment analysis, author recognition, forensic linguistics, ...

Applications: Sentiment Analysis

Fantastic... truly a wonderful family movie



I have a mixed feeling about this movie.



Well it is fun for sure but definitely not appropriate for kids 10 and below



My kids loved it!!



The movie is very funny and entertaining. Big A+



I got so boooored...



Disappointed. They showed all fun details in the trailer



Cute but not for adults



Applications: News Summarization

Columbia Newsblaster Articles

Summarizing all the news on the Web

Search for:
Offline summarization ▾

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

[View Today's Images](#)
[View Archive](#)

[About Newsblaster](#)
[About today's run](#)
[Newsblaster in Press](#)
[Academic Papers](#)

Article Sources:
[abcnews.go.com](#)
(71 articles)

Elon Musk unveils Dragon V2 reusable manned spacecraft
Summary from multiple countries, from articles in English
[[UPDATED](#)] (see summary with new information since yesterday)

In space there are currently two American astronauts on where the International Space Station living and working alongside three Russian cosmonauts tells more about the relationship. ([article 4](#)) A company that has flown unmanned capsules to the space station unveiled a spacecraft Thursday designed to ferry up to seven astronauts to low-Earth orbit that SpaceX CEO Elon Musk says will revolutionize access to space. ([article 3](#)) SpaceX unveiled its Dragon V2 spacecraft Thursday night, promising it will be able to carry seven astronauts to the International Space Station and back to Earth again, landing with the precision of a helicopter. ([article 5](#)) Lifting the vehicle's hatch, Musk settled into a reclined gold-and-black pilot's seat and pulled down a sleek, rounded glass control panel. ([article 2](#)) The cabin, designed to fly a crew of seven, looked more like a Star Trek movie set than the flight deck of NASA's now-retired space shuttle. ([article 2](#)) Dragon, which launches on a SpaceX Falcon 9 rocket, is one of three privately owned space taxis vying for NASA development funds and launch contracts. ([article 2](#)) The U.S. space agency turned over space station cargo runs and crew ferry flights after retiring its fleet of shuttles in 2011 and SpaceX already has a 1.6 billion contract for 12 station resupply missions ([article 2](#))

Other summaries about this story:

- [Summary from United States, from articles in English](#) (4 articles) [[compare](#)]
- [Summary from Canada, from articles in English](#) (1 articles) [[compare](#)]

Event tracking:

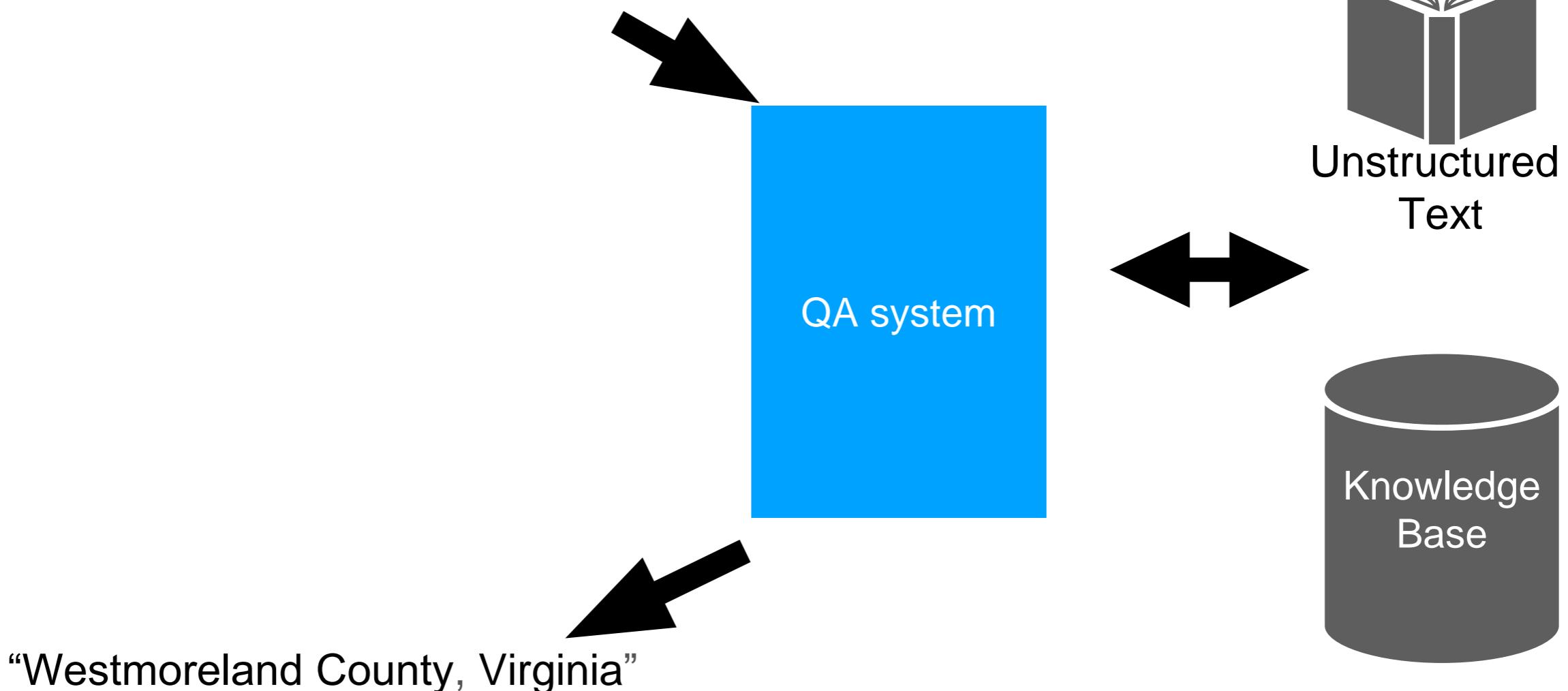
- [Track this story's development in time](#)

Story keywords

Space, spacecraft, astronauts, Musk, SpaceX

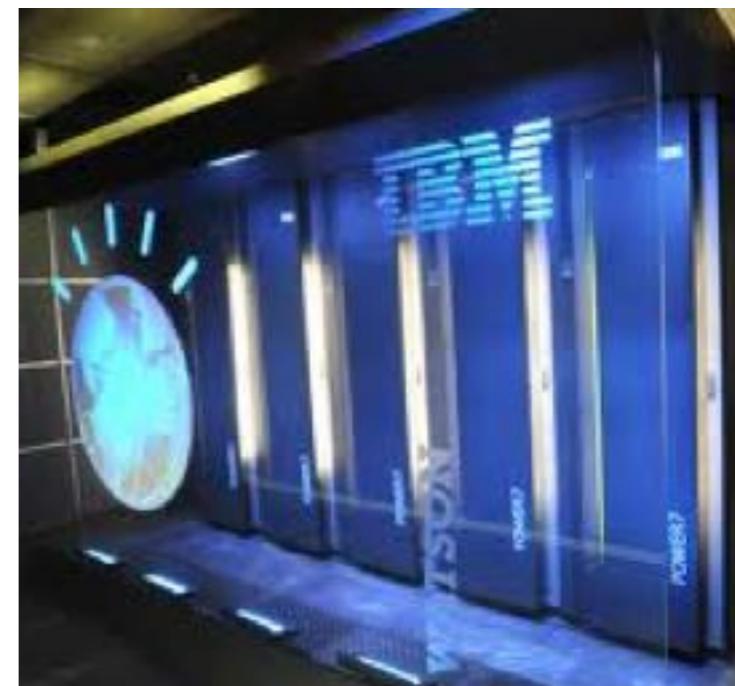
Application: Question Answering

“Where was George Washington born?”



Applications: Playing Jeopardy!

IBM Watson [2011]



William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" in

Who is Stoker?
(I FOR ONE WELCOME OUR
NEW COMPUTER OVERLORDS)

Who is Bram Stoker?
\$17,973

Combines information extraction & natural language understanding.

Applications: Machine Translation



Google <http://www.nytimes.com/2010/10/11/technology/11interior.html>

Translate From: English To: Spanish

Encontrar su camino en el centro comercial o en el aeropuerto, con un mapa Celular



Foto: Flickr, para The New York Times

FacBook ofrece un plan de pago y puede buscar tiendas y trazar un camino allí. Agitar el teléfono mostrará el barómetro más cercano.

Por VENDE G. Kupryoff
Publicado: 10 de octubre 2010

SAN FRANCISCO - Mapas de telefonía móvil han guiado la gente por las calles y callejones de todo el mundo. Pero cuando esas personas dentro de un edificio en expansión, se pueden perder.

[ENTRA PARA E-MAIL](#)
[IMPRIMIR](#)

Suscríbete a la tecnología

- Noticias de Tecnología
- Internet
- Smartphones
- Computadoras
- Empresas
- Negocios

MÁS POPULARES - TECNOLOGÍA

ENVIÉ UN CORREO ELECTRÓNICO

1. Bits: Malcolm Gladwell 'So' es un elemento de negocio
2. Gadgetwise: Breaking Up cables
3. Bits: cuál es el trato Beats I Tastemakers
4. Estado del Arte: La Revolución Pleasurable
5. Bits: ¿Por qué es Amazon lo que realmente necesita?
6. Bits: Las Fallas de iMessage
7. Bits: Google toma medida obvio Fallo
8. Bits: Cuentificación de privacidad de datos puede s
9. Q & A: Mantener en Wind
10. Bits: Dots, un juego altan

Machine Translation

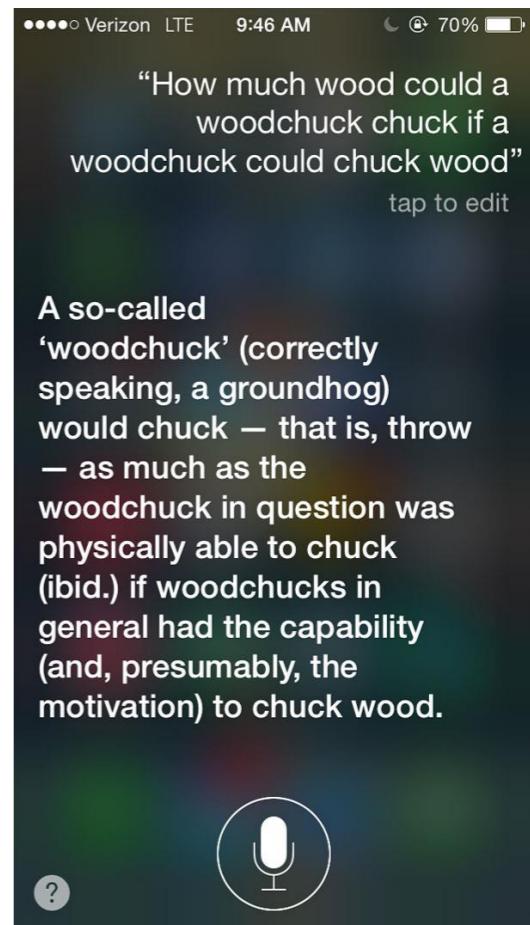
- One of the main research areas in NLP, and one of the oldest. Historical motivation: Translate Russian to English.
- MT is really difficult:
 - “Out of sight, out of mind” → “Invisible, imbecile”
 - “*The spirit is willing, but the flesh is weak*”
English → Russian → English
“*The vodka is good, but the meat is rotten*”
- Challenges: Word order, multiple translations for a word (need context), want to preserve meaning.

Machine Translation

- Until recently phrase-based translation was the predominant framework.
- Today neural network sequence-to-sequence models are used.
- Google Translate supports > 100 languages.

Applications: Virtual Assistants

- Siri (Apple), Google Now, Cortana (Microsoft), Alexa (Amazon).
- Subtasks: Speech recognition, language understanding (in context?), speech generation, ...

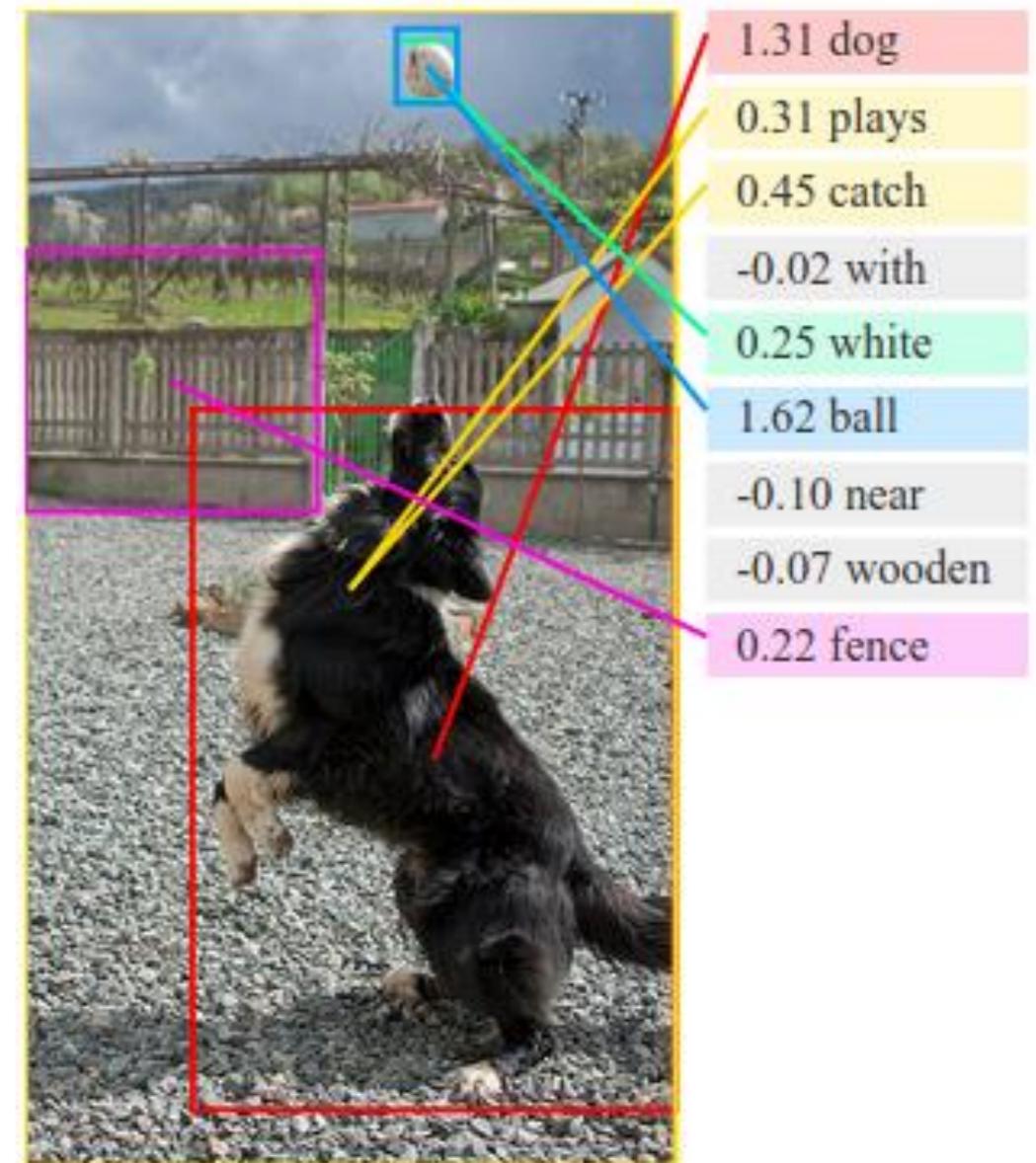


Applications: Image Captioning



“Man in black t-shirt is playing guitar.”

- Neural Networks for Object Detection and Language Generation.
- “Multi-modal” embeddings.
- Microsoft COCO data set.



What You Will Learn In This Course

- How can machines **understand** and **generate** natural language?
 - Theories about language (linguistics).
 - Algorithms.
 - Statistical / Machine Learning Methods.
 - Applications.

Course Overview

- Part I: Core NLP techniques.
 - Language modeling, part-of-speech tagging, syntactic parsing, word-sense disambiguation, semantic parsing, text similarity.
- Part II: Applications.
 - text classification, information retrieval, question answering, text generation, summarization, machine translation, image captioning, dialog systems.
- Machine Learning Techniques:
Supervised machine learning, bayesian models, sequence models (n-gram models, HMMs), neural networks, recurrent neural networks,...

Levels of Linguistic Representation

phonetics
phonology

sounds and sound
patterns of language



/bɔɪ/

morphology

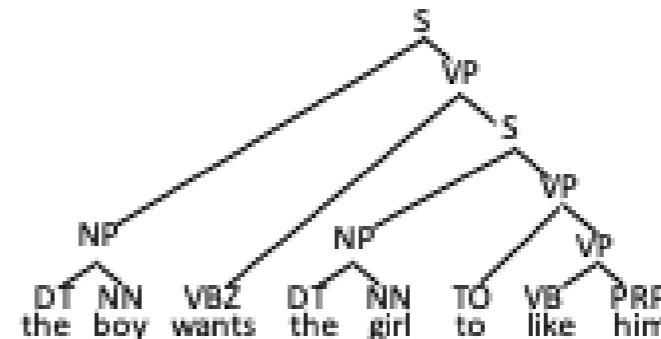
formation of words

in- + validate + -ed

DT | NN | VBZ | DT | NN | TO | VB | PRP | .
the | boy | want+s | the | girl | to | like | him | .

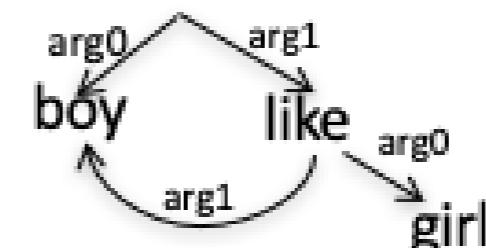
syntax

word order



semantics

word and sentence
meaning



pragmatics

influence of context and
situation

Natural Language Processing as Translation

- Most NLP techniques can be understood as translation tasks from one structure into another.
- For each translation step:
 - Construct search space of possible translations.
 - Find best paths through this space (decoding) according to some performance measure.
- Modern NLP relies on Machine Learning to figure out these translation steps.

NLP is hard: Ambiguity

- Unlike artificial languages, natural language is full of ambiguity.
- This can happen **on all levels of representation.**
 - “*Wreck a Nice Beach*” , “*Recognize Speech*”
 - “*inflammable*” = ***in* + *-flammable***
 - “*Enraged Cow Injures Farmer with Axe*”
 - “*Stolen Painting Found by Tree*”
 - “*Red Tape Holds Up New Bridges*”
 - “*Mouse*”



More Real Headlines

- *Ban on nude dancing on Governor's desk*
- *Kids Make Nutritious Snacks*
- *Drunk gets nine months in violin case*
- *Government head seeks arms*
- *Patient at death's door – doctors pull him through*
- *In America a woman has a baby every 15 minutes*

Syntactic Structure

- What is the **part-of-speech** of each word? (noun, verb, adjective, adverb, determiner, ...)
- What are the **constituents**:
 - Noun phrase: “*Enraged cow*”, “*The cat with the hat*”, “*Columbia University*”
- What are the **subjects and objects**:
 - “*Dog bites man*” vs. “*Man bites dog*”
- **Modification**:
 - “*John saw the man in the park with a telescope*”

Structural Ambiguity

- Interplay between constituent structure and modification.
- Prepositional Phrase (PP) attachment:

Enraged cow injures farmer with axe.

[*Enraged cow*] injures [*farmer with axe*]
NP NP

[*Enraged cow*] injures *farmer* [*with axe*]
NP NP PP

Representing Modification with Brackets

[Enraged cow] [injures [farmer [with axe]]]
NP NP PP

[Enraged cow] injures [farmer] [with axe]
NP NP PP

More PP attachment

[Ban] on [nude dancing] [on governor's desk]

NP

NP

NP

- What are the possible modifications? Which one is correct?

[[Ban] on [nude dancing]] [on governor's desk]

NP

PP

[Ban] on [[nude dancing] [on governor's desk]]

NP

PP

NP

Noun-Noun Modification

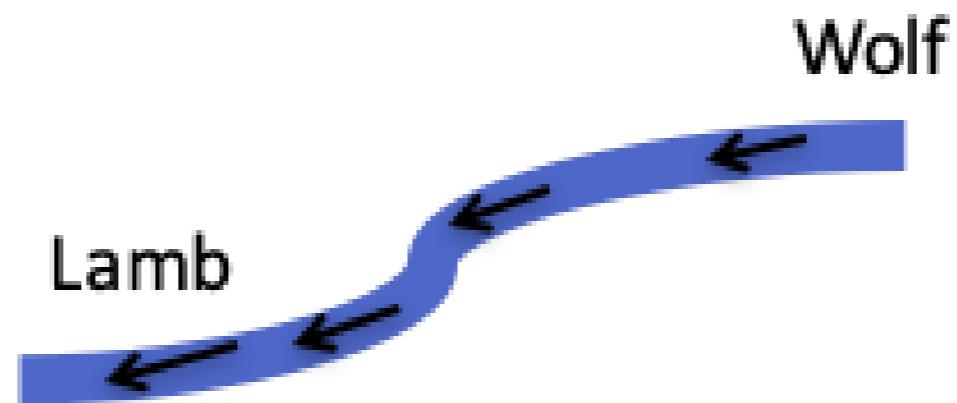
- What is the *semantic* relationship between nouns in a noun compound?
 - *Water fountain:* A fountain that **supplies** water.
 - *Water ballet:* A ballet that **takes place** in water.
 - *Water meter:* A device that **measures** water.
 - *Water barometer:* A barometer that **uses** water (instead of mercury) to measure air pressure.
 - *Water glass:* A glass that is meant to **hold** water.

Other tricky phenomena

- Need for semantic representation.

There was once a Wolf who saw a Lamb drinking at a river and wanted an excuse to eat it.

*For that purpose, **even though** he himself was **upstream**, he accused the Lamb of stirring up the water and keeping him from drinking. . .*



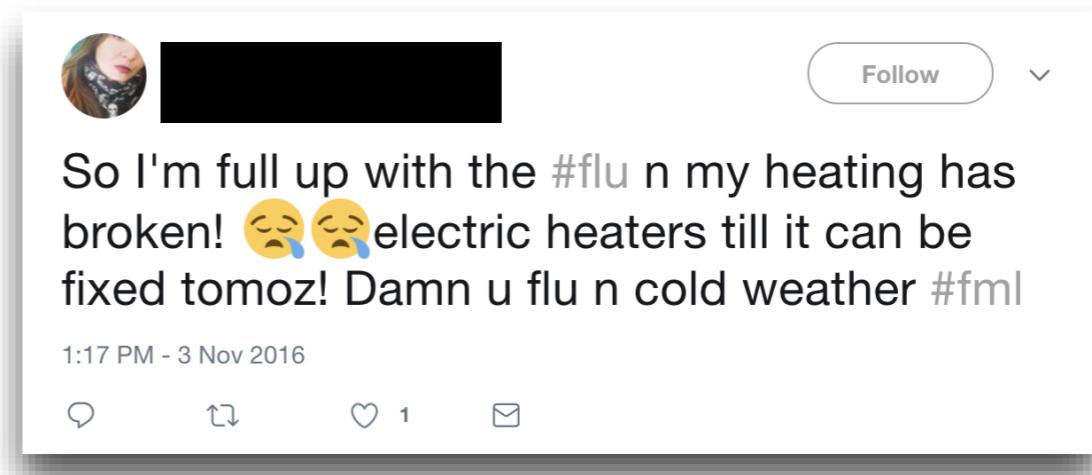
Other tricky issues: Language Variety

- Problem: Most NLP techniques were developed on English (specifically financial news written in American English in the 1980s), or other languages with many resources.
- Languages use different mechanisms to express meaning (morphology vs. word-order).



Other tricky issues: Domains and Language Change

- Non-standard English
- Idioms: *throw in the towel, get cold feet, kick the bucket*
- Neologisms (fixed lexicon doesn't work)
 - *noob, crowdsource, unfriend, retweet, bromance, ...*



Morphology

- Structure and formation of words.
- **Derivational** morphology: Create new words from old words (can also change the part-of-speech).
anti- + dis- + **establish** + -ment + -arian + -ism
- **Inflectional** morphology:
 - Convey information about number, person, tense, aspect, mood, voice, and the role a word plays in the sentence (case).
 - English has few morphological categories, but many languages are morphologically rich.

Morphology

- Morphological categories in English
 - Number (“*dog*”, “*dog +s*”)
 - Person (“*I run*”, “*She runs*”)
 - Tense (“*He waited*”)
 - Voice (“*The issue was decided*”)
- Other examples from other languages?

Acknowledgments

- Some slides and examples from Kathy McKeown, Dan Jurafsky, Dragomir Radev.

Natural Language Processing

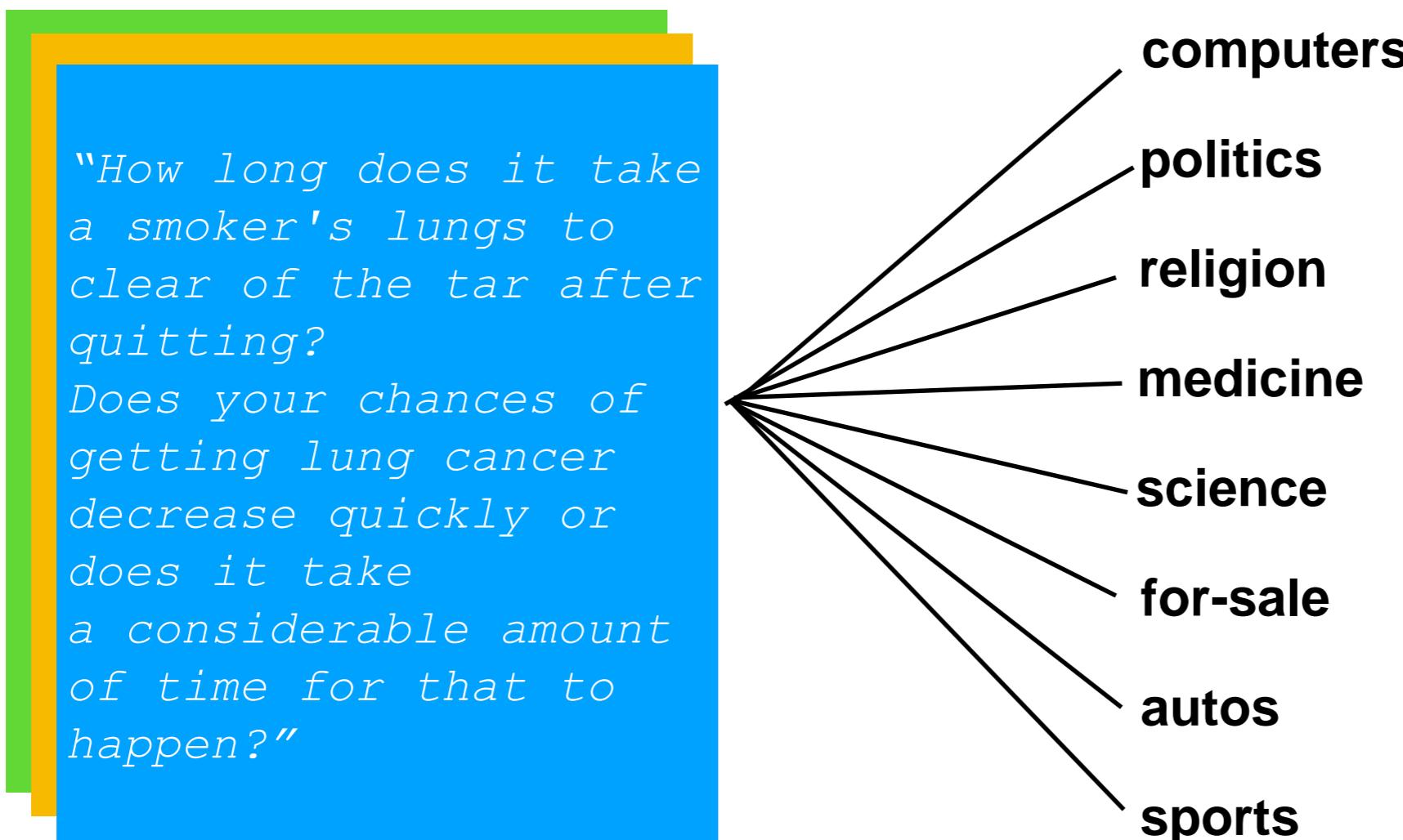
Lectures 2: Language Classification. Probability Review.
Machine Learning Background. Naive Bayes' Classifier.

1/31/2020

COMS W4705
Yassine Benajiba

Text Classification

- Given a representation of some document d , identify which class $c \in C$ the document belongs to.



From the 20-Newsgroups data set:

<http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html>

Text Classification

- Applications:
 - Spam detection.
 - Mood / Sentiment detection.
 - Author identification.
 - Identifying political affiliation.
 - Word Sense Disambiguation.
 - ...

Text Classification

- This is a machine learning problem.
 - How do we represent each document? (feature representation).
 - Can use different ML techniques.
 - **Supervised ML:** Fixed set of classes C .
Train a classifier from a set of labeled <document,class> pairs.
 - Discriminative vs. Generative models.
 - **Unsupervised ML:** Unknown set of classes C .
Topic modeling.

Types of Feedback

- **Supervised learning:** Given a set of input-output pairs, learn a function that maps inputs to outputs.
- **Unsupervised learning:** Learn patterns in the input without any explicit feedback.
One typical approach: clustering, identify clusters of input examples.
- **Semi-supervised learning:** Start with a few labeled input/output pairs, then use a lot of unlabeled data to improve.
- **Reinforcement learning:** Start with a *policy* determining the agent's actions. Feedback in the form of reward or punishment.

Supervised Learning

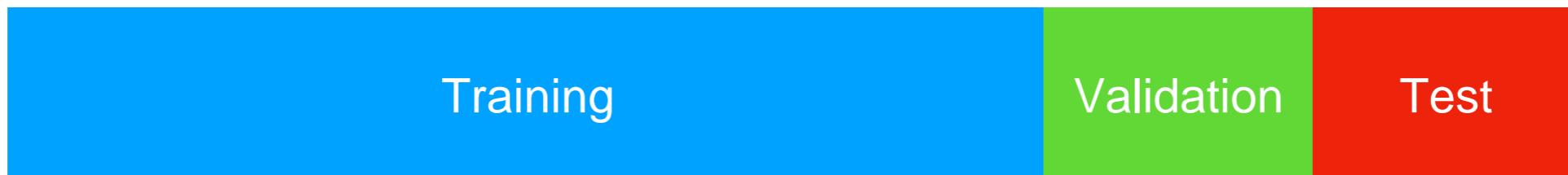
- Given: Training data consisting of training examples $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, where \mathbf{x}_i is an input example (a d -dimensional vector of attribute values) and y_i is the label.

example					label
1	x_{11}	x_{12}	...	x_{1d}	y_1
...
i	x_{i1}	x_{i2}	...	x_{id}	y_i
...
n	x_{n1}	x_{n2}	...	x_{nd}	y_n

- Goal: learn a hypothesis function $h(x)$ that approximates the true relationship between x and y . This function should: 1) ideally be consistent with the training data. 2) generalize to unseen examples.
- In NLP y_i typically form a finite, discrete set.

Running Machine Learning Experiments

- When running machine learning experiments we typically split the labeled data in three sections:



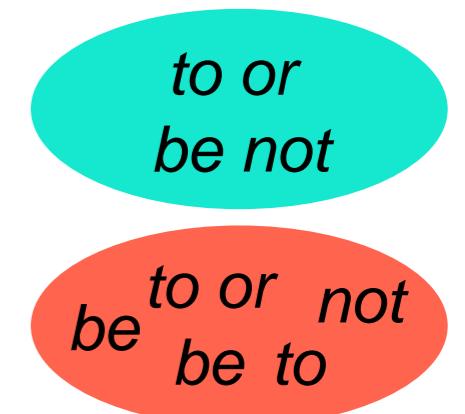
- For example: 80% Training, 10% Validation (development), 10% Test or 90/5/5
- Validation set is used to tune model parameters (for example smoothing parameters), but cannot be used for training. This can help with overfitting.
- Test set is used to assess the performance of the final model and provide an estimation of the test error.

Note: Never train or tune parameters on the test set!

Representing Documents

to be, or not to be

- Set-of-words representation.
- Bag-of-words representation (Multi-set).
- Vector-space model: Each word corresponds to one dimension in vector space. Entries are either:
 - Binary (Word appears / does not appear)
 - Raw or normalized frequency counts.
 - Weighted frequency counts
 - Probabilities.



be	2
:	:
not	1
:	:
or	1
:	:
to	2

What is a Word?

- e.g., are “*Cat*”, “*cat*” and “*cats*” the same word?
- “*September*” and “*Sept*”?
- “*zero*” and “*oh*”?
- Is “_” a word? “.”? “*”? “(“?
- How many words are there in “*don’t*” ? “*Gonna*” ? “*I.B.M.*”?
- In Japanese and Chinese text -- how do we identify a word?
- ...

Text Normalization

- Every NLP task needs to do some text normalization.
 - Segmenting / tokenizing words in running text.
 - Normalizing word forms (lemmatization or stemming, possibly replacing named-entities).
 - Sentence splitting.

Linguistic Terminology

- **Sentence:** Unit of written language.
- **Utterance:** Unit of spoken language.
- Word **Form:** the inflected form as it actually appears in the corpus. “*produced*”
- Word **Stem:** The part of the word that never changes between morphological variations. “*produc*”
- **Lemma:** an abstract base form, shared by word forms, having the same **stem**, part of speech, and word sense – stands for the **class** of words with **stem**. “*produce*”
- **Type:** number of distinct words in a corpus (vocabulary size).
- **Token:** Total number of word occurrences.

Tokenization

- Tokenization: The process of segmenting text (a sequence of characters) into a sequence of tokens (words).

"Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing."

mr. o'neill thinks that the boys' stories about Chile's capital are n't
amusing .

- Simple (but weak) approach: Separate off punctuation. Then split on whitespaces.
- Typical implementations use regular expressions (finite state automata).

Tokenization Issues

- Dealing with punctuation (some may be part of a word)
“Ph.D.”, “O'Reilly”, “pick-me-up”
- Which tokens to include (punctuation might be useful for parsing, but not for text classification)?
- Language dependent: Some languages don't separate words with whitespaces.
de: “*Lebensversicherungsgesellschaftsangestellter*”

zh: 日文章鱼怎么说? - *Japanese Octopus how say?*

日文章鱼怎么说? - *Sun article fish how say?*

Lemmatization

- Converting Lemmas into their base form.

"Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing."

mr. o'neill think that the boy story about chile's capital are n't
amusing .

PER PER think that the boy story about LOC 's capital are n't
amusing .

Probabilities in NLP

- Ambiguity is everywhere in NLP. There is often *uncertainty* about the “correct” interpretation. Which is more likely:
 - Speech recognition: “*recognize speech*” vs. “*wreck a nice beach*”
 - Machine translation: “*l'avocat general*”: “*the attorney general*” vs. “*the general avocado*”
 - Text classification: is a document that contains the word “*rice*” more likely to be about politics or about agriculture?
What if it also includes several occurrences of the word “*stir*”?
- Probabilities make it possible to combine evidence from multiple sources systematically to (using Bayesian statistics)

Bayesian Statistics

- Typically, we observe some evidence (for example, words in a document) and the goal is to infer the “correct” interpretation (for example, the topic of a text).
- Probabilities express the degree of belief we have in the possible interpretations.
 - **Prior probabilities:** Probability of an interpretation prior to seeing any evidence.
 - **Conditional (Posterior) probability:** Probability of an interpretation after taking evidence into account.

Probability Basics

- Begin with a **sample space** Ω
 - Each $\omega \in \Omega$ is a possible basic outcome / “possible world” (e.g. the 6 possible rolls of a die).
- A **probability distribution** assigns a probability to each basic outcome.

$$P(\omega) \leq 1.0 \text{ for every } \omega \in \Omega$$

$$\sum_{\omega \in \Omega} P(\omega) = 1.0$$

- E.g: six-sided die

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1.0$$

Events

- An *event* A is any subset of Ω .

$$P(A) = \sum_{\omega \in A} P(\omega)$$

- Example:

$$P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$$

Random Variables

- A random variable is a function from basic outcomes to some range, e.g. real numbers or booleans.

$$Odd(1) = \text{true}$$

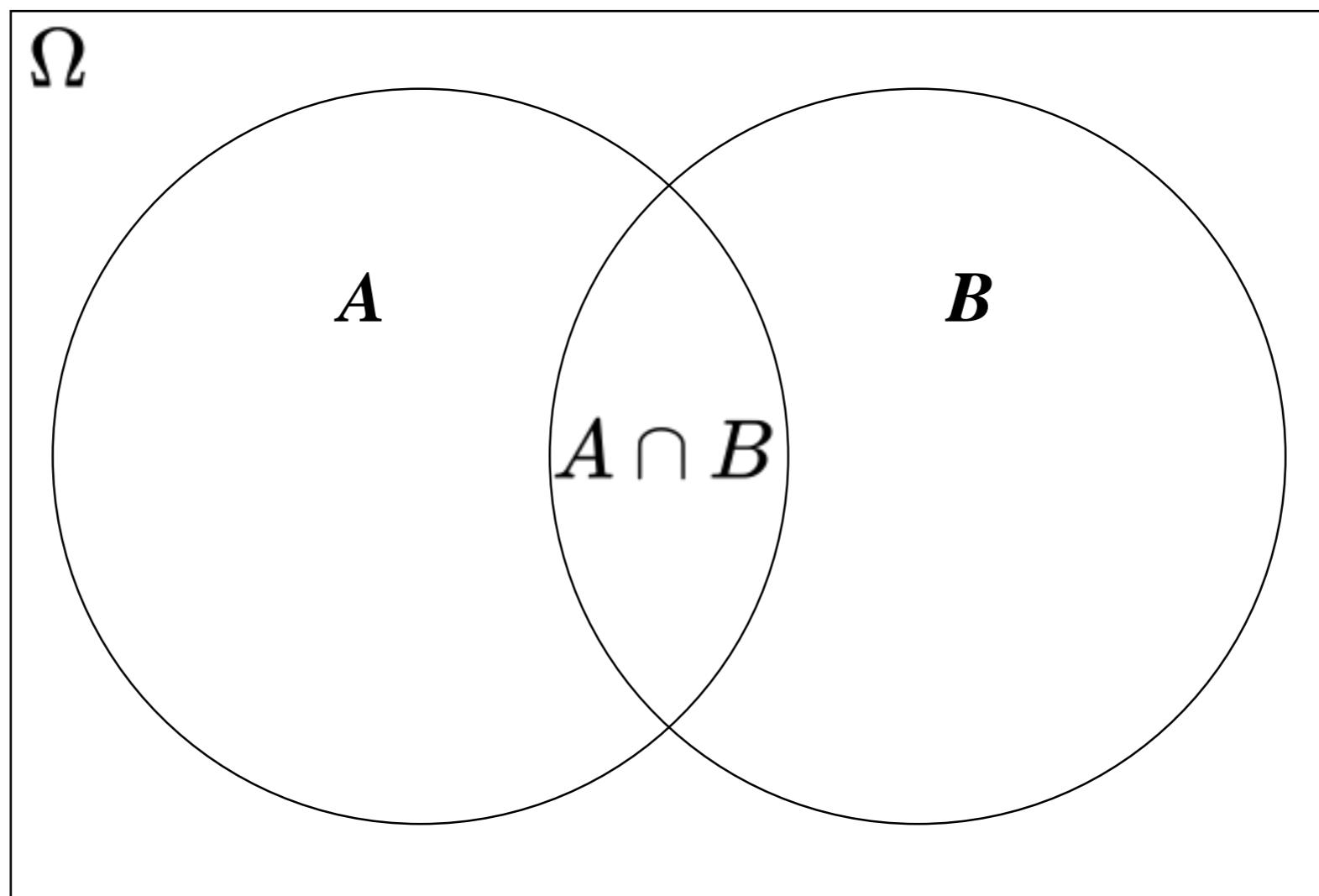
- A distribution P induces a probability distribution for any random variable.

$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

- E.g $P(Odd = \text{true}) = P(1) + P(3) + P(5) = 1/2$

Joint and Conditional Probability

Joint probability: $P(A \cap B)$ also written as $P(A, B)$



Conditional probability: $P(A|B) = \frac{P(A, B)}{P(B)}$

Rules for Conditional Probability

- Product rule: $P(A, B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A)$

- Chain rule (generalization of product rule):

$$P(A_n, \dots, A_1) = P(A_n | A_{n-1}, \dots, A_1) \cdot P(A_{n-1}, \dots, A_1)$$

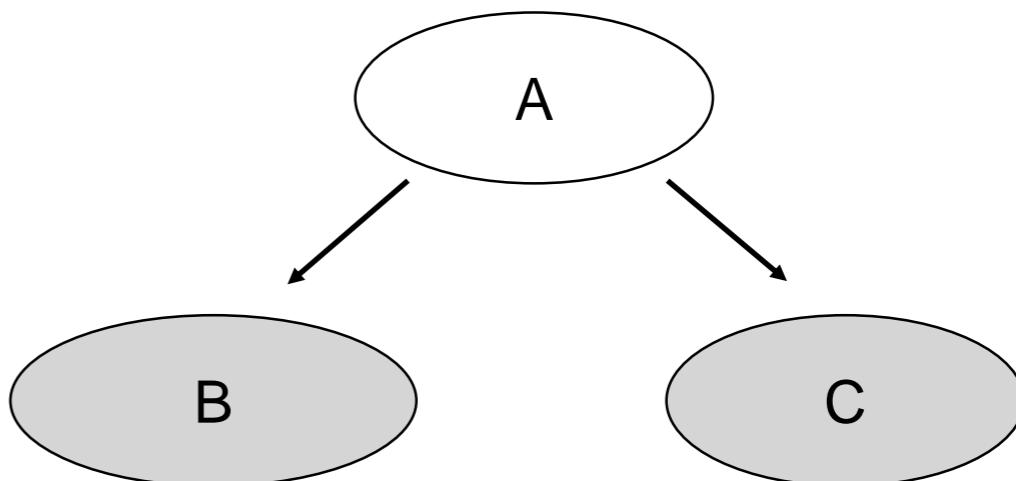
- **Bayes' Rule:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Independence

- Two events are independent if $P(A) = P(A|B)$
or equivalently $P(A, B) = P(A) \cdot P(B)$ (if $P(B) > 0$)
- Two events are **conditionally independent** if:
$$P(B, C|A) = P(B|A)P(C|A)$$

or equivalently
$$P(B|A, C) = P(B|A)$$
 and $P(C|A, B) = P(C|A)$



Probabilities and Supervised Learning

- Given: Training data consisting of training examples
 $\text{data} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n),$
Goal: Learn a mapping h from x to y .
- We would like to learn this mapping using $P(y|x)$.
- Two approaches:
 - Discriminative algorithms learn $P(y|x)$ directly.
 - Generative algorithms use Bayes rule

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

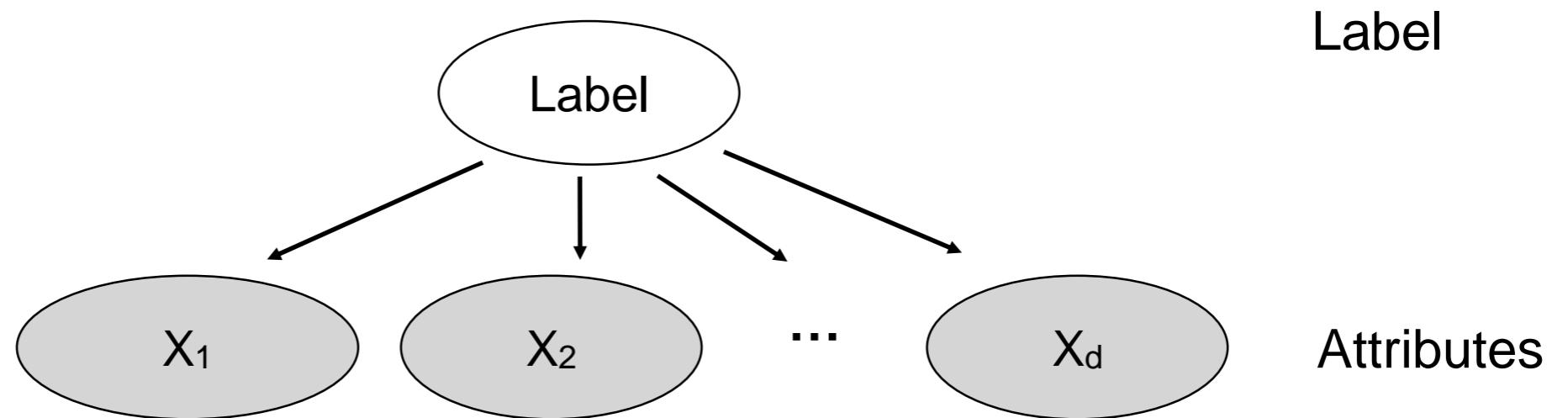
Discriminative Algorithms

- Model conditional distribution of the label given the data $P(y|x)$
- Learns decision boundaries that separate instances of the different classes.
- To predict a new example, check on which side of the decision boundary it falls.
- Examples:
support vector machine (SVM), decision trees, random forests, neural networks, log-linear models.

Generative Algorithms

- Assume the observed data is being “generated” by a “hidden” class label.
- Build a **different model** for each class.
- To predict a new example, check it under each of the models and see which one matches best.
- Estimate $P(x|y)$ and $P(y)$. Then use bases rule
$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$
- Examples:
Naive Bayes, Hidden Markov Models, Gaussian Mixture Models, PC

Naive Bayes

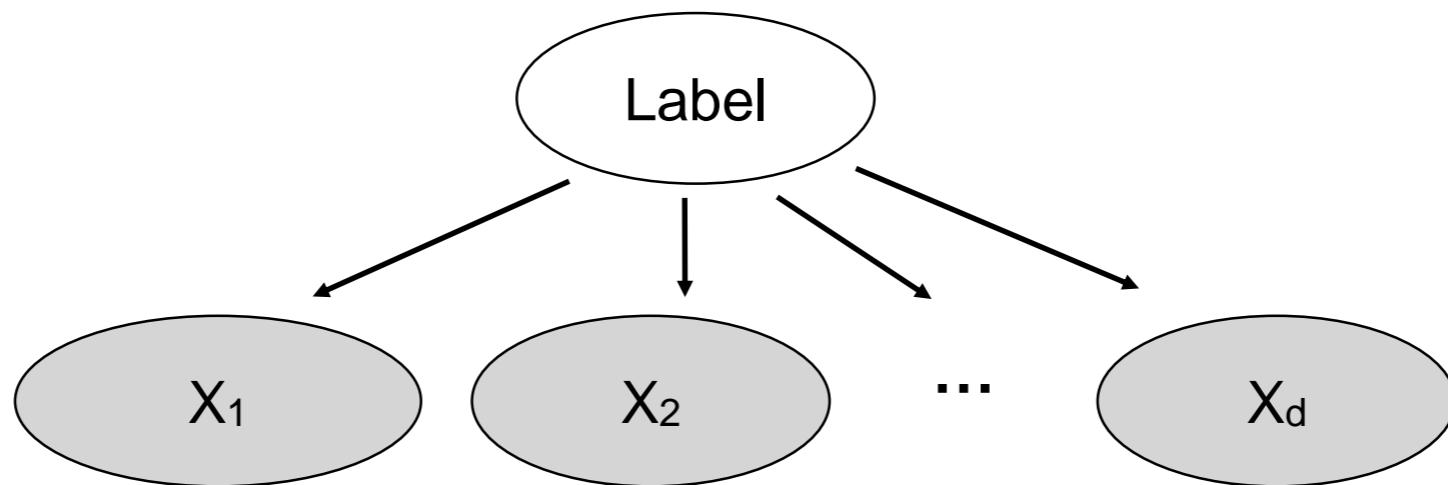


$$\mathbf{P}(Label, X_1, \dots, X_d) = \mathbf{P}(Label) \prod_i P(X_i | Label)$$

$$\mathbf{P}(Label | X_1, \dots, X_d) = \frac{\mathbf{P}(Label) \prod_i P(X_i | Label)}{\prod_i P(X_i)}$$

$$= \alpha [\mathbf{P}(Label) \prod_i P(X_i | Label)]$$

Naive Bayes Classifier



$$\mathbf{P}(Label|X_1, \dots, X_d) = \alpha [\mathbf{P}(Label) \prod_i P(X_i|Label)]$$

$$y^* = \arg \max_y P(y) \prod_i P(x_i|y)$$

Note that the normalizer α does no longer matter for the argmax because α is independent of the class label.

Training the Naive Bayes' Classifier

- Goal: Use the training data to estimate $P(\text{Label})$ and $P(X_i|\text{Label})$ from training data.
- Estimate the prior and posterior probabilities using **Maximum Likelihood Estimates (MLE)**:

$$P(y) = \frac{\text{Count}(y)}{\sum_{y' \in Y} \text{Count}(y')}$$

$$P(x_i|y) = \frac{\text{Count}(x_i, y)}{\sum_{x'} \text{Count}(x', y)} = \frac{\text{Count}(x_i, y)}{\text{Count}(y)}$$

- I.e. we just count how often each token in the document appears together with each class label.

Why the Independence Assumption Matters

- Without the independence assumption we would have to estimate $\mathbf{P}(X_1, \dots, X_d | Label)$
- There would be many combinations of x_1, \dots, x_d that are never seen (sparse data).
- The independence assumption allows us to estimate each $\mathbf{P}(X_1 | label)$ independently.

Is this a safe assumption for documents?
Are the words really independent of each other?

Training the Naive Bayes' Classifier

- Ways to improve this model?
- Some issues to consider...
 - What if there are words that do not appear in the training set? What if it appears only once?
 - What if the plural of a word never appears in the training set?
 - How are extremely common words (e.g., "the", "a") handled?

Acknowledgments

- Some slides and examples from:
 - Kathy McKeown, Dragomir Radev

Natural Language Processing

Lecture 3: n-gram language models

1/31/2020

COMS W4705
Yassine Benajiba

Probability of a Sentence

Probability of a Sentence

“But it must be recognized that the notion of ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.”

Noam Chomsky (1969)

Language Modeling

- Task: predict the next word given the context.
- Used in speech recognition, handwritten character recognition, spelling correction, text entry UI, machine translation,...

Language Modeling

- Stocks plunged this ...
- Let's meet in Times ...
- I took the subway to ...

From a NYT story

- *Stocks plunged this*
 - *Stocks plunged this morning, despite a cut interest rates by the ...*
 - *Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall ...*
 - *Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall Street began*

Human Word Prediction

- Clearly at least some of us have the ability to predict the future.
- How does this work?
 - Domain knowledge
 - Syntactic knowledge (guess correct part of speech)
 - Lexical knowledge

Probability of the Next Word

- Idea: We do not need to model domain, syntactic, and lexical knowledge perfectly.
- Instead, we can rely on the notion of **probability of a sequence** (letters, words...).

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

Applications

- Speech recognition: $P(\text{"recognize speech"}) > P(\text{"wreck a nice beach"})$
- Text generation: $P(\text{"three houses"}) > P(\text{"three house"})$
- Spelling correction $P(\text{"my cat eats fish"}) > P(\text{"my xat eats fish"})$
- Machine Translation $P(\text{"the blue house"}) > P(\text{"the house blue"})$
- Other uses
 - OCR
 - Summarization
 - Document classification
 - Essay scoring

Language Models

- This model can also be used to describe the probability of an entire sentence, not just the last word.
- Use the chain rule:

$$P(w_1, \dots, w_n) =$$

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

$$P(w_n | w_1, \dots, w_{n-1}) P(w_{n-1} | w_{n-2}, \dots, w_1) P(w_{n-2}, \dots, w_1)$$

...

Markov Assumption

- $P(w_n | w_1, w_2, w_3, \dots, w_{n-1})$ is difficult to estimate.
- The longer the sequence becomes, the less likely $w_1 w_2 w_3 \dots w_{n-1}$ will appear in training data.
- Instead, we make the following simple independence assumption (Markov assumption):
 - The probability to see w_n depends only on the previous $k-1$ words.

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

$$\approx P(w_n | w_{n-k+1}, \dots, w_{n-1})$$

bi-gram language model

- Using the Markov assumption and the chain rule:

$$P(w_1, \dots, w_n) \approx$$

$$P(w_1) \cdot P(w_2 | w_1) \cdot P(w_3 | w_2) \cdots P(w_n | w_{n-1})$$

- More consistent to use only bigrams:

$$P(w_1 | start) \cdot P(w_2 | w_1) \cdot P(w_3 | w_2) \cdots P(w_n | w_{n-1})$$

n-grams

- The sequence w_n is a unigram.
- The sequence w_{n-1}, w_n is a bigram.
- The sequence w_{n-2}, w_{n-1}, w_n is a trigram....
- The sequence w_{n-2}, w_{n-1}, w_n is a quadrigram...

Variable-Length Language Models

- We typically don't know what the length of the sentence is.
- Instead, we use a special marker STOP that indicates the end of a sentence.
- We typically just augment the sentence with START and STOP markers to provide the appropriate context.

START i want to eat Chinese food END

$$P(i/START) \cdot P(want/i) \cdot P(to/want) \cdot P(eat/to) \cdot P(Chinese/eat) \cdot P(food/Chinese) \cdot P(END/food)$$

trigram example

$$P(i/START, START) \cdot P(want/START,i) \cdot P(to/i,want) \cdot P(eat/want,to) \cdot \\ P(Chinese/to,eat) \cdot P(food/eat,Chinese) \cdot P(END/Chinese,food)$$

Bigram example from the Berkeley Restaurant Project (BeRP)

Eat on	0.16	Eat Thai	0.03
Eat some	0.06	Eat breakfast	0.03
Eat lunch	0.06	Eat in	0.02
Eat dinner	0.05	Eat Chinese	0.02
Eat at	0.04	Eat Mexican	0.02
Eat a	0.04	Eat tomorrow	0.01
Eat Indian	0.04	Eat dessert	0.007

Bigram example from the Berkeley Restaurant Project (BeRP)

START I	0.25	Want some	0.04
START I'd	0.06	Want Thai	0.01
START Tell	0.04	To eat	0.26
START I'm	0.02	To have	0.14
I want	0.32	To spend	0.09
I would	0.29	To be	0.02
I don't	0.08	British food	0.60
I like	0.01	Raj's	0.15

Bigram example from the Berkeley Restaurant Project (BeRP)

- Assume $P(\text{END} \mid \text{food}) = 0.2$

$$\begin{aligned} P(\text{I want to eat British food}) &= \\ P(\text{I} \mid \text{START}) \cdot P(\text{want} \mid \text{I}) \cdot P(\text{to} \mid \text{want}) \cdot P(\text{eat} \mid \text{to}) \cdot \\ P(\text{British} \mid \text{eat}) \cdot P(\text{food} \mid \text{British}) \cdot P(\text{END} \mid \text{food}) &= \\ .25 \cdot .32 \cdot .65 \cdot .26 \cdot .001 \cdot .60 \cdot .2 &= .0000016 \end{aligned}$$

$$\begin{aligned} P(\text{I want to eat Chinese food}) &= \\ P(\text{I} \mid \text{START}) \cdot P(\text{want} \mid \text{I}) \cdot P(\text{to} \mid \text{want}) \cdot P(\text{eat} \mid \text{to}) \cdot \\ P(\text{Chinese} \mid \text{eat}) \cdot P(\text{food} \mid \text{Chinese}) \cdot P(\text{END} \mid \text{food}) &= \\ .25 \cdot .32 \cdot .65 \cdot .26 \cdot .02 \cdot .60 \cdot .2 &= .000032 \end{aligned}$$

log probabilities

- Probabilities can become very small (a few orders of magnitude per token).
- We often work with log probabilities in practice.

$$p(w_1 \dots w_n) = \prod_{i=1}^n p(w_i | w_{i-1})$$

$$\log p(w_1 \dots w_n) = \sum_{i=1}^n \log p(w_i | w_{i-1})$$

What do ngrams capture?

- Probabilities seem to capture *syntactic facts* and *world knowledge*.
- *eat* is often followed by a NP.
- *British* food is not too popular, but *Chinese* is.

Estimating n-gram probabilities

- We can estimate n-gram probabilities using maximum likelihood estimates.

$$p(w|u) = \frac{\text{count}(u, w)}{\text{count}(u)}$$

- Or for trigrams:

$$p(w|u, v) = \frac{\text{count}(u, v, w)}{\text{count}(u, v)}$$

Bigram Counts from BeRP

	I	Want	To	Eat	Chinese	Food	lunch
I	8	1087	0	13	0	0	0
Want	3	0	786	0	6	8	6
To	3	0	10	860	3	0	12
Eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
Food	19	0	17	0	0	0	0

Counts to Probabilities

	I	Want	To	Eat	Chinese	Food	lunch
I	8	1087	0	13	0	0	0
Want	3	0	786	0	6	8	6
To	3	0	10	860	3	0	12
Uni	0	0	2	0	19	2	52
Eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
Food	19	$\frac{count(I, want)}{count(I)}$	0	$1087/3437$	0	0.32	0
Lunch	4	0	0	0	0	1	0

Corpora

- Large digital collections of text or speech. Different languages, domains, modalities. Annotated or un-annotated.
- English:
 - Brown Corpus
 - BNC, ANC
 - Wall Street Journal
 - AP newswire
 - DARPA/NIST text/speech corpora
(Call Home, ATIS, switchboard, Broadcast News,...)
 - MT: Hansards, Europarl

Google Web 1T 5-gram Corpus

File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens: 1,024,908,267,229

Number of sentences: 95,119,665,584

Number of unigrams: 13,588,391

Number of bigrams: 314,843,401

Number of trigrams: 977,069,902

Number of fourgrams: 1,313,818,354

Number of fivegrams: 1,176,470,663

Google Web 1T 5-gram Corpus

- 3-gram examples:

ceramics collectables collectibles 55

ceramics collectables fine 130

ceramics collected by 52

ceramics collectible pottery 50

ceramics collectibles cooking 45

ceramics collection , 144

ceramics collection . 247

ceramics collection </S> 120

ceramics collection and 43

ceramics collection at 52

ceramics collection is 68

ceramics collection of 76

Google Web 1T 5-gram Corpus

- 4-gram examples:

serve as the incoming 92
serve as the incubator 99
serve as the independent 794
serve as the index 223
serve as the indication 72
serve as the indicator 120
serve as the indicators 45
serve as the indispensable 111
serve as the indispensable 40
serve as the individual 234
serve as the industrial 52
serve as the industry 607
serve as the info 42
serve as the informal 102

Data sparsity in n-gram models

- Sparsity is a problem all over NLP: Test data contains language phenomena not encountered during training.
- For n-gram models there are two issues:
 - We may not have seen all tokens.
 - We may not have seen all ngrams (even though the individual tokens are known).
 - Token has not been encountered in this context before.

$$P(\text{lunch} \mid \text{I}) = 0.0$$

Unseen Tokens

- Typical approach to unseen tokens:
 - Start with a specific lexicon of known tokens.
 - Replace all tokens in the training and testing corpus that are not in the lexicon with an *UNK* token.
- Practical approach:
 - Lexicon contains all words that appear more than k times in the training corpus.
 - Replace all other tokens with UNK.

Unseen Contexts

- Two basic approaches:
 - Smoothing / Discounting: Move some probability mass from seen trigrams to unseen trigrams.
 - Back-off: Use n-1-..., n-2-... grams to compute n-gram probability.
- Other techniques:
 - Class-based backoff, use back-off probability for a specific word class / part-of-speech.

Zipf's Law

- Problem: n-grams (and most other linguistic phenomena) follow a *Zipfian* distribution.
- A few words occur very frequently.
- Most words occur very rarely. Many are seen only once.

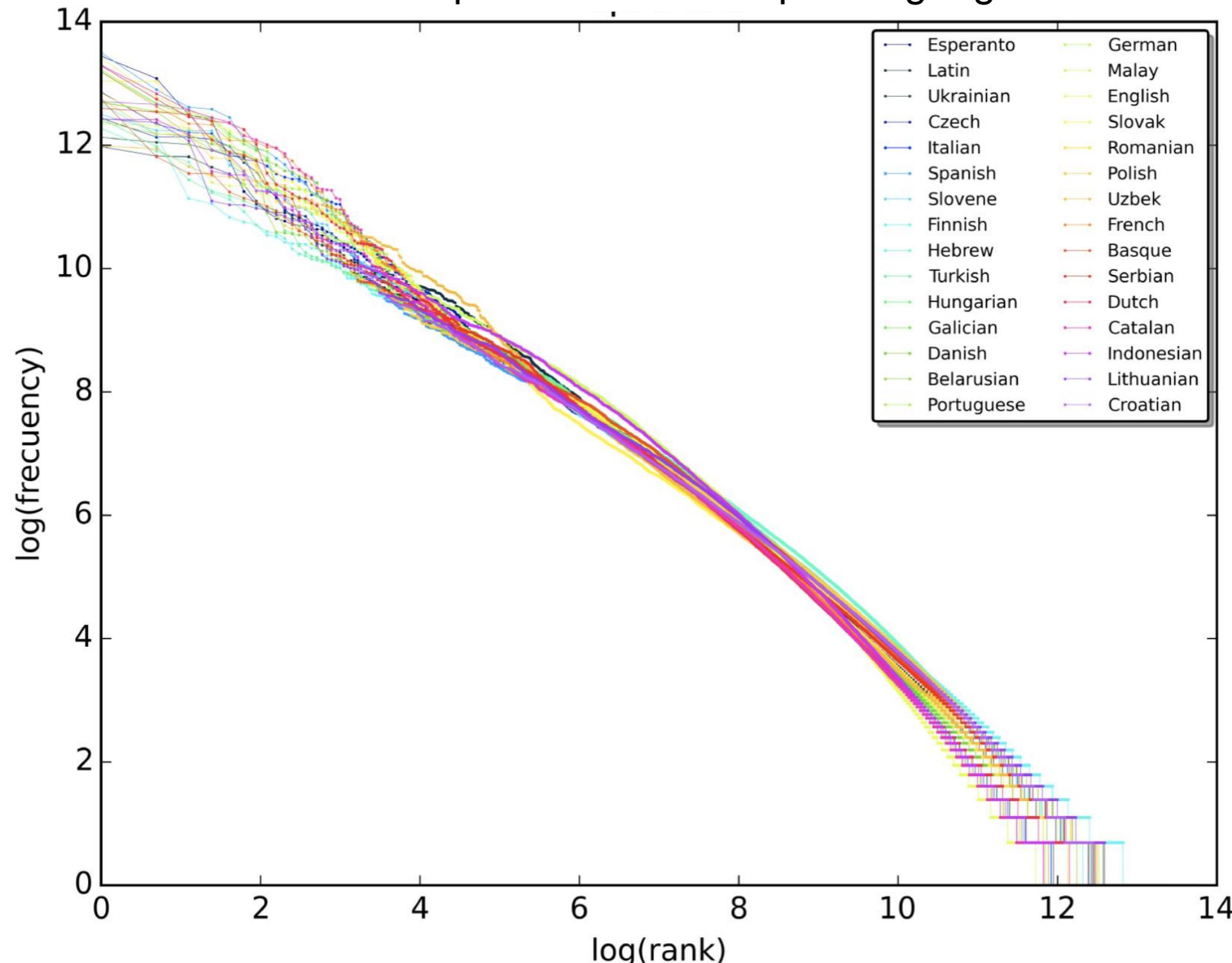
Zipf's law: a word's frequency is approximately inversely proportional to its rank in the word distribution list.

Zipf's Law



Zipf's Law

Wikipedia 10m words per language

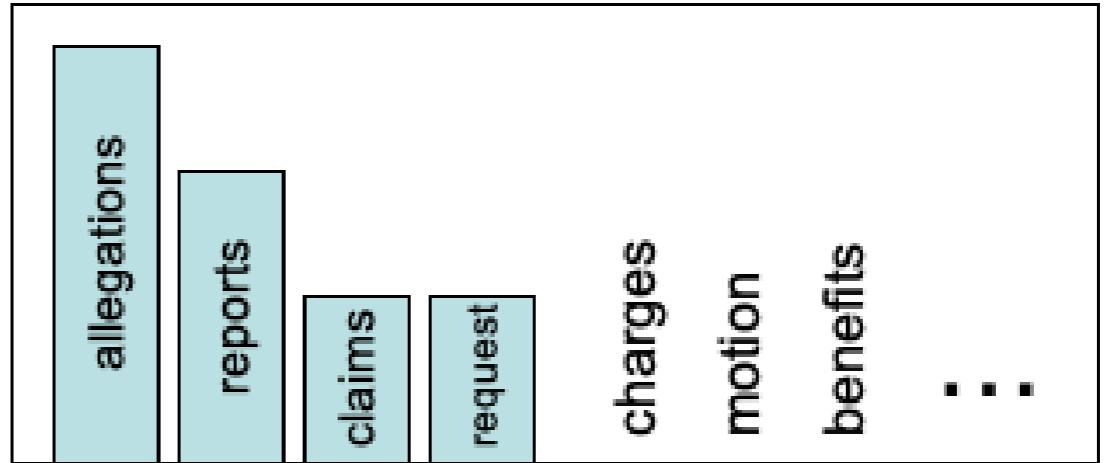


Smoothing

- Smoothing flattens spiky distributions.

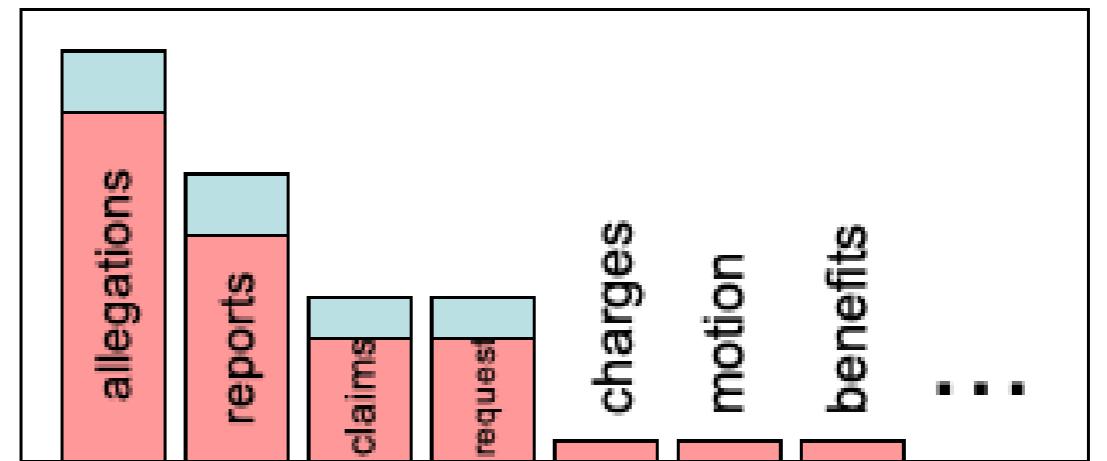
- before $P(w \mid \text{We denied the})$

3 allegations
2 reports
1 claims
1 request
7 total



- after $P(w \mid \text{We denied the})$

2.5 allegations
1.5 reports
0.5 claims
0.4 request
2 UNK
7 total



Smoothing is like Robin Hood: Steal from the rich, give to the poor.

Additive Smoothing

- Classic approach: Laplacian, a.k.a. additive smoothing.

$$P(w_i) = \frac{\text{count}(w_i) + 1}{N + V}$$

- N is the number of tokens, V is the number of types (i.e. size of the vocabulary)

$$P(w|u) = \frac{\text{count}(u, w) + 1}{\text{count}(u) + V}$$

- Inaccurate in practice.

Linear Interpolation

- Use denser distributions of shorter ngrams to “fill in” sparse ngram distributions.

$$p(w|u, v) = \lambda_1 \cdot p_{mle}(w|u, v) + \lambda_2 \cdot p_{mle}(w|v) + \lambda_3 \cdot p_{mle}(w)$$

- Where $\lambda_1, \lambda_2, \lambda_3 > 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.
- Works well in practice (but not a lot of theoretical justification why).
- Parameters can be estimated on development data (for example, using Expectation Maximization).

Discounting

- Idea: set aside some probability mass, then fill in the missing mass using back-off.
- $\text{count}^*(v, w) = \text{count}(v, w) - \beta$ where $0 < \beta < 1$.
- Then for all seen bigrams: $p(w|v) = \frac{\text{count}^*(v, w)}{\text{count}(v)}$
- For each context v the missing probability mass is

$$\alpha(v) = 1 - \sum_{w:c(v,w)>0} \frac{\text{count}^*(v, w)}{\text{count}(v)}$$

- We can now divide this held-out mass between the unseen words (evenly or using back-off).

Katz' Backoff

- Divide the held-out probability mass proportionally to the unigram probability of the unseen words in context v .

$$p(w|v) = \begin{cases} \frac{\text{count}^*(v,w)}{\text{count}(v)} & \text{if } \text{count}(v,w) > 0 \\ \alpha(v) \times \frac{p_{mle}(w)}{\sum_{u:\text{count}(v,u)=0} p_{mle}(u)} & \text{otherwise.} \end{cases}$$

Katz' Backoff for Trigrams

- For trigrams: recursively compute backoff-probability for unseen bigrams. Then distribute the held-out probability mass proportionally to that bigram backoff-probability.

$$p(w|u, v) = \begin{cases} \frac{\text{count}^*(u, v, w)}{\text{count}(u, v)} & \text{if } \text{count}(u, v, w) > 0 \\ \alpha(u, v) \times \frac{p_{BO}(w|v)}{\sum_{z: \text{count}(v, z) = 0} p_{BO}(z|v)} & \text{otherwise.} \end{cases}$$

- where: $\alpha(u, v) = 1 - \sum_{w: \text{count}(u, v, w) > 0} \frac{\text{count}^*(u, v, w)}{\text{count}(u, v)}$
- Often combined with Good-Turing smoothing.

Evaluating n-gram models

- Extrinsic evaluation: Apply the model in an application (for example language classification). Evaluate the application.
- Intrinsic evaluation: measure how well the model approximates unseen language data.
 - Can compute the probability of each sentence according to the model. Higher probability -> better model.
 - Typically we compute *Perplexity instead.*

Perplexity

- Perplexity (per word) measures how well the ngram model predicts the sample.
- Given a corpus of ‘m’ sentences ‘ s_i ’, where ‘M’ is total number of tokens in the corpus
 - Perplexity is defined as 2^{-l} , where
$$l = \frac{1}{M} \sum_{i=1}^m \log_2 p(s_i)$$
.
 - Lower perplexity = better model. Intuition:
 - Assume we are predicting one word at a time.
 - With uniform distribution, all successor words are equally likely. Perplexity is equal to vocabulary size.
 - Perplexity can be thought of as “effective vocabulary size”.

Natural Language Processing

Lecture 4: Sequence Labeling with Hidden Markov
Models. Part-of-Speech Tagging.

2/6/2020

COMS W4705
Yassine Benajiba

Garden-Path Sentences

- *The horse raced past the barn.*
- *The horse raced past the barn fell.*
- *The old dog the footsteps of the young.*
- *The cotton clothing is made of grows in Mississippi.*

Garden-Path Sentences

- Why does this happen?

- **raced** can be a past tense verb or a past participle (indicating passive voice).
 - The verb interpretation is more likely before *fell* is read.

Garden-Path Sentences

- Why does this happen?

past participle
VBN VBD

[The horse **raced** past the barn] fell

NP

- **raced** can be a past tense verb or a past participle (indicating passive voice).
 - Once *fell* is read, the verb interpretation is impossible.

Garden-Path Sentences

- Why does this happen?

adjective
JJ NN
[The old dog] [the footsteps of the young]
NP NP

- **dog** can be a noun or a verb (plural, present tense)

Garden-Path Sentences

- Why does this happen?

NN VB
[The old] dog [the footsteps of the young]
NP NP

- **dog** can be a noun or a verb (plural, present tense)

Parts-of-Speech

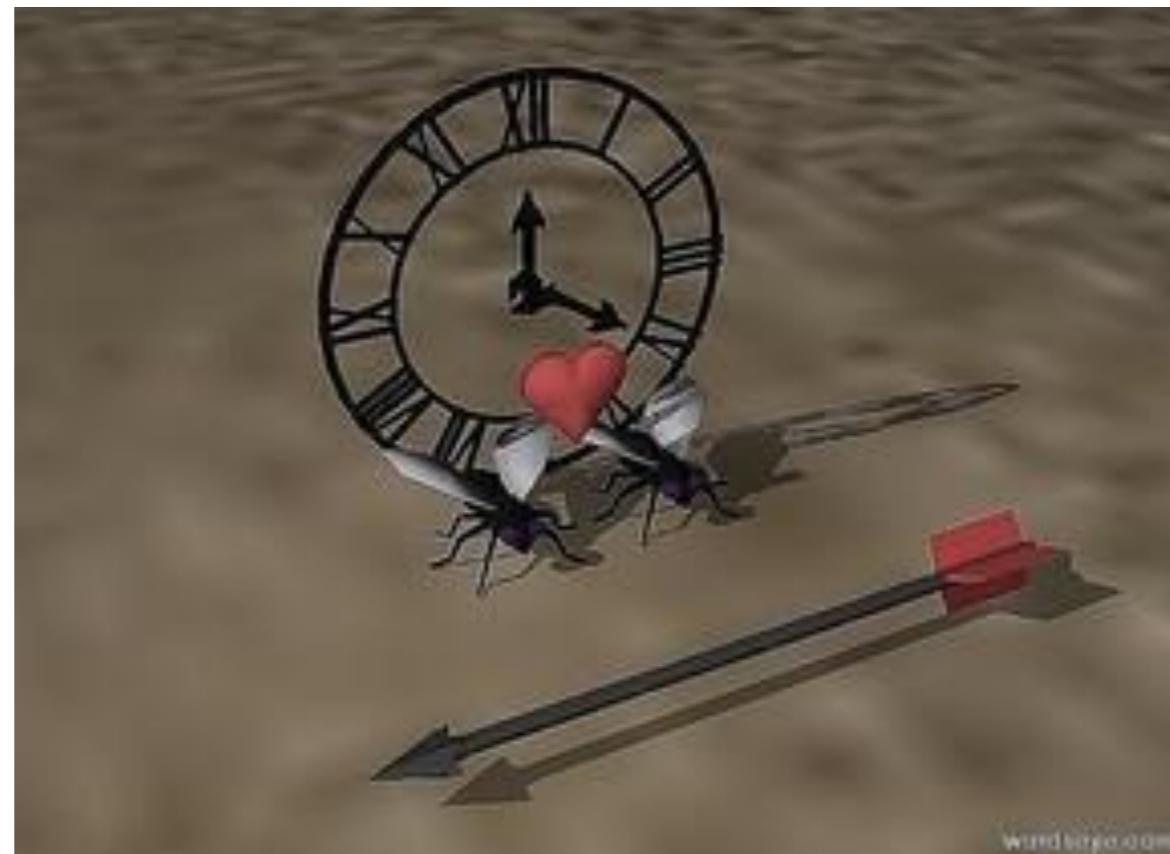
- Classes of words that behave alike:
 - Appear in similar contexts.
 - Perform a similar grammatical function in the sentence.
 - Undergo similar morphological transformations.
- ~9 traditional parts-of-speech:
 - noun, pronoun, determiner, adjective, verb, adverb, preposition, conjunction, interjection

Syntactic Ambiguities and Parts-of-Speech

- | N / V? | N / V? | V / Preposition |
|---------------|---------------|------------------------|
| • <i>Time</i> | <i>flies</i> | <i>like</i> |
| | | <i>an arrow.</i> |

Syntactic Ambiguities and Parts-of-Speech

- [Time flies] like *an arrow.*
 N N V
 NP



Why do we need P.O.S.?

- Interacts with most levels of linguistic representation.
- Speech processing:
 - *object, object*
 - *content, content*
- Syntactic parsing
- ...
- P.O.S. tag-set should contain morphological and maybe syntactic information.

Penn Treebank Tagset

CC	Coordinating conjunction	PRP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential <i>there</i>	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	<i>to</i>
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WP	Wh-pronoun
PDT	Predeterminer	WP\$	Possessive wh-pronoun
POS	Possessive ending	WRB	Wh-adverb
PRP	Personal pronoun		plus punctuation symbols

P.O.S. Tagsets

- Tagset is language specific.
- Some languages capture more morphological information which should be reflected in the tag set.
- “Universal Part Of Speech Tags?”
 - Petrov et al. 2011: Mapping of 25 language specific tag-sets to a common set of 12 universal tags

Part-of-Speech Tagging

- Goal: Assign a part-of-speech label to each word in a sentence.

<i>DT</i>	<i>NN</i>	<i>VBD</i>	<i>DT</i>	<i>NNS</i>	<i>IN</i>	<i>DT</i>	<i>NN</i>	.
<i>the</i>	<i>koala</i>	<i>put</i>	<i>the</i>	<i>keys</i>	<i>on</i>	<i>the</i>	<i>table</i>	.

- This is an example of a **sequence labeling** task.
- Think of this as a translation task from a sequence of words $(w_1, w_2, \dots, w_n) \in V^*$, to a sequence of tags $(t_1, t_2, \dots, t_n) \in T^*$.

Determining Part-of-Speech

- *A blue seat / A child seat*: noun or adj?
- Syntactic tests:
 - *A very **blue** seat*
 - *This seat is **blue***
- Morphological Tests
 - *bluer*
 - **childer*

Determining Part-of-Speech

- Preposition or Particle?

- *He threw **out** the garbage.*

out is a particle

*He threw the garbage **out**.*

- *He threw the garbage **out** the door.*

out is a preposition

He threw the garbage the door **out*

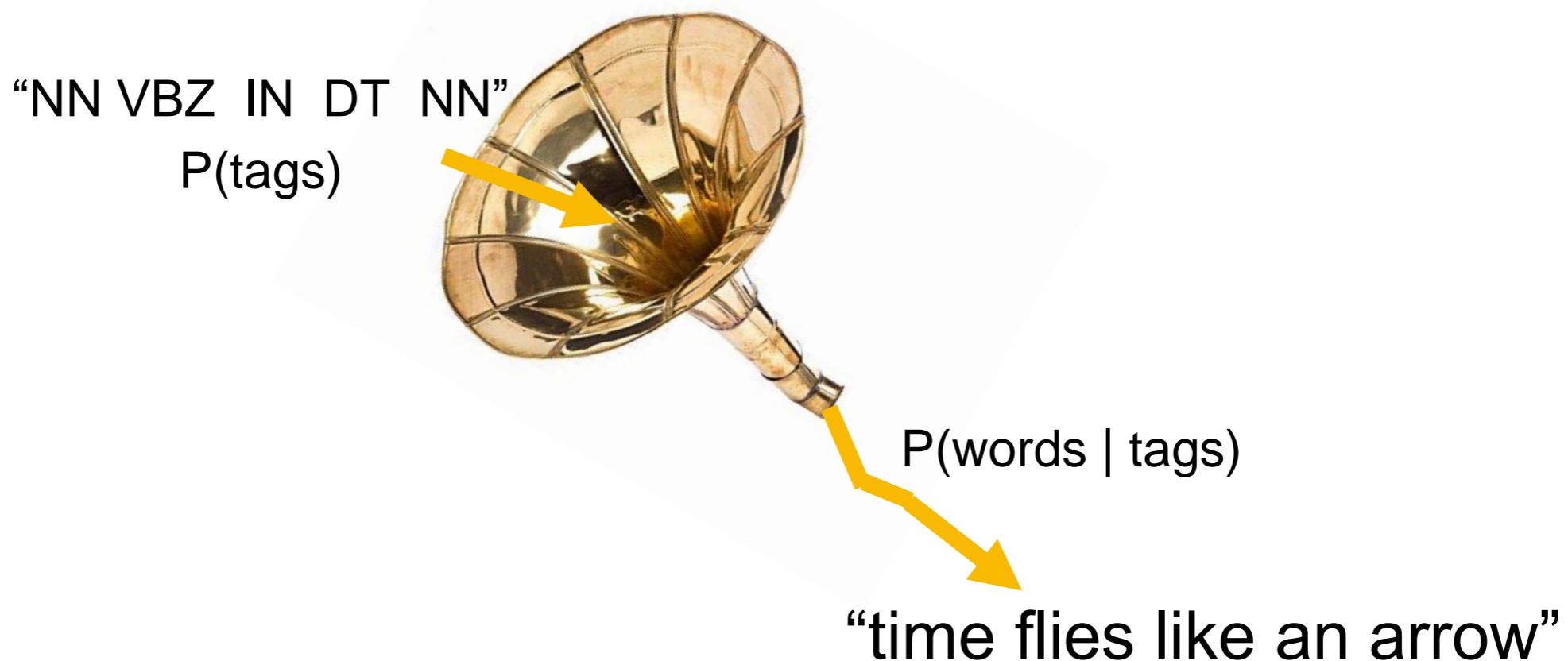
Part-of-Speech Tagging

- Goal: Translate from a sequence of words $(w_1, w_2, \dots, w_n) \in V^*$, to a sequence of tags $(t_1, t_2, \dots, t_n) \in T^*$.
- NLP is full of translation problems from one structure to another. Basic solution:
 - For each translation step:
 1. Construct search space of possible translations.
 2. Find best paths through this space (decoding) according to some performance measure.

Bayesian Inference for Sequence Labeling

- Recall Bayesian Inference (Generative Models): Given some observation, infer the value of some *hidden variable*. (see Naive Bayes')
- We can apply this approach to sequence labeling:
 - Assume each word w_i in the observed sequence $(w_1, w_2, \dots, w_n) \in V^*$ was generated by some hidden variable t_i .
 - Infer the most likely sequence of hidden variables given the sequence of observed words.

Noisy Channel Model



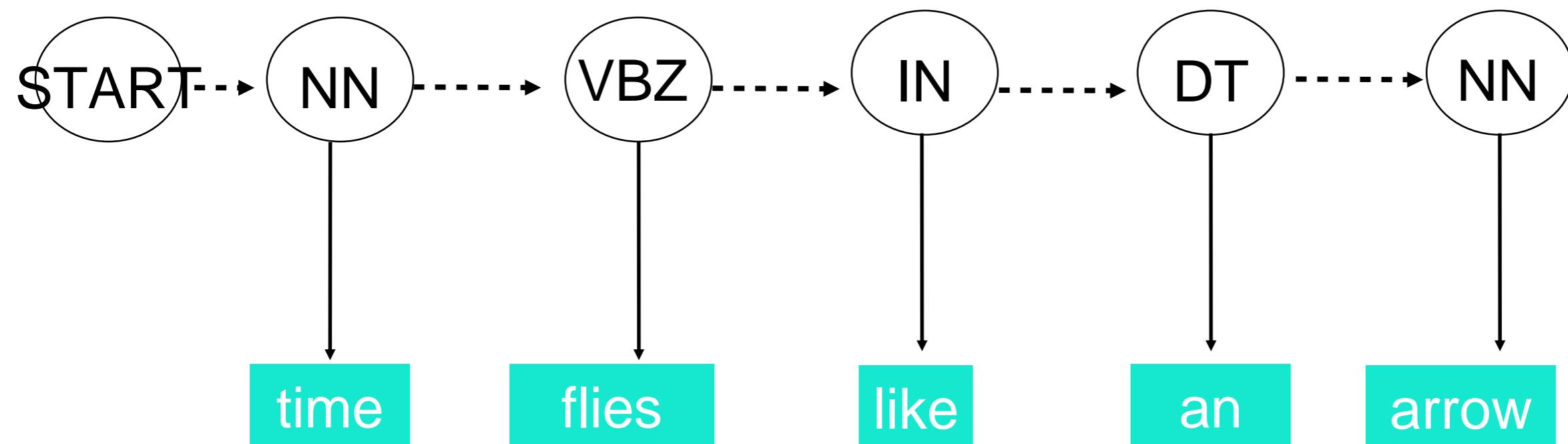
- Goal: figure out what the original input to the the channel was. Use Bayes' rule:

$$\arg \max_{tags} P(tags|words) = \arg \max_{tags} \frac{P(tags) \cdot P(word|tags)}{P(words)}$$

- This model is used widely (speech recognition, MT)

Hidden Markov Models (HMMs)

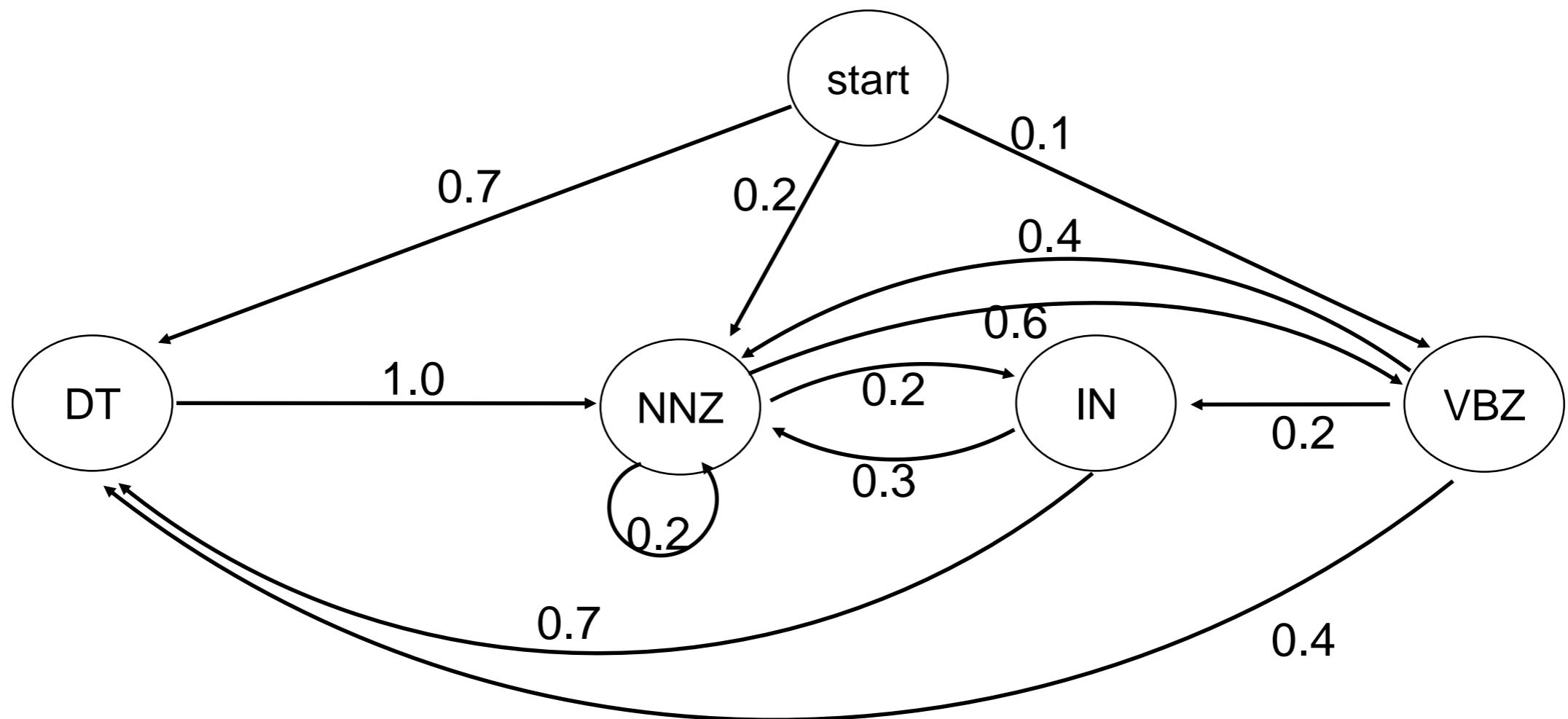
- Generative (Bayesian) probability model.
Observations: sequences of words.
Hidden states: sequence of part-of-speech labels.



- Hidden sequence is generated by an n-gram language model (typically a bi-gram model)

$$t_0 = \text{START} \quad P(t_1, t_2, \dots, t_n) = \prod_{i=1}^n P(t_i | t_{i-1})$$

Markov Chains



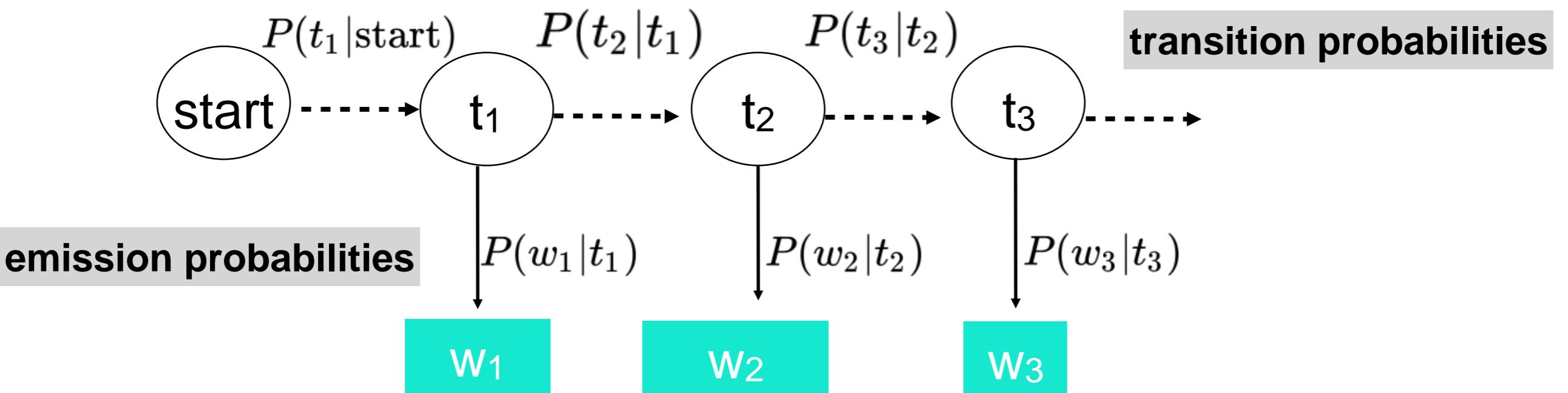
- A **Markov chain** is a sequence of random variables X_1, X_2, \dots
- The domain of these variables is a set of states.
- **Markov assumption:** Next state depends only on current state.

$$P(X_{n+1} | X_1, X_2, \dots, X_n) = P(X_{n+1} | X_n)$$

- This is a special case of a weighted finite state automaton (WFSA).

Hidden Markov Models (HMMs)

- There are two types of probabilities:
Transition probabilities and Emission Probabilities.



$$P(t_1, t_2, \dots, t_n, w_1, w_2, \dots, w_n) =$$

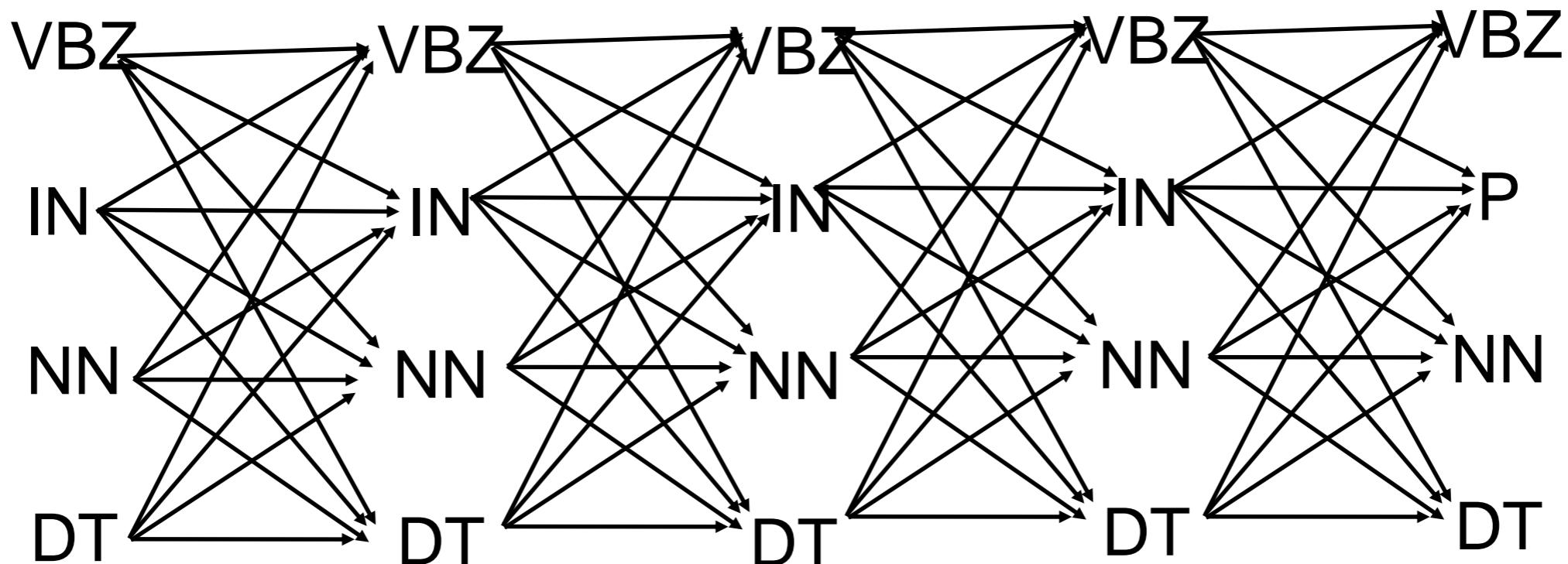
$$P(t_1|\text{start})P(w_1|t_1)P(t_2|t_1)P(w_2|t_2)\cdots P(t_n|t_{n-1})P(w_n|t_n)$$

$$= \prod_{i=1}^n P(t_i|t_{i-1})P(w_i|t_i)$$

Important Tasks on HMMs

- **Decoding:** Given a sequence of words, find the *most likely* probability sequence.
(Bayesian inference using *Viterbi algorithm*).
- **Evaluation:** Given a sequence of words, find the *total probability for this word sequence* given an HMM.
Note that we can view the HMM as another type of language model. (Forward algorithm)
- **Training:** Estimate emission and transition probabilities from training data. (MLE, Forward-Backward a.k.a Baum-Welch algorithm)

Decoding HMMs



time

flies

like

an

arrow

Goal: Find the path with the highest total probability (given the words)

$$\arg \max_{t_1, \dots, t_n} \prod_{i=1}^n P(t_i | t_{i-1}) P(w_i | t_i)$$

There are d^n paths for n words and d tags.

Emission Probabilities

- $P(\text{time} \mid \text{VB}) = 0.2$
 $P(\text{flies} \mid \text{VB}) = 0.3$
 $P(\text{like} \mid \text{VB}) = 0.5$
- $P(\text{time} \mid \text{NN}) = 0.3$
 $P(\text{flies} \mid \text{NN}) = 0.2$
 $P(\text{arrow} \mid \text{NN}) = 0.5$
- $P(\text{like} \mid \text{P}) = 1.0$
- $P(\text{an} \mid \text{DT}) = 1.0$

Viterbi Algorithm

.1 x .2 = .02 VBZ

0 IN

.2 x .3 = .06 NN

0 DT

time

VBZ

IN

NN

DT

flies

VBZ

IN

NN

DT

like

VBZ

IN

NN

DT

an

VBZ

IN

NN

DT

arrow

- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm

.02 VBZ

VBZ

VBZ

IN

.06 NN

NN

NN

DT

time

flies

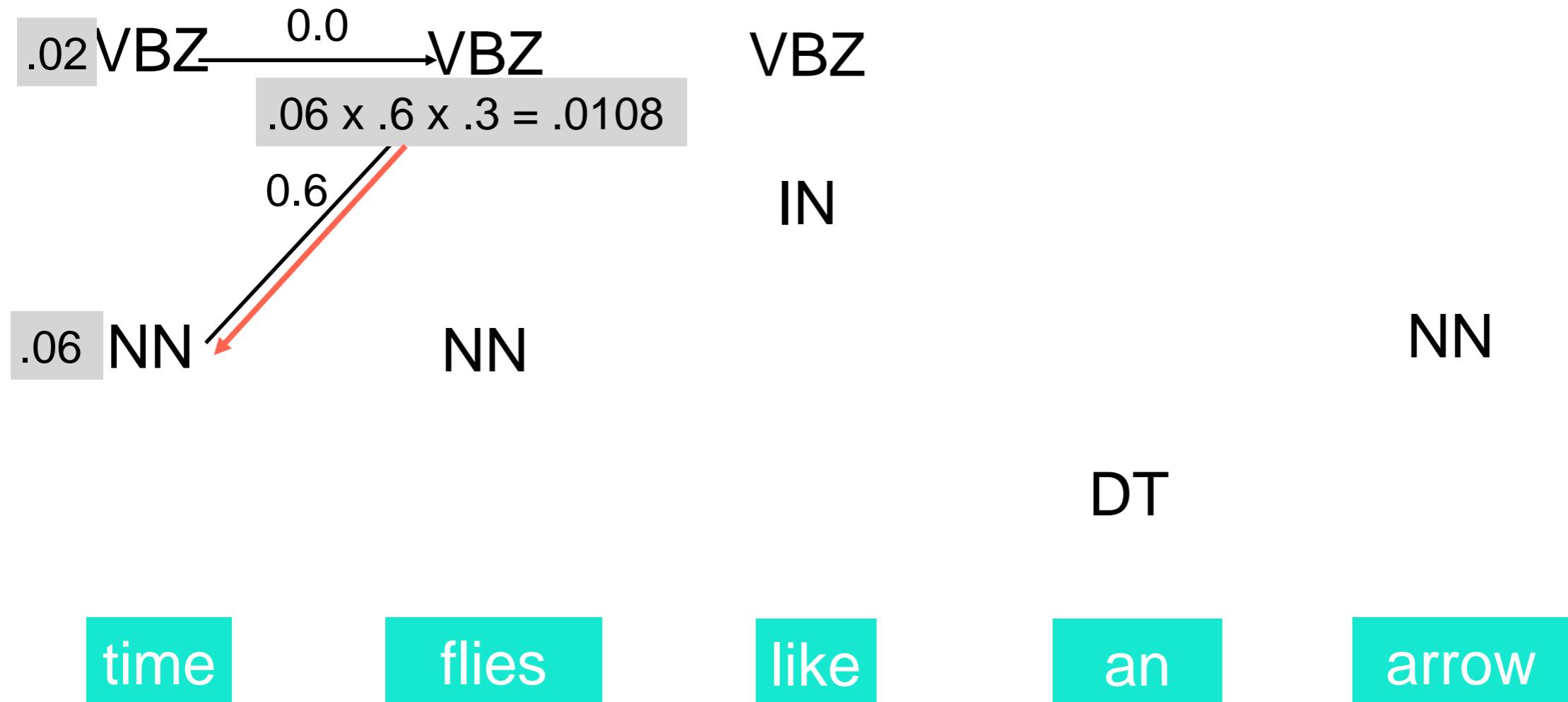
like

an

arrow

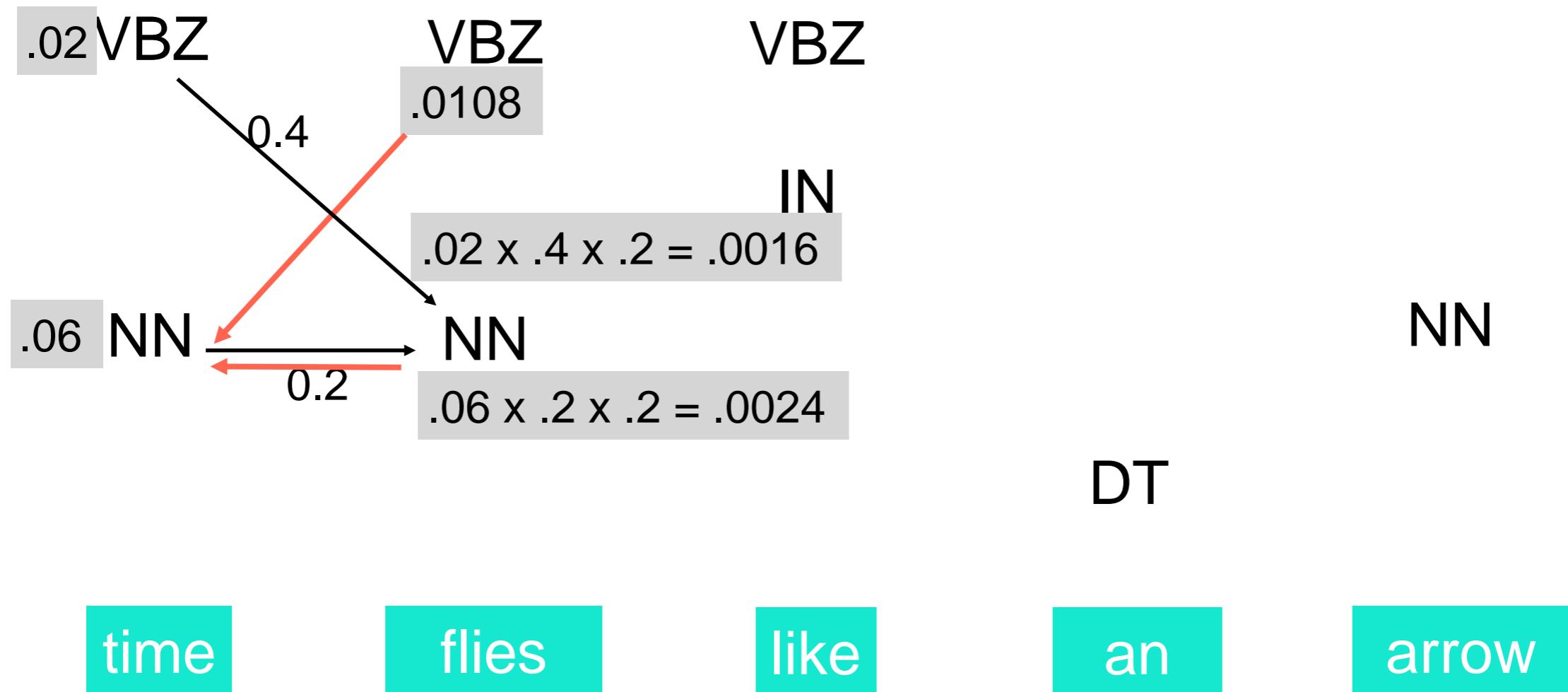
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



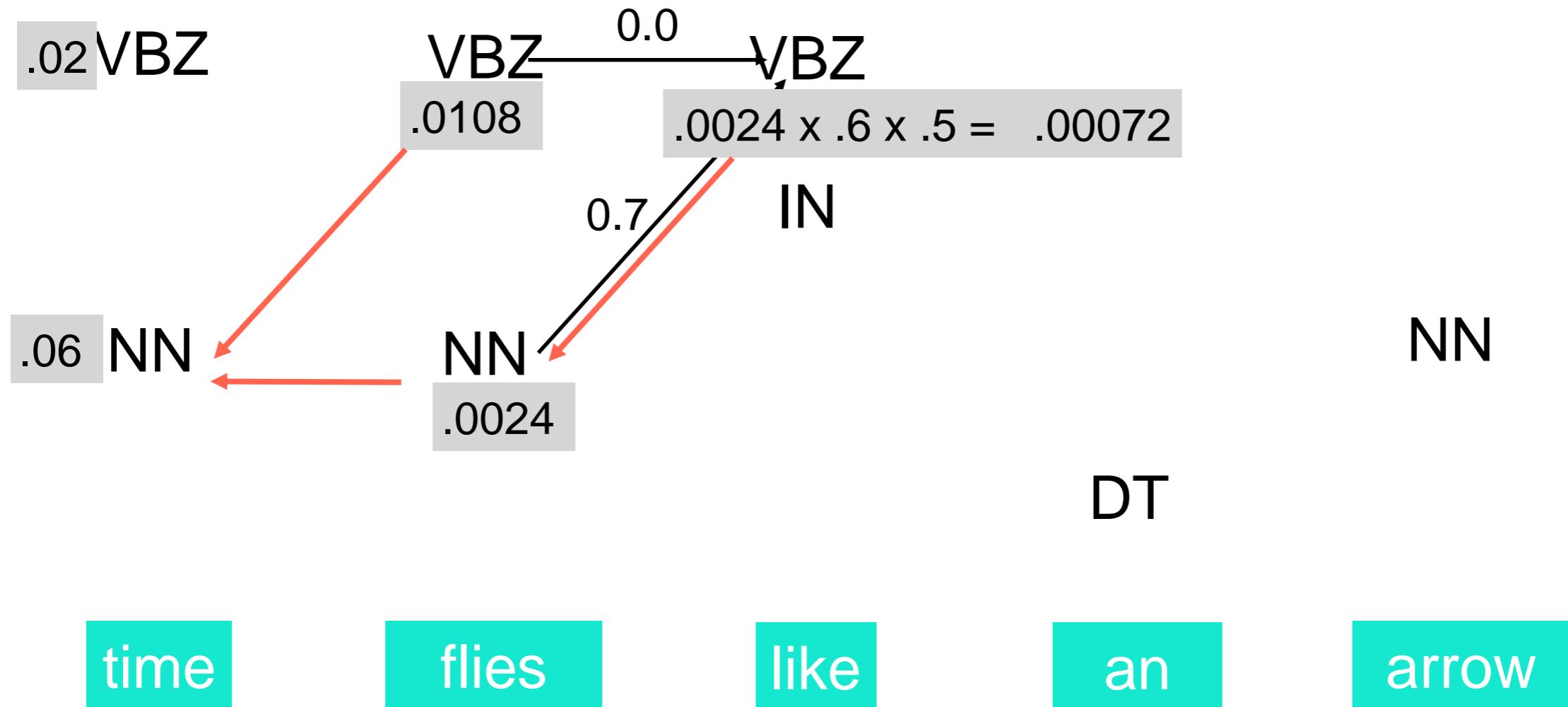
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



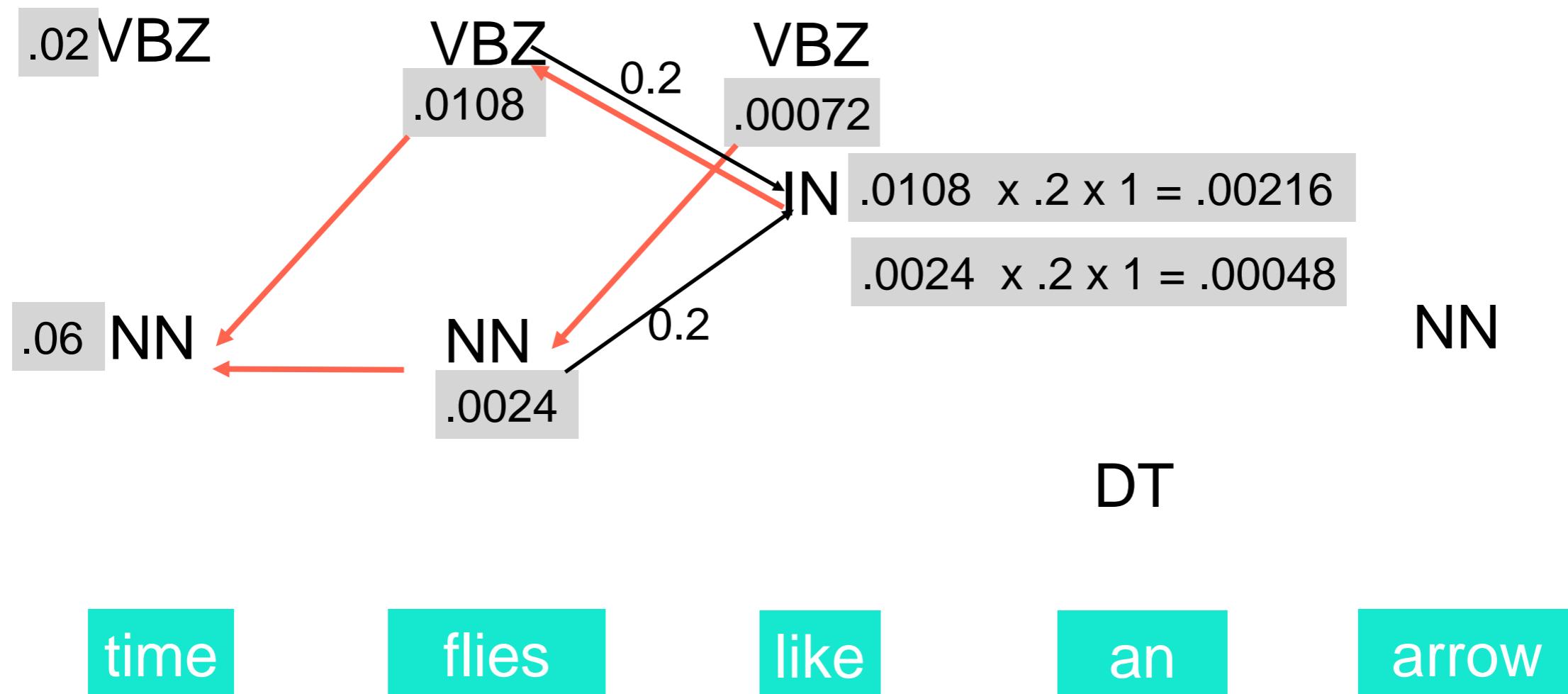
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



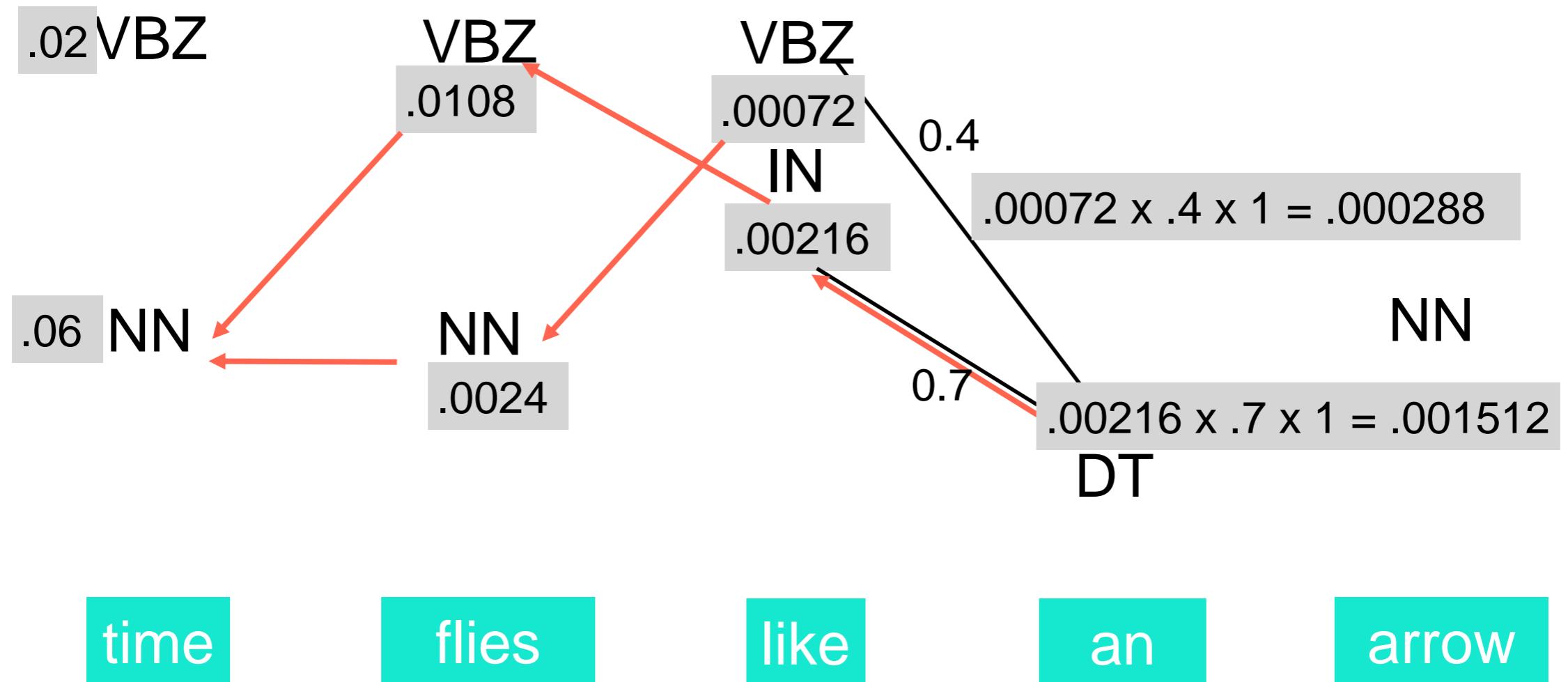
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



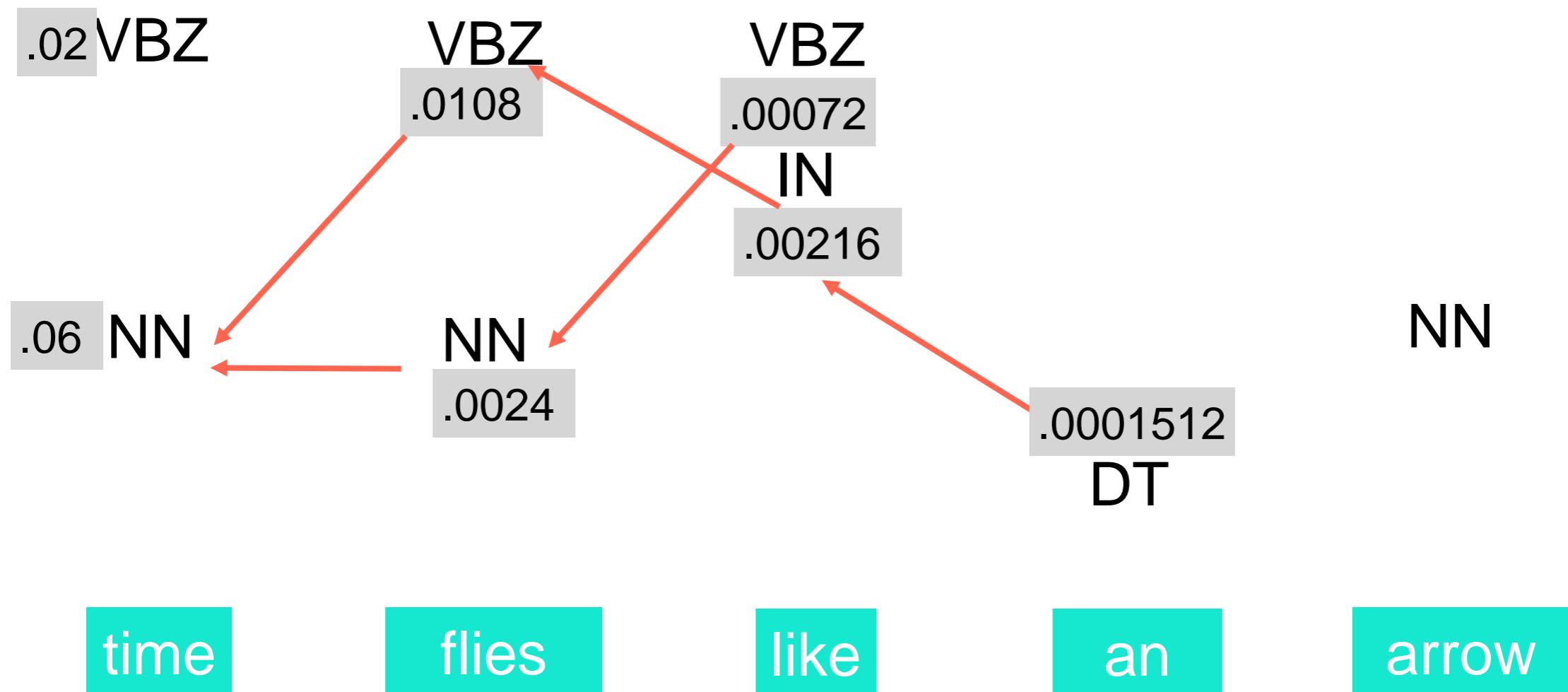
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



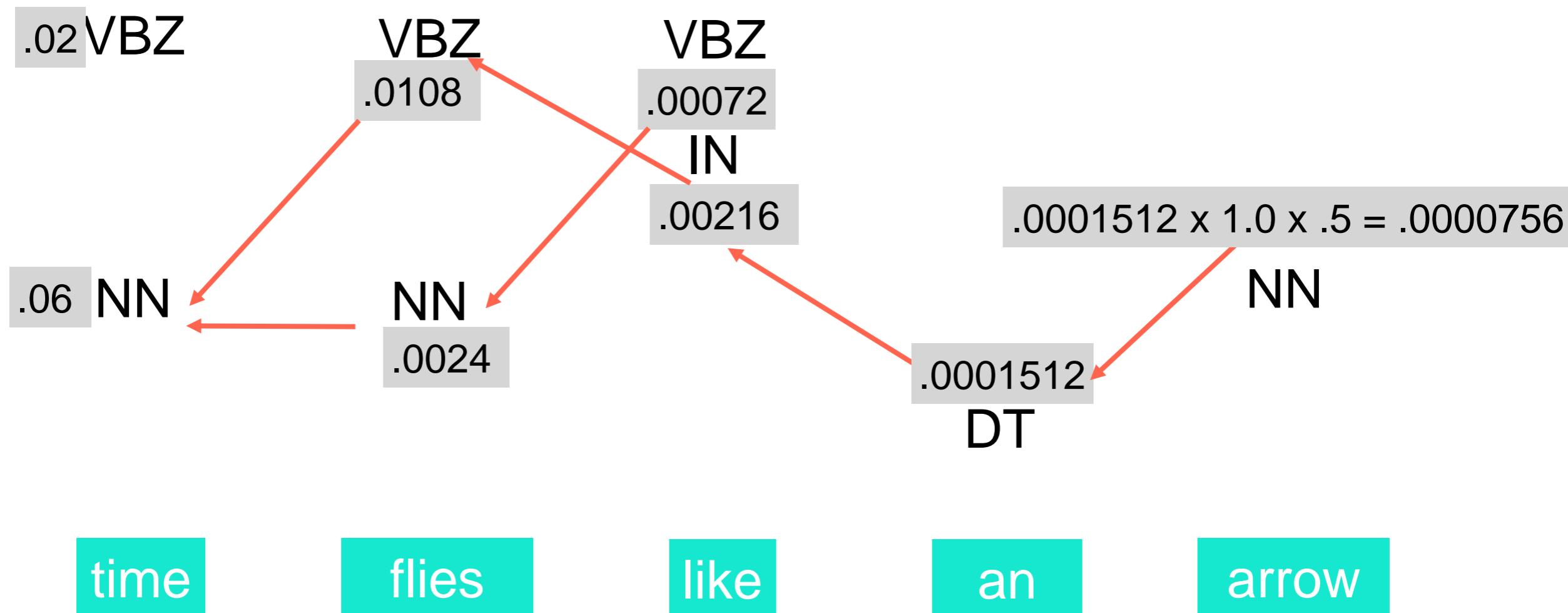
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



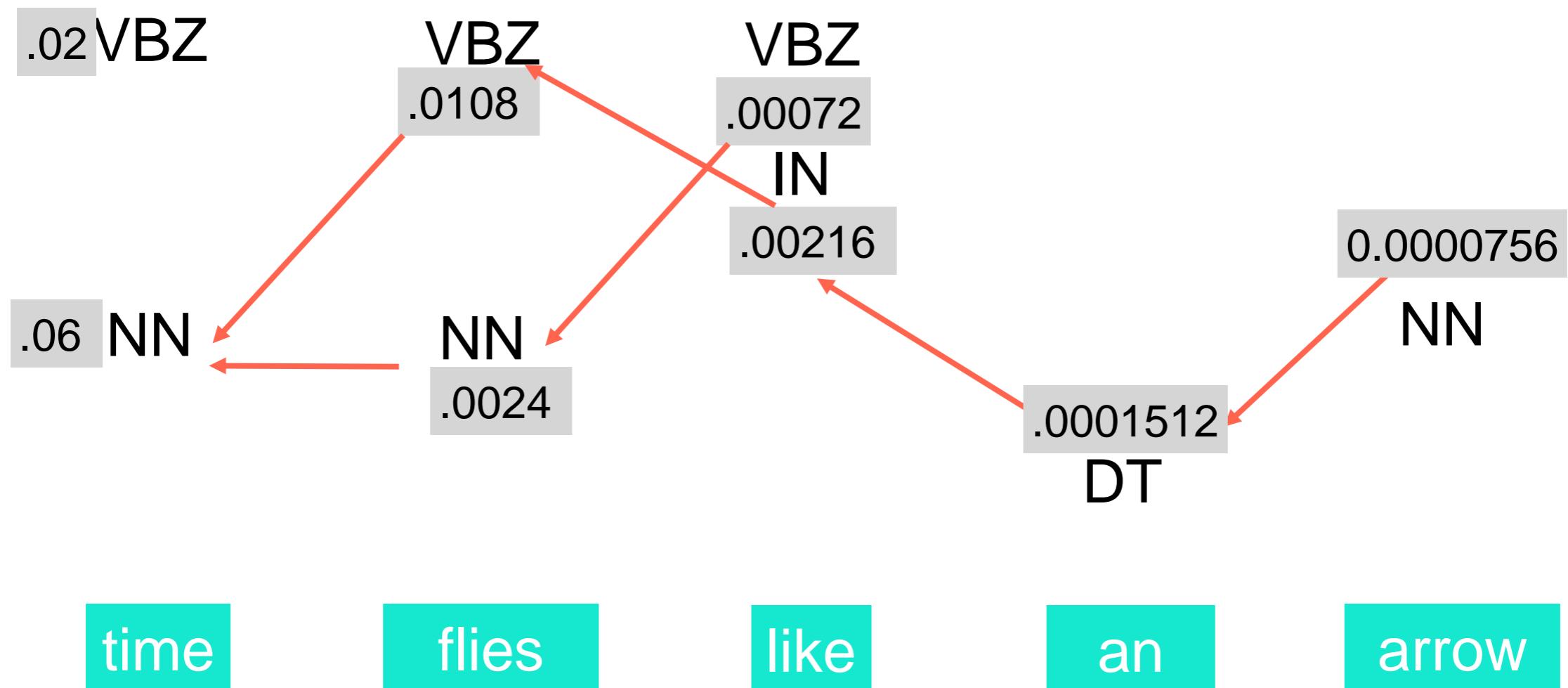
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



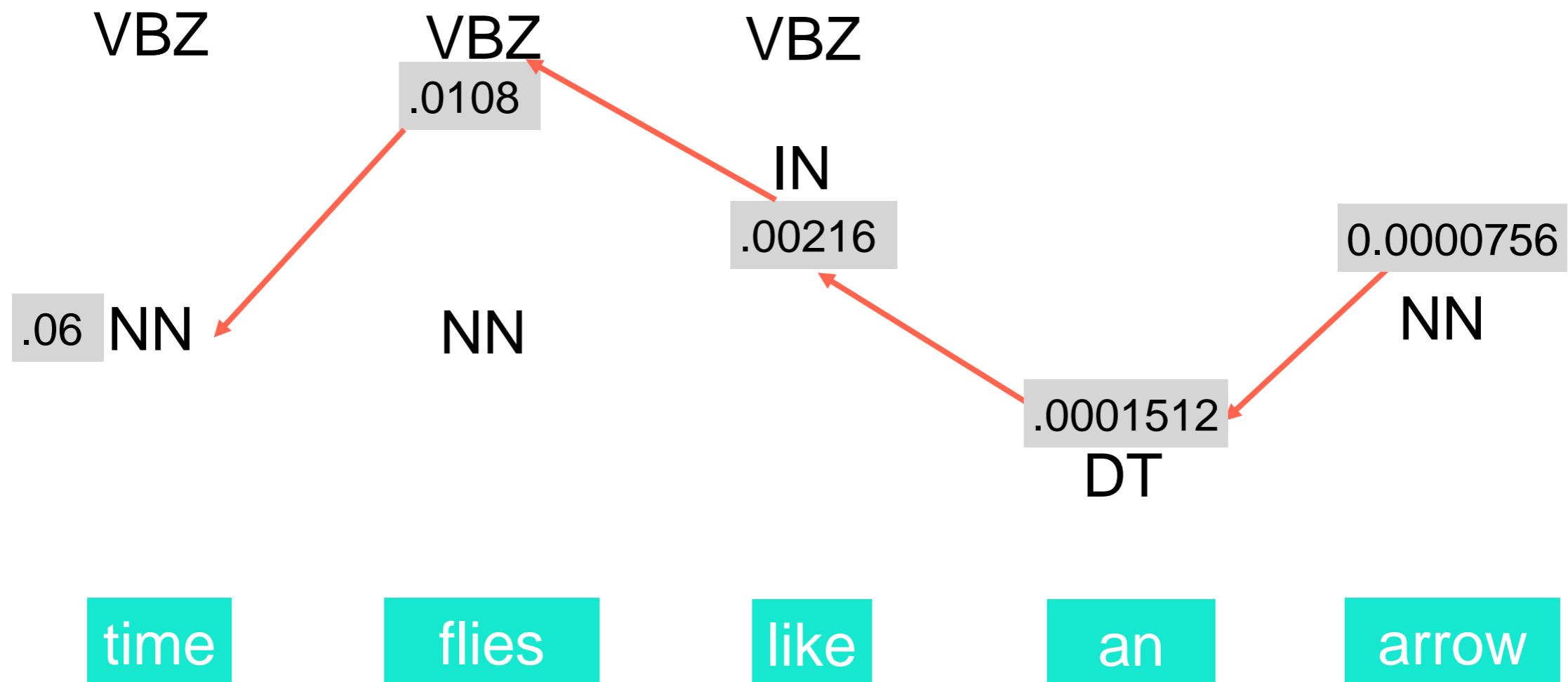
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



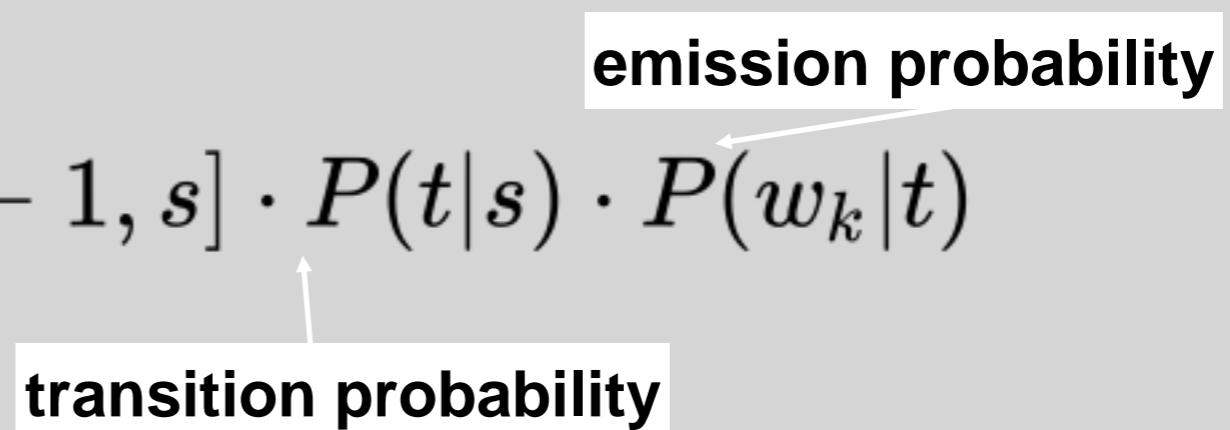
- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm



- Idea: Because of the Markov assumption, we only need the probabilities for X_n to compute the probabilities for X_{n+1} . This suggests a **dynamic programming** algorithm.

Viterbi Algorithm

- **Input:** Sequence of observed words w_1, \dots, w_n
 - Create a table π , such that each entry $\pi[k, t]$ contains the score of the highest-probability sequence ending in tag t at time k .
 - initialize $\pi[0, \text{start}] = 1.0$ and $\pi[0, t] = 0.0$ for all tags $t \in T$.
 - *for* $k=1$ to n :
 - *for* $t \in T$:
 - $\pi[k, t] \leftarrow \max_s \pi[k - 1, s] \cdot P(t|s) \cdot P(w_k|t)$
 - *return* $\max_s \pi[n, s]$
- 

Trigram Language Model

- Instead of using a unigram context $P(t_i | t_{i-1})$ use a bigram context $P(t_i | t_{i-2} t_{i-1})$
 - Think of this as having states that represent *pairs of tags*.
- So the HMM probability for a given tag and word sequence is:
$$\prod_{i=1}^n P(t_i | t_{i-2} t_{i-1}) P(w_i | t_i)$$
- Need to handle data sparseness when estimating transition probabilities (for example using backoff or linear interpolation)

More POS tagging tricks

- It is also often useful in practice to add an end-of-sentence marker (just like we did for n-gram language models).

$$P(t_1, \dots, t_n, w_1, \dots, w_n) = \left[\prod_{i=1}^n P(t_i | t_{i-2} t_{i-1}) P(w_i | t_i) \right] P(t_{n+1} | t_n)$$

where $t_{-1} = t_0 = \text{START}$ and $t_{n+1} = \text{STOP}$.

- Another useful trick is to replace words with “pseudo words” representing an entire class.
 - For example: replace {"01", "85", "90", ...} with *twoDigitNumber*
replace {"1985", "2018", ...} with *fourDigitNumber*
replace {"1", "1.0", "234.3", ...} with *otherNum*
replace {"IBM", "DNC", ...} with *allCaps* etc.

Using a smoothed trigram HMM model with these tricks, we can build a tagger that is close to the state-of-the art (~97% accuracy on the Penn Treebank).

HMMs as Language Models

- We can also use an HMM as language models (language generation, MT, ...), i.e. **evaluate** $P(w_1, \dots, w_n)$ for a given sentence.

What is the advantage over a plain word n-gram model?

- Problem: There are many tag-sequences that could have generated w_1, \dots, w_n .

$$P(w_1, \dots, w_n, t_1, \dots, t_n) = \prod_{i=1}^n P(t_i | t_{i-1}) P(w_i | t_i)$$

- This is an example of **spurious ambiguity**.
- Need to compute:
$$P(w_1, \dots, w_n) = \sum_{t_1, \dots, t_n} P(w_1, \dots, w_n, t_1, \dots, t_n)$$

$$= \sum_{t_1, \dots, t_n} \left[\prod_{i=1}^n P(t_i | t_{i-1}) P(w_i | t_i) \right]$$

Forward Algorithm

- **Input:** Sequence of observed words w_1, \dots, w_n
- Create a table π , such that each entry $\pi[k, t]$ contains the score of the highest-probability sequence ending in tag t at time k .
- initialize $\pi[0, \text{start}] = 1.0$ and $\pi[0, t] = 0.0$ for all tags $t \in T$.
- *for* $k=1$ to n :
 - *for* $t \in T$:
 - $\pi[k, t] \leftarrow \sum_s \pi[k - 1, s] \cdot P(t|s) \cdot P(w_k|t)$
 - *return* $\sum_s \pi[n, s]$

Named Entity Recognition as Sequence Labeling

- Use 3 tags:
 - O - outside of named entity
 - I - inside named entity
 - B - first word (beginning) of named entity

... O O B I O
identification of tetronic acid in ...

- Other encodings are possible (for example, NE-type specific)
- This can also be used for phrase chunking.

Natural Language Processing

Lecture 5: Introduction to Syntax and
Formal Languages.

2/14/2020

COMS W4705
Yassine Benajiba

Sentences: the good, the bad, and the ugly

- Some good sentences:
 - *the boy likes a girl*
 - *the small girl likes a big girl*
 - *a very small nice boy sees a very nice boy*
- Some bad sentences:
 - *the boy the girl likes*
 - *small boy likes nice girl*
- Ugly word salad: *very like nice the girl boy*

Syntax as an Interface

- Syntax can be seen as the interface between morphology (structure of words) and semantics.
- Why treat syntax separately from semantics?
 - Can judge if a sentence is grammatical or not, even if it doesn't make sense semantically.

Colorless green ideas sleep furiously.

**Sleep ideas furiously colorless green.*

Key Concepts of Syntax

- Constituency and Recursion.
- Dependency.
- Grammatical Relations.
- Subcategorization.
- Long-distance dependencies.

Constituents

- A constituent is a group of words that behave as a single unit (within a hierarchical structure).
- Noun-Phrase examples:
 - [they], [the woman], [three parties from Brooklyn], [a high-class spot such as Mindy's], [the horse raced past the barn]
 - Noun phrases can appear before verbs (among other things) and they must be complete:
 - **from arrive...*
 - **the is*
 - **spot sat....*

Constituency Tests

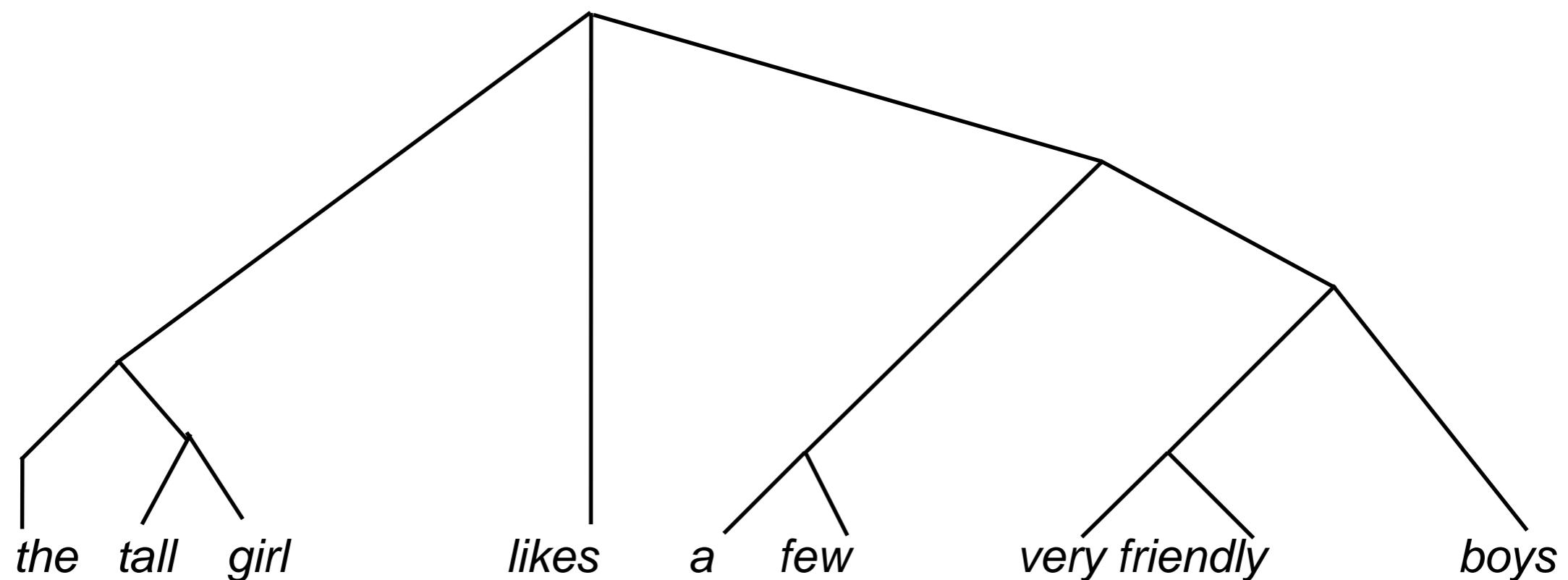
- *On September seventeenth, I'd like to fly to New York.*
- *I'd like to fly to New York on September seventeenth.*
- *I'd like to fly on September seventeenth to New York.*
- **On I'd like to fly to New York September seventeenth.*
- **On September I'd like to fly seventeenth to New York.*

More Constituency Tests

- There is a great number of constituency tests. They typically involve moving constituents around or replacing them.
- Topicalization:
 - *I won't eat **that pizza*** ***That pizza**, I won't eat* ****pizza** I won't eat **that***
- Pro-form Substitution:
 - *I don't know **the man who sent flowers.*** *I don't know **him.***
I don't know **him flowers.*
- Wh-question test.
 - ***Where** would you like to fly on September seventeenth?*
 - ***When** would you like to fly to New York?*

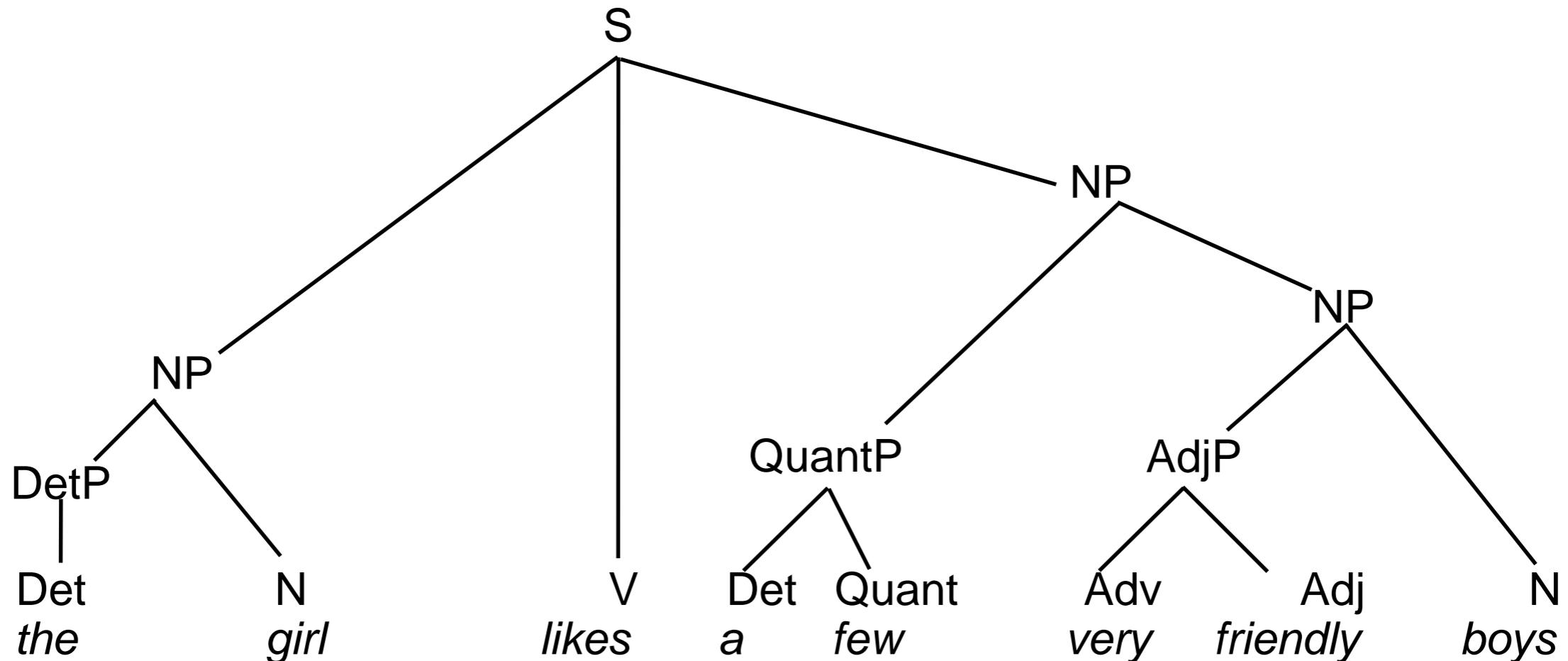
Sentence Structure as Trees

- [the tall girl likes a few very friendly boys]
- [[the tall girl] likes [a few very friendly boys]]
- [[the] tall girl] likes [[a few] [very friendly] boys]]



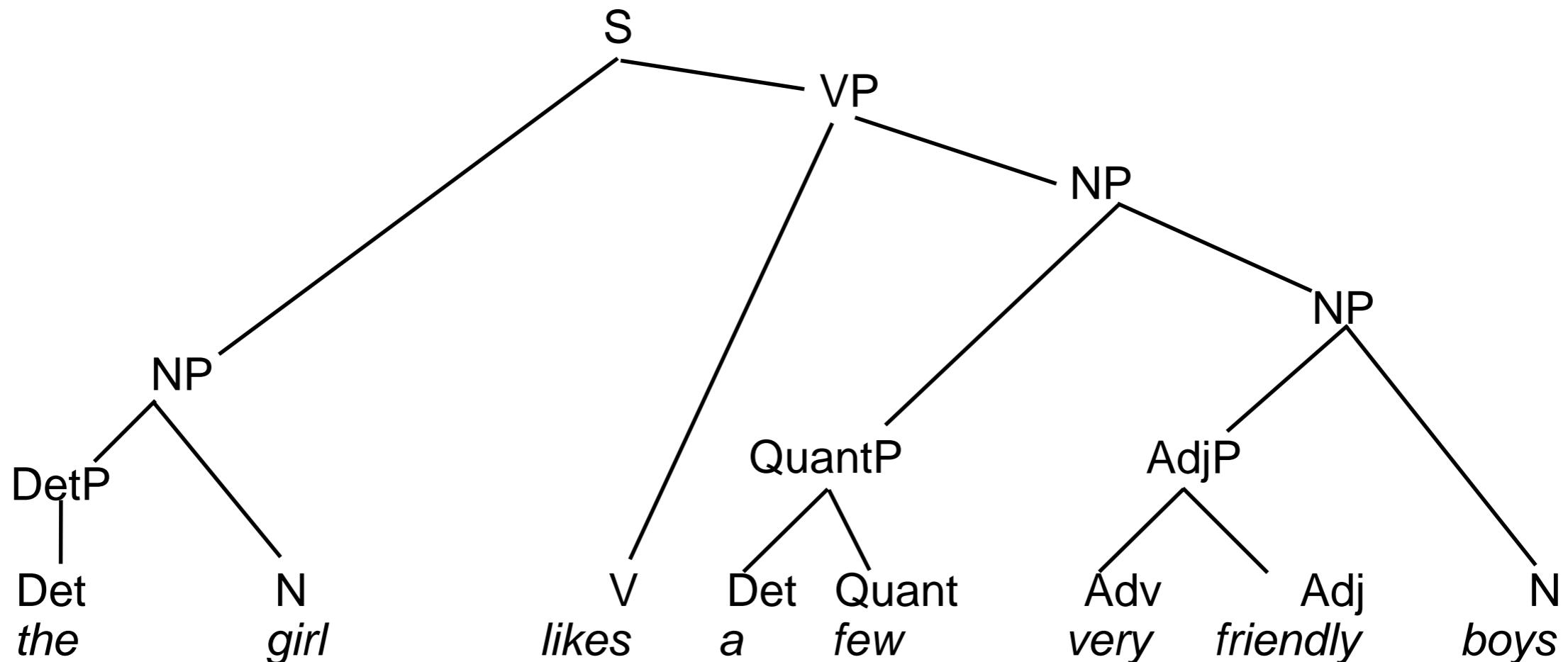
Constituent Labels

- Choose constituents so each one has one non-bracketed word: the **head**.
- Category of Constituent: XP, where X is the part-of-speech of the head
NP, VP, AdjP, AdvP, DetP



Constituent Labels

- Choose constituents so each one has one non-bracketed word: the **head**.
- Category of Constituent: XP, where X is the part-of-speech of the head
NP, VP, AdjP, AdvP, DetP



Review: Constituency

The students easily completed the difficult NLP homework.

Which constituents can you identify? What tests could you use?

Recursion in Language

- One of the most important attributes of Natural Languages is that they are **recursive**.
 - *He made pie [with apples [from the orchard [near the farm [in ...]]]]*
 - *[The mouse [the cat [the dog chased]] ate] died.*
- There are infinitely many sentences in a language, but in predictable structures.
- How do we model the set of sentences in a language and their structure?

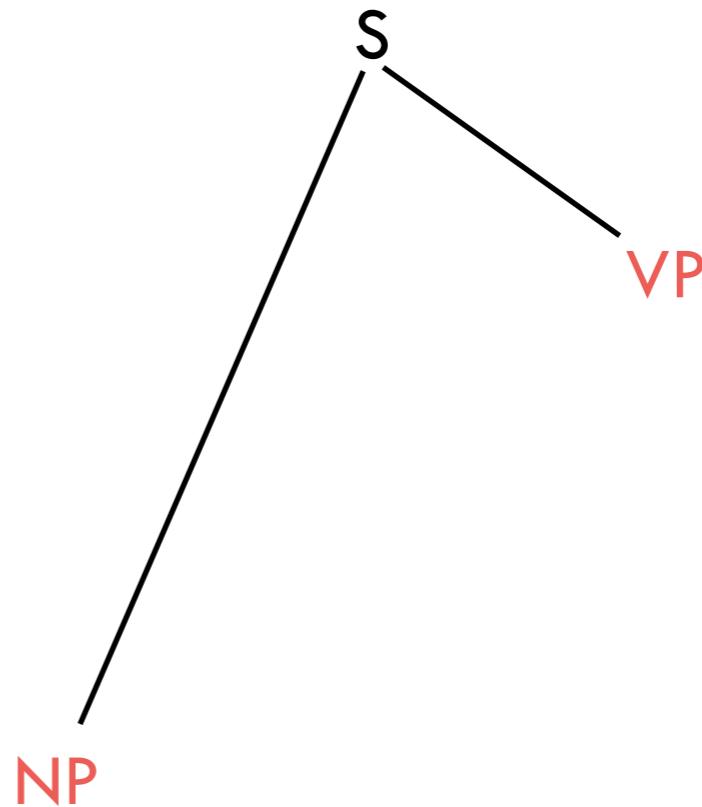
Context Free Grammars (CFG)

$S \rightarrow NP\ VP$	$V \rightarrow saw$
$VP \rightarrow V\ NP$	$P \rightarrow with$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow tail$
$NP \rightarrow NP\ PP$	$N \rightarrow student$

S

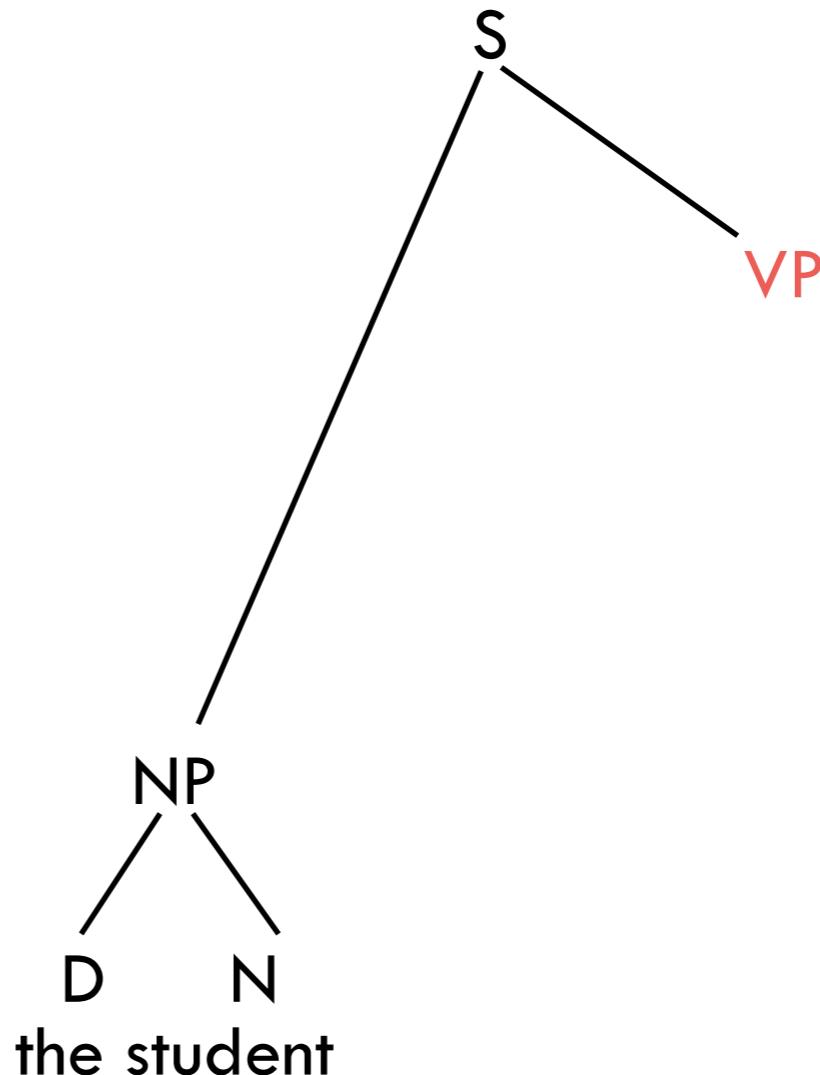
Context Free Grammars (CFG)

$S \rightarrow NP\ VP$	$V \rightarrow \text{saw}$
$VP \rightarrow V\ NP$	$P \rightarrow \text{with}$
$VP \rightarrow VP\ PP$	$D \rightarrow \text{the}$
$PP \rightarrow P\ NP$	$N \rightarrow \text{cat}$
$NP \rightarrow D\ N$	$N \rightarrow \text{tail}$
$NP \rightarrow NP\ PP$	$N \rightarrow \text{student}$



Context Free Grammars (CFG)

$S \rightarrow NP\ VP$	$V \rightarrow \text{saw}$
$VP \rightarrow V\ NP$	$P \rightarrow \text{with}$
$VP \rightarrow VP\ PP$	$D \rightarrow \text{the}$
$PP \rightarrow P\ NP$	$N \rightarrow \text{cat}$
$NP \rightarrow D\ N$	$N \rightarrow \text{tail}$
$NP \rightarrow NP\ PP$	$N \rightarrow \text{student}$



Context Free Grammars (CFG)

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

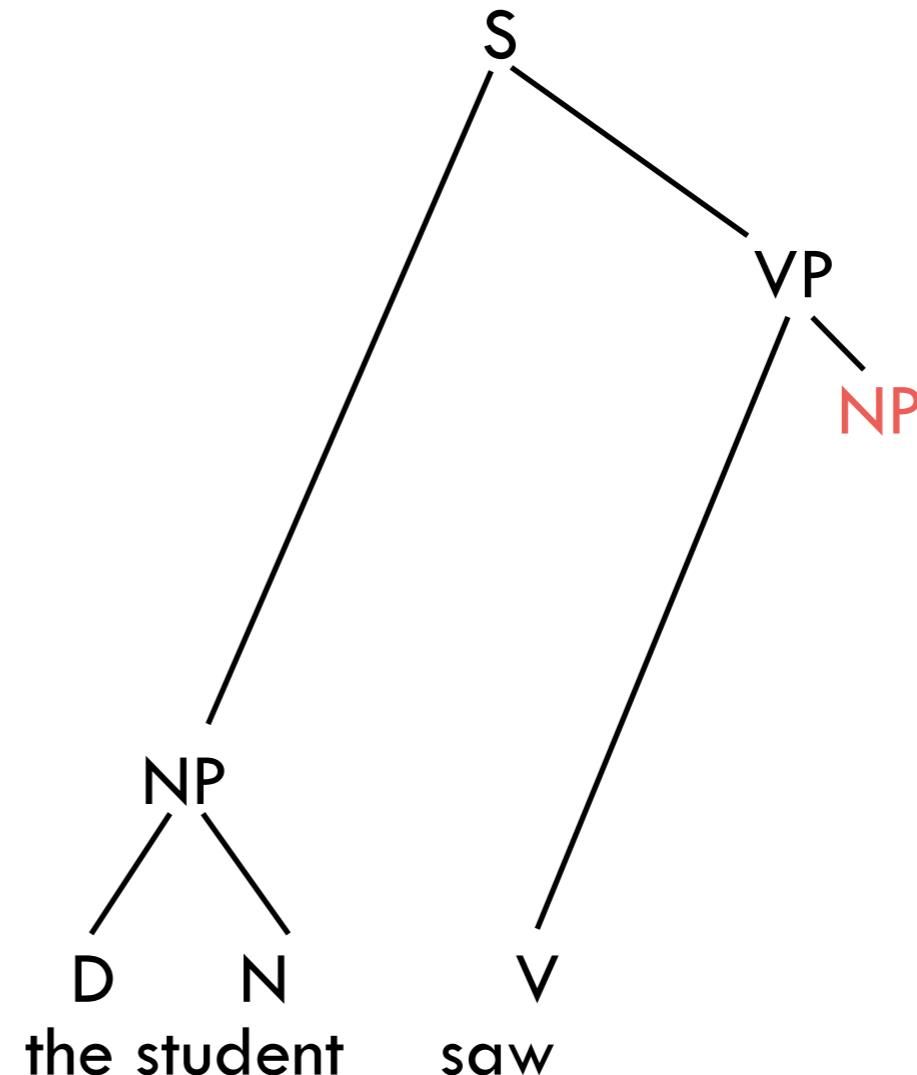
$P \rightarrow with$

$D \rightarrow the$

$N \rightarrow cat$

$N \rightarrow tail$

$N \rightarrow student$



Context Free Grammars (CFG)

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

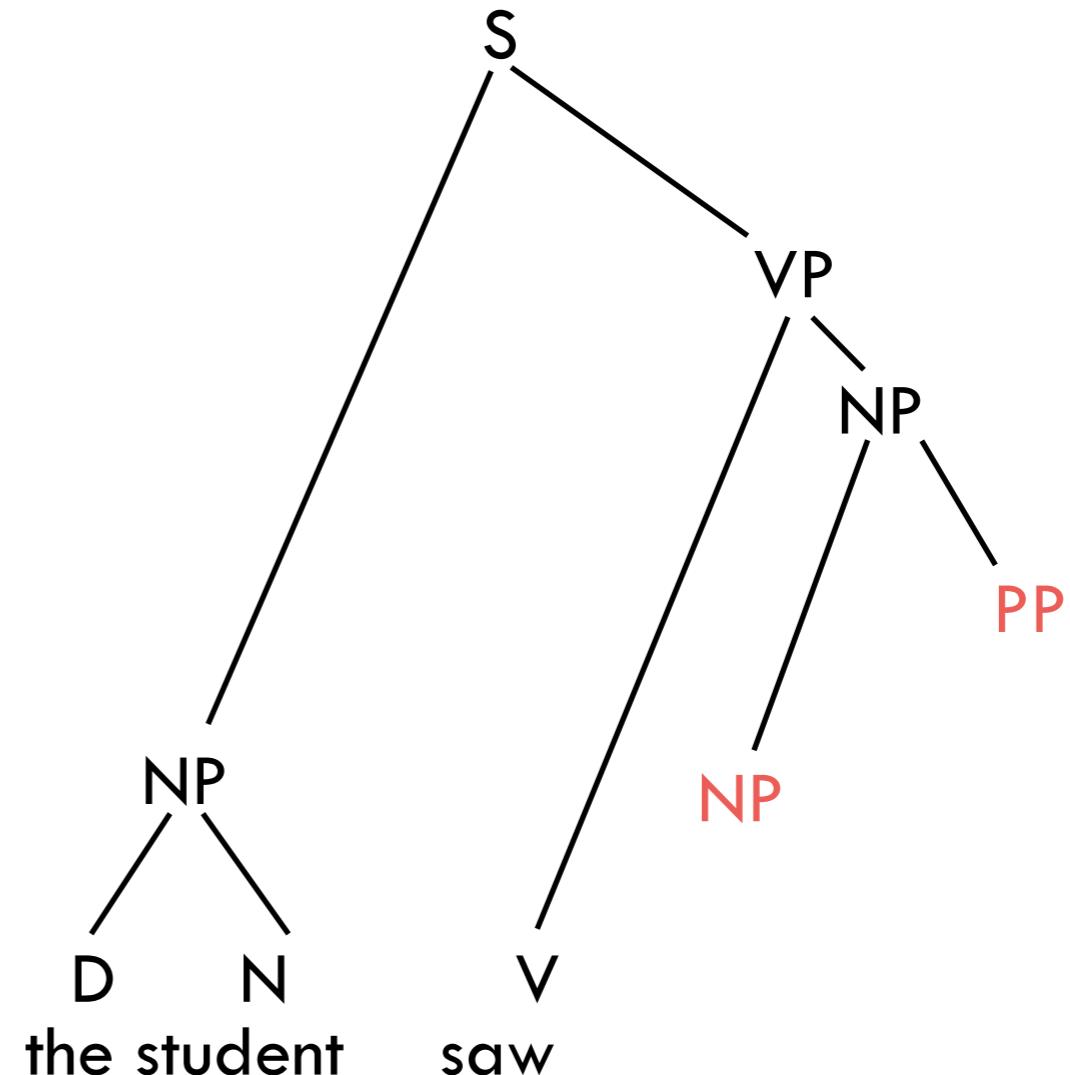
$P \rightarrow with$

$D \rightarrow the$

$N \rightarrow cat$

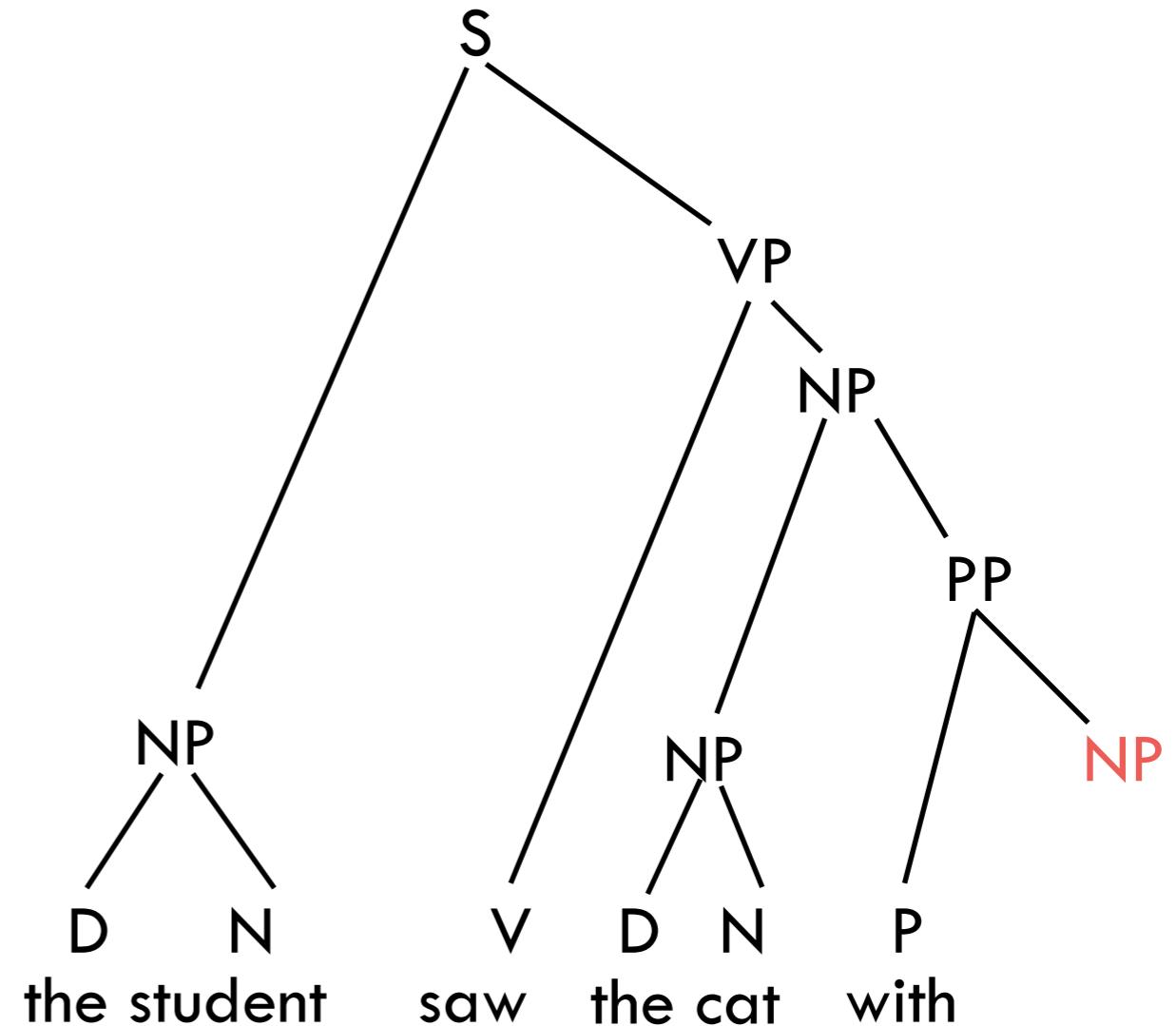
$N \rightarrow tail$

$N \rightarrow student$



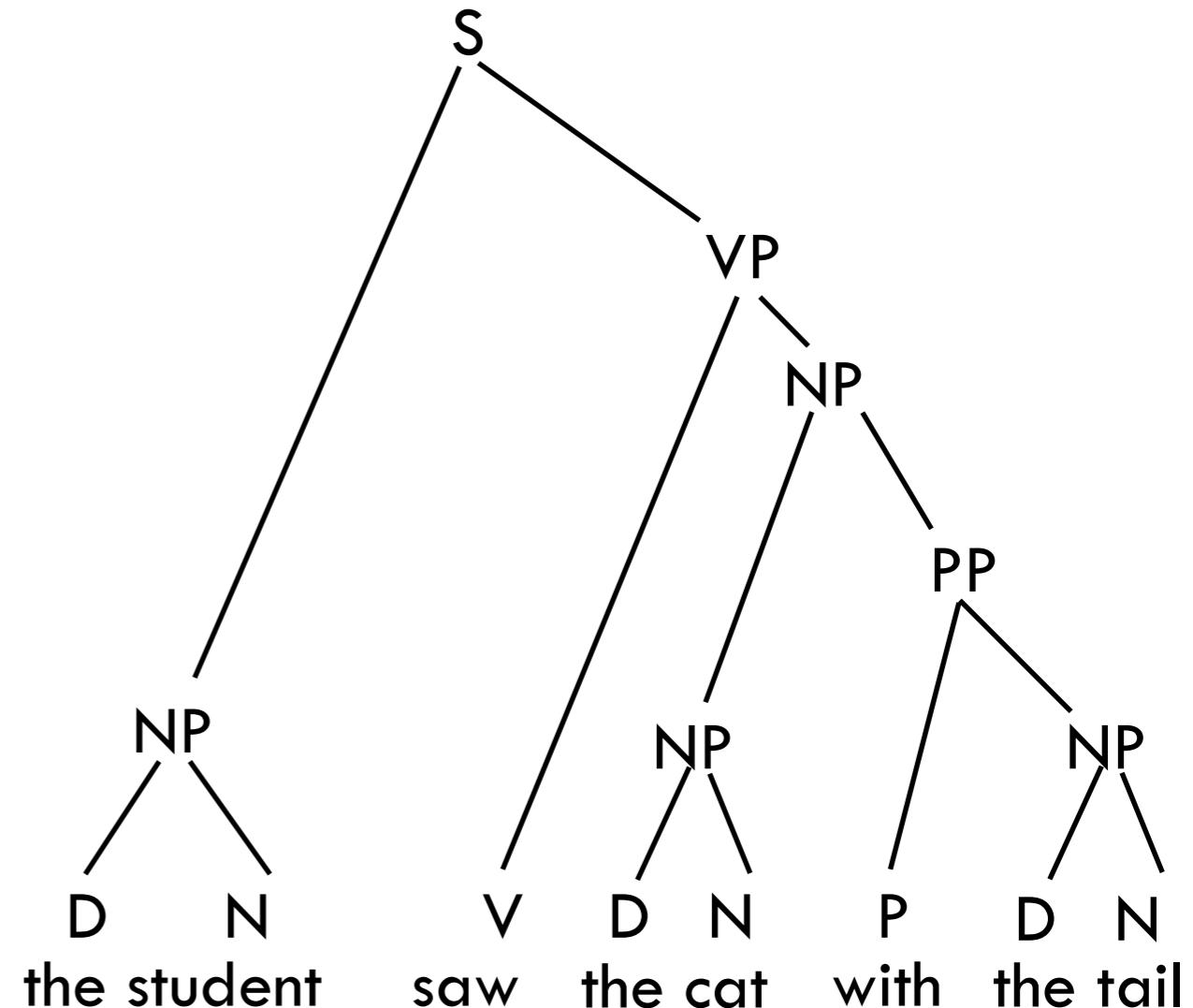
Context Free Grammars (CFG)

$S \rightarrow NP\ VP$	$V \rightarrow saw$
$VP \rightarrow V\ NP$	$P \rightarrow with$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow tail$
$NP \rightarrow NP\ PP$	$N \rightarrow student$



Context Free Grammars (CFG)

$S \rightarrow NP\ VP$	$V \rightarrow saw$
$VP \rightarrow V\ NP$	$P \rightarrow with$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow tail$
$NP \rightarrow NP\ PP$	$N \rightarrow student$



Context Free Grammars

- A context free grammar is defined by:
 - Set of **terminal symbols** Σ .
 - Set of **non-terminal symbols** N .
 - A **start symbol** $S \in N$.
 - Set R of **productions** of the form $A \rightarrow \beta$, where $A \in N$ and $\beta \in (\Sigma \cup N)^*$, i.e. β is a string of terminals and non-terminals.

Language of a CFG

- Given a CFG $G=(N, \Sigma, R, S)$:
 - Given a string $\alpha A \gamma$, where $A \in N$, we can derive $\alpha \beta \gamma$ if there is a production $A \rightarrow \beta \in R$.
 - $\alpha \Rightarrow \beta$ means that G can derive β from α in a single step.
 - $\alpha \Rightarrow^* \beta$ means that G can derive β from α in a finite number of steps.
 - The **language of G** is defined as the set of all terminal strings that can be derived from the start symbol.

$$L(G) = \{\beta \in T^*, \text{ s.t. } S \Rightarrow^* \beta\}$$

Derivations and Derived Strings

- CFG is a string rewriting formalism, so the **derived objects** are strings.
- A derivation is a sequence of rewriting steps.
- CFGs are **context free**: applicability of a rule depends only on the nonterminal symbol, not on its context.
 - Therefore, the order in which multiple non-terminals in a partially derived string are replaced does not matter. We can represent identical derivations in a **derivation tree**.
 - The derivation tree implies a parse tree.

Recursion in CFGs

$S \rightarrow NP\ VP$
 $VP \rightarrow V\ NP$
 $VP \rightarrow VP\ PP$
 $PP \rightarrow P\ NP$
 $NP \rightarrow D\ N$
 $NP \rightarrow NP\ PP$

$V \rightarrow \text{saw}$
 $P \rightarrow \text{with}$
 $D \rightarrow \text{the}$
 $N \rightarrow \text{cat}$
 $N \rightarrow \text{tail}$
 $N \rightarrow \text{student}$

Parse Tree:

NP

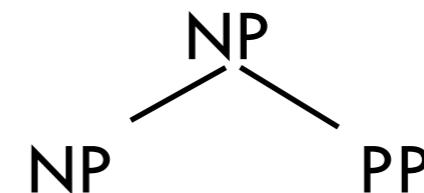
Derived String:

NP

Recursion in CFGs

$S \rightarrow NP\ VP$	$V \rightarrow saw$
$VP \rightarrow V\ NP$	$P \rightarrow with$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow tail$
$NP \rightarrow NP\ PP$	$N \rightarrow student$

Parse Tree:



Derived String:

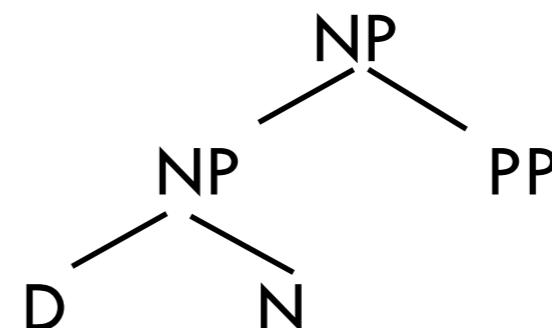
NP PP

Recursion in CFGs

$S \rightarrow NP\ VP$
 $VP \rightarrow V\ NP$
 $VP \rightarrow VP\ PP$
 $PP \rightarrow P\ NP$
 $NP \rightarrow D\ N$
 $NP \rightarrow NP\ PP$

$V \rightarrow \text{saw}$
 $P \rightarrow \text{with}$
 $D \rightarrow \text{the}$
 $N \rightarrow \text{cat}$
 $N \rightarrow \text{tail}$
 $N \rightarrow \text{student}$

Parse Tree:



Derived String:

the student PP

Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

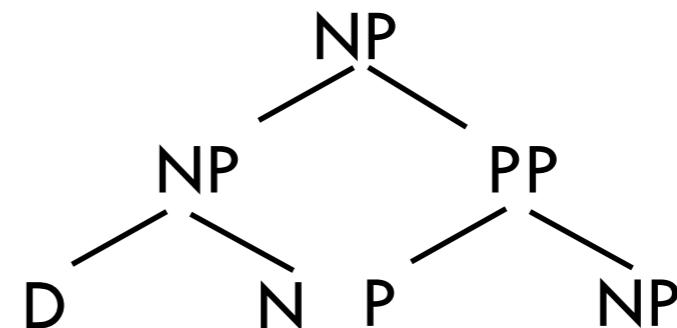
$D \rightarrow the$

$N \rightarrow cat$

$N \rightarrow tail$

$N \rightarrow student$

Parse Tree:



Derived String:

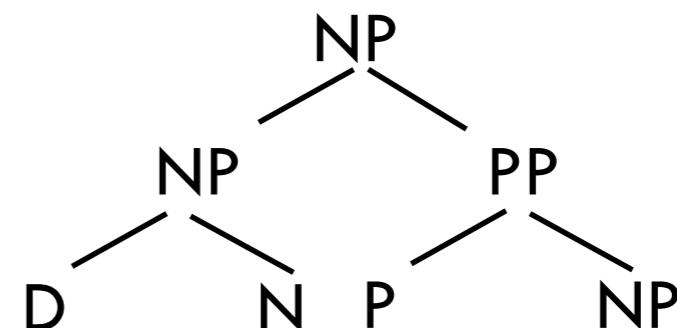
the student P NP

Recursion in CFGs

$S \rightarrow NP\ VP$
 $VP \rightarrow V\ NP$
 $VP \rightarrow VP\ PP$
 $PP \rightarrow P\ NP$
 $NP \rightarrow D\ N$
 $NP \rightarrow NP\ PP$

$V \rightarrow saw$
 $P \rightarrow with$
 $D \rightarrow the$
 $N \rightarrow cat$
 $N \rightarrow tail$
 $N \rightarrow student$

Parse Tree:



Derived String:

the student with NP

Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

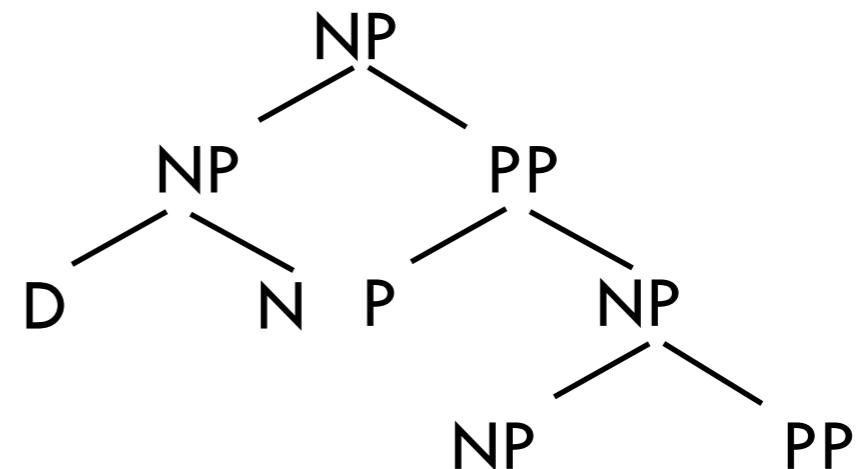
$D \rightarrow the$

$N \rightarrow cat$

$N \rightarrow tail$

$N \rightarrow student$

Parse Tree:



Derived String:

the student with NP PP

Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

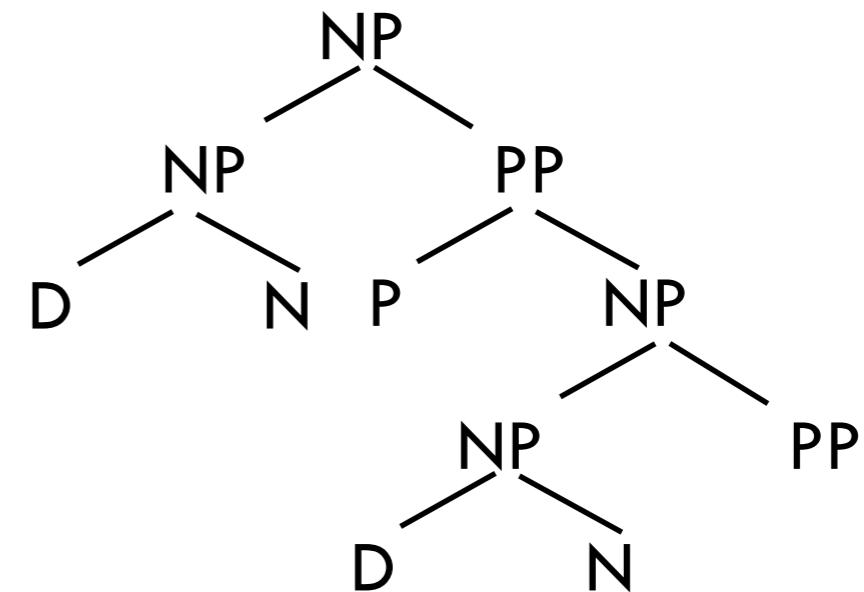
$D \rightarrow the$

$N \rightarrow cat$

$N \rightarrow tail$

$N \rightarrow student$

Parse Tree:



Derived String:

the student with the cat PP

Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

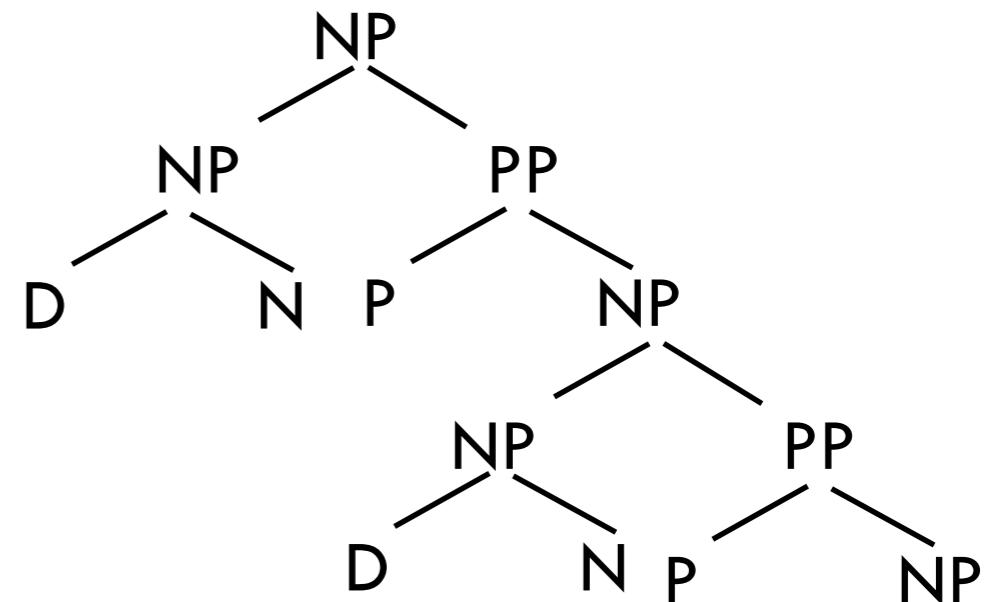
$D \rightarrow the$

$N \rightarrow cat$

$N \rightarrow tail$

$N \rightarrow student$

Parse Tree:



Derived String:

the student with the cat with NP

Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

$D \rightarrow the$

$N \rightarrow cat$

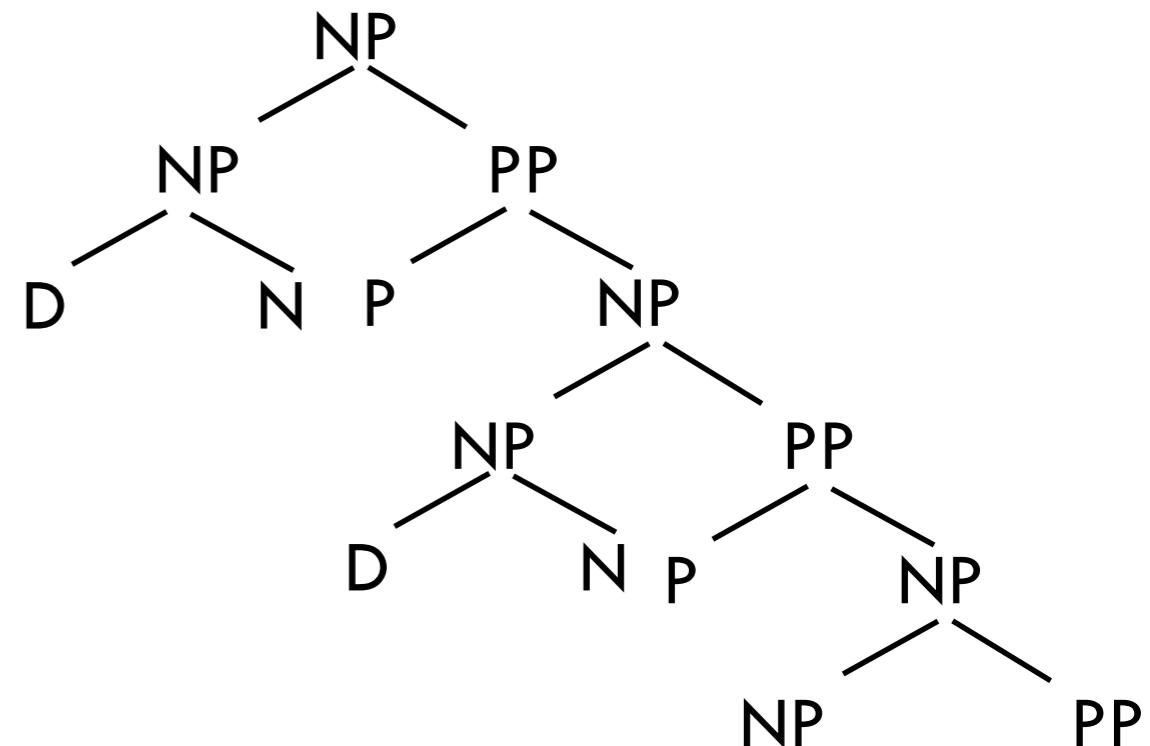
$N \rightarrow tail$

$N \rightarrow student$

Derived String:

the student with the cat with NP PP

Parse Tree:



Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

$D \rightarrow the$

$N \rightarrow cat$

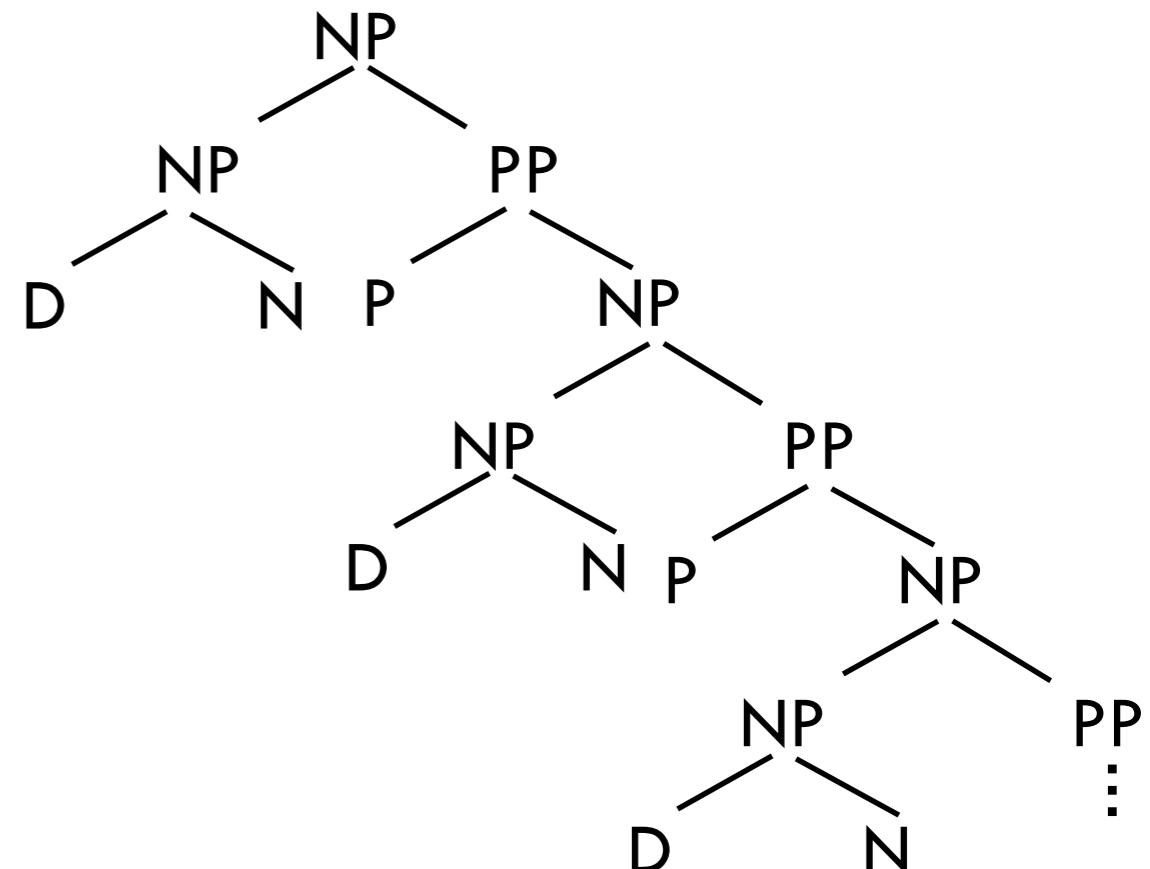
$N \rightarrow tail$

$N \rightarrow student$

Derived String:

the student with the cat with the tail PP

Parse Tree:



Recursion in CFGs

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow P\ NP$

$NP \rightarrow D\ N$

$NP \rightarrow NP\ PP$

$V \rightarrow saw$

$P \rightarrow with$

$D \rightarrow the$

$N \rightarrow cat$

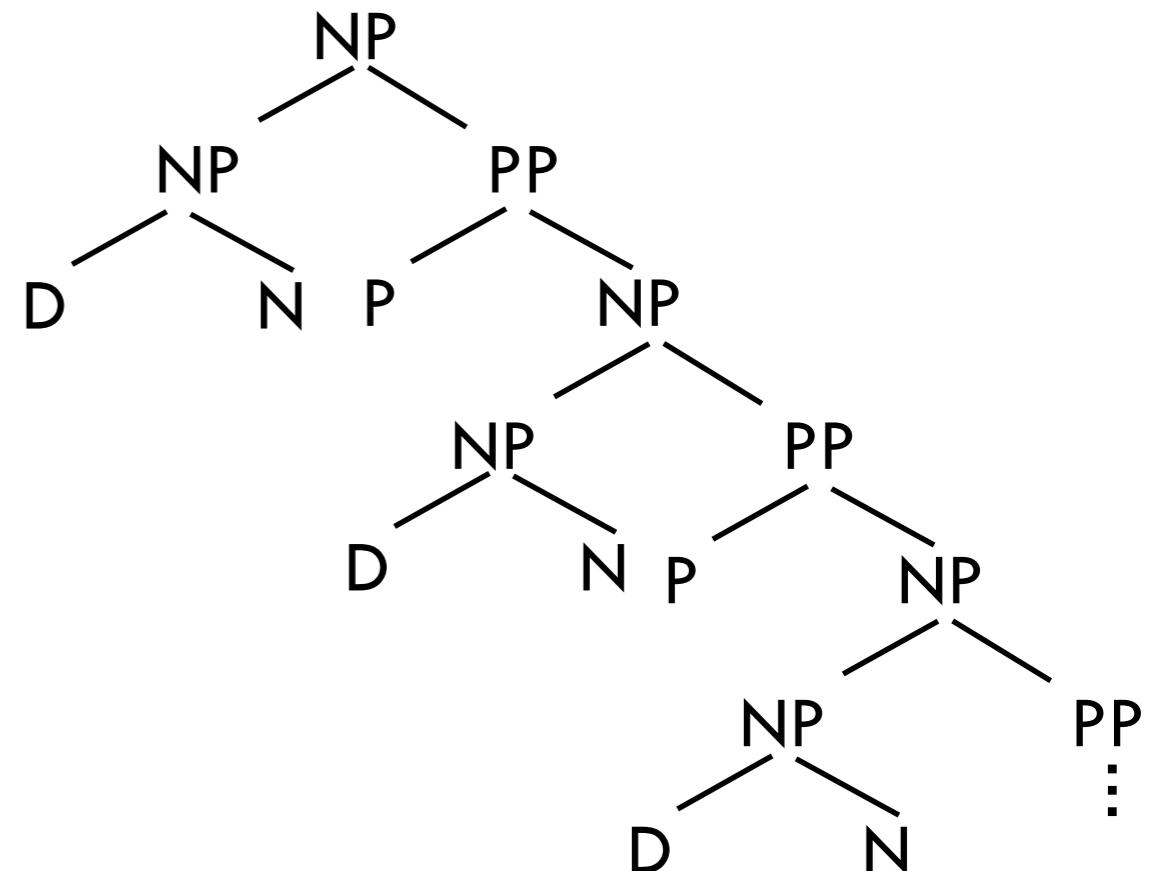
$N \rightarrow tail$

$N \rightarrow student$

Derived String:

the student with the cat with the tail PP

Parse Tree:



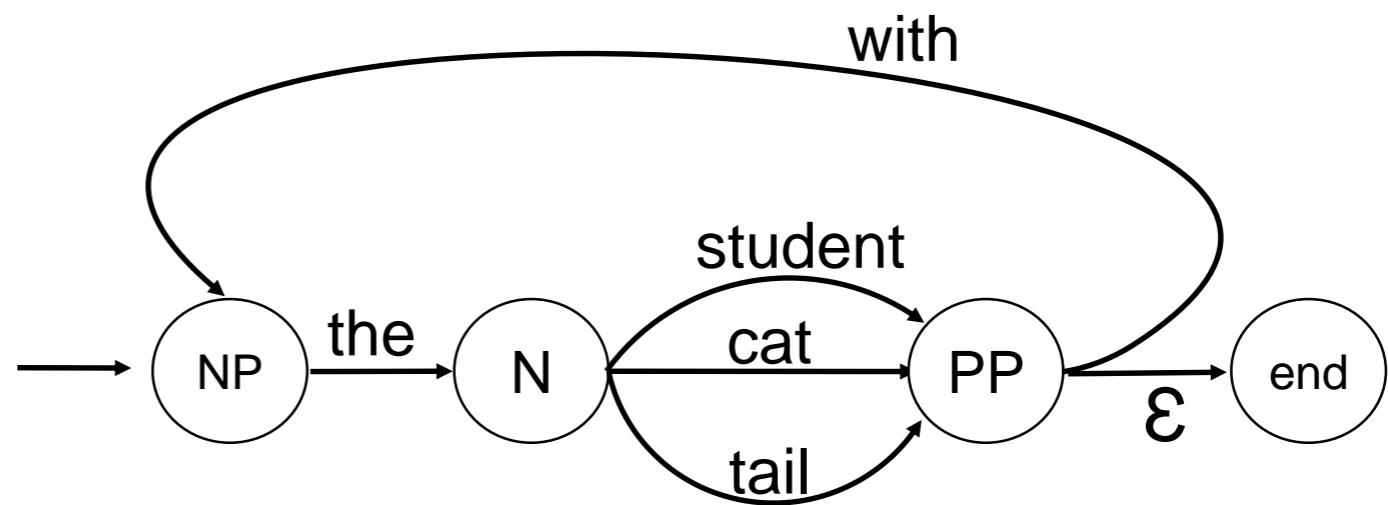
Regular Grammars

- A regular grammar is defined by:
 - Set of **terminal symbols** Σ .
 - Set of **non-terminal symbols** N .
 - A **start symbol** $S \in N$.
 - Set R of **productions** of the form $A \rightarrow aB$, or $A \rightarrow a$ where $A, B \in N$ and $a \in \Sigma$.

Finite State Automata

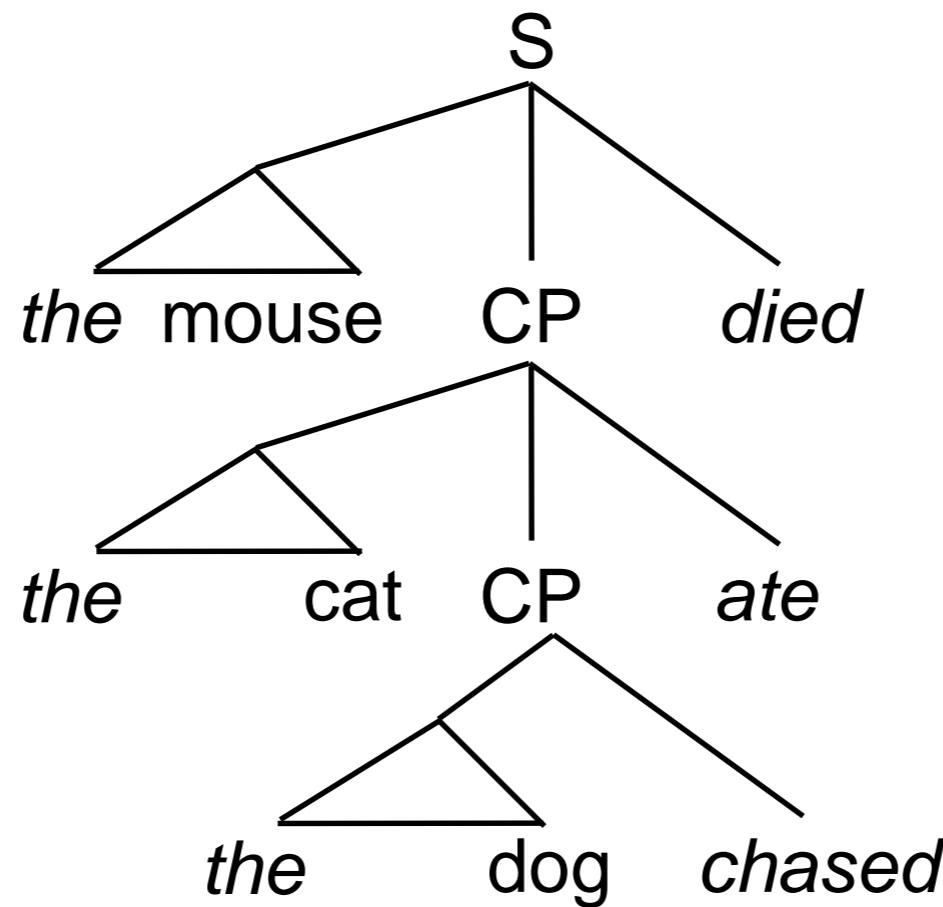
- Regular grammars can be implemented as finite state automata.

$NP \rightarrow \text{the } N$
 $N \rightarrow \text{student } PP$
 $N \rightarrow \text{cat } PP$
 $N \rightarrow \text{tail } PP$
 $PP \rightarrow \text{with } NP$
 $PP \rightarrow \epsilon$



- The set of all regular languages is strictly smaller than the set of context-free languages.

Center Embeddings

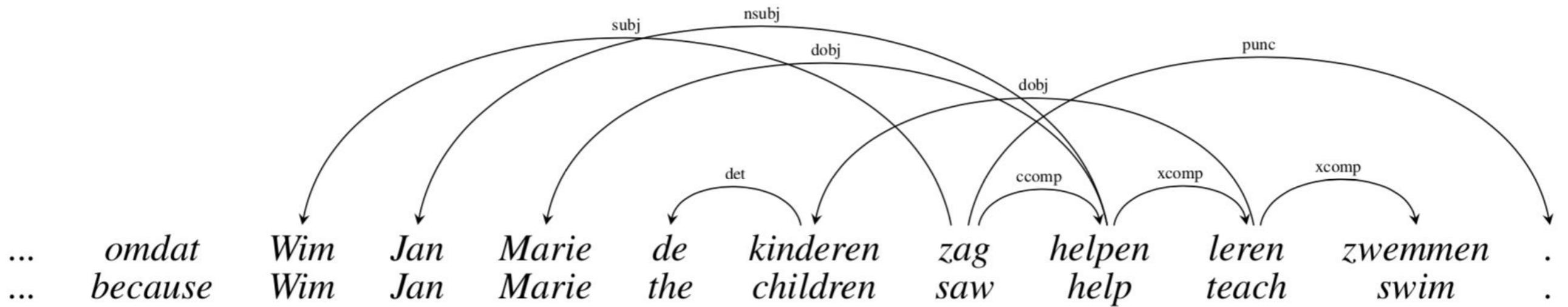


- Problem: Regular grammars cannot capture long-distance dependencies.
- This example follows the pattern $a^n b^n$. Can show that is language is not regular (using the “pumping lemma”).

Linguistically, this is not a perfect analysis.

Is Natural Language Context Free?

- Probably not. An example from Dutch:



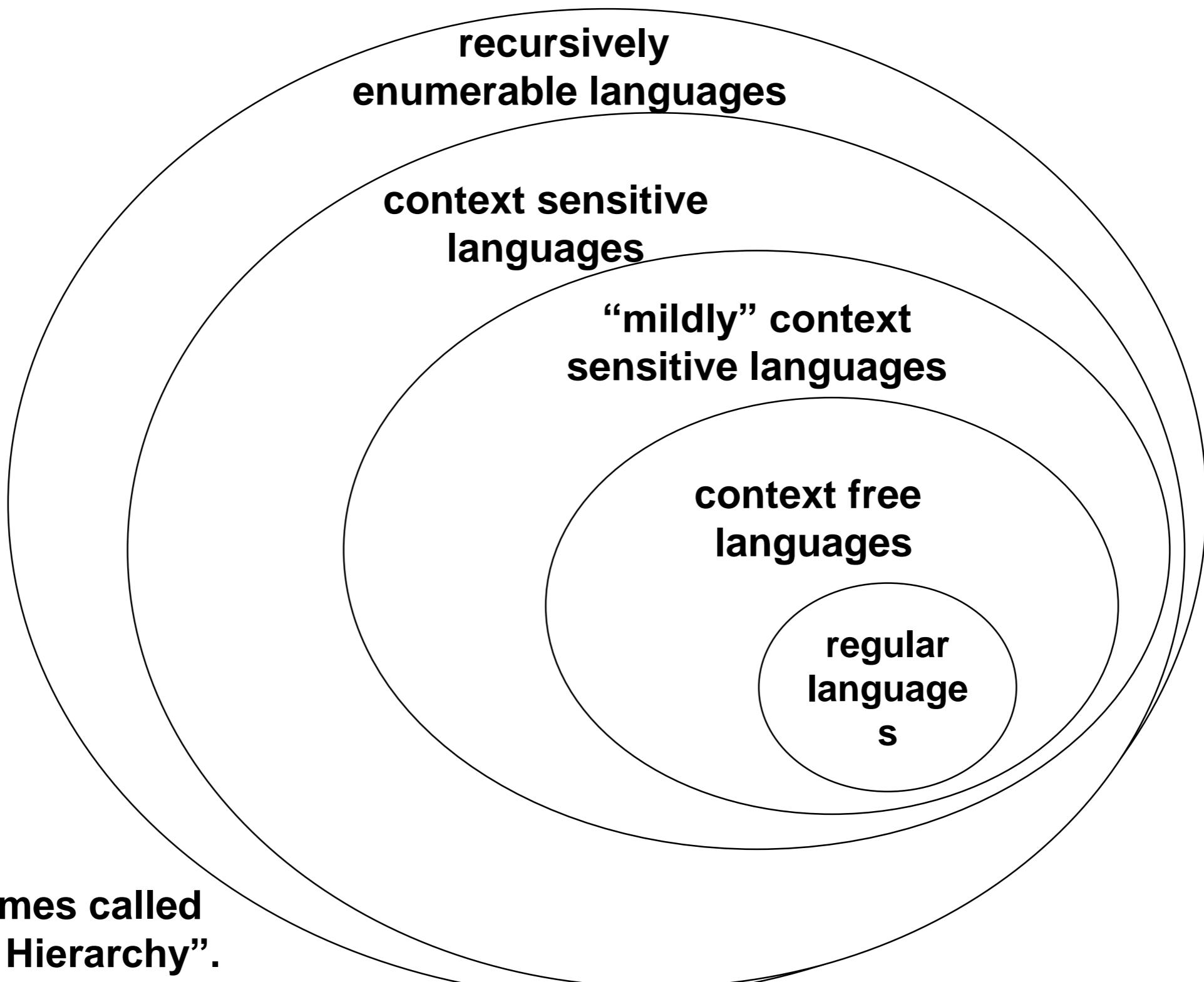
"...because Wim saw Jan help Marie teach the children to swim"

- Context Free Grammars cannot describe crossing dependencies. For example, it can be shown that

$$a^n b^m c^n d^m$$

is not a context free language.

Complexity Classes



This is sometimes called
the “Chomsky Hierarchy”.

Formal Grammar and Parsing

- Formal Grammars are used in linguistics, NLP, programming languages.
- We want to build a compact model that describes a complete language.
- Need efficient algorithms to determine if a sentence is in the language or not (**recognition problem**).
- We also want to recover the structure imposed by the grammar (**parsing problem**).

Acknowledgments

- Some slides by Kathy McKeown and Owen Rambow.

Natural Language Processing

Lecture 6: Parsing with Context Free Grammars I.
CKY algorithm

2/14/2019

COMS W4705
Yassine Benajiba

Formal Grammar and Parsing

- Formal Grammars are used in linguistics, NLP, programming languages.
- We want to build a compact model that describes a complete language.
- Need efficient algorithms to determine if a sentence is in the language or not (**recognition problem**).
- We also want to recover the structure imposed by the grammar (**parsing problem**).

Syntactic Parsing

- Formalisms like CFGs and Finite State Automata define the (possibly infinite) set of legal strings of a language.
- **Parsing algorithms** determine if an input string is part of this language or not. For CFGs, they assign each string one or more syntactic analyses.

Two Approaches to Parsing

- Bottom-up: Start at the words (terminal symbols) and see which subtrees you can build. Then combine these subtrees into larger trees. (Driven by the input sentence.)
CKY algorithm - requires Grammars in Chomsky Normal Form.
- Top-down: Start at the start symbol (S), try to apply production rules that are compatible with the input.
(Driven by the grammar - next week)
Earley algorithm
- Both approaches can be seen as a kind of search problem (next week).

Chomsky Normal Form

- A CFG $G=(N, \Sigma, R, S)$ is in Chomsky Normal Form (CNF) if the rules take one of the following forms:
 - $A \rightarrow B C$, where $A \in N, B \in N, C \in N$.
 - $A \rightarrow b$, where $A \in N, b \in \Sigma$.

$S \rightarrow NP\ VP$	$V \rightarrow saw$
$VP \rightarrow V\ NP$	$P \rightarrow with$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow tail$
$NP \rightarrow NP\ PP$	$N \rightarrow student$

Any CFG can be converted to an equivalent grammar in CNF that expresses the same language.

Cocke-Kasami-Younger (CKY) Algorithm - Motivation

- A nonterminal A covers a sub-span $[i,j]$ of the input string s if the rules in the grammar can derive $s[i,j]$ from A .
Let $\pi[i,j]$ be the set of nonterminals that cover $[i,j]$.
- The string is recognized by the grammar if $S \in \pi[0,n]$.
- Approach: Compute $\pi[i,j]$ for all sub-spans bottom-up, using dynamic-programming.

$\pi[0,8] = \{S\}$								
		$\pi[2,8] = \{VP\}$						
$\pi[0,5] = \{S\}$								
			$\pi[2,5] = \{VP\}$			$\pi[5,8] = \{NP\}$		
$\pi[0,2] = \{NP\}$			$\pi[3,5] = \{NP\}$			$\pi[6,8] = \{NP\}$		
$\pi[0,1] = \{D\}$	{N}	{V,N}	{D}	{N}	{P}	{D}	$\pi[7,8] = \{D\}$	
$s =$	the	student	saw	the	cat	with	the	tail
0	1	2	3	4	5	6	7	8

CKY Data Structure

- Use a 2-dimensional “parse table” to represent $\pi[i,j]$.

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

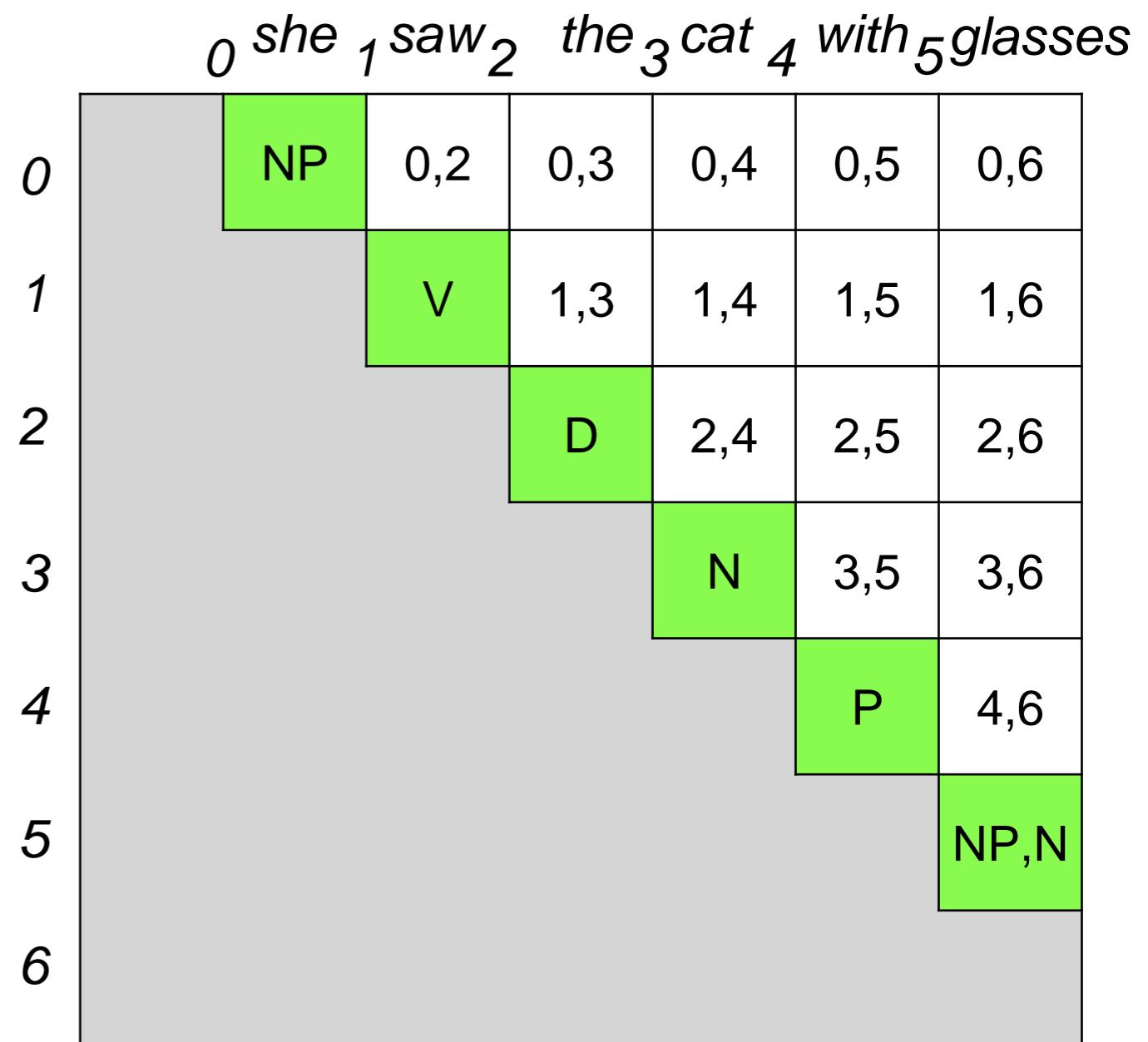
		0	she	1	saw	2	the	3	cat	4	with	5	glasses
		0	1,2	0,2	0,3	0,4	0,5	0,6					
0	1		0,1										
2	3			1,2	1,3	1,4	1,5	1,6					
4	5				2,3	2,4	2,5	2,6					
6	7					3,4	3,5	3,6					
8	9						4,5	4,6					
10	11							5,6					

CKY Initialization

- For $i=0 \dots \text{length}(s)-1$:

$$\pi[i, i+1] = \{A \mid A \rightarrow s[i:i+1] \in R\}$$

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$



CKY - finding the split

- CKY requires grammar to be in CNF.
- Assume subspan $[i,j]$ is covered by nonterminal A.
 - Then this nonterminal was recognized by some production of the form $A \rightarrow B C$, where $A \in N, B \in N, C \in N$ (grammar is in CNF).
 - Span $[i,j]$ can be split into two parts: $[i,k]$, which is covered by B, and $[k,j]$ which is covered by C.



CKY - Recursive Definition

- To compute $\pi[i, j]$, try all possible split points k , such that $i < k < j$.
 - For each k , check if the nonterminals in $\pi[i, k]$ and $\pi[k, j]$ match any of the rules in the grammar.
- Recursive definition for $\pi[i, j]$:
$$\pi[i, j] = \bigcup_{k=i+1\dots j-1} \{A | A \rightarrow B \text{ } C \in R \text{ and } B \in \pi[i, k] \text{ and } C \in \pi[k, j]\}$$

CKY Full Algorithm

- **Input:** Grammar $G=(N, \Sigma, R, S)$, input string s of length n .
- **for** $i=0\dots n-1$: initialization
 $\pi[i, i+1] = \{A \mid A \rightarrow s[i]\}$
- **for** $length=2\dots n$: main loop
for $i=0\dots (n-length)$:
 $j = i+length$
for $k=i+1\dots j-1$:
 $M = \{A \mid A \rightarrow BC \text{ and } C \in R \text{ and } B \in \pi[i, k] \text{ and } C \in \pi[k, j]\}$
 $\pi[i, j] = \pi[i, j] \cup M$
- **if** $S \in \pi[0, i+1]$ **return** True, otherwise False

CKY Algorithm

```

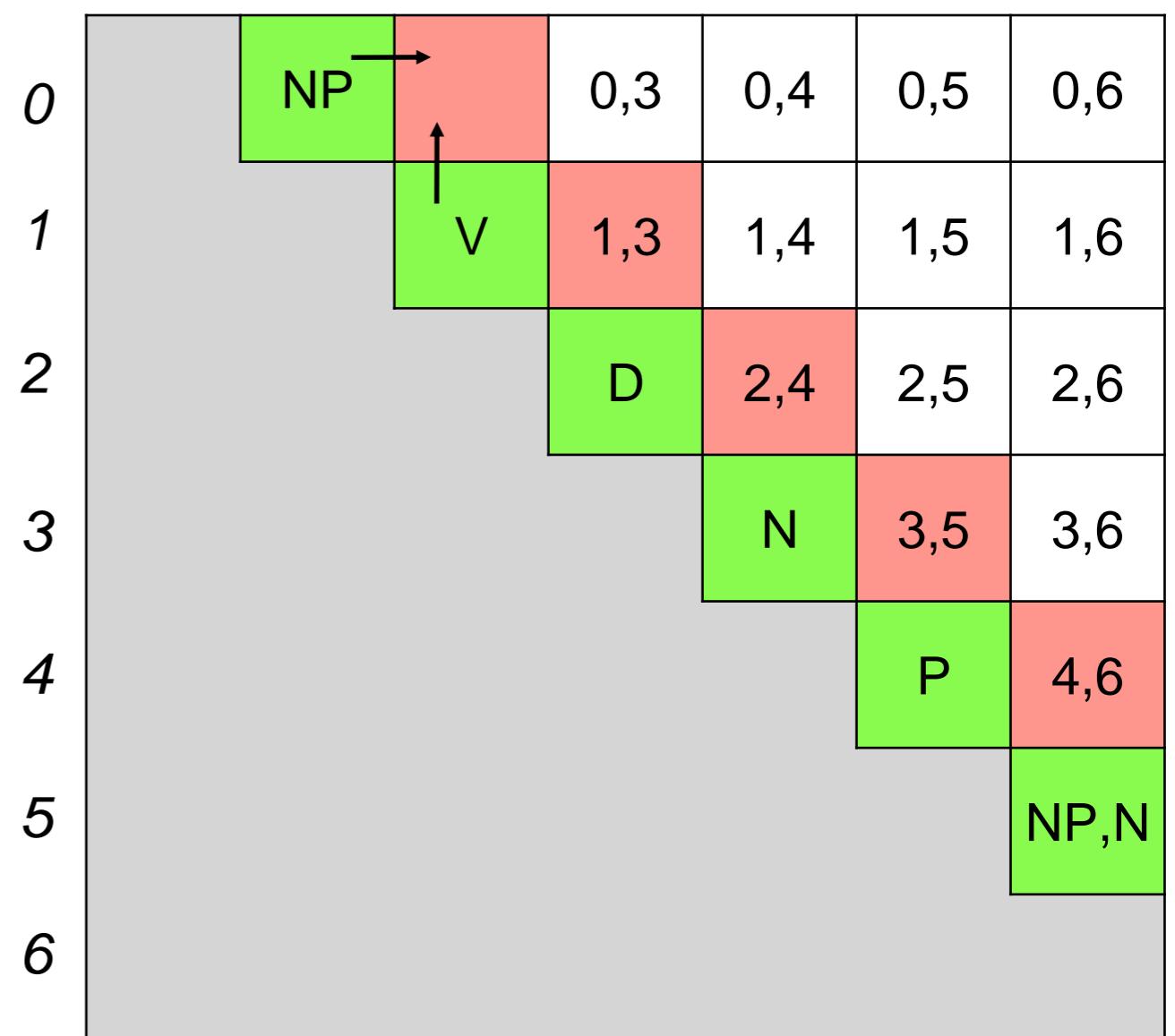
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=2$

$i=0, k=1, j=2$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

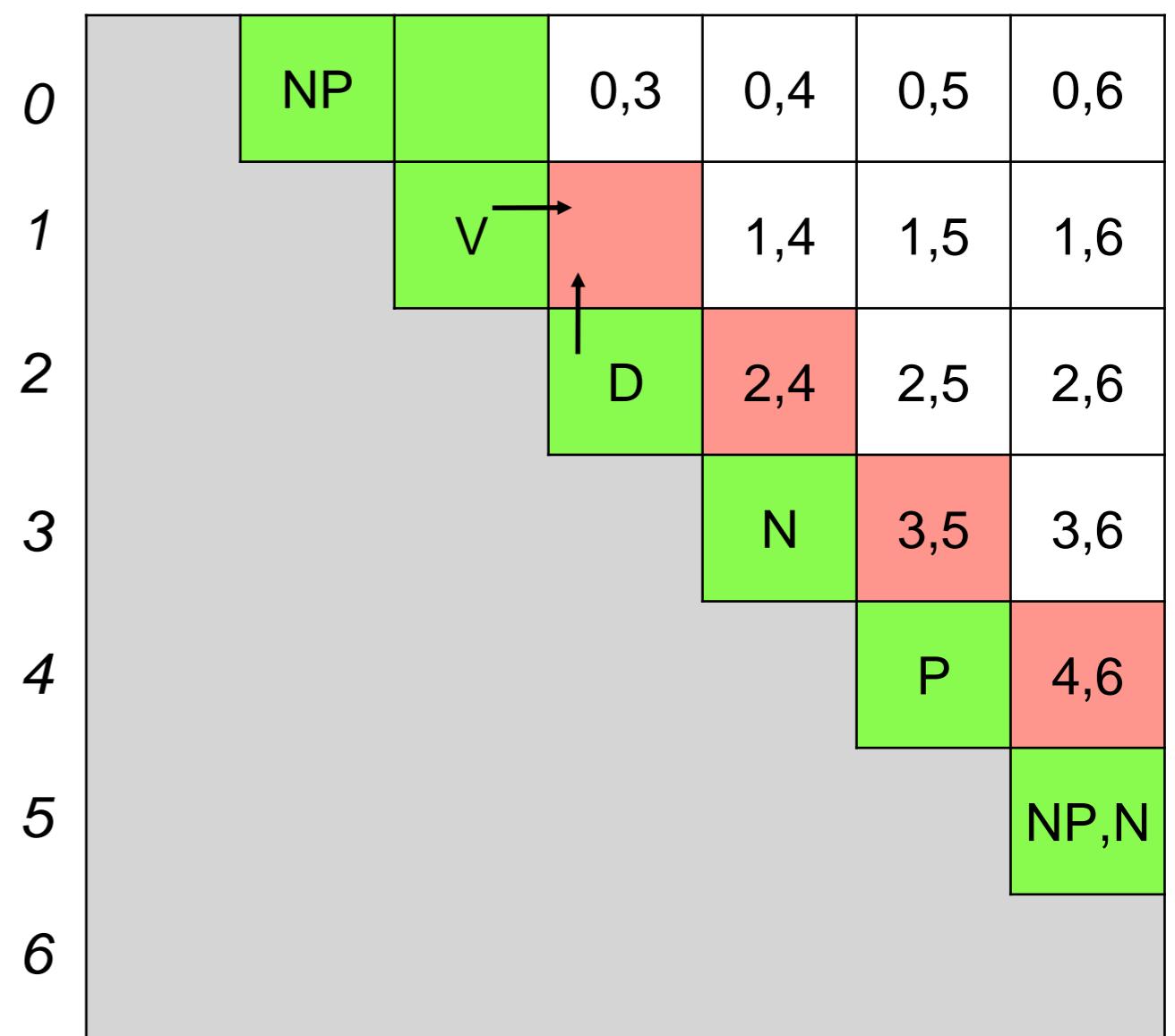
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=2$

$i=1, k=2, j=3$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

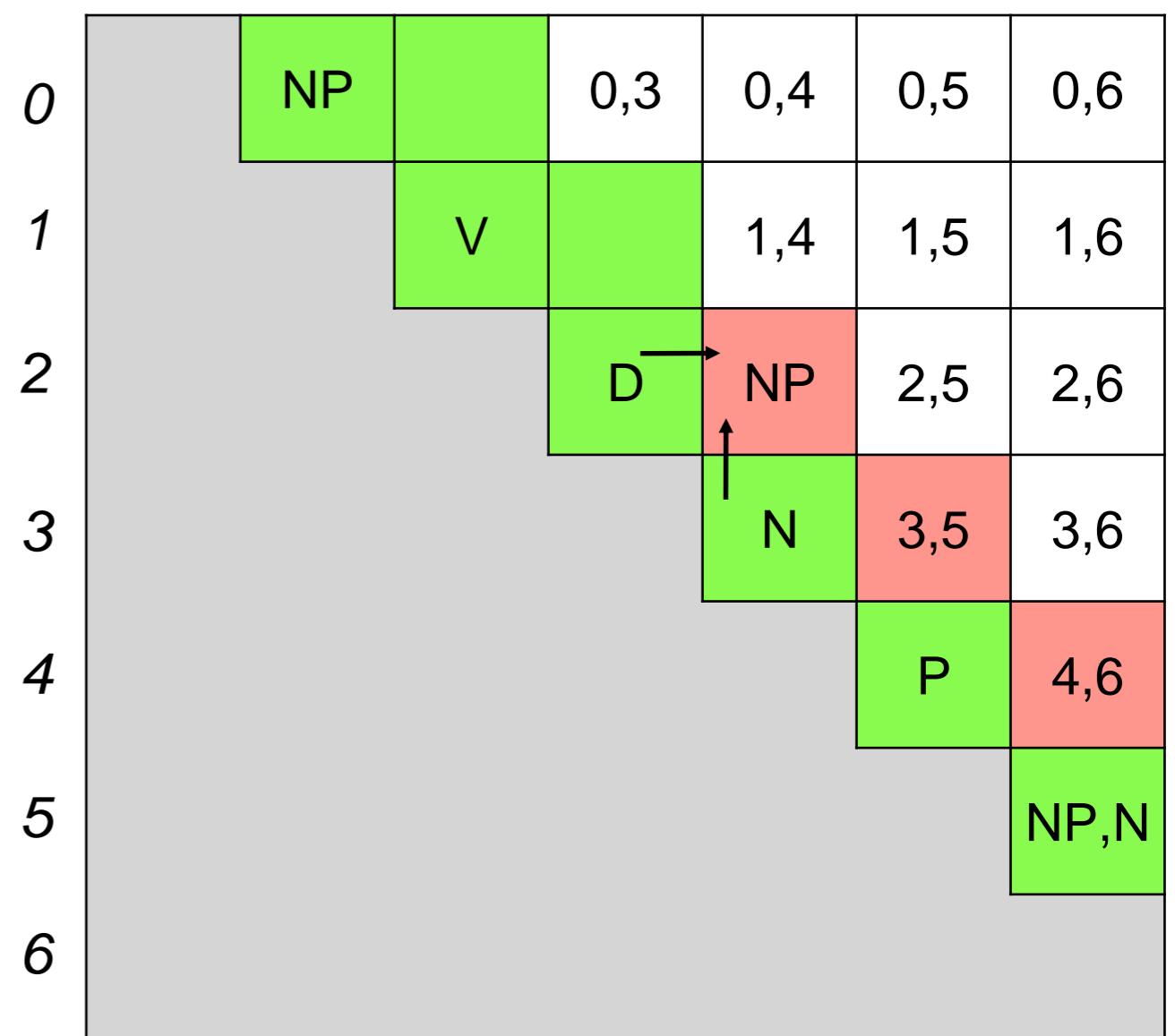
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=2$

$i=2, k=3, j=4$

$0 she_1 saw_2 the_3 cat_4 with_5 glasses$



CKY Algorithm

```

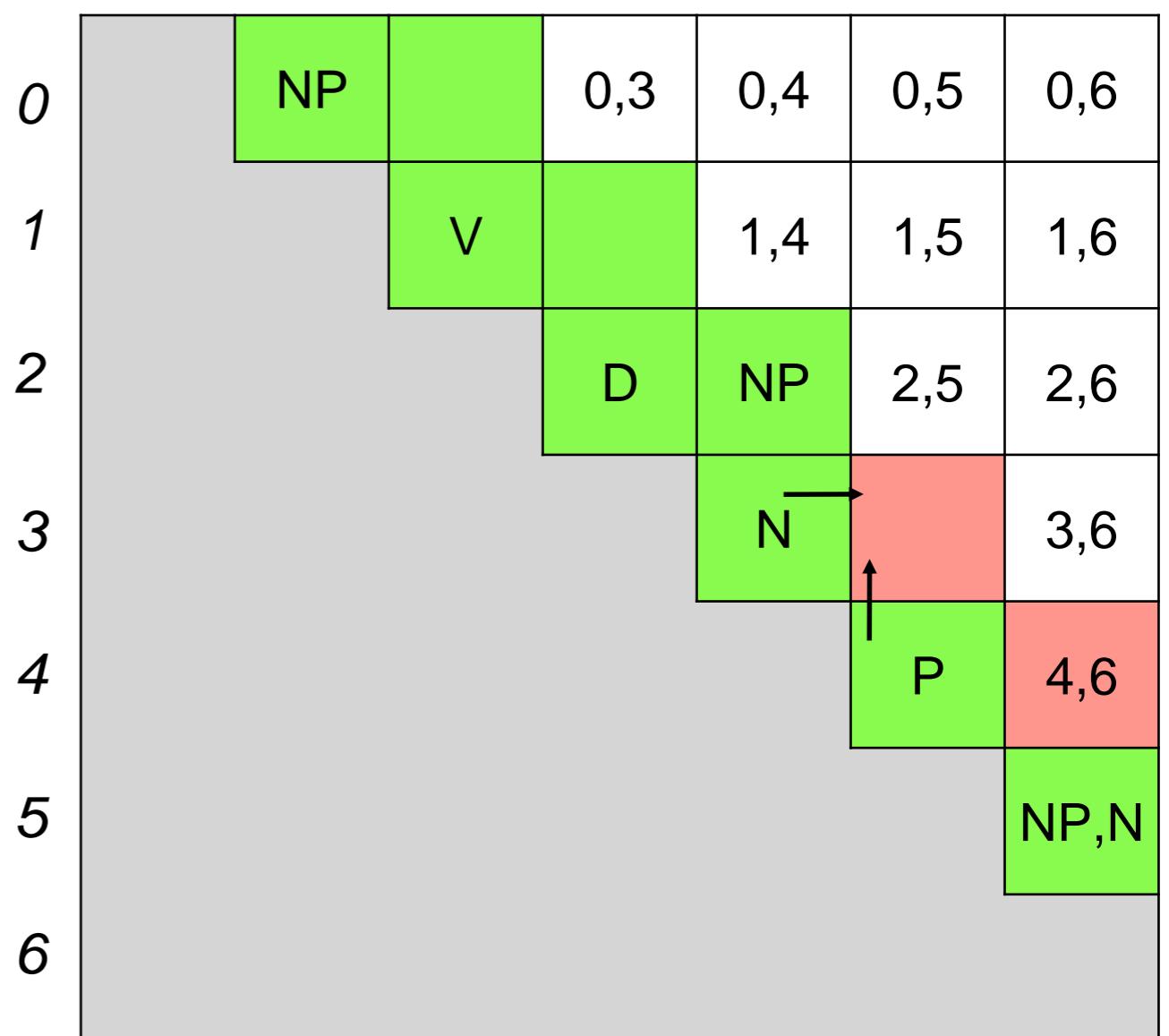
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=2$

$i=3, k=4, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

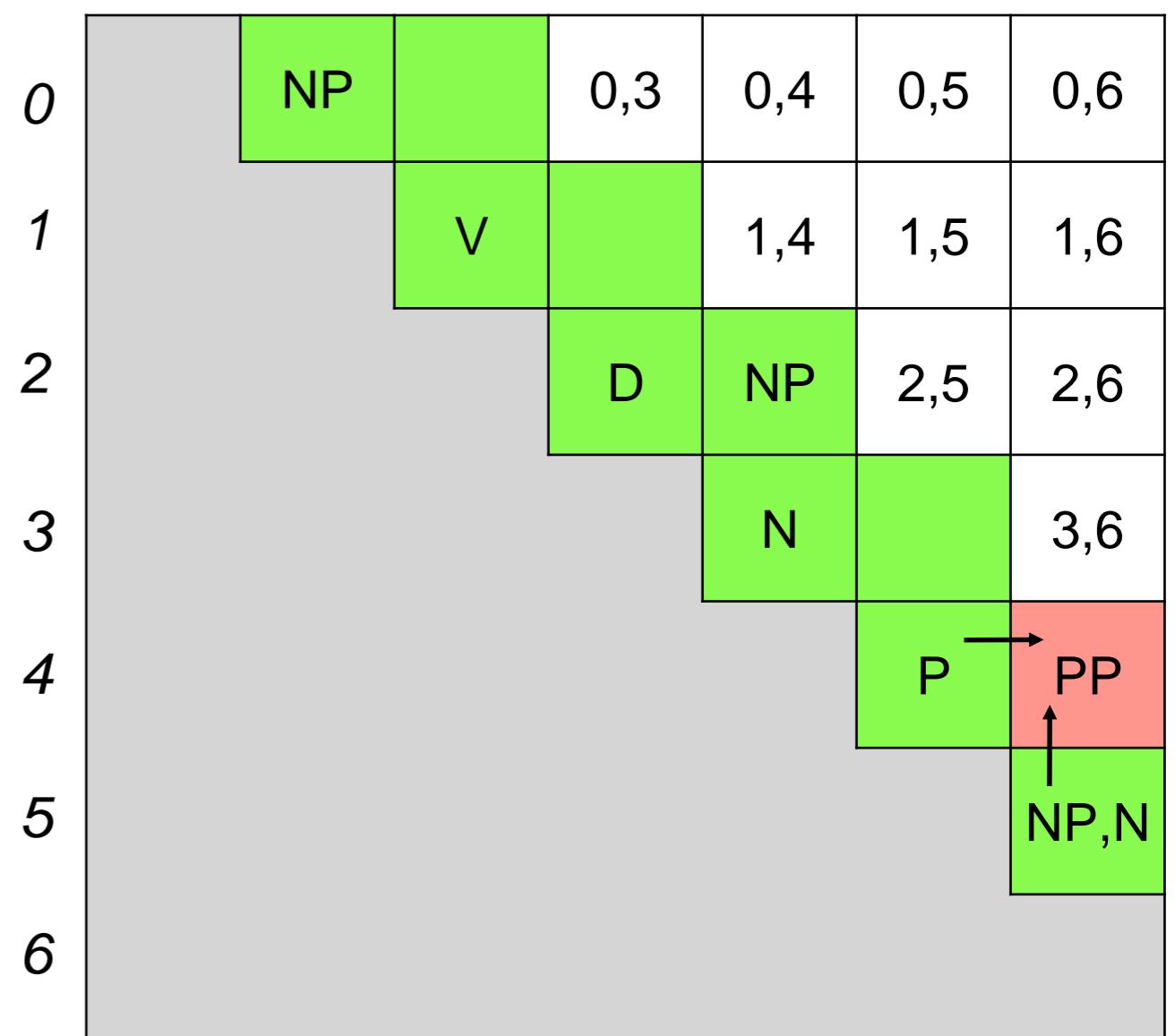
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=2$

$i=4, k=5, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

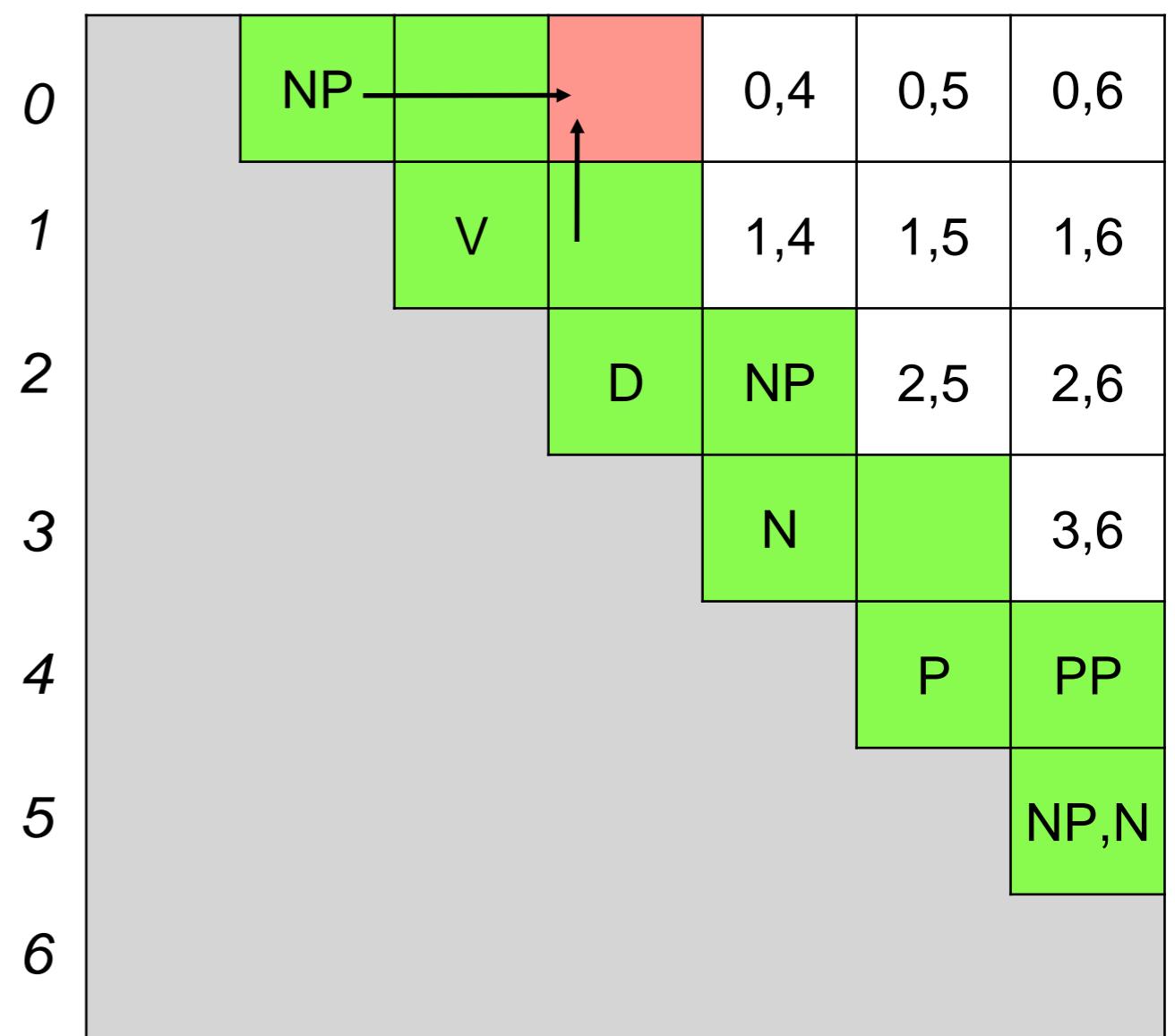
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=0, k=1, j=3$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

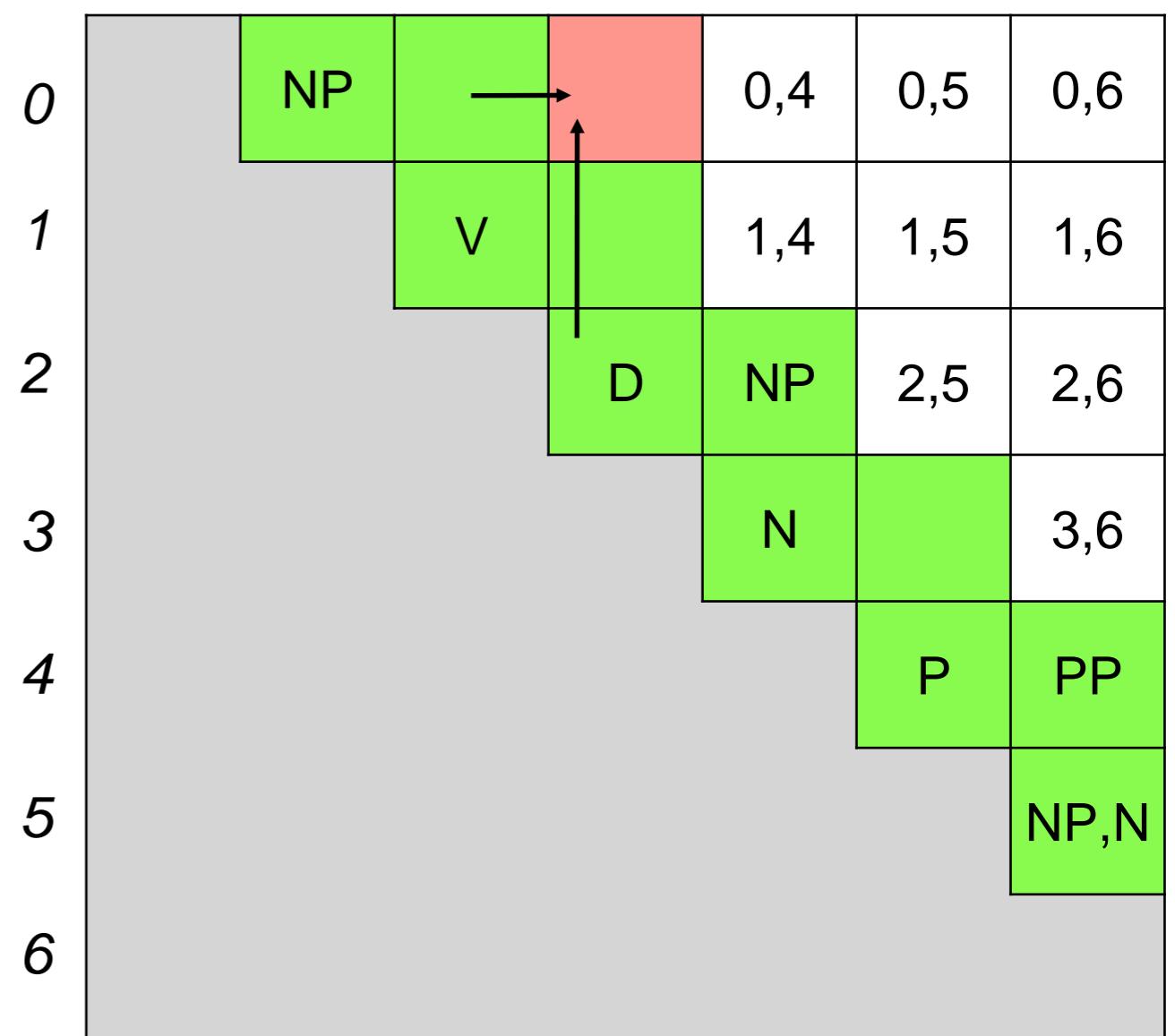
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=0, k=2, j=3$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

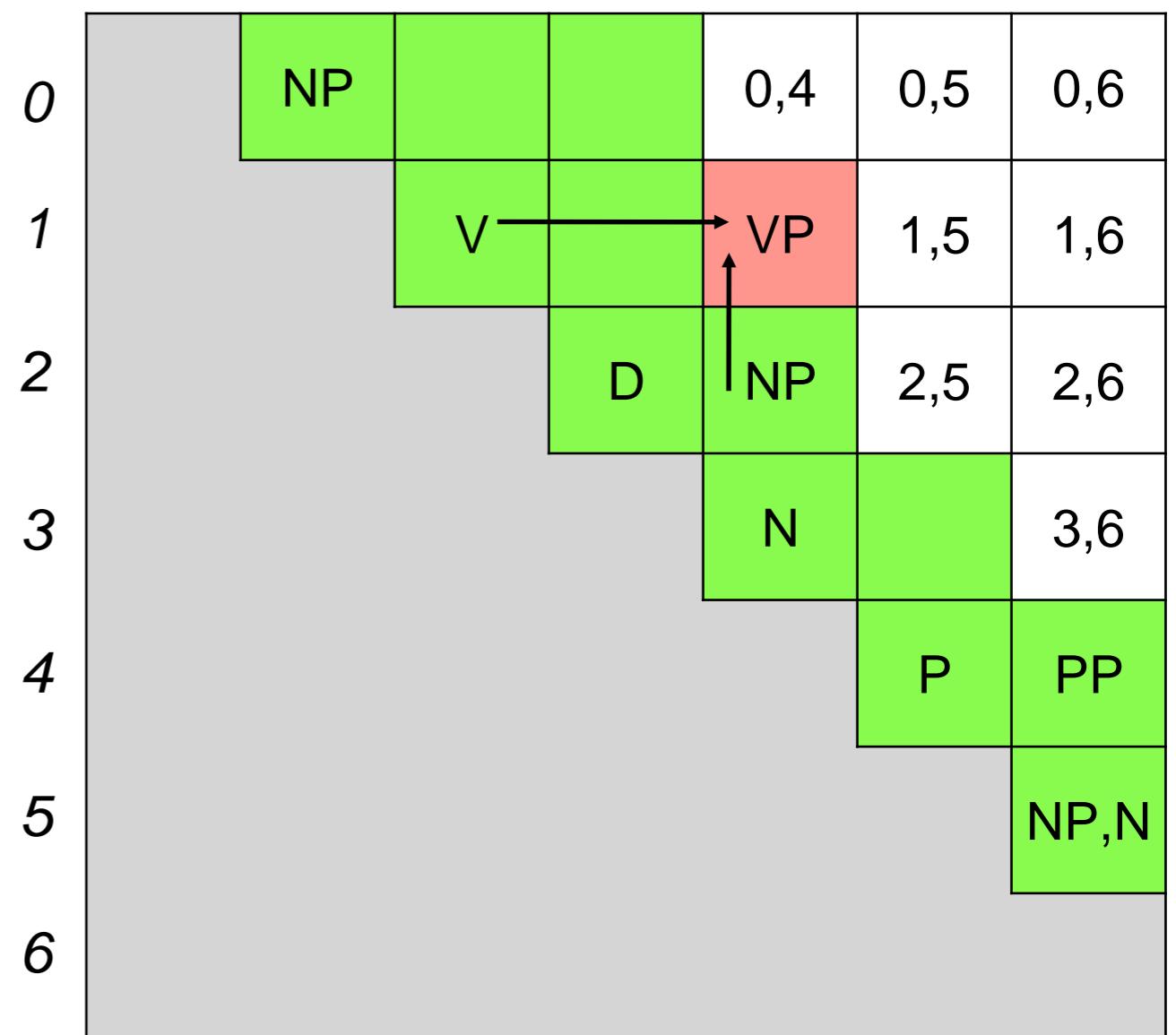
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=1, k=2, j=4$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

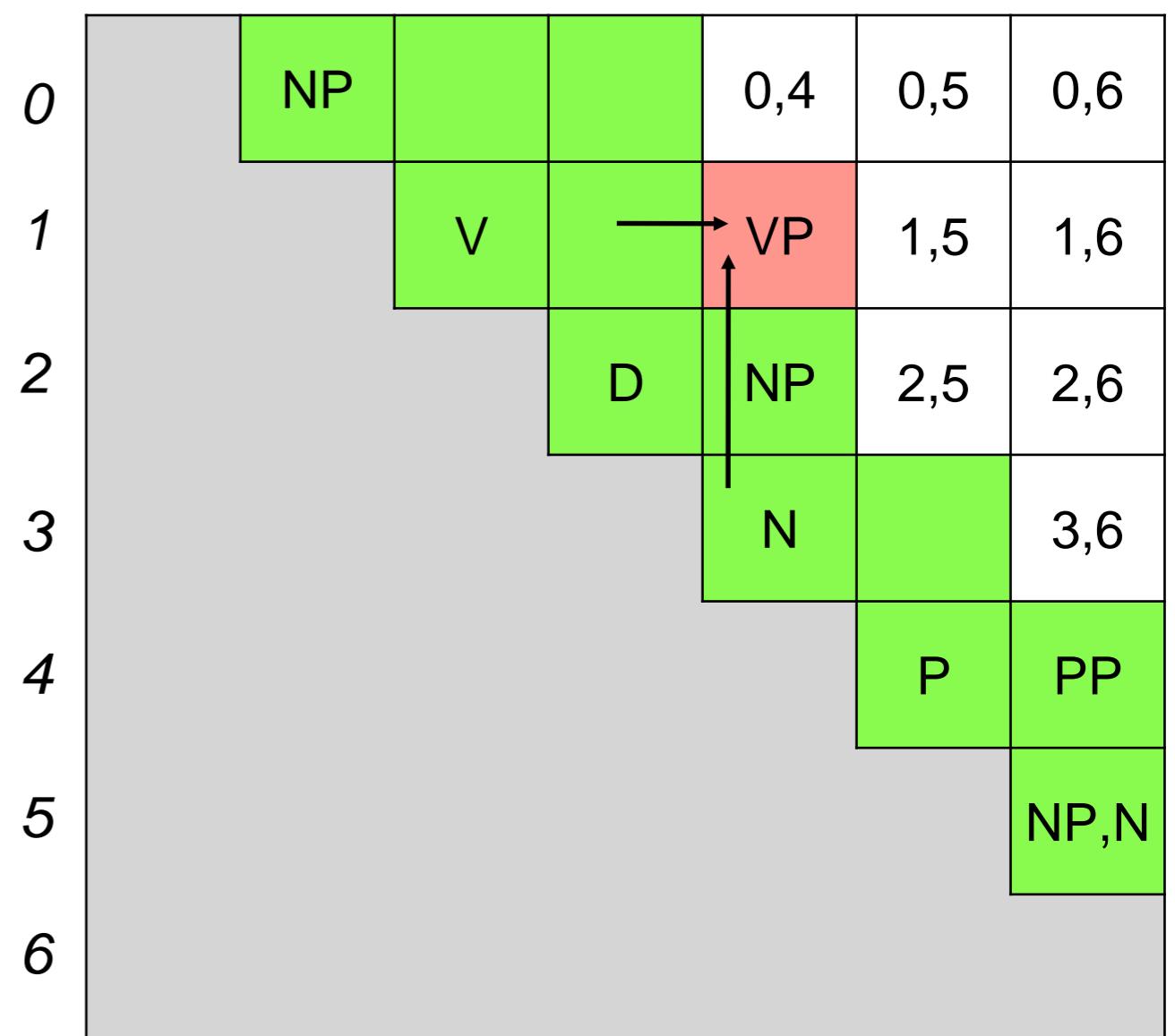
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=1, k=3, j=4$

$0 she_1 saw_2 the_3 cat_4 with_5 glasses$



CKY Algorithm

```

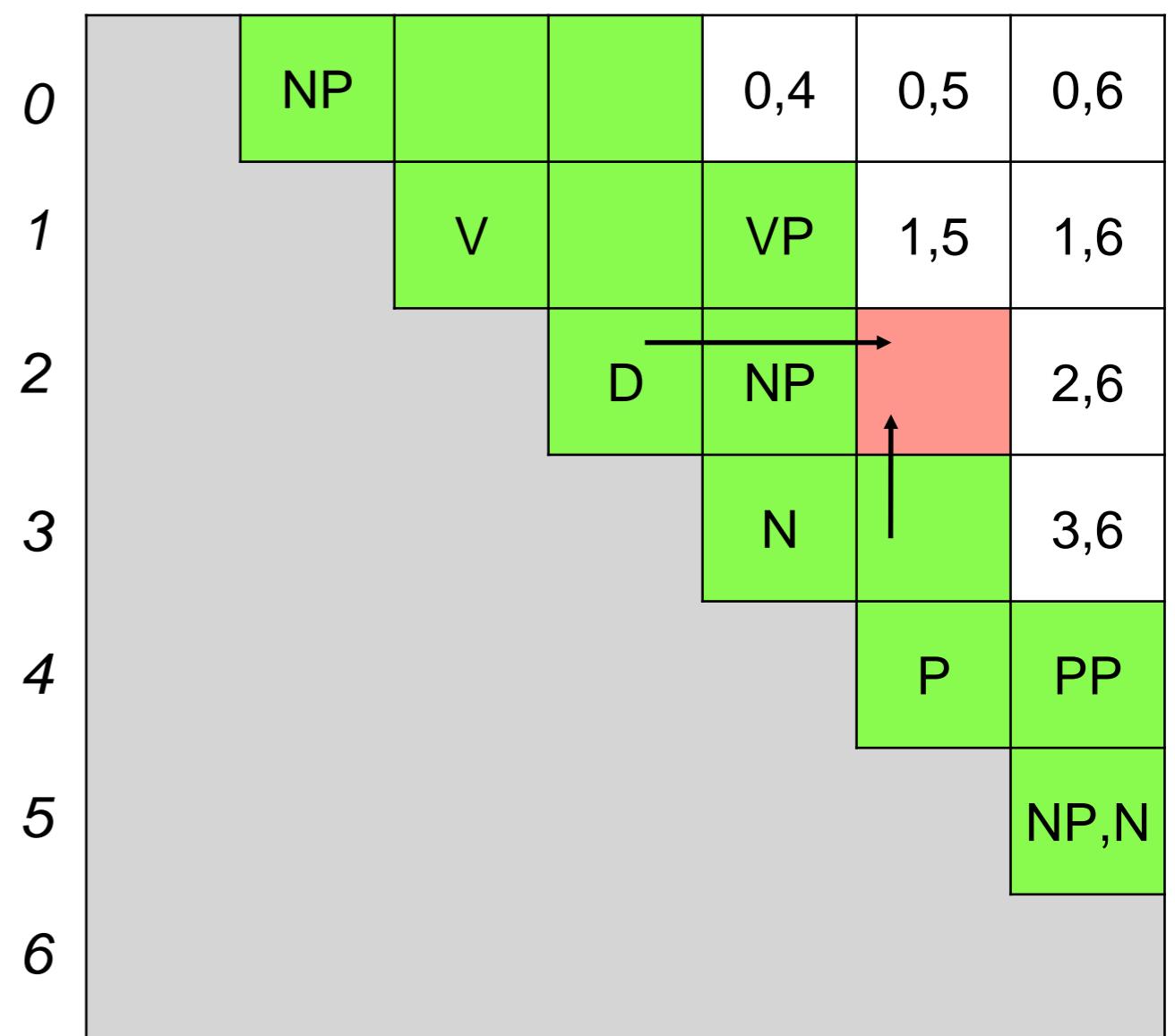
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=2, k=3, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

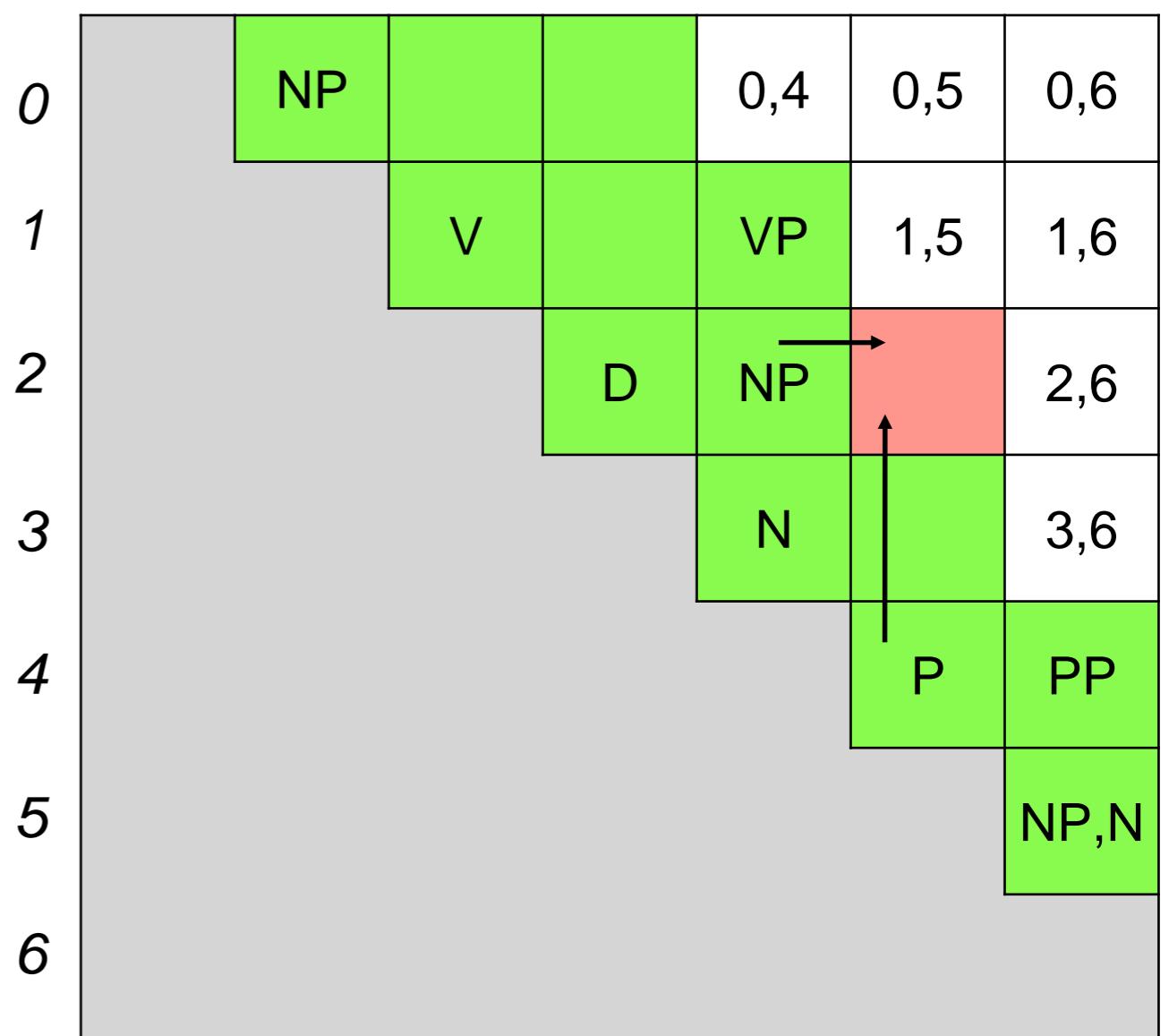
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=2, k=4, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

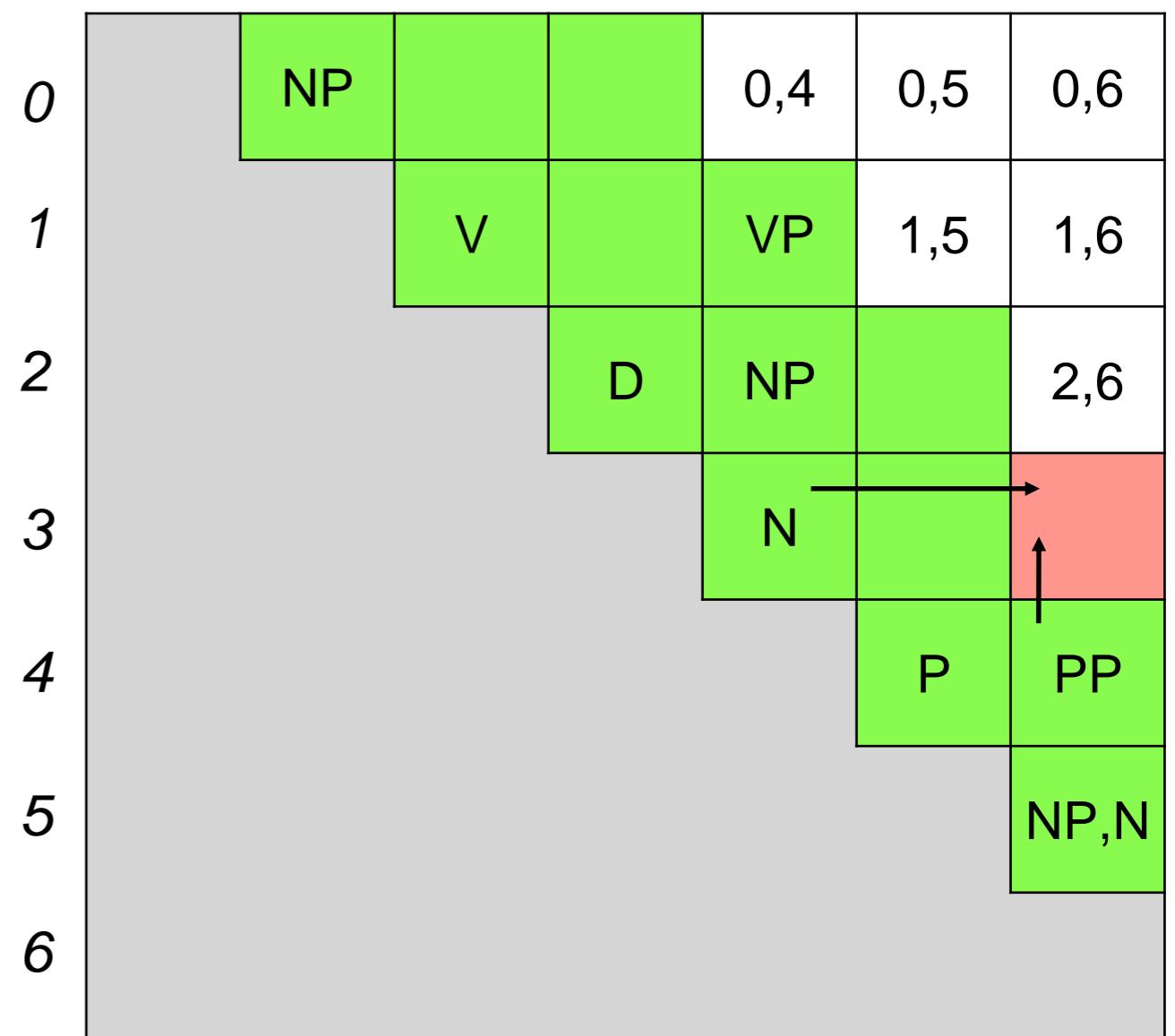
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=3, k=4, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

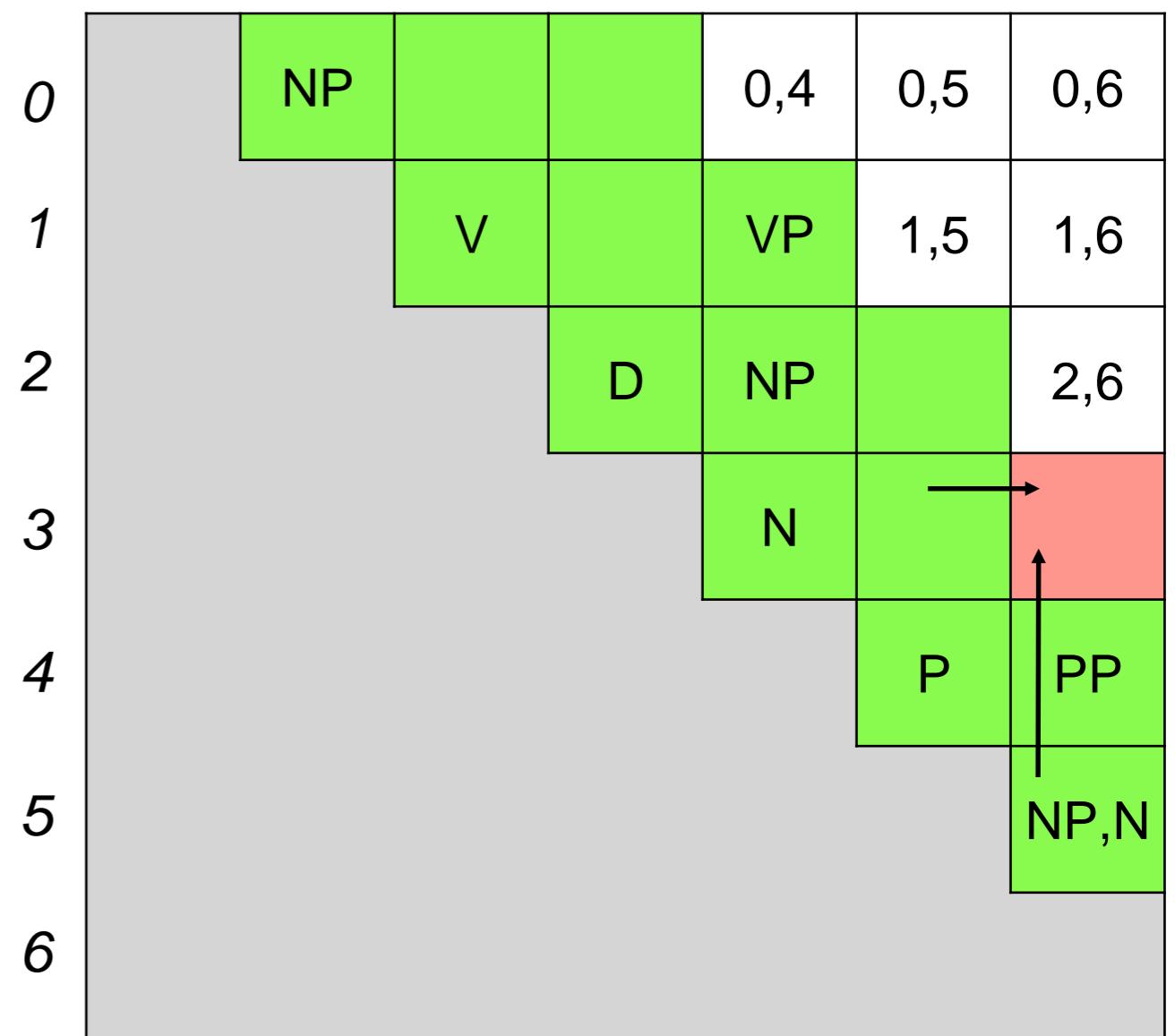
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=3$

$i=3, k=5, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

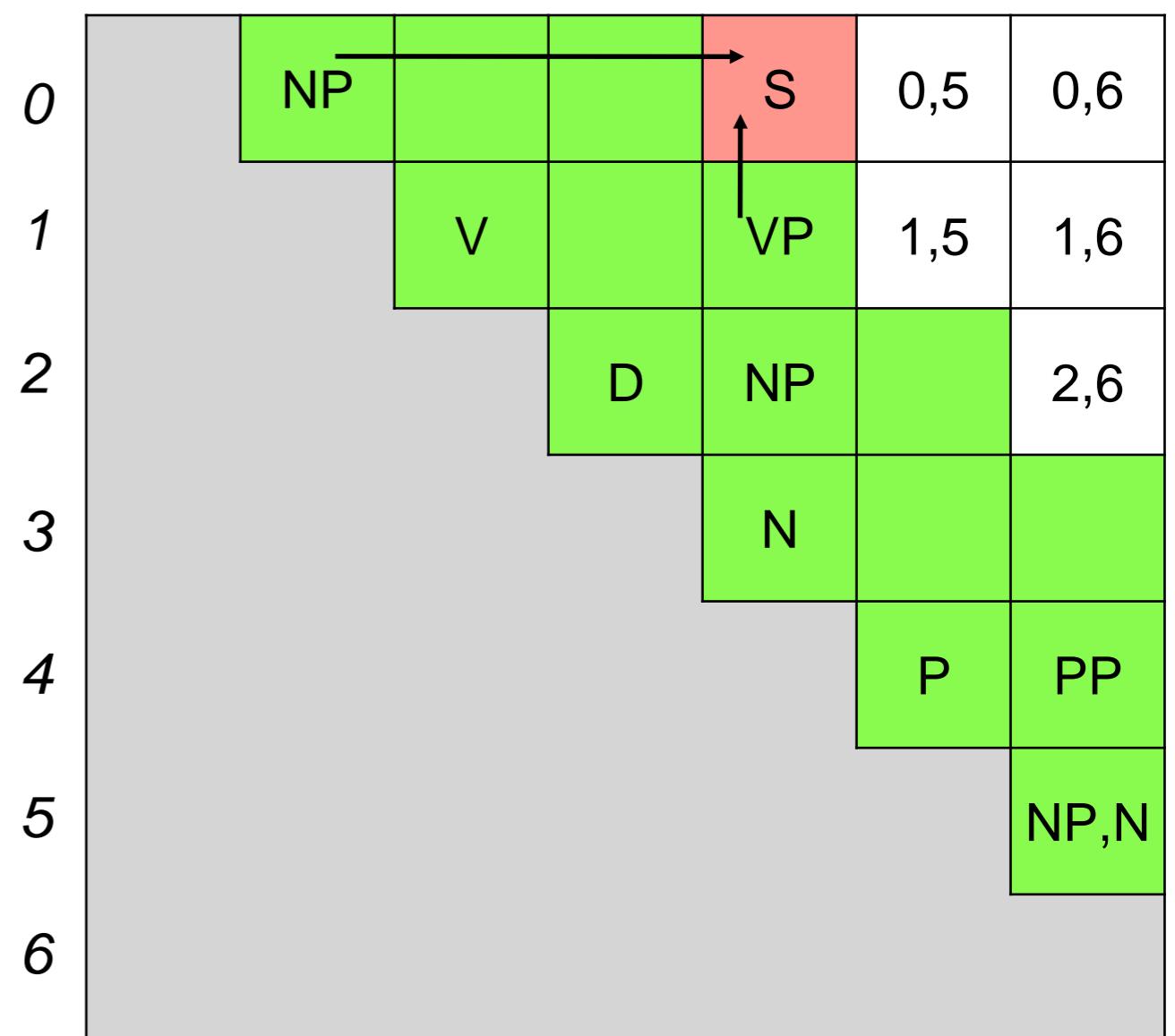
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

S → NP VP	NP → she
VP → V NP	NP → glasses
VP → VP PP	D → the
PP → P NP	N → cat
NP → D N	N → glasses
NP → NP PP	V → saw
	P → with

length=4

i=0,k=1,j=4

0 *she* 1 *saw* 2 *the* 3 *cat* 4 *with* 5 *glasses*



CKY Algorithm

```

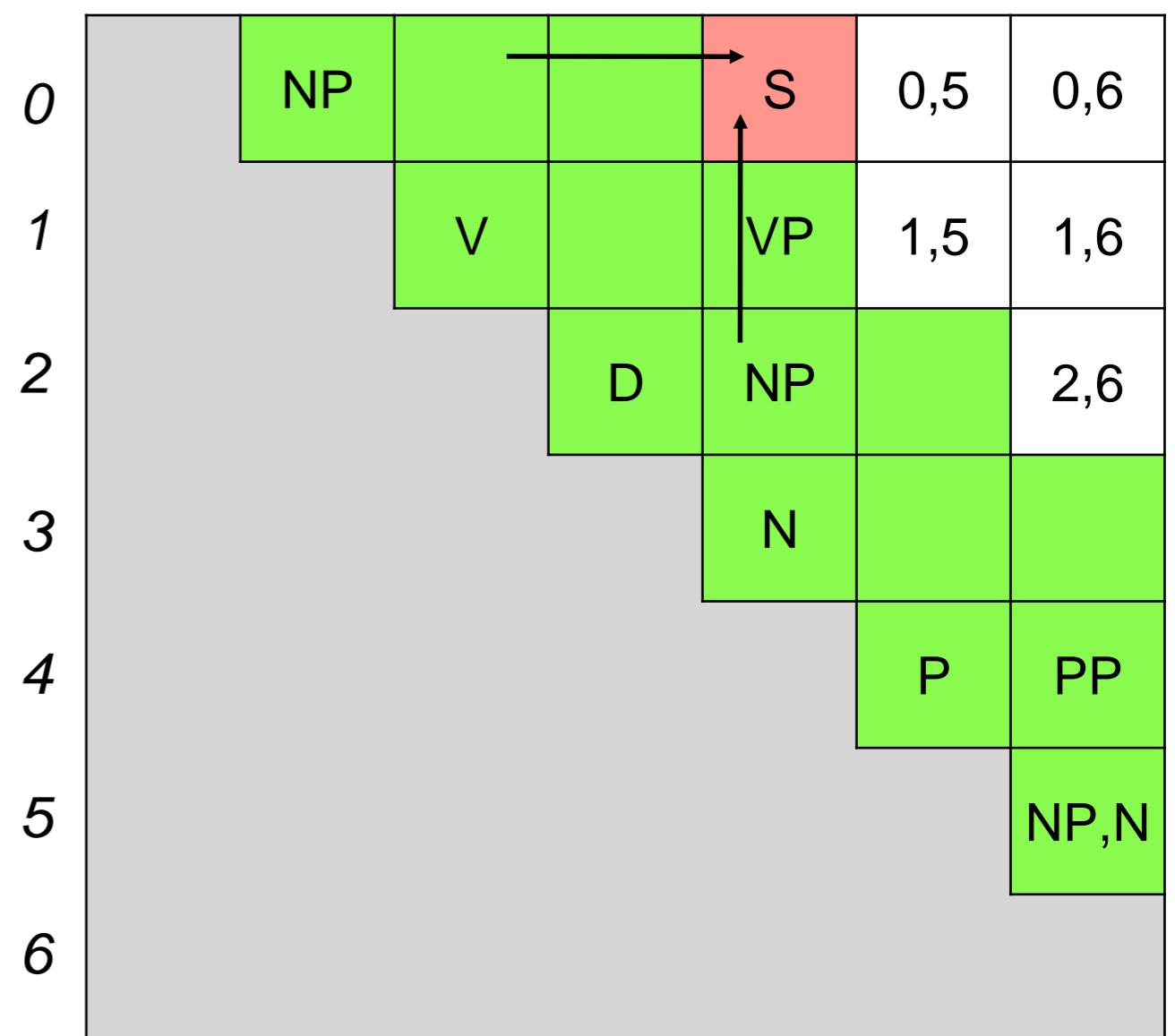
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=0, k=2, j=4$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for  $i=0 \dots (n-length)$ :
     $j = i+length$ 
    for  $k=i+1 \dots j-1$ :
        ....

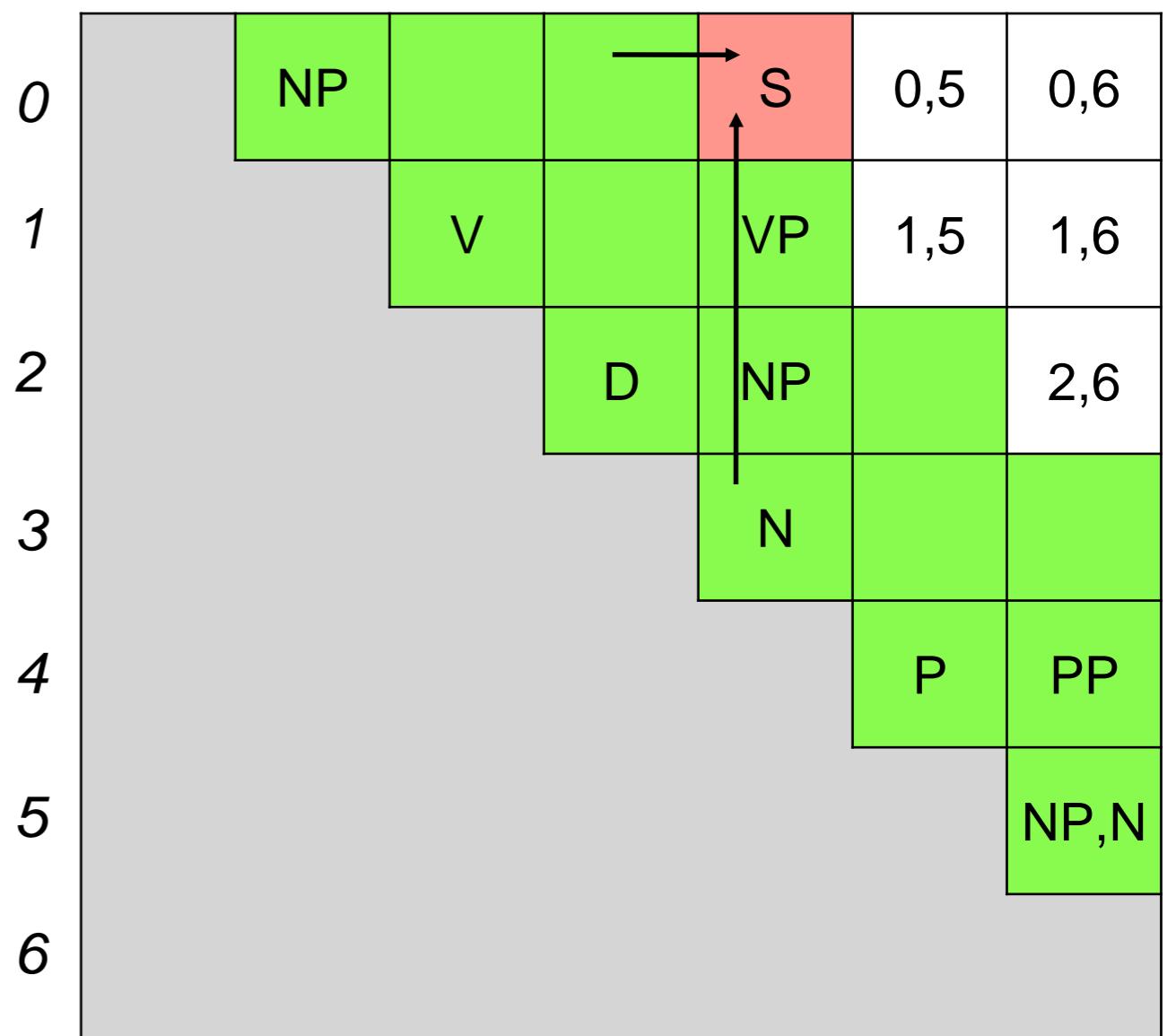
```

S → NP VP	NP → she
VP → V NP	NP → glasses
VP → VP PP	D → the
PP → P NP	N → cat
NP → D N	N → glasses
NP → NP PP	V → saw
	P → with

length=4

$i=0, k=3, j=4$

0 she 1 saw 2 the 3 cat 4 with 5 glasses



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

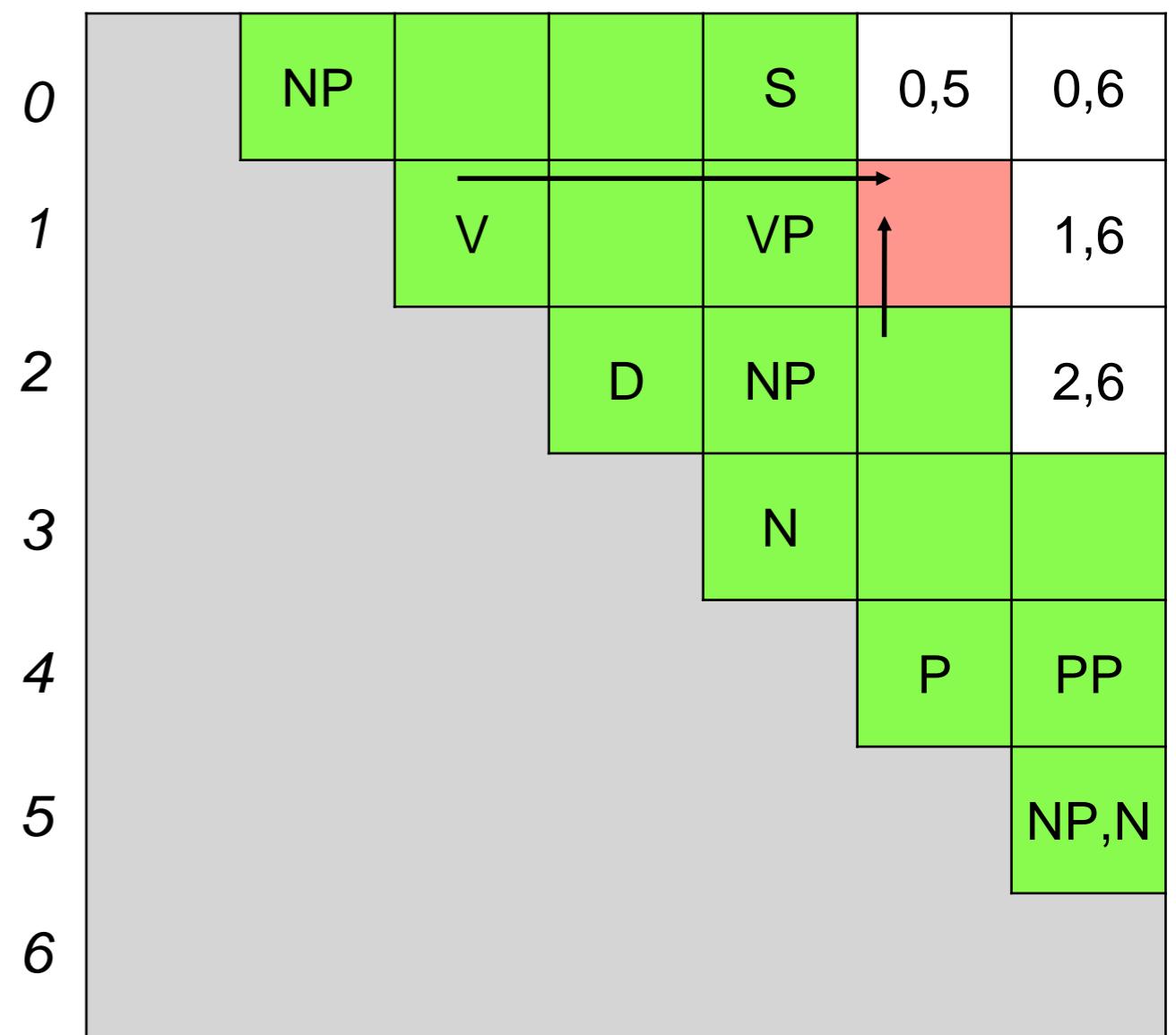
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=1, k=2, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

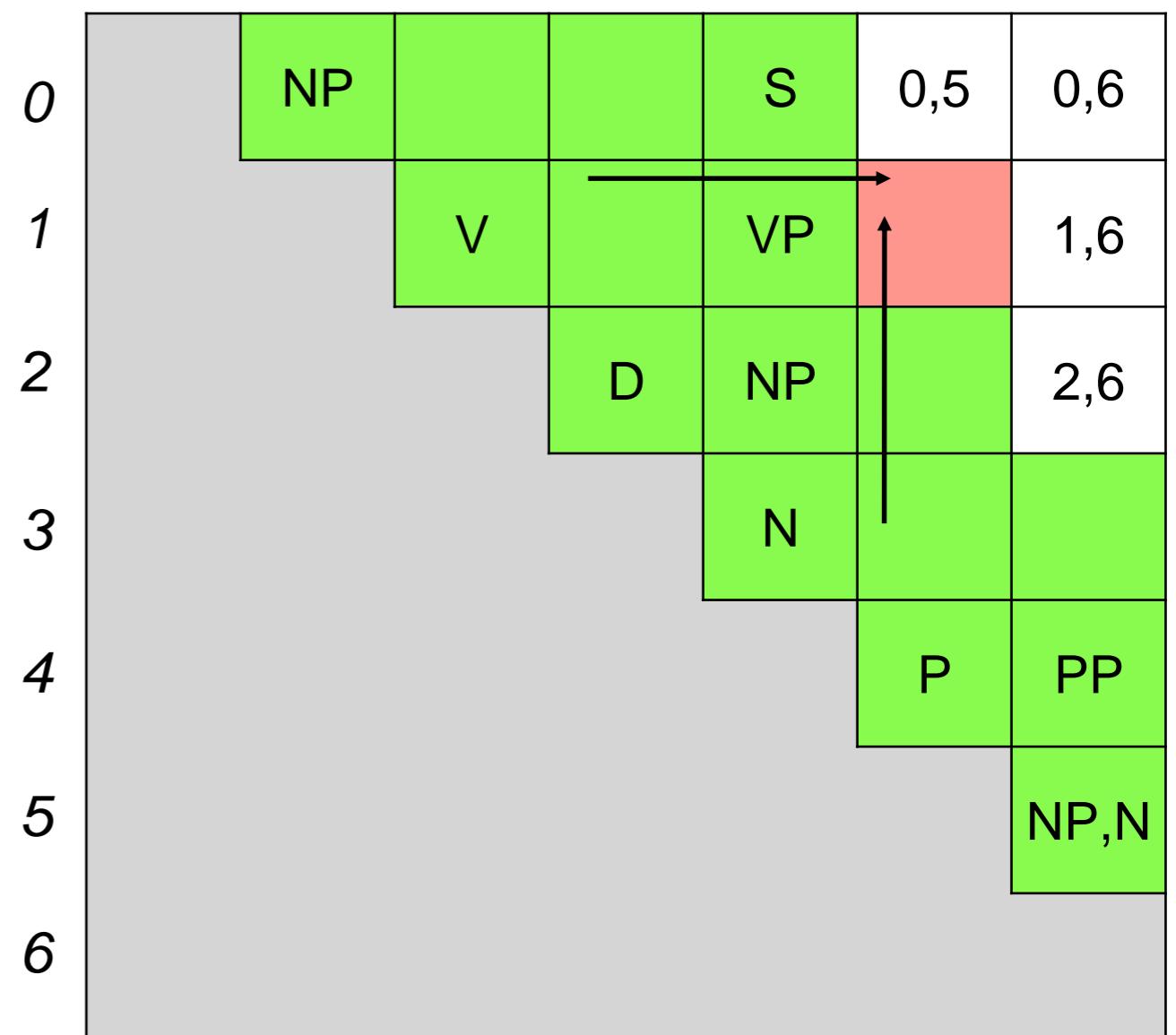
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=1, k=3, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

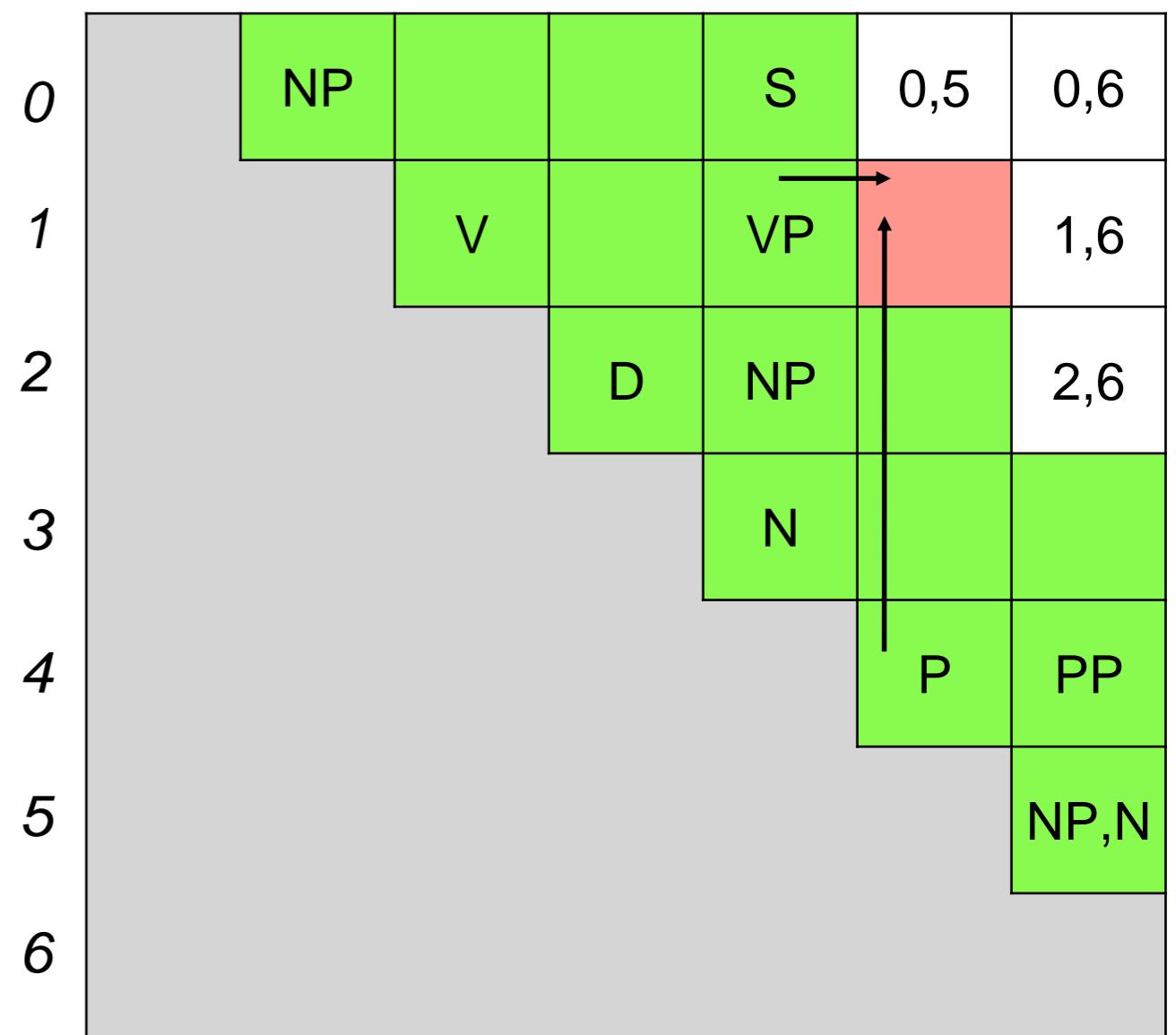
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=1, k=4, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

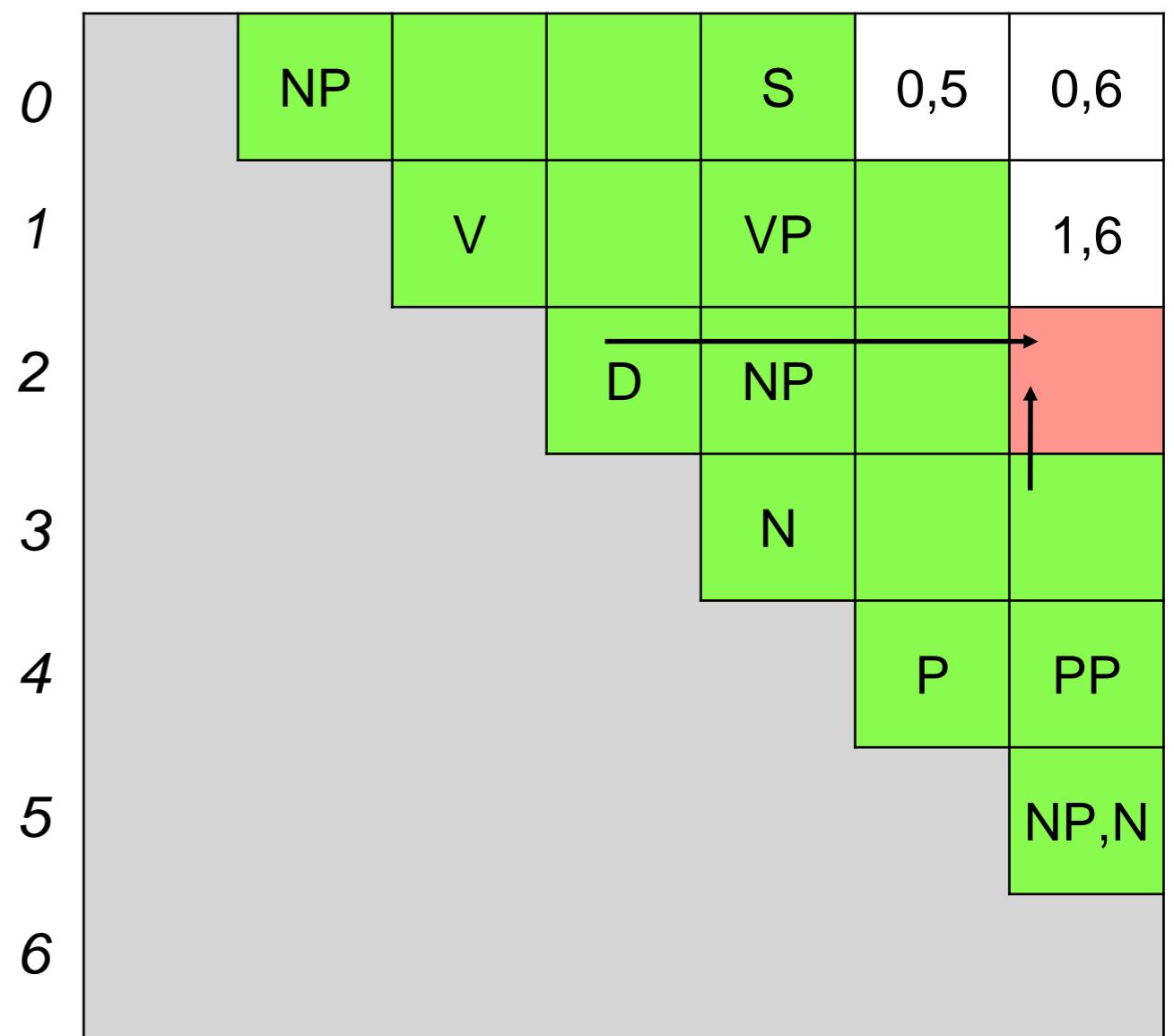
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=2, k=3, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

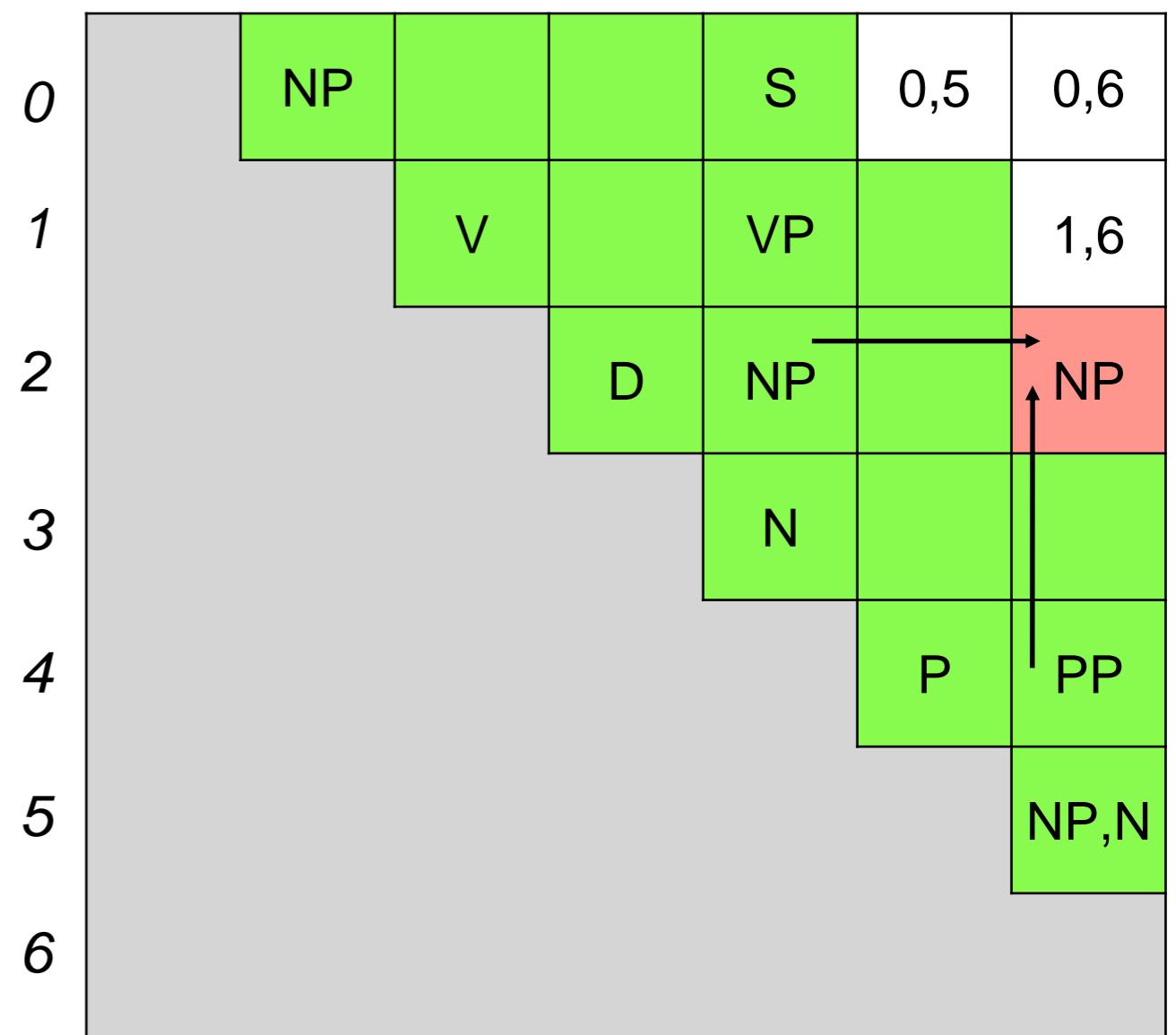
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=2, k=4, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

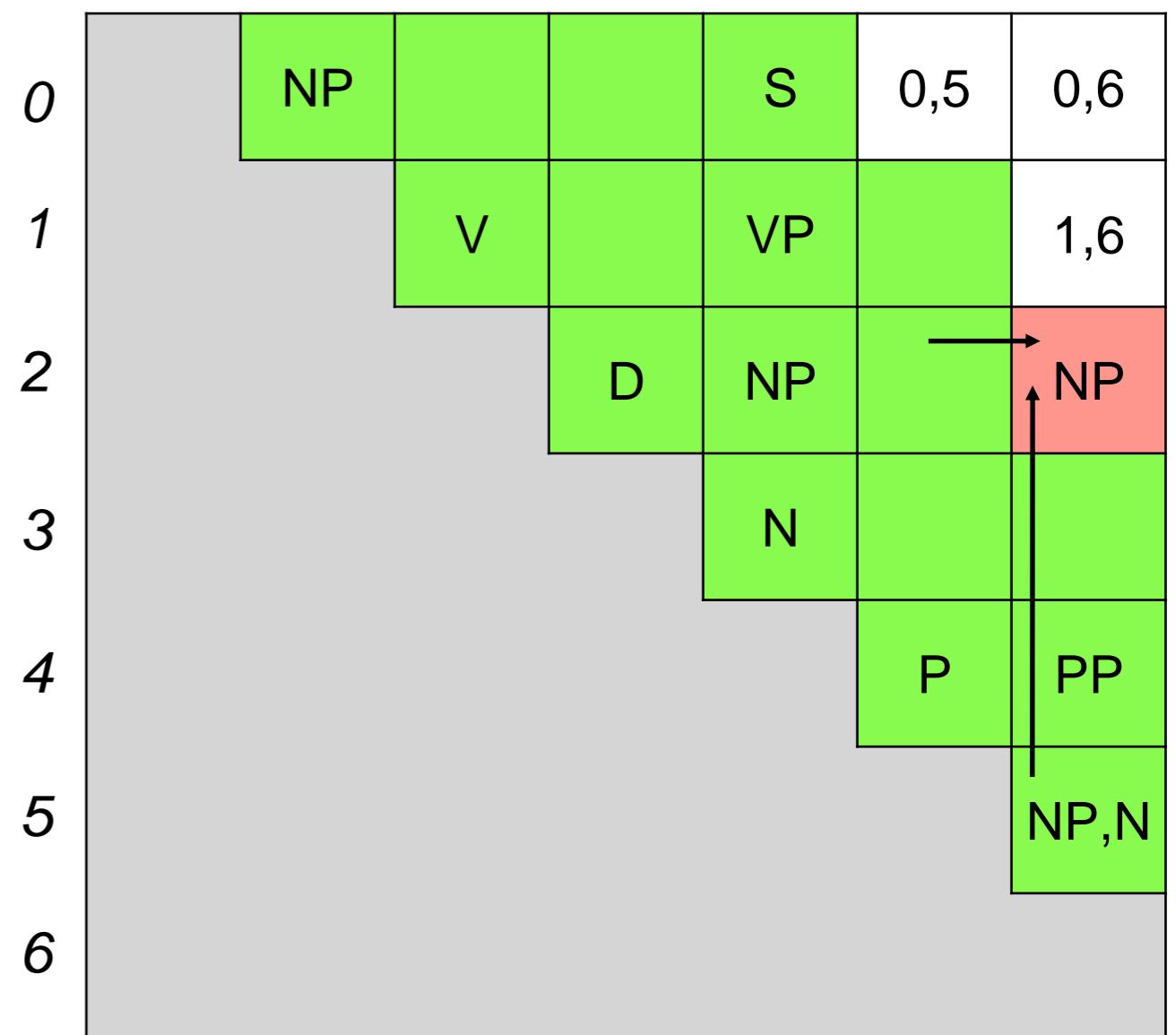
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=4$

$i=2, k=5, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

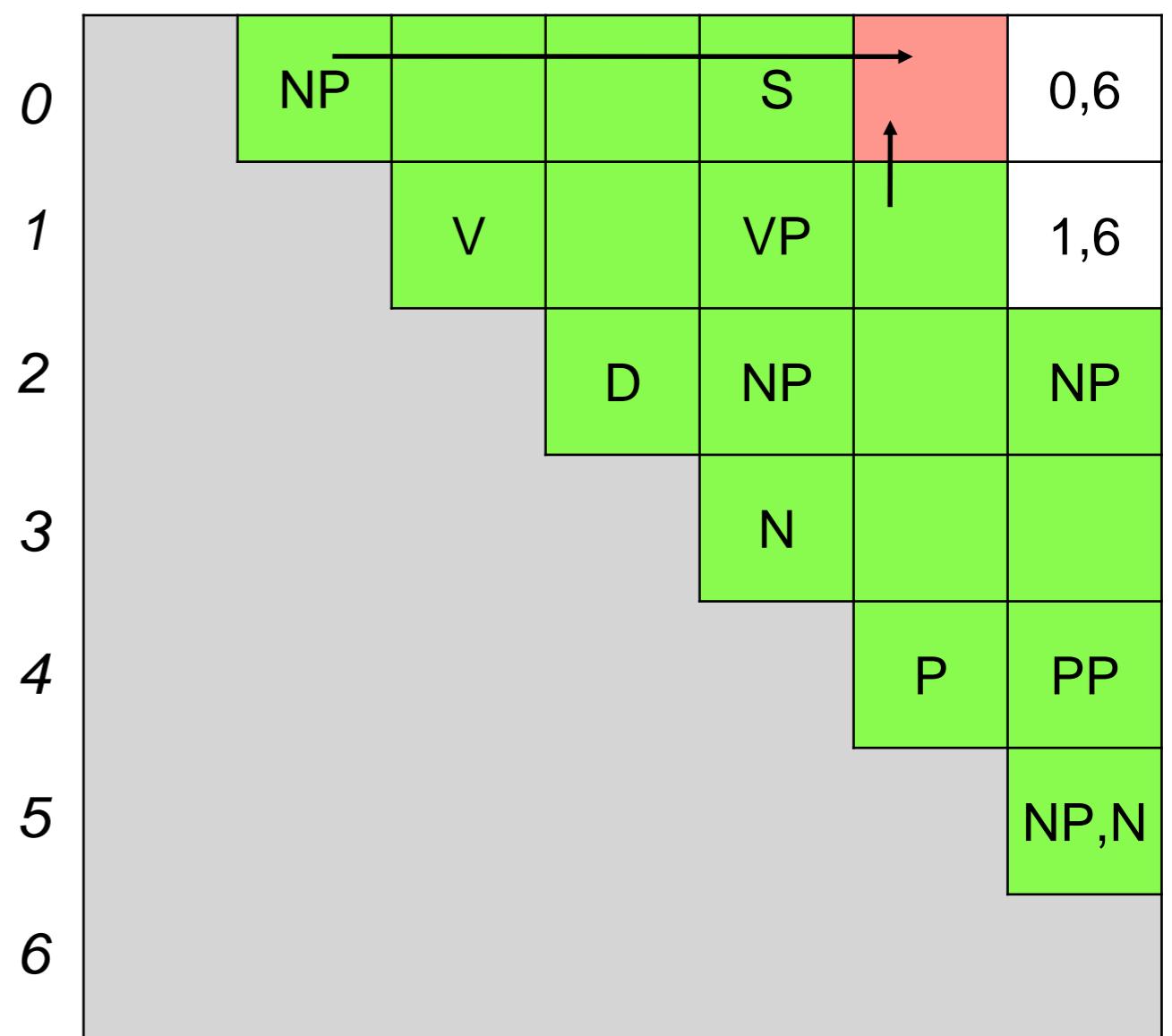
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=1, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

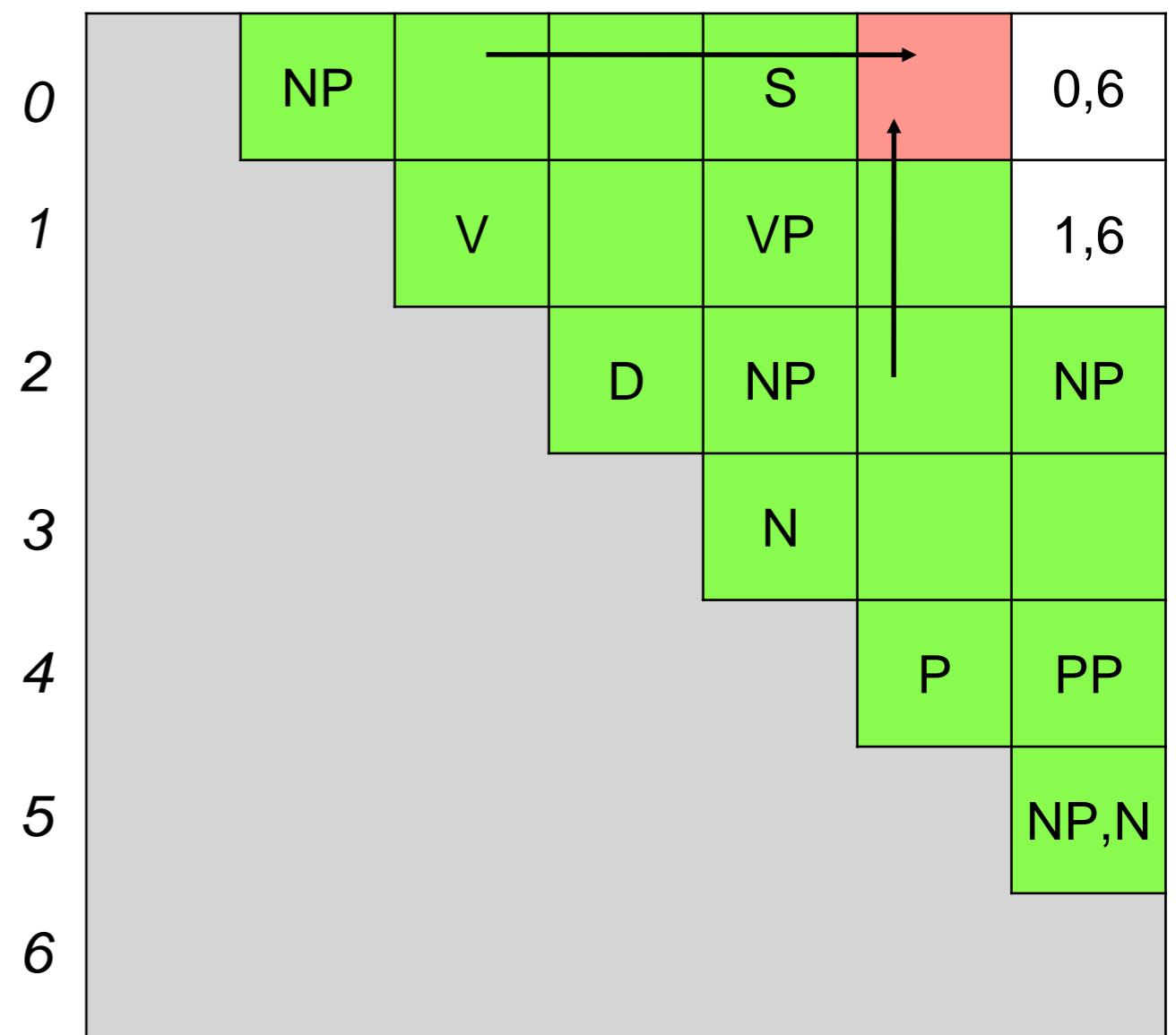
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=2, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

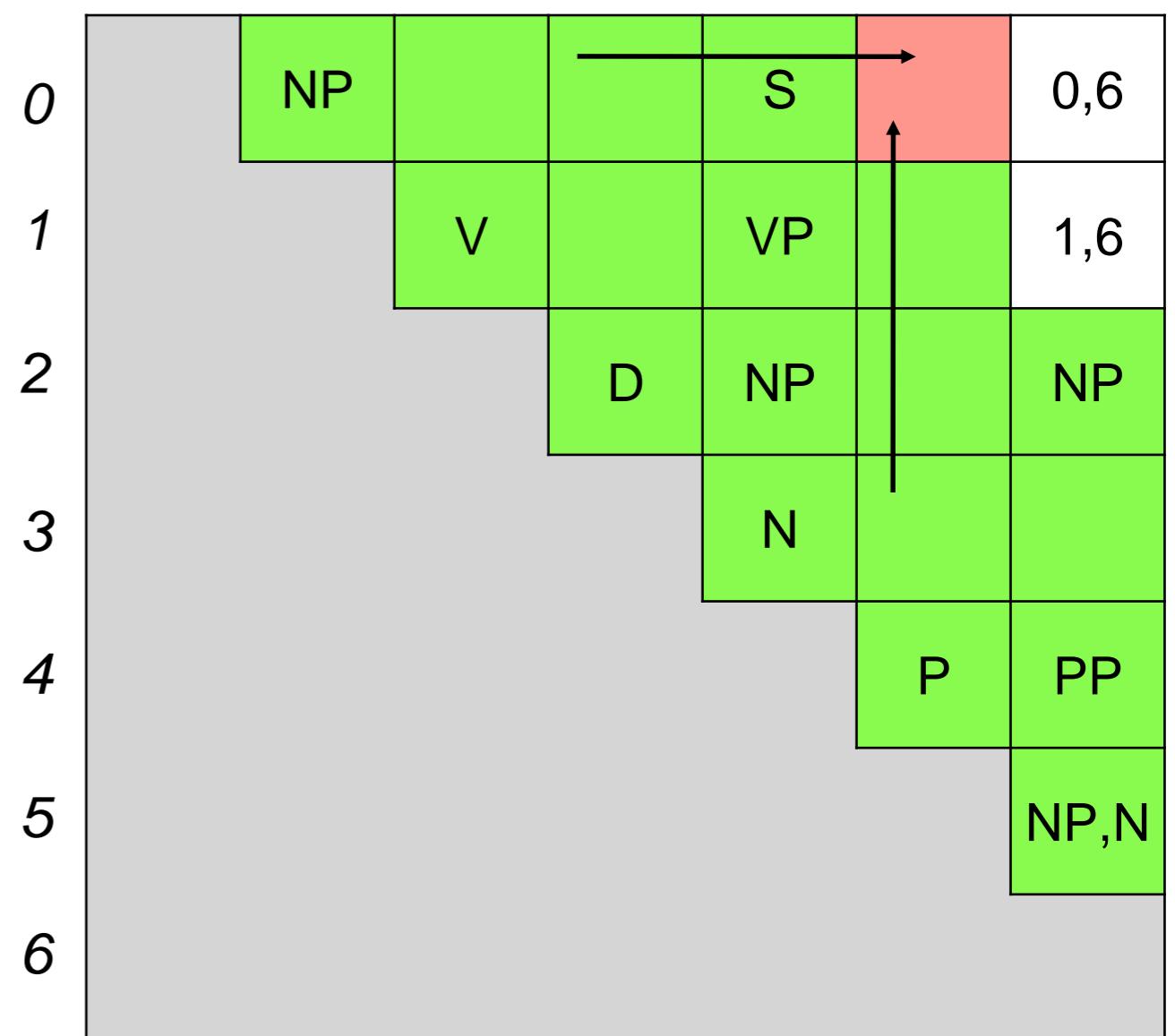
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=3, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

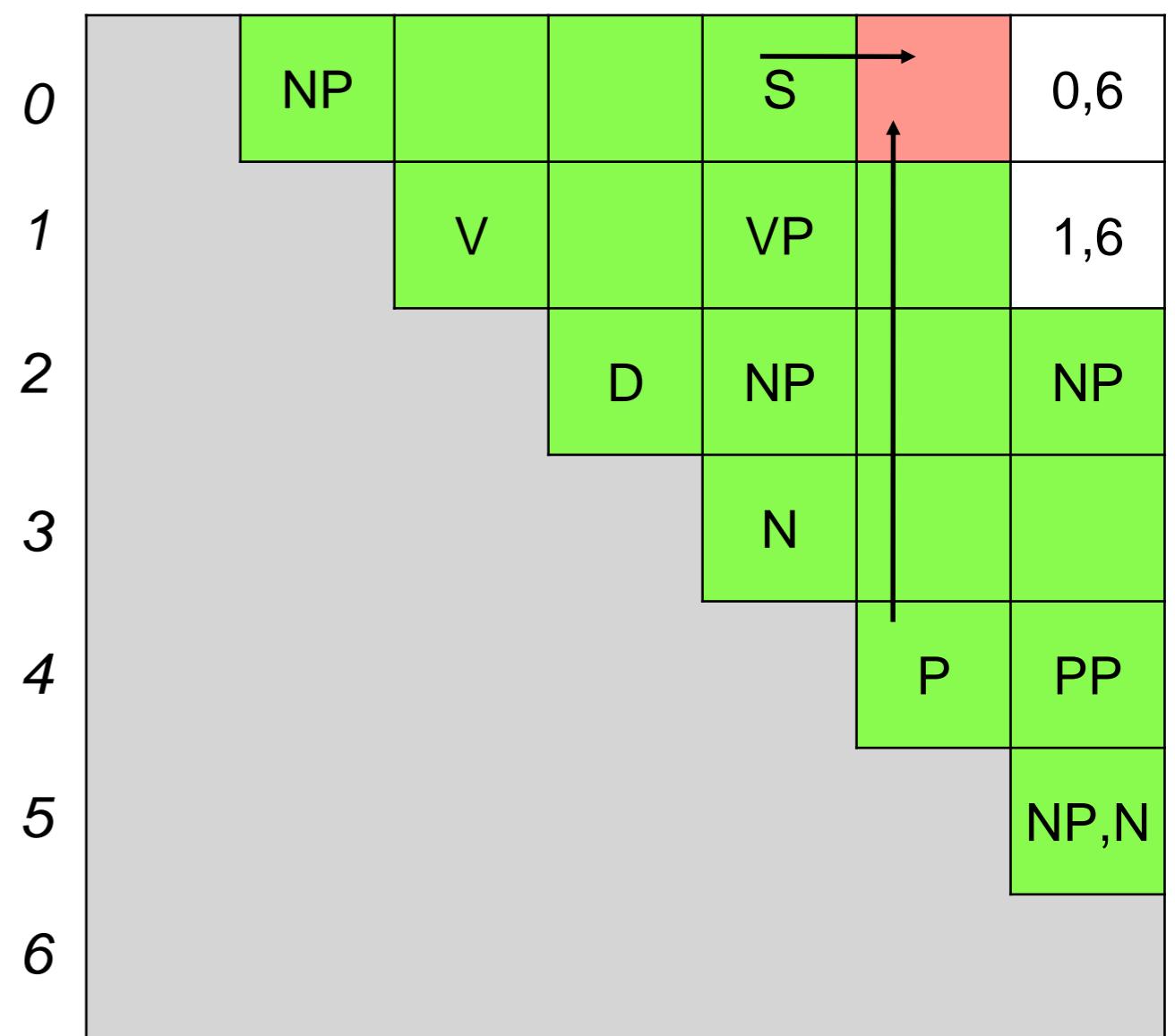
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=4, j=5$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

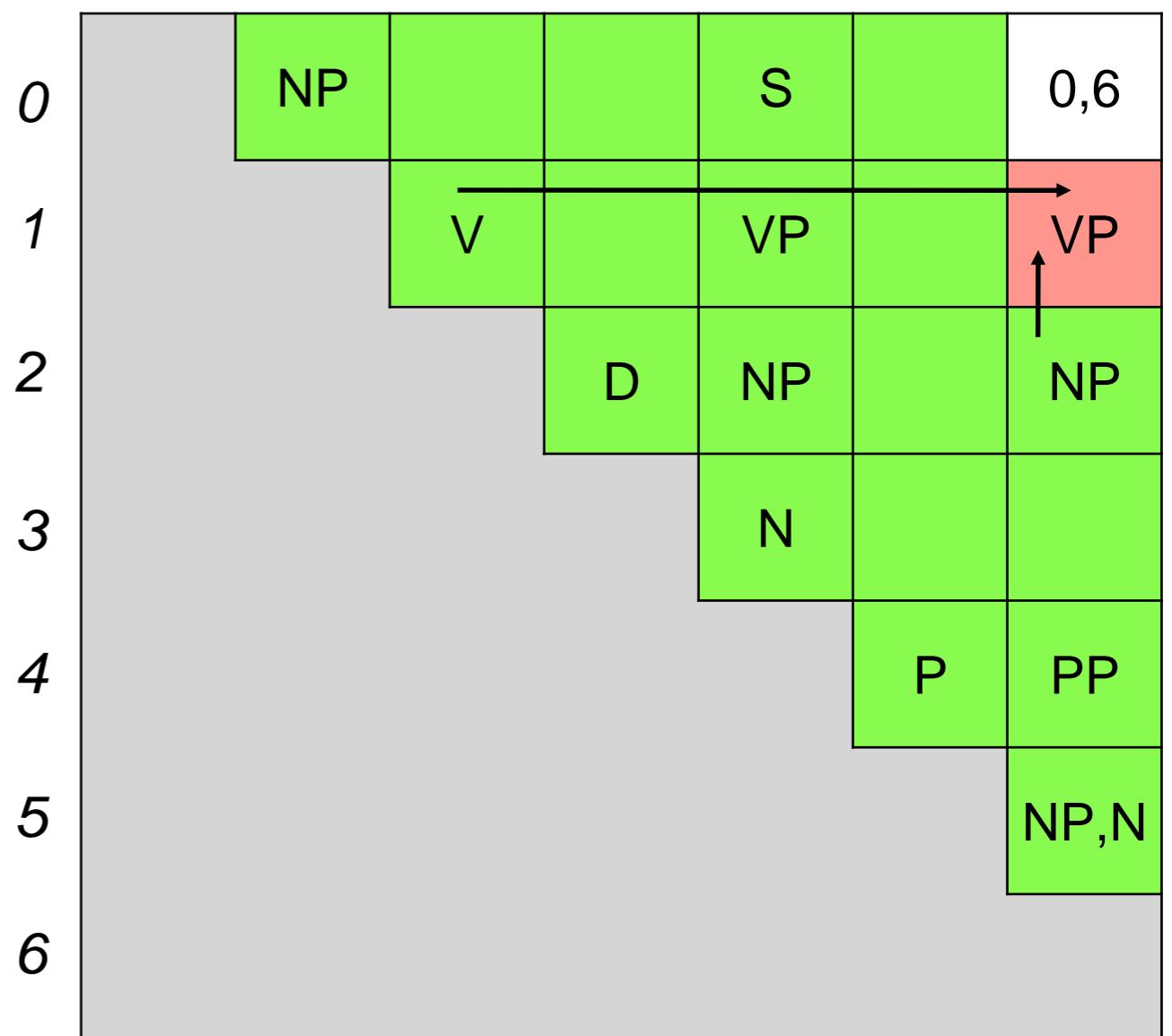
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=1, k=2, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....

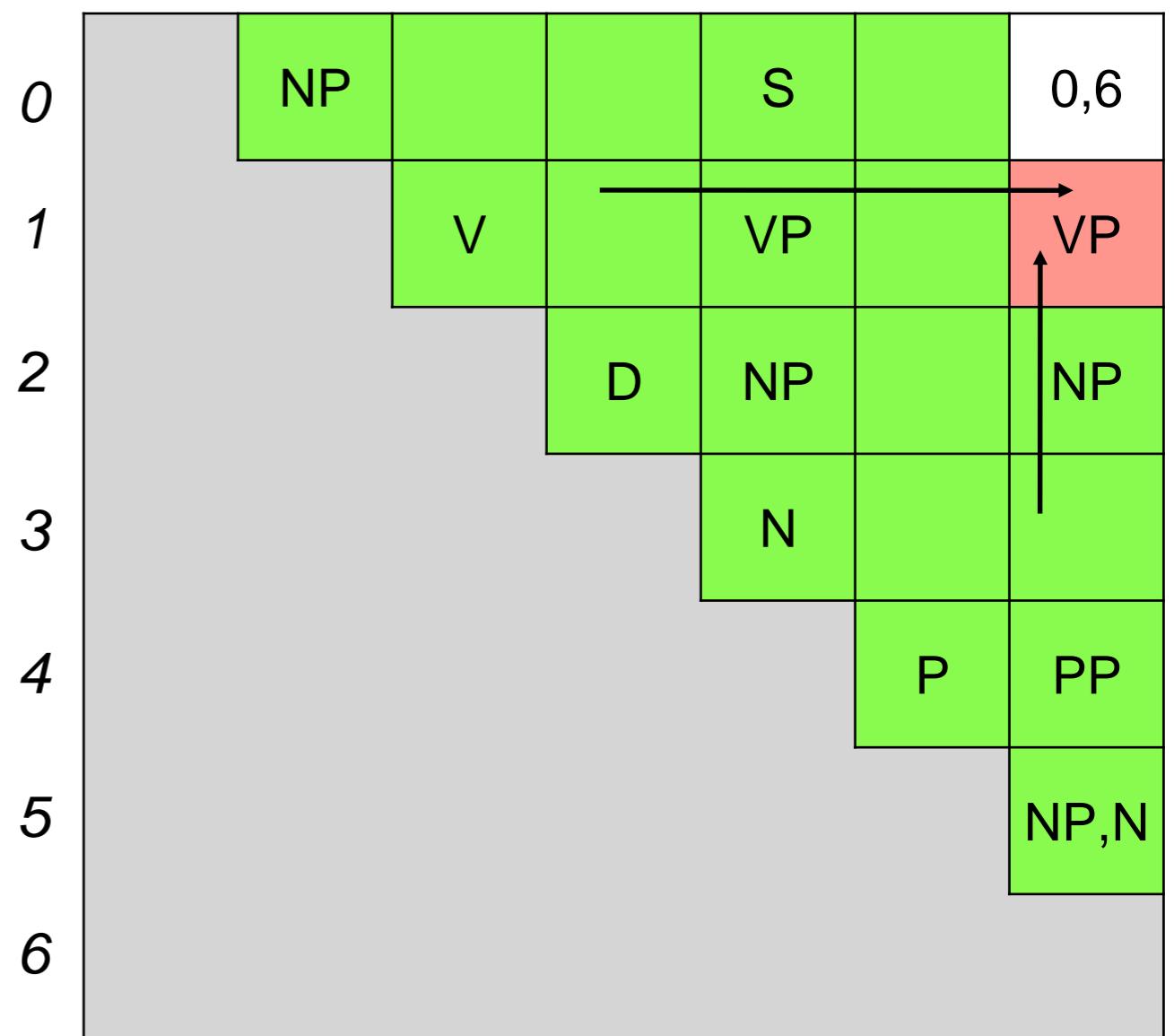
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=1, k=3, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

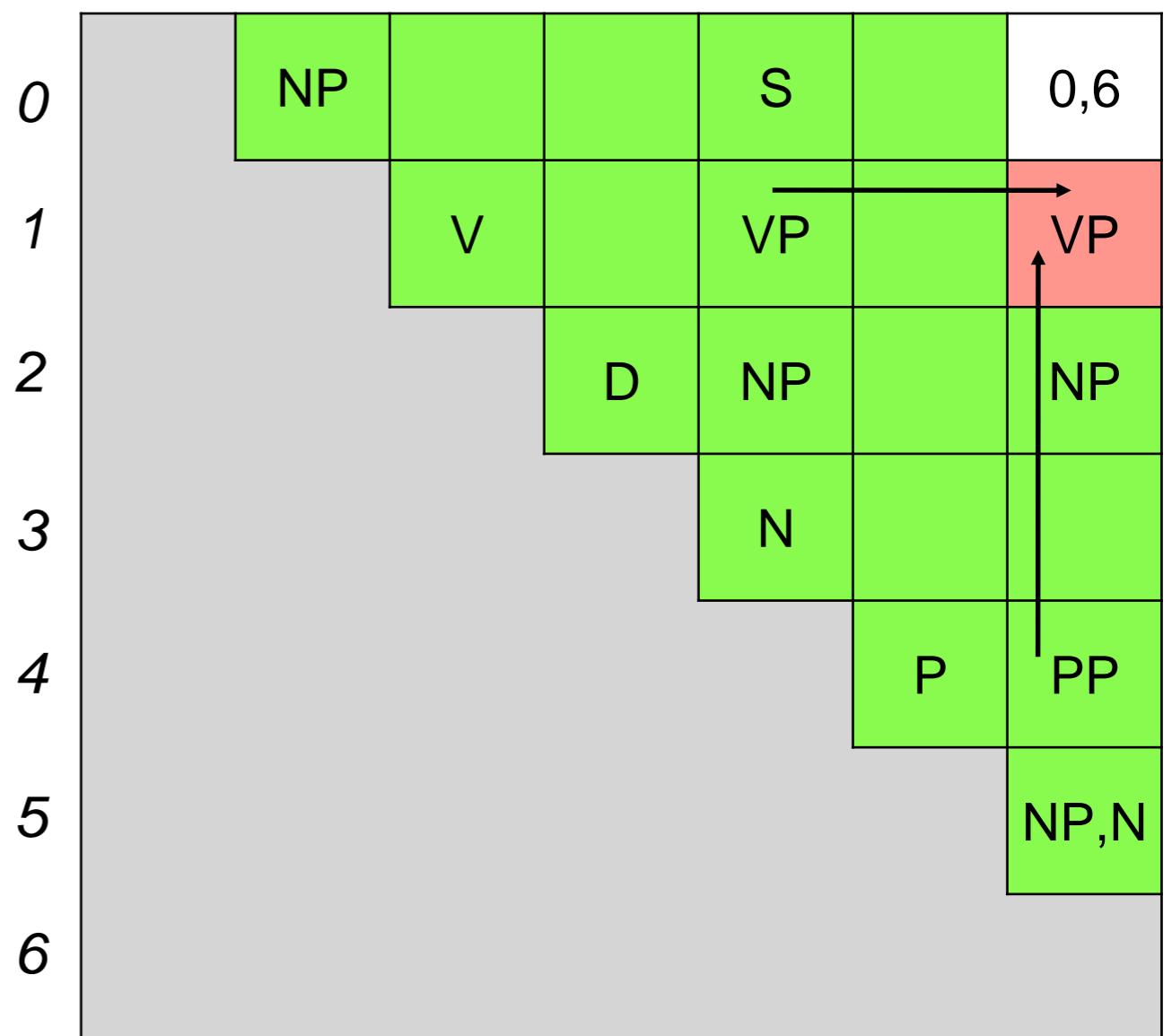
$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

! We can build VP over [1,6] in two ways!

$length=5$

$i=1, k=4, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

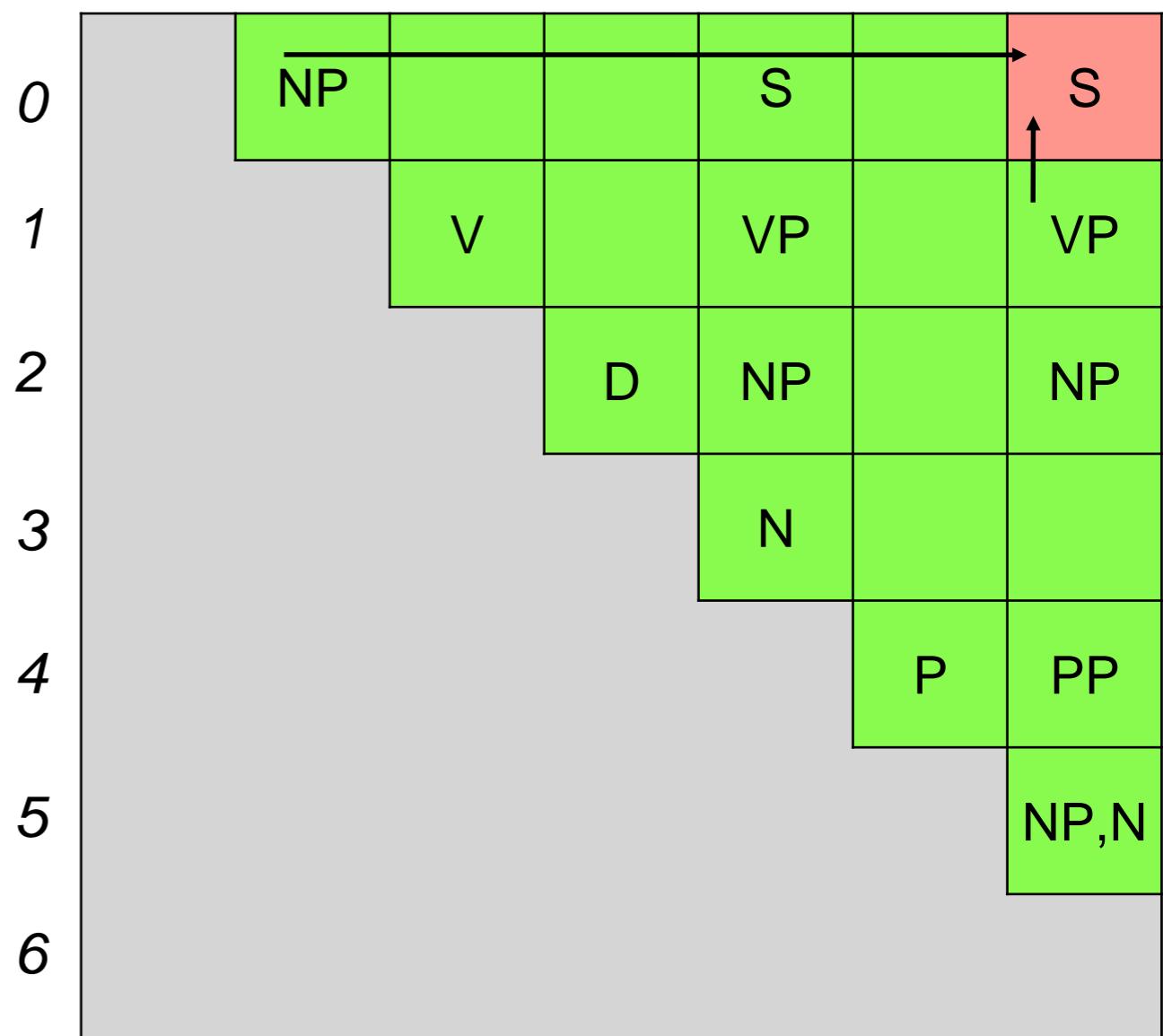
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=1, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

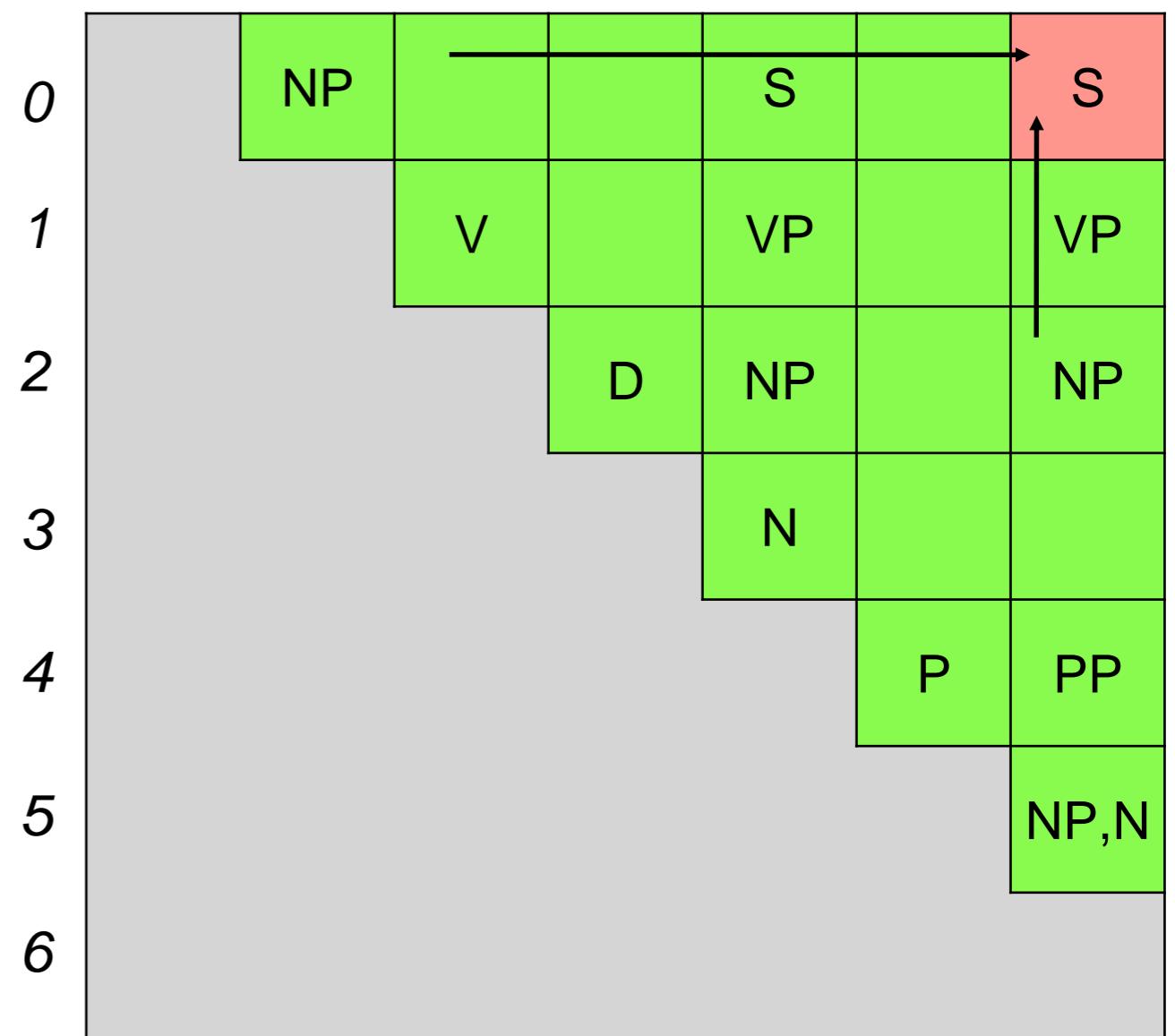
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=2, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

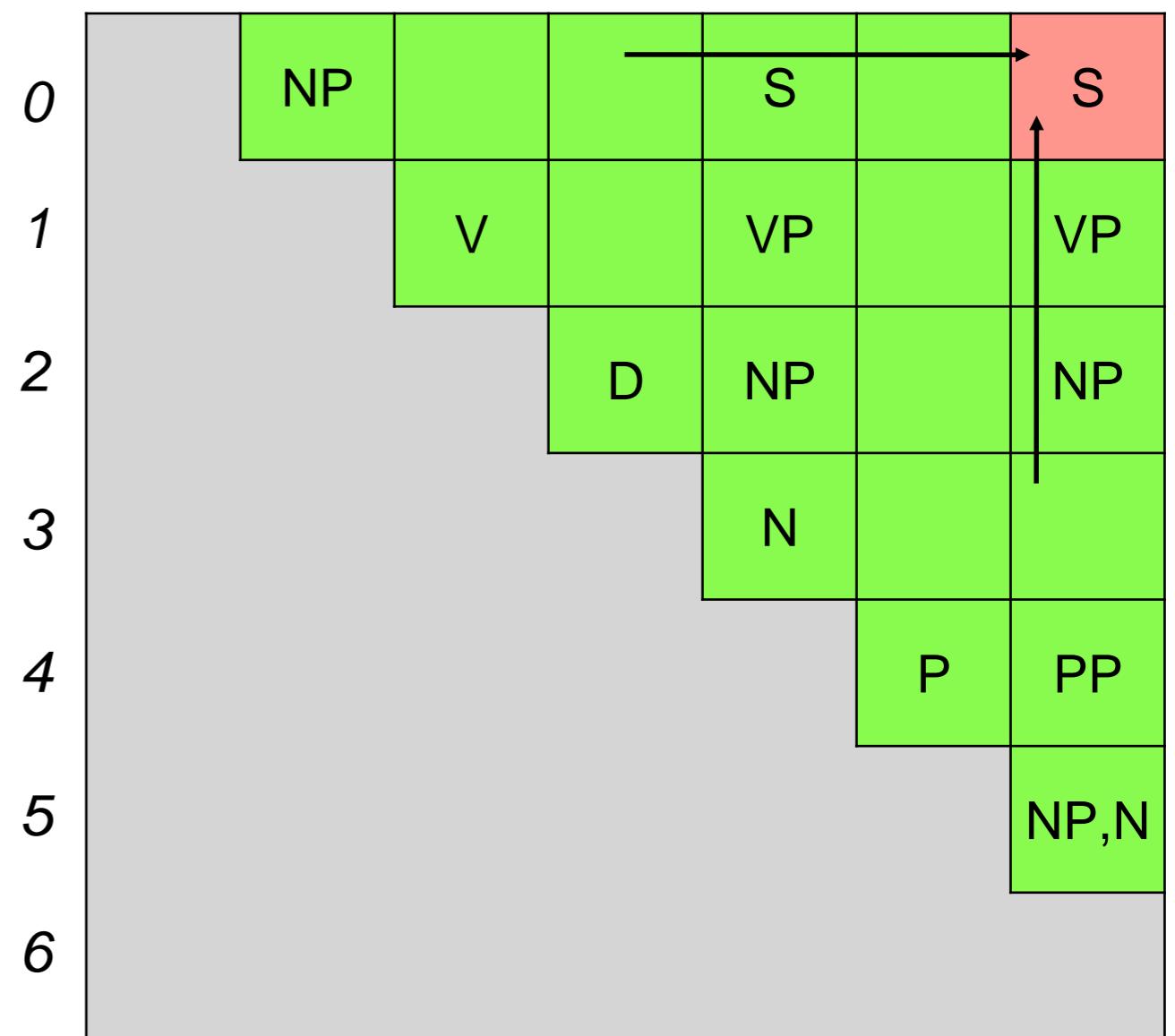
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=3, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

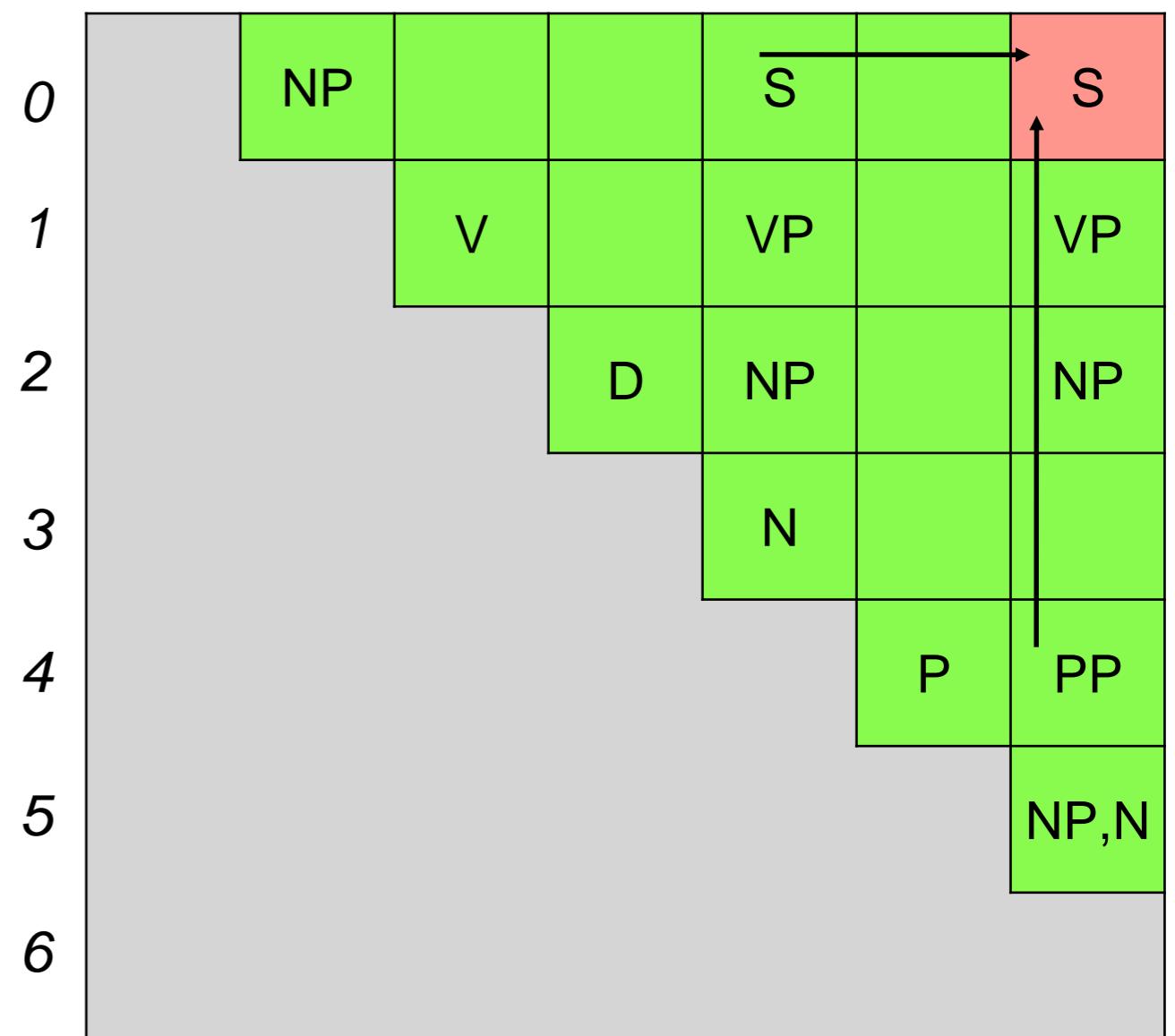
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=4, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$



CKY Algorithm

```

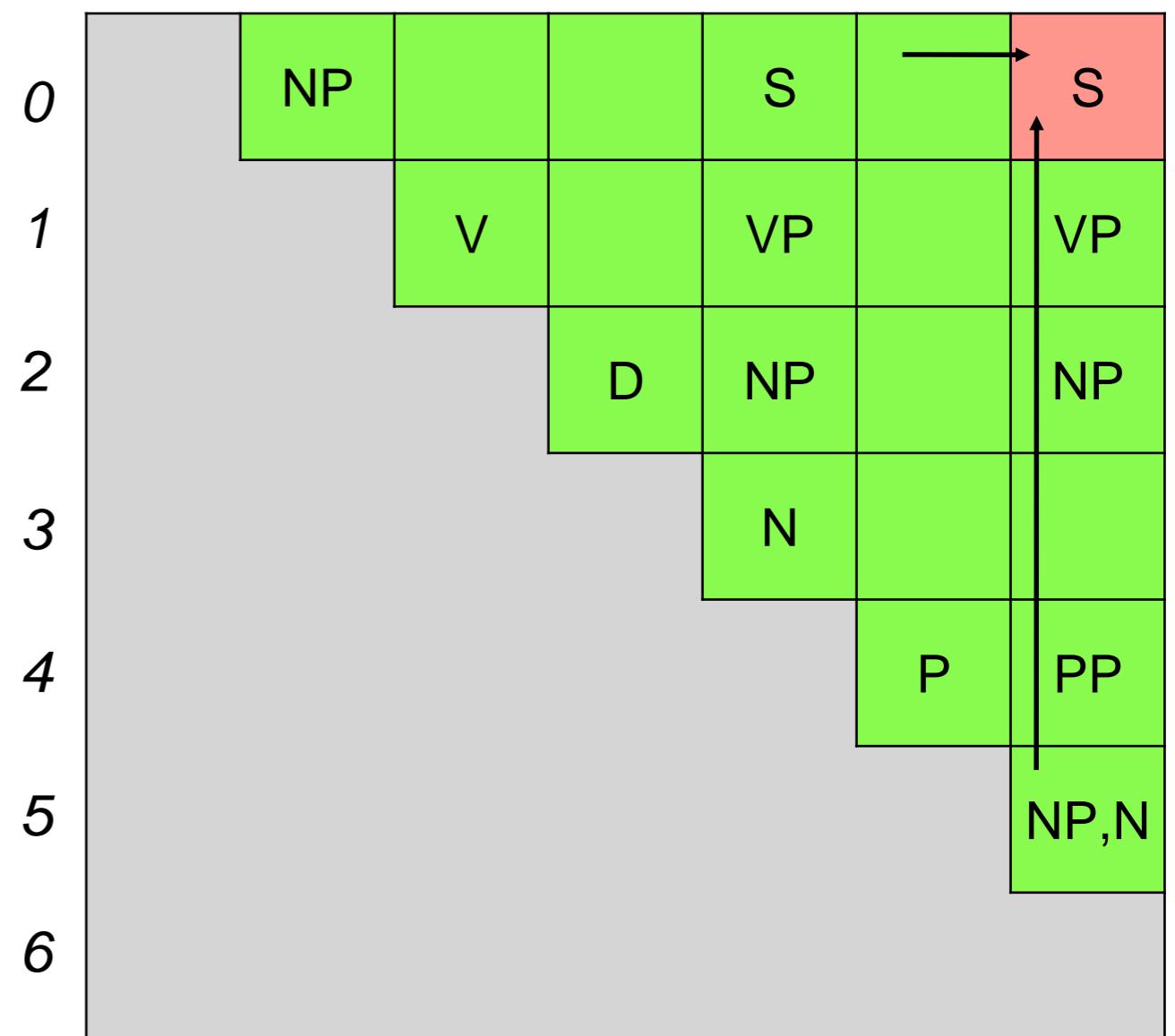
for i=0...(n-length):
    j = i+length
    for k=i+1...j-1:
        ....
    
```

$S \rightarrow NP VP$	$NP \rightarrow she$
$VP \rightarrow V NP$	$NP \rightarrow glasses$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow glasses$
$NP \rightarrow NP PP$	$V \rightarrow saw$
	$P \rightarrow with$

$length=5$

$i=0, k=5, j=6$

$0 \text{ she } 1 \text{ saw } 2 \text{ the } 3 \text{ cat } 4 \text{ with } 5 \text{ glasses}$

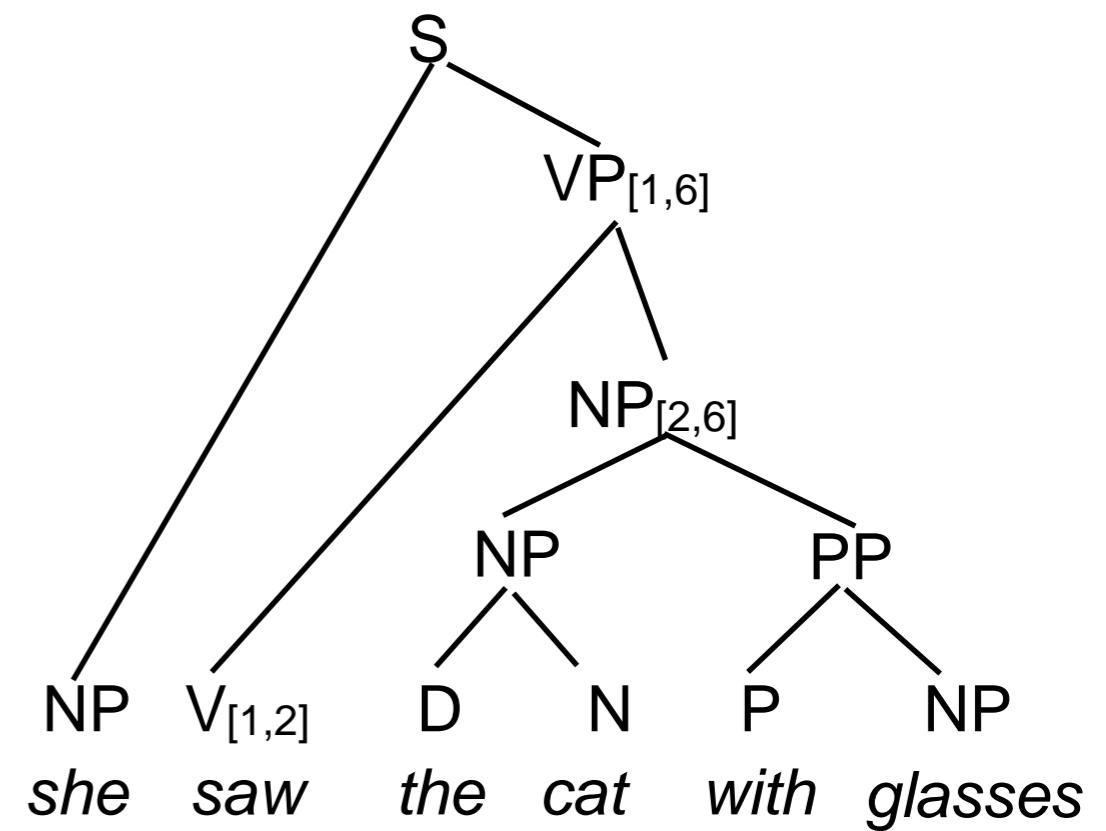
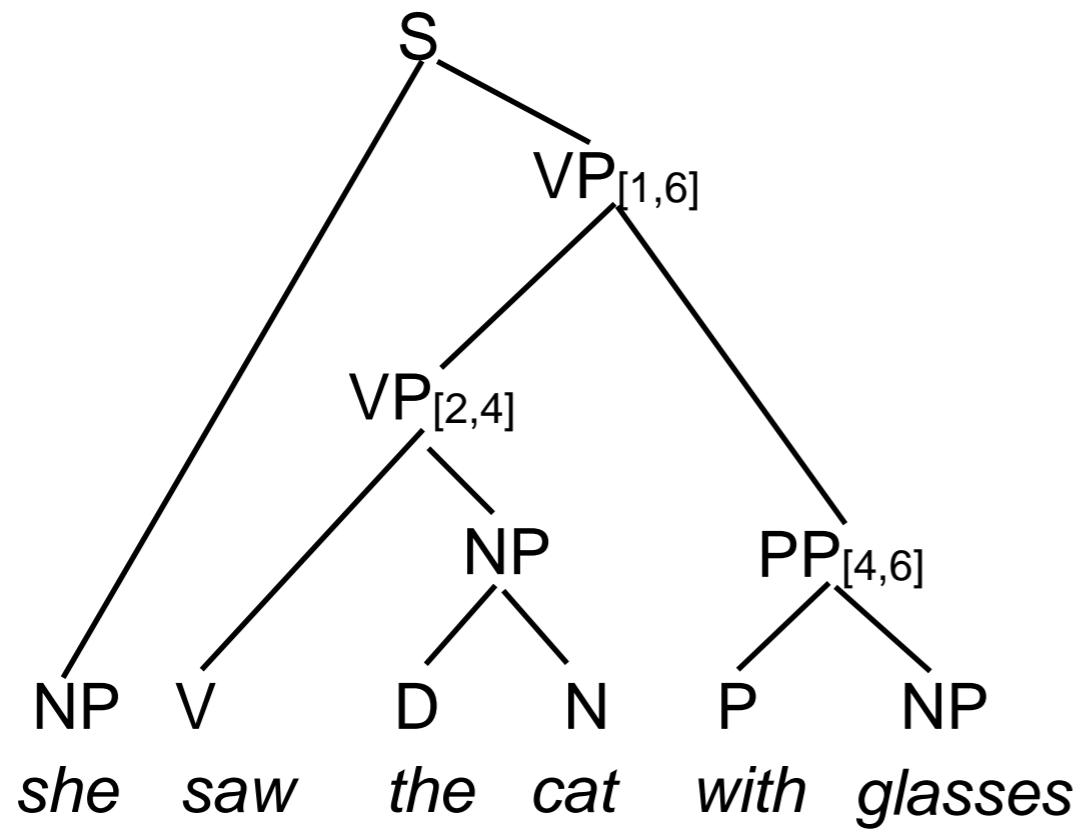


CKY Runtime

- **Input:** Grammar $G=(N, \Sigma, R, S)$, input string s of length n .
- for $i=0\dots n-1$: O(N x |R|)
 $\pi[i, i+1] = \{A \mid A \rightarrow s[i]\}$
- for $length=2\dots n$: O(N)
 for $i=0\dots (n-length)$: O(N) Total : O(N³ x |R|)
 $j = i+length$ O(N)
 for $k=i+1\dots j-1$:
 $M = \{A \mid A \rightarrow BC \text{ and } C \in R \text{ and } B \in \pi[i, k] \text{ and } C \in \pi[k, j]\}$
 $\pi[i, j] = \pi[i, j] \cup M$
- if $S \in \pi[0, i+1]$ return True, otherwise False

Syntactic Ambiguity

$S \rightarrow NP\ VP$	$NP \rightarrow she$
$VP \rightarrow V\ NP$	$NP \rightarrow glasses$
$VP \rightarrow VP\ PP$	$D \rightarrow the$
$PP \rightarrow P\ NP$	$N \rightarrow cat$
$NP \rightarrow D\ N$	$N \rightarrow glasses$
$NP \rightarrow NP\ PP$	$V \rightarrow saw$
	$P \rightarrow with$

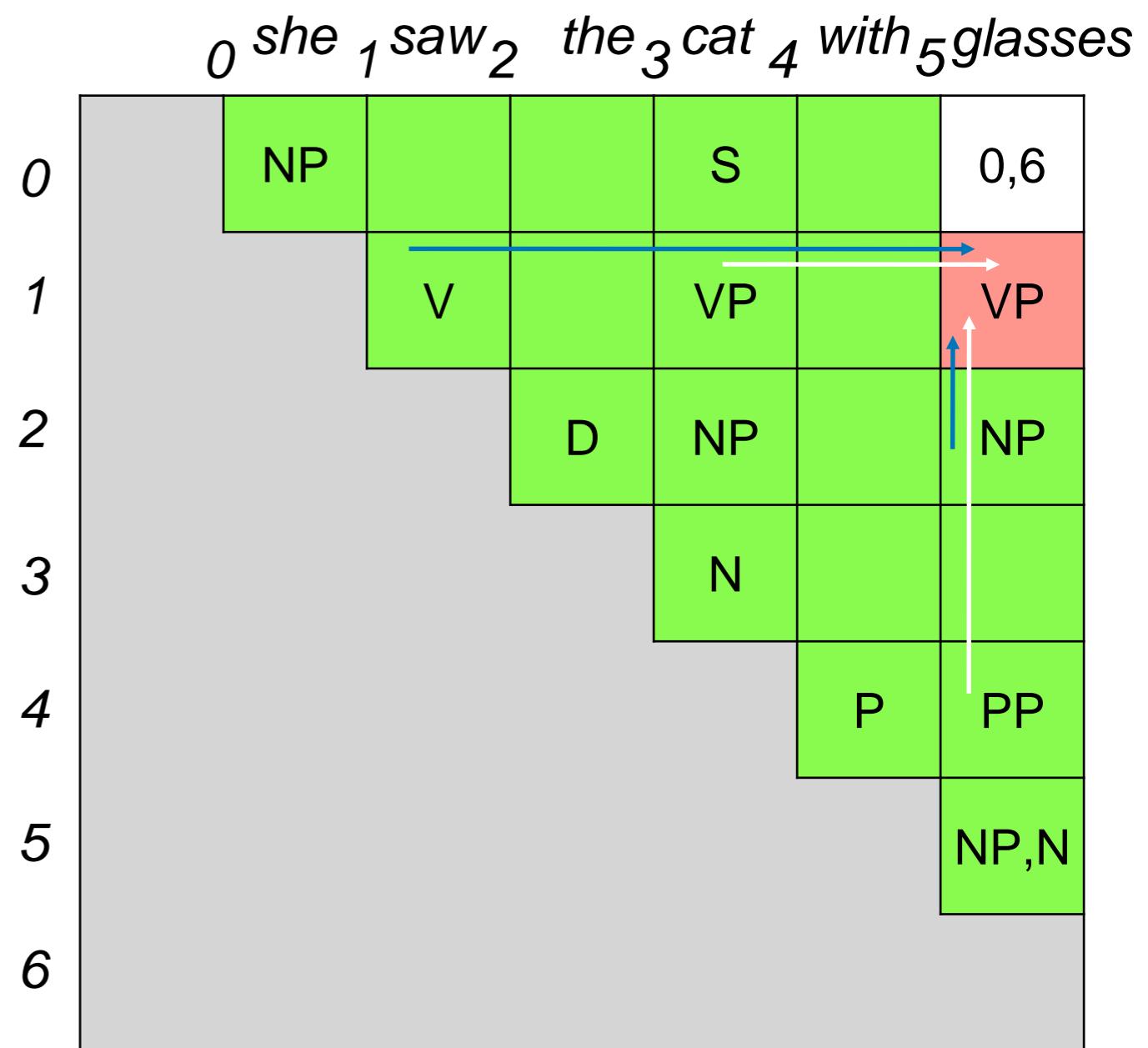


Backpointers

- The CKY algorithm presented so far determines if a sentence is recognized by a grammar.

- Also want to retrieve the parse trees!
- Instead of a set of nonterminals, store a list of instantiated rules and backpointers.

$$\left\{ \begin{array}{l} VP_{[1,6]} \rightarrow V_{[1,2]} \quad NP_{[2,6]} \\ VP_{[1,6]} \rightarrow VP_{[1,4]} \quad PP_{[4,6]} \end{array} \right\}$$



Retrieving Parse-Trees

- Start at the $[0,n]$ entry and recursively follow the backpointers.
Return a set of subtrees from the recursion.

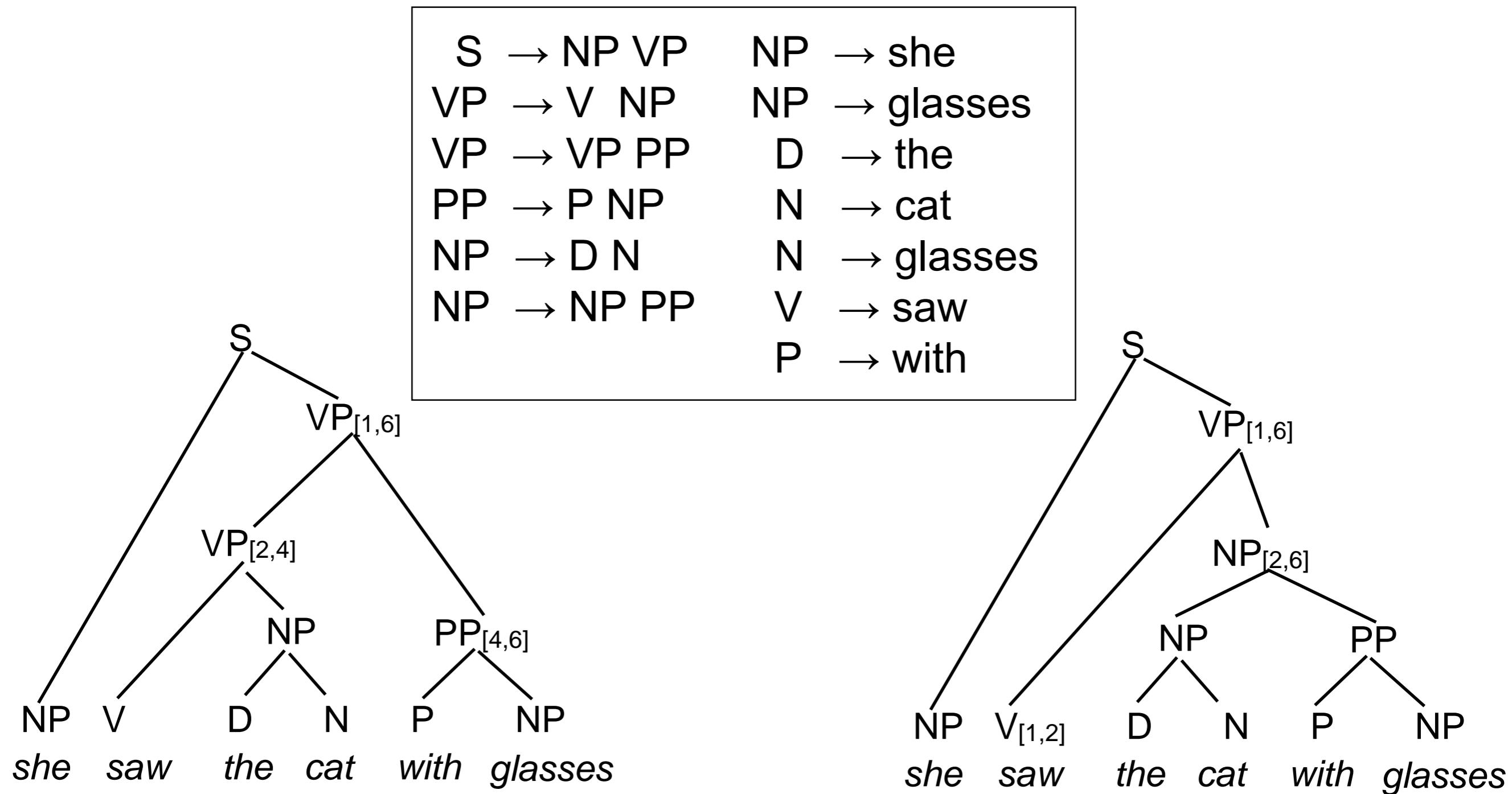
Natural Language Processing

Lecture 7: Parsing with Context Free Grammars II.
CKY for PCFGs. Earley Parser.

2/21/2020

COMS W4705
Yassine Benajiba

Recall: Syntactic Ambiguity



Which parse tree is “better”? More probable?

Probabilities for Parse Trees

- Let \mathcal{T}_G be the set of all parse trees generated by grammar G.
- We want a model that assigns a probability to each parse tree, such that $\sum_{t \in \mathcal{T}_G} P(t) = 1$.
- We can use this model to select the most probable parse tree compatible with an input sentence.
 - This is another example of a generative model!

Selecting Parse Trees

- Let $\mathcal{T}_G(s)$ be the set of trees generated by grammar G whose *yield* (sequence of leafs) is string s .
- The most likely parse tree produced by G for string s is

$$\arg \max_{t \in \mathcal{T}_G(s)} P(t)$$

- How do we define $P(t)$?
- How do we learn such a model from training data (annotated or un-annotated).
- How do we find the highest probability tree for a given sentence? (*parsing/decoding*)

Probabilistic Context Free Grammars (PCFG)

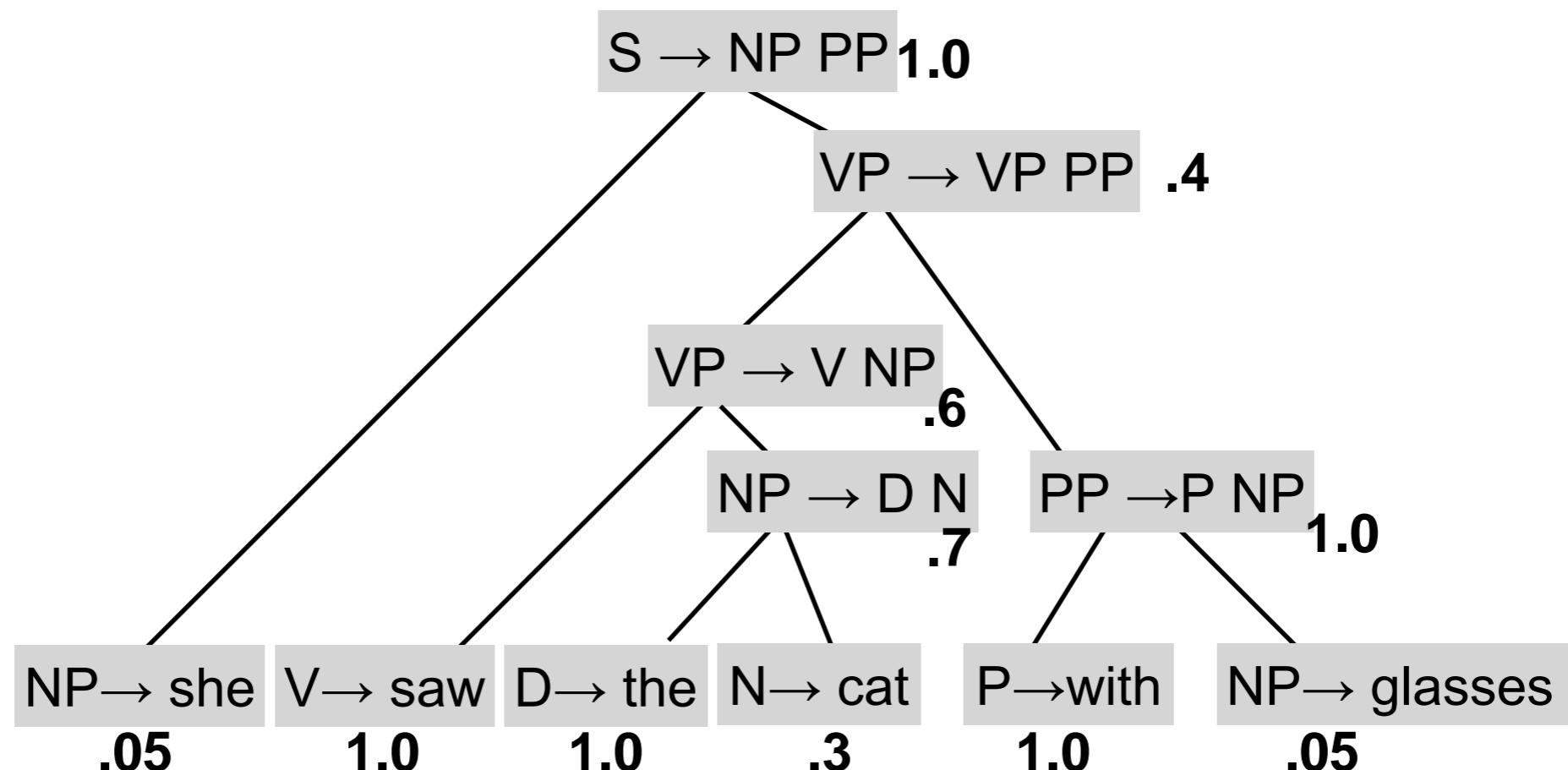
- A PCFG consists of a Context Free Grammar $G=(N, \Sigma, R, S)$ and a probability $P(A \rightarrow \beta)$ for each production $A \rightarrow \beta \in R$.
 - The probabilities for all rules with the same left-hand-side sum up to 1:
$$\sum_{A \rightarrow \beta: A=X} P(A \rightarrow \beta) = 1 \text{ for all } X \in N$$
 - Think of this as the conditional probability for $A \rightarrow \beta$, given the left-hand-side nonterminal A .

PCFG Example

S → NP VP [1.0]	NP → she [0.05]
VP → V NP [0.6]	NP → glasses [0.05]
VP → VP PP [0.4]	D → the [1.0]
PP → P NP [1.0]	N → cat [0.3]
NP → D N [0.7]	N → glasses [0.7]
NP → NP PP [0.2]	V → saw [1.0]
	P → with [1.0]

Parse Tree Probability

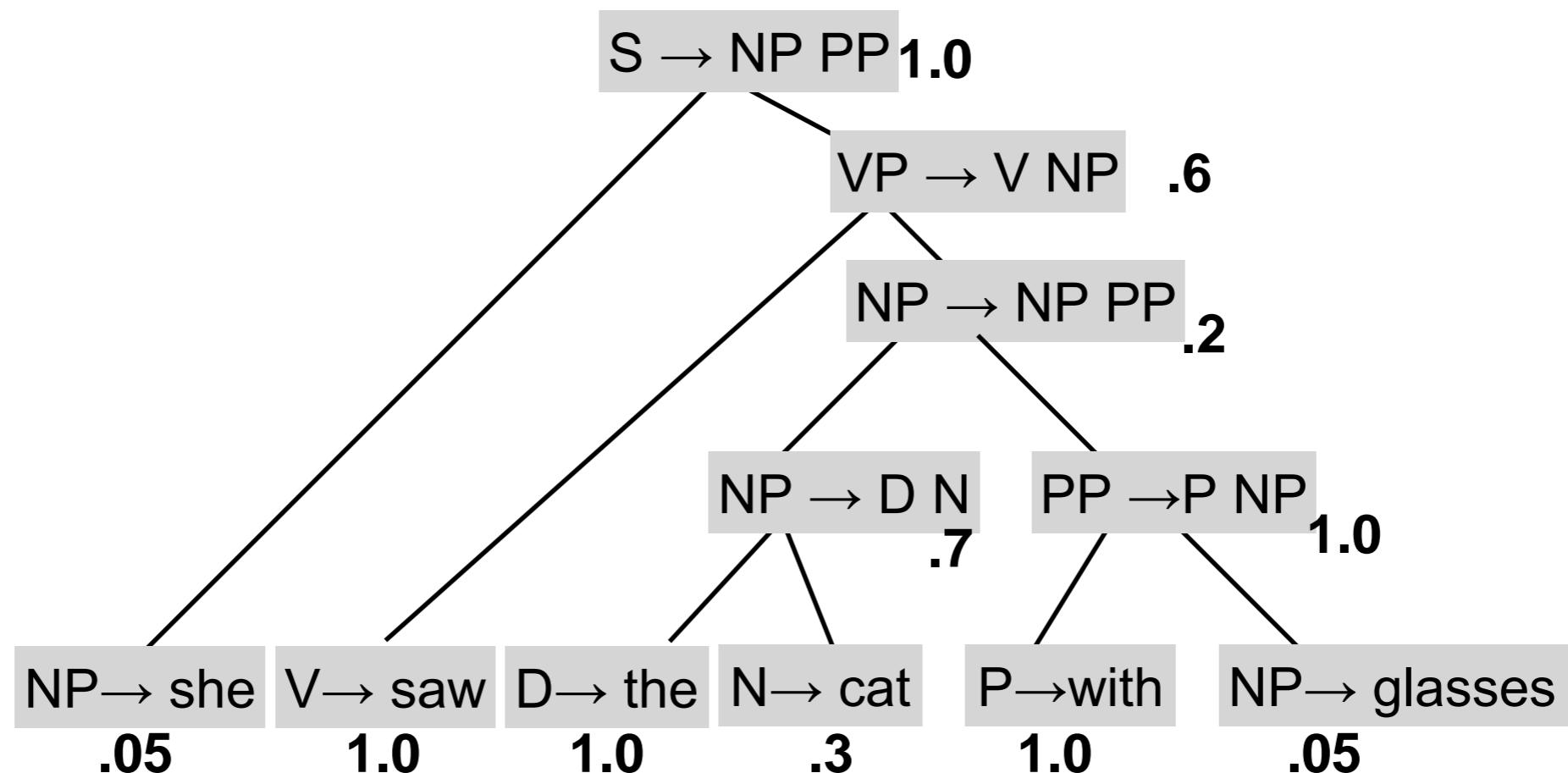
- Given a parse tree $t \in \mathcal{T}_G$, containing rules $A_1 \rightarrow \beta_1, \dots, A_n \rightarrow \beta_n$ the probability of t is $P(t) = \prod_{i=1}^n P(A_i \rightarrow \beta_i)$



$$1 \times .05 \times .4 \times .6 \times 1 \times 0.7 \times 1 \times 0.3 \times 1 \times 1 \times .05 = .000126$$

Parse Tree Probability

- Given a parse tree $t \in \mathcal{T}_G$, containing rules $A_1 \rightarrow \beta_1, \dots, A_n \rightarrow \beta_n$ the probability of t is $P(t) = \prod_{i=1}^n P(A_i \rightarrow \beta_i)$



$$1 \times 0.05 \times 0.6 \times 1 \times 0.2 \times 0.7 \times 1 \times 0.3 \times 1 \times 1 \times 0.05 = 0.000063 < 0.000126$$

Estimating PCFG probabilities

- Supervised training: We can estimate PCFG probabilities from a *treebank*, a corpus manually annotated with constituency structure using maximum likelihood estimates:

$$P(A \rightarrow \beta) = \frac{\text{count}(A \rightarrow \beta)}{\text{count}(A)}$$

- Unsupervised training:
 - What if we have a grammar and a corpus, but no annotated parses?
 - Can use the **inside-outside** algorithm for parsing and do EM estimation of the probabilities (not discussed in this course)

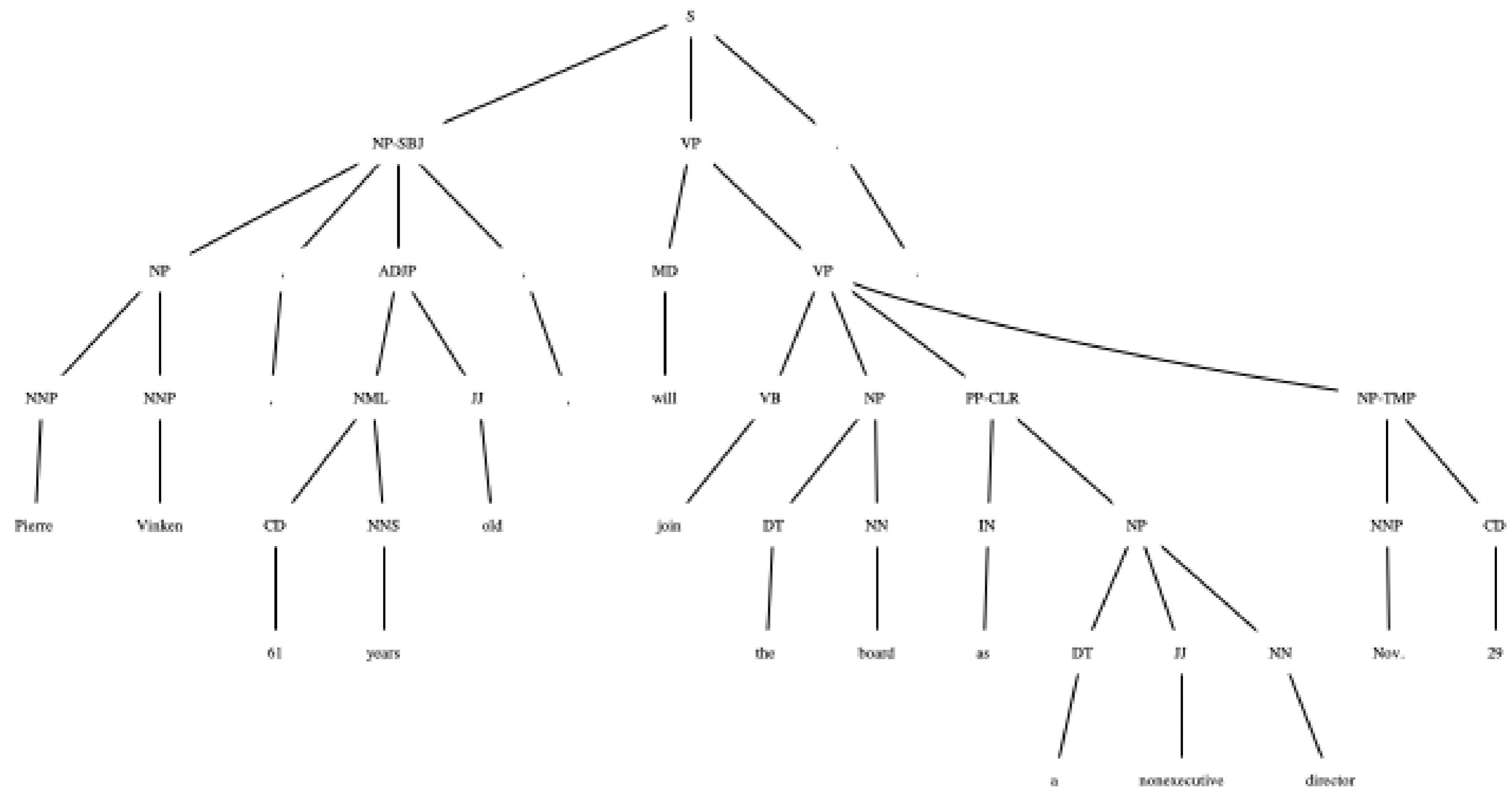
The Penn Treebank

- Syntactically annotated corpus of newspaper text (1989 Wall Street Journal Articles).
- The source text is naturally occurring but the treebank is not:
 - Assumes a specific linguistic theory (although a simple one).
 - Very flat structure (NPs, Ss, VPs).

PTB Example

```
( (S (NP-SBJ (NP (NNP Pierre) (NNP Vinken))  
      (, ,)  
      (ADJP (NML (CD 61) (NNS years))  
      (JJ old))  
      (, ,))  
      (VP (MD will))  
      (VP (VB join)  
          (NP (DT the) (NN board))  
          (PP-CLR (IN as)  
              (NP (DT a) (JJ nonexecutive) (NN director)))  
          (NP-TMP (NNP Nov.) (CD 29))))  
      (..))) )
```

PTB Example

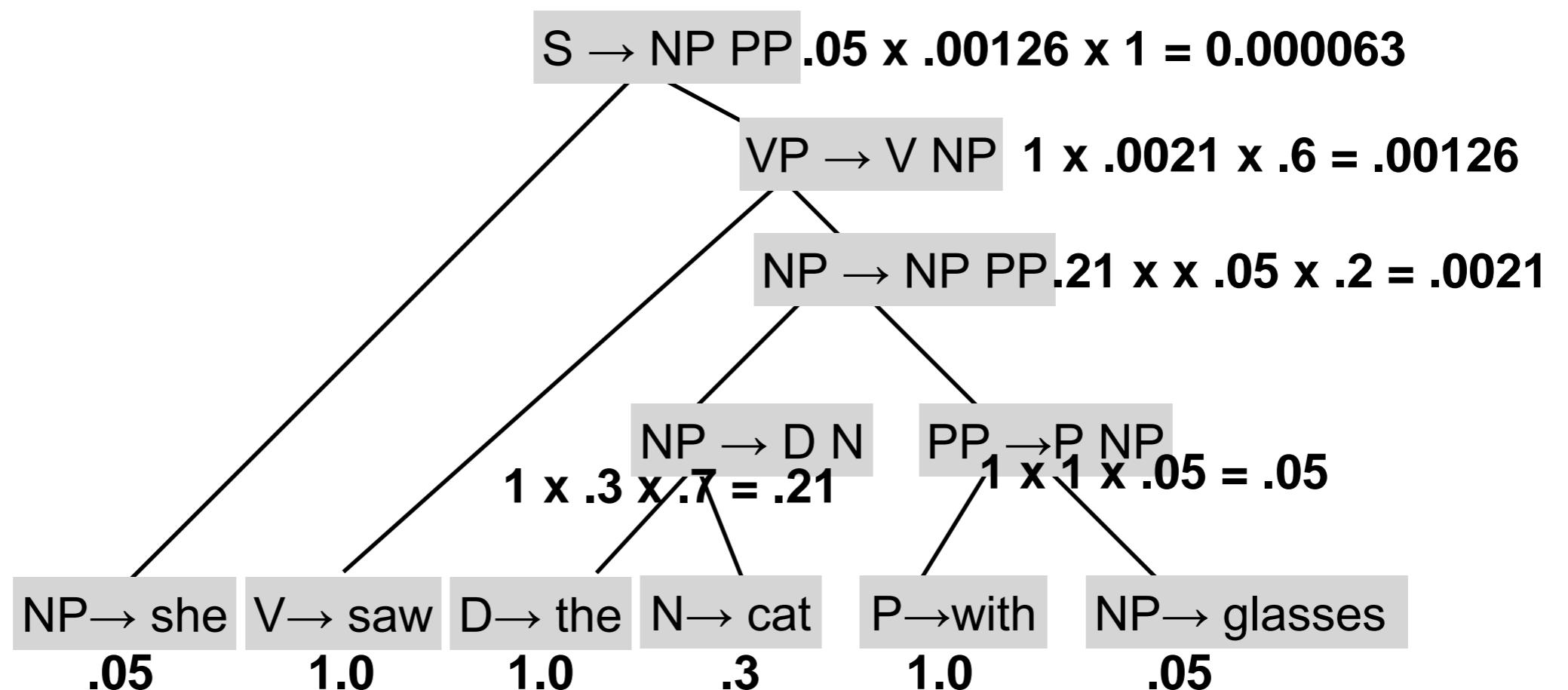


Parsing with PCFG

- We want to use PCFG to answer the following questions:
 - What is the total probability of the sentence under the PCFG?
 - What is the most probable parse tree for a sentence under the PCFG? (*decoding/parsing*)
- We can modify the CKY algorithm.
Basic idea: Compute these probabilities bottom-up using dynamic programming.

Computing Probabilities

Bottom-Up



CKY for PCFG Parsing

- Let $T_G(s, A)$ be the set of trees generated by grammar G starting at nonterminal A , whose *yield* is string s
- Use a chart π so that $\pi[i, j, A]$ contains the probability of the highest probability parse tree for string $s[i, j]$ starting in nonterminal A .

$$\pi[i, j, A] = \max_{t \in T_G(s[i, j], A)} P(t)$$

- We want to find $\pi[0, \text{length}(s), S]$ -- the probability of the highest-scoring parse tree for s rooted in the start symbol S .

CKY for PCFG Parsing

- To compute $\pi[0, \text{length}(s), S]$ we can use the following recursive definition:

Base case:
$$\pi[i, i + 1, A] = \begin{cases} P(A \rightarrow s_i) & \text{if } A \rightarrow s_i \in R \\ 0 & \text{otherwise} \end{cases}$$

$$\pi[i, j, A] = \max_{\substack{k=i+1 \dots j-1, \\ A \rightarrow BC \in R}} P(A \rightarrow BC) \cdot \pi[i, k, B] \cdot \pi[k, j, C]$$

- Then fill the chart using dynamic programming.

CKY for PCFG Parsing

- **Input:** PCFG $G=(N, \Sigma, R, S)$, input string s of length n .
- for $i=0\dots n-1$: initialization
$$\pi[i, i + 1, A] = \begin{cases} P(A \rightarrow s_i) & \text{if } A \rightarrow s_i \in R \\ 0 & \text{otherwise} \end{cases}$$
- for $length=2\dots n$: main loop
 - for $i=0\dots(n-length)$:
 - $j = i+length$
 - for $k=i+1\dots j-1$:
 - for $A \in N$:
$$\pi[i, j, A] = \max_{\substack{k=i+1\dots j-1, \\ A \rightarrow BC \in R}} P(A \rightarrow BC) \cdot \pi[i, k, B] \cdot \pi[k, j, C]$$

Use **backpointers** to retrieve the highest-scoring parse tree (see previous lecture).

Probability of a Sentence

- What if we are interested in the probability of a sentence, **not** of a single parse tree (for example, because we want to use the PCFG as a language model).
- Problem: Spurious ambiguity. Need to sum the probabilities of **all** parse trees for the sentence.
- How do we have to change CKY to compute this?

$$\pi[i, j, A] = \sum_{\substack{k=i+1 \dots j-1, \\ A \rightarrow BC \in R}} P(A \rightarrow BC) \cdot \pi[i, k, B] \cdot \pi[k, j, C]$$

Earley Parser

- CKY parser starts with words and builds parse trees bottom-up; requires the grammar to be in CNF.
- The Earley parser instead starts at the start symbol and tries to “guess” derivations top-down.
 - It discards derivations that are incompatible with the sentence.
 - The early parser sweeps through the sentence left-to-right only once. It keeps partial derivations in a table (“chart”).
 - Allows arbitrary CFGs, no limitation to CNF.

Parser States

- Earley parser keeps track of partial derivations using **parser states / items**.
- State represent hypotheses about constituent structure based on the grammar, taking into account the input.
- Parser states are represented as **dotted rules with spans**.
 - The constituents to the left of the · have already been seen in the input string s (corresponding to the span)

$S \rightarrow \cdot N P \ V P \ [0,0]$ “According to the grammar, there may be an NP starting in position 0.“

$N P \rightarrow D \ A \cdot N \ [0,2]$ “There is a determiner followed by an adjective in s[0,2]“

$N P \rightarrow N P \ P P \cdot [3,8]$ “There is a complete NP in s[3,8], consisting of an NP and PP“

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$NP \rightarrow \cdot D N [0,0]$

$D \rightarrow \cdot the [0,0]$

Three parser operations:

1. Predict new subtrees top-down.

0 *the* 1 *student* 2 *saw* 3 *the* 4 *cat* 5 *with* 6 *the* 7 *tail*

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$NP \rightarrow \cdot D N [0,0]$

$D \rightarrow the \cdot [0,1]$

Three parser operations:

1. Predict new subtrees top-down.
2. Scan input terminals.

0 *the* 1 *student* 2 *saw* 3 *the* 4 *cat* 5 *with* 6 *the* 7 *tail*

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$NP \rightarrow \cdot D N [0,0]$

$D \rightarrow the \cdot [0,1]$

passive state

Three parser operations:

1. Predict new subtrees top-down.
2. Scan input terminals.

0 *the* 1 *student* 2 *saw* 3 *the* 4 *cat* 5 *with* 6 *the* 7 *tail*

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$S \rightarrow \cdot NP VP [0,0]$

$NP \rightarrow D \cdot N [0,1]$

$D \rightarrow the \cdot [0,1]$

passive state

Three parser operations:

1. Predict new subtrees top-down.
2. Scan input terminals.
3. Complete with passive states.

0 the 1 student 2 saw 3 the 4 cat 5 with 6 the 7 tail

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$S \rightarrow \cdot NP VP [0,0]$

$NP \rightarrow D \cdot N [0,1]$

$D \rightarrow the \cdot [0,1]$

$N \rightarrow \cdot cat [1,1]$

$N \rightarrow \cdot tail [1,1]$

$N \rightarrow \cdot student [1,1]$

Three parser operations:

1. Predict new subtrees top-down.

2. Scan input terminals.

3. Complete with passive states.

0 the 1 student 2 saw 3 the 4 cat 5 with 6 the 7 tail

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$NP \rightarrow D \cdot N [0,1]$

$D \rightarrow the \cdot [0,1]$

$N \rightarrow \cdot cat [1,1]$

$N \rightarrow \cdot tail [1,1]$

$N \rightarrow student \cdot [1,2]$

Three parser operations:

1. Predict new subtrees top-down.

2. Scan input terminals.

3. Complete with passive states.

0 the 1 student 2 saw 3 the 4 cat 5 with 6 the 7 tail

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$NP \rightarrow \cdot NP PP [0,0]$

$NP \rightarrow D N \cdot [0,2]$

$D \rightarrow the \cdot [0,1]$

$N \rightarrow \cdot cat [1,1]$

$N \rightarrow \cdot tail [1,1]$

$N \rightarrow student \cdot [1,2]$

Three parser operations:

1. Predict new subtrees top-down.

2. Scan input terminals.

3. Complete with passive states.

0 the 1 student 2 saw 3 the 4 cat 5 with 6 the 7 tail

Earley Parser (sketch)

$S \rightarrow NP VP$	$V \rightarrow saw$
$VP \rightarrow V NP$	$P \rightarrow with$
$VP \rightarrow VP PP$	$D \rightarrow the$
$PP \rightarrow P NP$	$N \rightarrow cat$
$NP \rightarrow D N$	$N \rightarrow tail$
$NP \rightarrow NP PP$	$N \rightarrow student$

$S \rightarrow NP \cdot VP [0,2]$

$NP \rightarrow NP \cdot PP [0,2]$

$NP \rightarrow D N \cdot [0,2]$

$D \rightarrow the \cdot [0,1]$

$N \rightarrow \cdot cat [1,1]$

$N \rightarrow \cdot tail [1,1]$

$N \rightarrow student \cdot [1,2]$

Three parser operations:

1. Predict new subtrees top-down.

2. Scan input terminals.

3. Complete with passive states.

0 the 1 student 2 saw 3 the 4 cat 5 with 6 the 7 tail

Earley Algorithm

- Keep track of parser states in a table (“chart”). $Chart[k]$ contains a set of all parser states that end in position k .

- **Input:** Grammar $G=(N, \Sigma, R, S)$, input string s of length n .
- **Initialization:** For each production $S \rightarrow \alpha \in R$
add a state $S \rightarrow \cdot \alpha [0,0]$ to $Chart[0]$.
- for $i = 0$ to n :
 - for each $state$ in $Chart[i]$:
 - if $state$ is of form $A \rightarrow \alpha \cdot s[i] \beta [k,i]$:
 $\text{scan}(state)$
 - elif $state$ is of form $A \rightarrow \alpha \cdot B \beta [k,i]$:
 $\text{predict}(state)$
 - else: // $state$ is of form $A \rightarrow \alpha \cdot [k,i]$
 $\text{complete}(state)$

Earley Algorithm - Scan

- The scan operation can only be applied to a state if the dot is in front of a terminal symbol that matches the next input terminal.

- function **scan**(state): // *state* is of form $A \rightarrow \alpha \cdot s[i] \beta [k, i]$

- Add a new state $A \rightarrow \alpha s[i] \cdot \beta [k, i+1]$
to Chart[i+1]

Earley Algorithm - Predict

- The predict operation can only be applied to a state if the dot is in front of a non-terminal symbol.
- function **predict(state)**: // *state* is of form $A \rightarrow \alpha \cdot B \beta [k,i]$:
 - Add a new state $B \rightarrow \cdot \gamma [i,i]$ to *Chart[i]*
- Note that this modifies *Chart[i]* **while** the algorithm is looping through it.
- No duplicate states are added (*Chart[i]* is a set)

Earley Algorithm - Complete

- The complete operation may only be applied to a passive item.
- function **complete**(state): *// state is of form $A \rightarrow \alpha \cdot [k,j]$*
 - for each state $B \rightarrow \beta \cdot A \gamma [i,k]$ add a new state $B \rightarrow \beta A \cdot \gamma [i,j]$ to Chart[j]
- Note that this modifies Chart[i] **while** the algorithm is looping through it.
- Note that it is important to make a copy of the old state before moving the dot.
- This operation is similar to the combination operation in CKY!

Earley Algorithm - Runtime

- The runtime depends on the number of items in the chart (each item is “visited” exactly once).
- We proceed through the input exactly once, which takes $O(N)$.
 - For each position on the chart, there are $O(N)$ possible split points where the dot could be.
 - Each complete operation can produce $O(N)$ possible new items (with different starting points).
- Total: $O(N^3)$

Earley Algorithm - Some Observations

- How do we recover parse trees?
 - What happens in case of ambiguity?
 - Multiple ways to Complete the same state.
 - Keep back-pointers in the parser state objects.
 - Or use a separate data structure (CKY-style table or hashed states)
 - How do we make the algorithm work with PCFG?
 - Easy to compute probabilities on Complete. Follow back pointer with max probability.

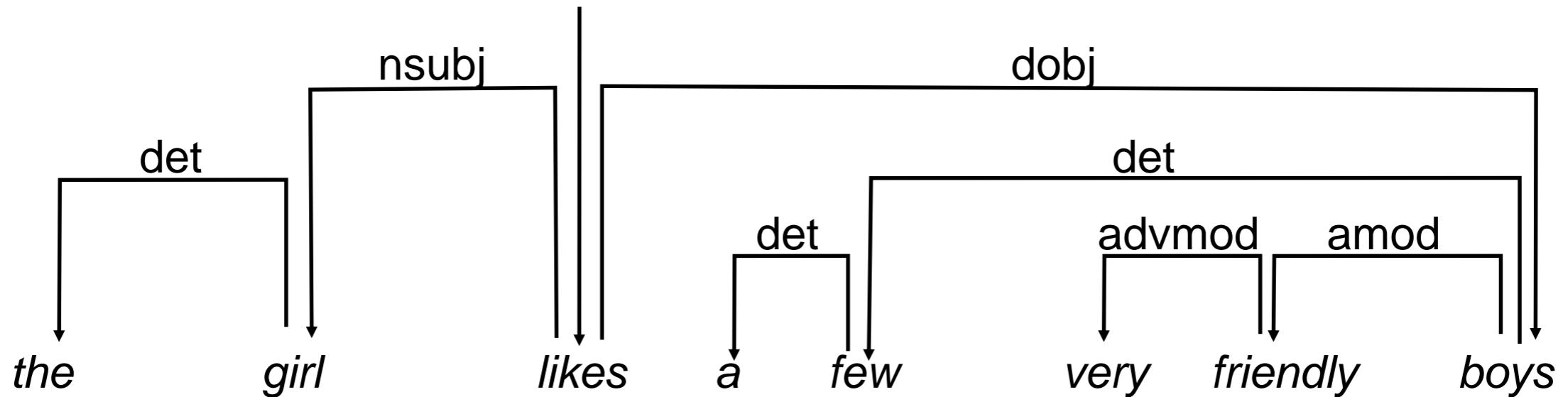
Natural Language Processing

Lecture 8: Dependency Parsing

2/28/2020

COMS W4705
Yassine Benajiba

Dependency Structure



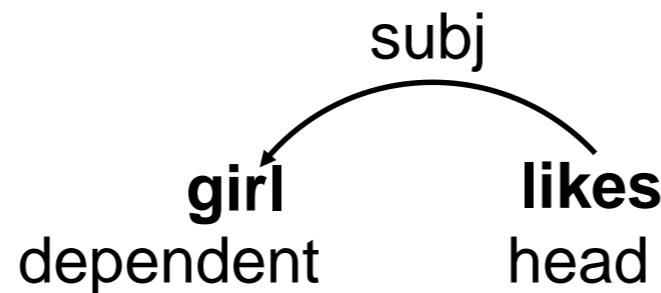
- The edges can be labeled with **grammatical relations** between words (typed dependencies):
 - Arguments (Subject, Object, Indirect Object, Prepositional Object)
 - Adjunct (Temporal, Locative, Causal, Manner...) / Modifier
 - Function words

Dependency Structure

- Long history in linguistics (Starting with Panini's Grammar of Sanskrit, 4th century BCE).
 - Modern dependency grammar originates with Tesniere and Mel'čuk.
- Different from phrase structure (but related via the concept of constituency and heads)
 - Focus is on grammatical relationships between words (Subject, Object, ...)
- Tighter connection to natural language semantics.

Dependency Relations

- Each dependency relation consists of a head and a dependent.



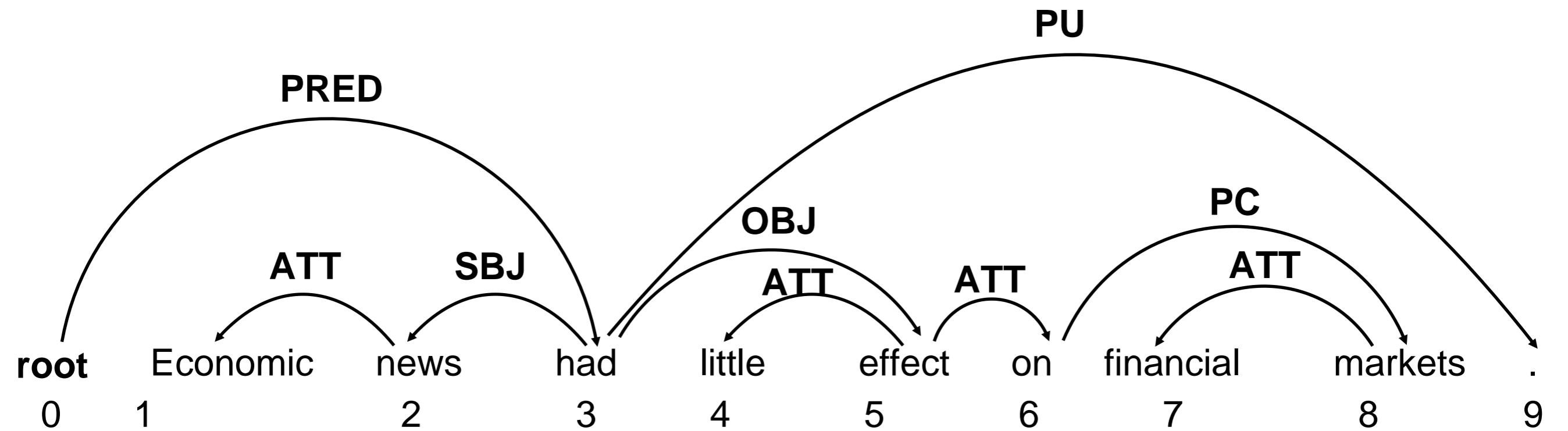
- Represent individual edges as `subj(likes-02, girl-01)`
- or as a triple `(likes, nsubj, girl)`
- And the entire sentence structure as a set of edges:

`root(likes-2), subj(likes-2, girl-1), det(the-0, girl-1), obj(likes-2, boys-7),
det(boys-7, few-4), det(few-4, a-3), amod(boys-7, friendly-6), advmod(friendly-6, very-5)`

Heads and Dependents

- How do we identify the grammatical relation between head H and Dependent D (in a particular constituent C)?
 - H determines the syntactic category of C and can often replace C.
 - H determines the semantic category of C; D gives semantic specification.
 - H is obligatory; D may be optional.
 - H selects D and determines whether D is obligatory or optional.
 - The form of D depends on H (agreement or government).
 - The linear position of D is specified with reference to H.

Another Example



Dependency structure $G = (V_s, A)$

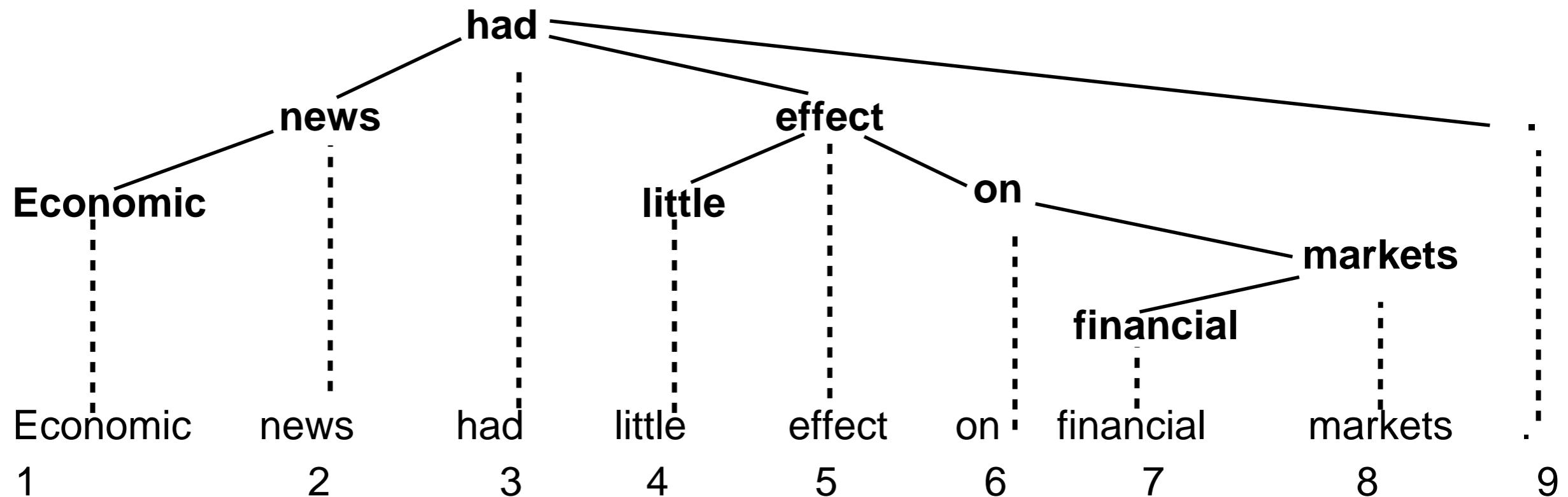
set of nodes

$V_s = \{\text{root}, \text{Economic}, \text{news}, \text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .\}$

set of edges/
arcs

$A = \{(\text{root}, \text{PRED}, \text{had}), (\text{had}, \text{SBJ}, \text{news}), (\text{had}, \text{OBJ}, \text{effect}), (\text{had}, \text{PU}, .), (\text{news}, \text{ATT}, \text{Economic}), (\text{effect}, \text{ATT}, \text{little}), (\text{effect}, \text{ATT}, \text{on}), (\text{on}, \text{PC}, \text{markets}), (\text{markets}, \text{ATT}, \text{financial})\}$

Another Example



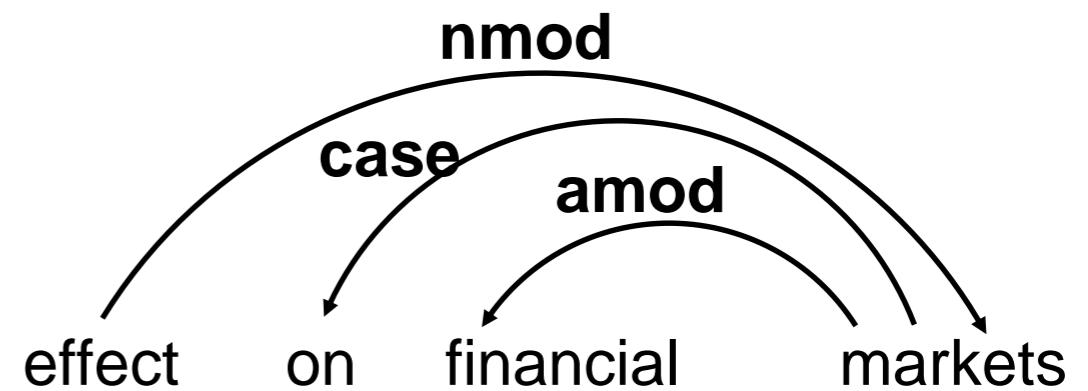
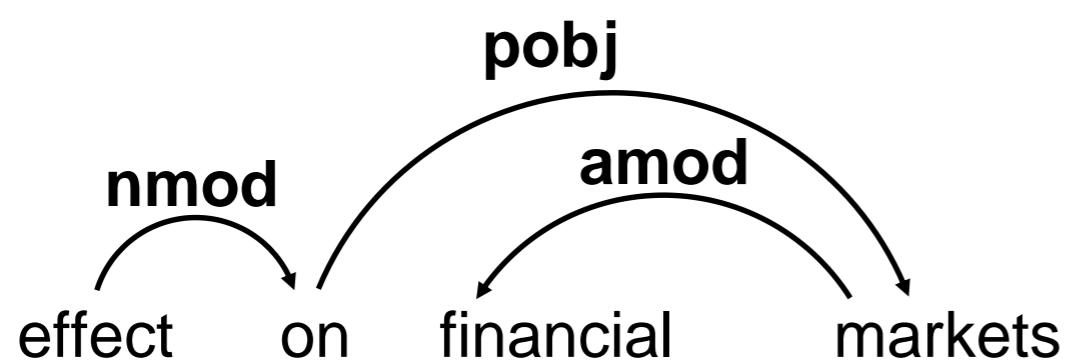
$$G = (V, A)$$

$$V = \{\text{root}, \text{Economic}, \text{news}, \text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, \cdot\}$$

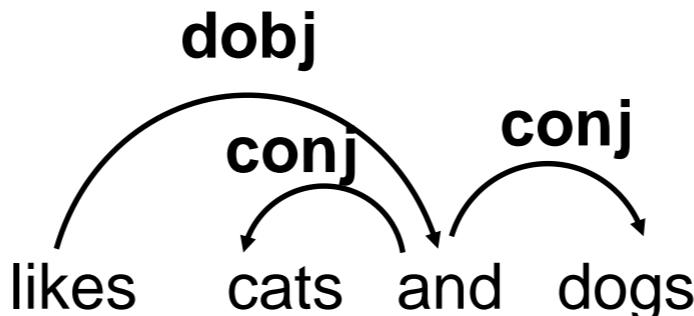
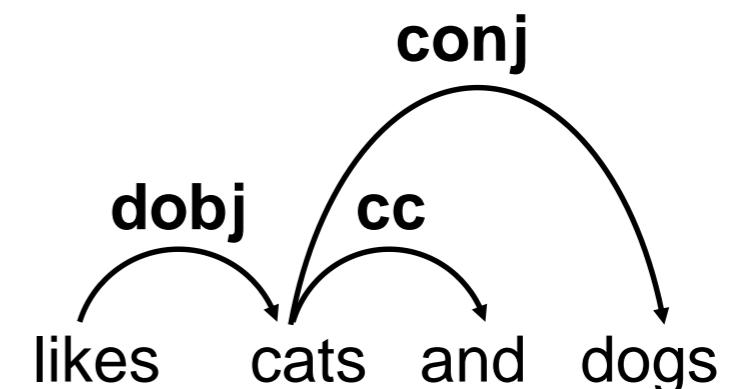
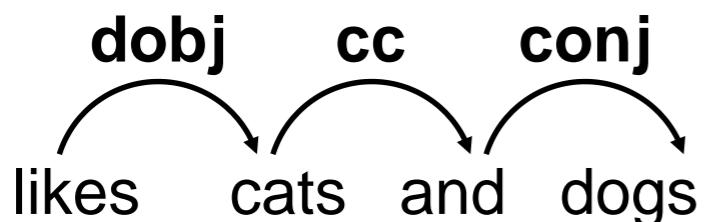
$$A = \{(\text{root}, \text{PRED}, \text{had}), (\text{had}, \text{SBJ}, \text{news}), (\text{had}, \text{OBJ}, \text{effect}), (\text{had}, \text{PU}, \cdot), (\text{news}, \text{ATT}, \text{Economic}), (\text{effect}, \text{ATT}, \text{little}), (\text{effect}, \text{ATT}, \text{on}), (\text{on}, \text{PC}, \text{markets}), (\text{markets}, \text{ATT}, \text{financial})\}$$

Different Dependency Representations

- How to deal with prepositions?



- How to deal with conjunctions?



Inventory of Relations

"Universal Dependencies" (Marneffe *et al.* 2014)

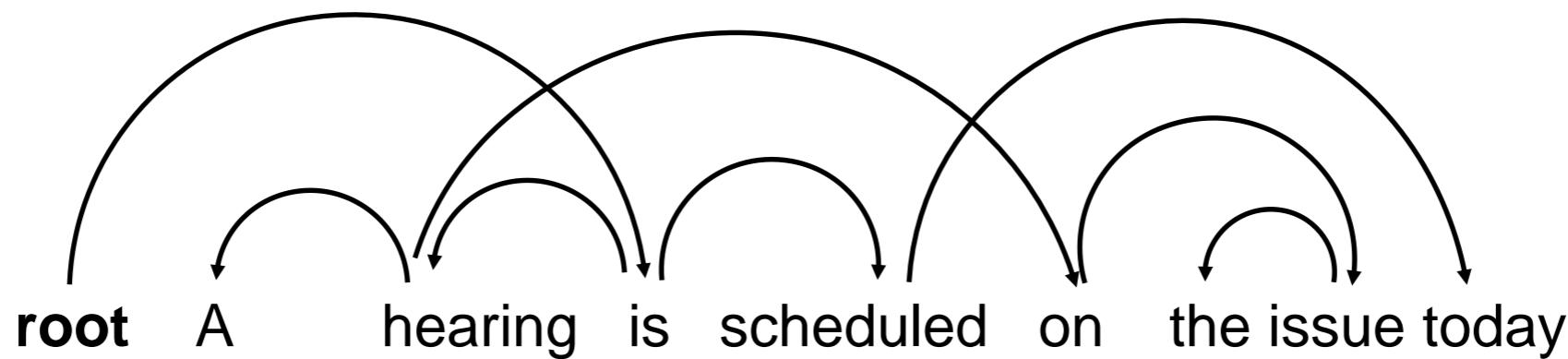
	Nominals	Clauses	Modifier words	Function Words
Core arguments	<u>nsubj</u> <u>obj</u> <u>iobj</u>	<u>csubj</u> <u>ccomp</u> <u>xcomp</u>		
Non-core dependents	<u>obl</u> <u>vocative</u> <u>expl</u> <u>dislocated</u>	<u>advcl</u>	<u>advmod</u> * <u>discourse</u>	<u>aux</u> <u>cop</u> <u>mark</u>
Nominal dependents	<u>nmod</u> <u>appos</u> <u>nummod</u>	<u>acl</u>	<u>amod</u>	<u>det</u> <u>clf</u> <u>case</u>
Coordination	MWE	Loose	Special	Other
<u>conj</u> <u>cc</u>	<u>fixed</u> <u>flat</u> <u>compound</u>	<u>list</u> <u>parataxis</u>	<u>orphan</u> <u>goeswith</u> <u>reparandum</u>	<u>punct</u> <u>root</u> <u>dep</u>

Dependency Trees

- Dependency structure is typically assumed to be a tree.
 - Root node 0 must not have a parent.
 - All other nodes must have exactly one parent.
 - The graph needs to be connected.
 - Nodes must not form a cycle.

Projectivity

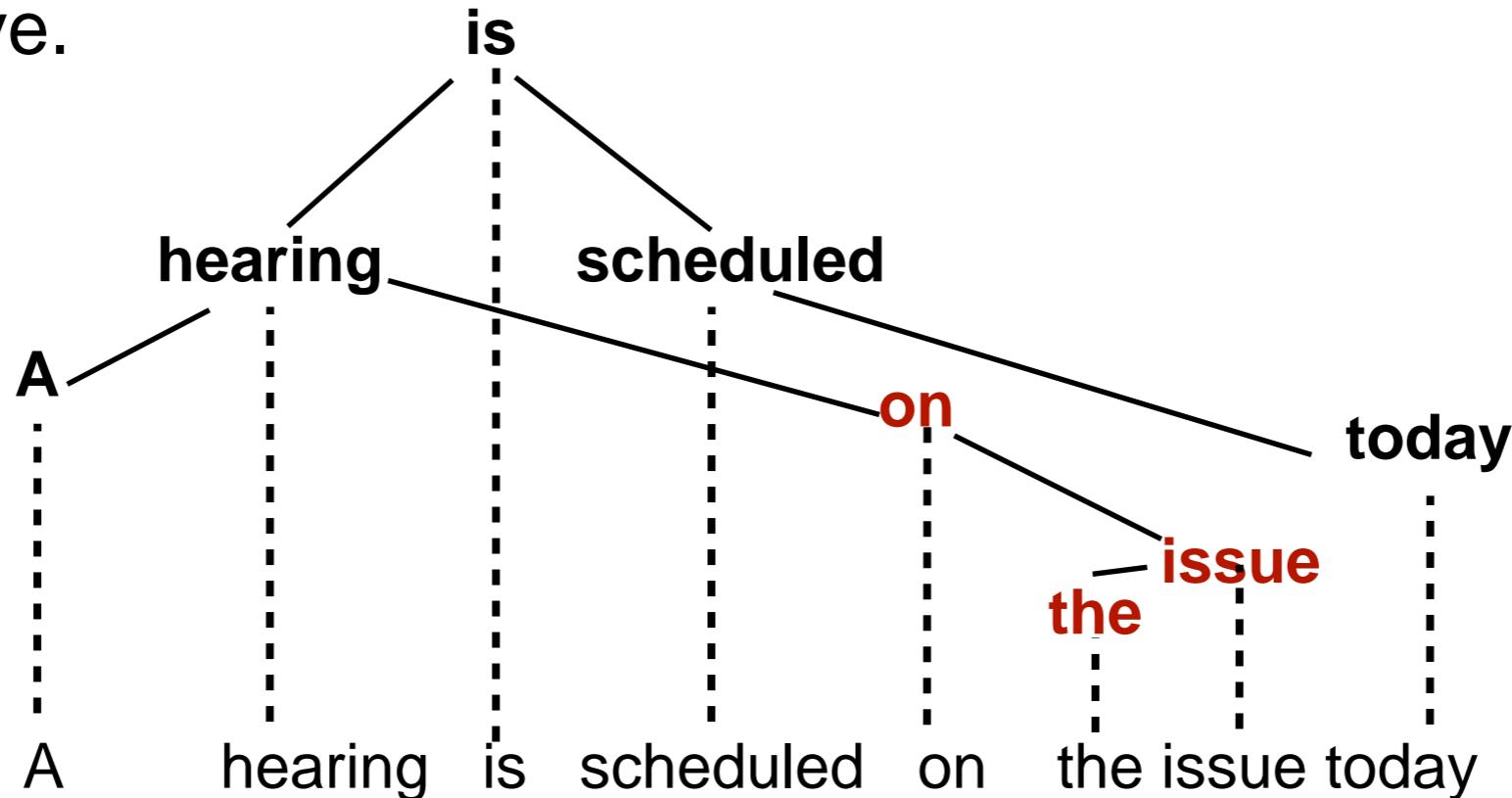
- Words in a sentence appear in a linear order.
- If dependency edges cross, the dependency structure is non-projective.



- Non-projective structures appear more frequently in some languages than others (Hungarian, German, ...)
- Some approaches to dependency parsing cannot handle non-projectivity.

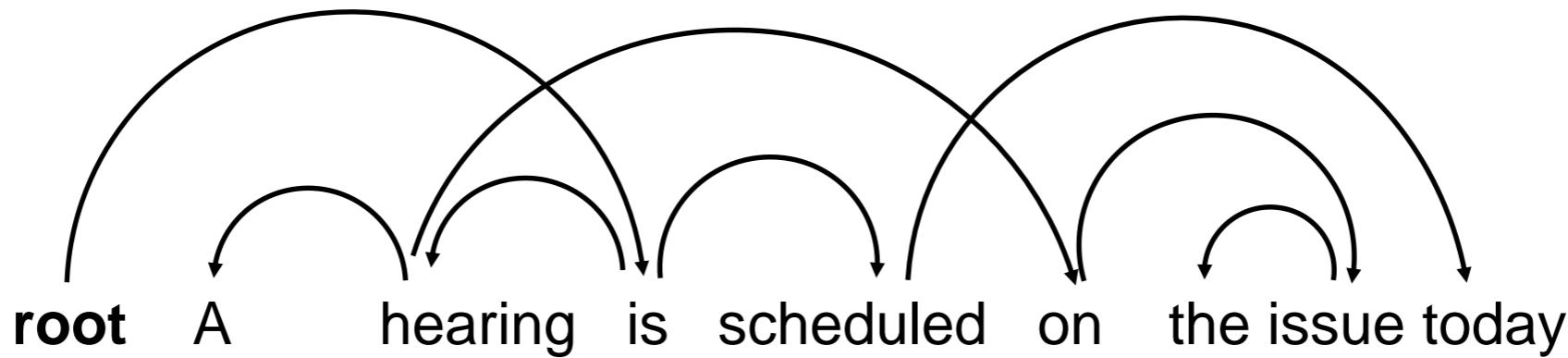
Projectivity

- Words in a sentence stand in a linear order.
- If dependency edges cross, the dependency structure is non-projective.



- Non-projective structures appear more frequently in some languages than others (Hungarian, German, ...)
- Some approaches to dependency parsing cannot handle non-projectivity.

Projectivity



An edge (i, r, j) in a dependency tree is projective if there is a directed path from i to k for all $i < k < j$ (if $i < j$)
or all $j < k < i$ ($j < i$).

Dependency Parsing

- Input:
 - a set of nodes $V_s = \{w_0, w_1, \dots, w_m\}$ corresponding to the input sentence $s = w_1, \dots, w_m$ (0 is the special **root** node)
 - an inventory of labels $R = \{\text{PRED}, \text{SBJ}, \text{OBJ}, \text{ATT}, \dots\}$
- Goal: Find a set of labeled, directed edges between the nodes, such that the resulting graph forms a **correct dependency tree** over V_s .

↑
structural constraints

Dependency Parsing

- What information could we use?
 - bi-lexical affinities
 - *financial markets, meeting... scheduled*
 - dependency distance (prefer close words?)
 - Intervening words
 - *had little effect, little gave effect*
 - subcategorization/valency of heads.

Subcategorization/Valency

- Verbs may take a different number of arguments of different syntactic types in different positions:
 - *The baby slept.* **The baby slept the house.*
 - *He pretended to sleep.* **He pretended the cat.*
 - *Godzilla destroyed the city.* **Godzilla destroyed.*
 - *Jenny gave the book to Carl.* **Jenny gave the book.*
 - ... examples for *ask, promise, bet, load,...*

Dependency Parsing

- As with other NLP problems, we can think of dependency parsing as a kind of search problem:
 - Step 1: Define the space of possible analyses for a sentence
 - Step 2: Select the best analysis from this search space.
- Need to define the search space, search algorithm, and a way to determine the "best" parse.

Dependency Parsing

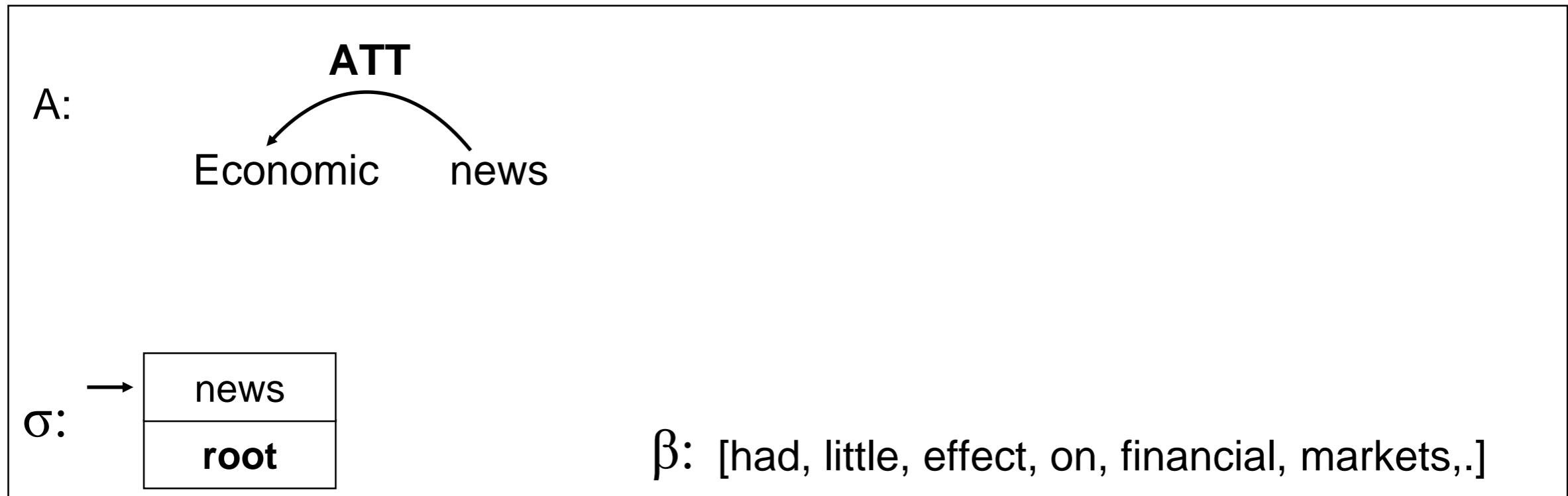
- Approaches to Dependency Parsing:
 - Grammar-based
 - Data-based
 - Dynamic Programming (e.g. Eisner 1996,)
 - Graph Algorithms (e.g. McDonald 2005, MST Parser)
 - **Transition-based (e.g. Nivre 2003, MaltParser)**
 - Constraint satisfaction (Karlsson 1990)

Transition-Based Dependency Parsing

- Defines the search space using parser states (configurations) and operations on these states (transitions).
- Start with an initial configuration and find a sequence of transitions to the terminal state.
- Uses a greedy approach to find the best sequence of transitions.
 - Uses a discriminative model (classifier) to select the next transition.

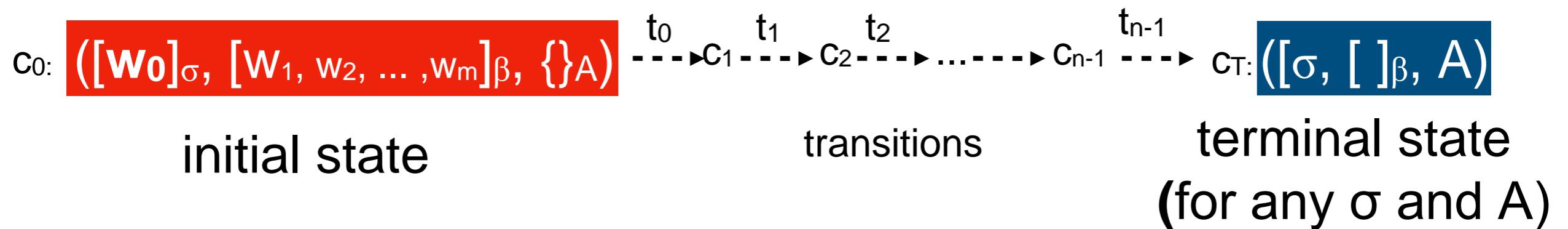
Transition-Based Parsing - States

- A parser state (configuration) is a triple $c = (\sigma, \beta, A)$
 - σ - is a **stack** of words $w_i \in V_s$
 - β - is a **buffer** of words $w_i \in V_s$
 - A - is a set of dependency arcs (w_i, r, w_j)



$([\text{root}, \text{news}]_\sigma, [\text{had, little, effect, on, financial, markets,..}]_\beta, \{ (\text{news}, \text{ATT}, \text{Economic}) \}_A$

Transition-Based Parsing - initial and terminal state



- Start with initial state c_0 .
- Apply sequence of transitions, t_0, \dots, t_{n-1} .
- Once a terminal state C_T is reached, return final parse A from state C_T .

Transition-Based Parsing - Transitions ("Arc-Standard")

- **Shift:**

Move next word from the buffer to the stack

$$(\sigma, w_i | \beta, A) \Rightarrow (\sigma | w_i, \beta, A)$$

- **Left-Arc (for relation r):**

Build an edge from the next word on the buffer to the top word on the stack.

$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma, w_j | \beta, A \cup \{w_j, r, w_i\})$$

- **Right-Arc (for relation r)**

Build an edge from the top word on the stack to the next word on the buffer.

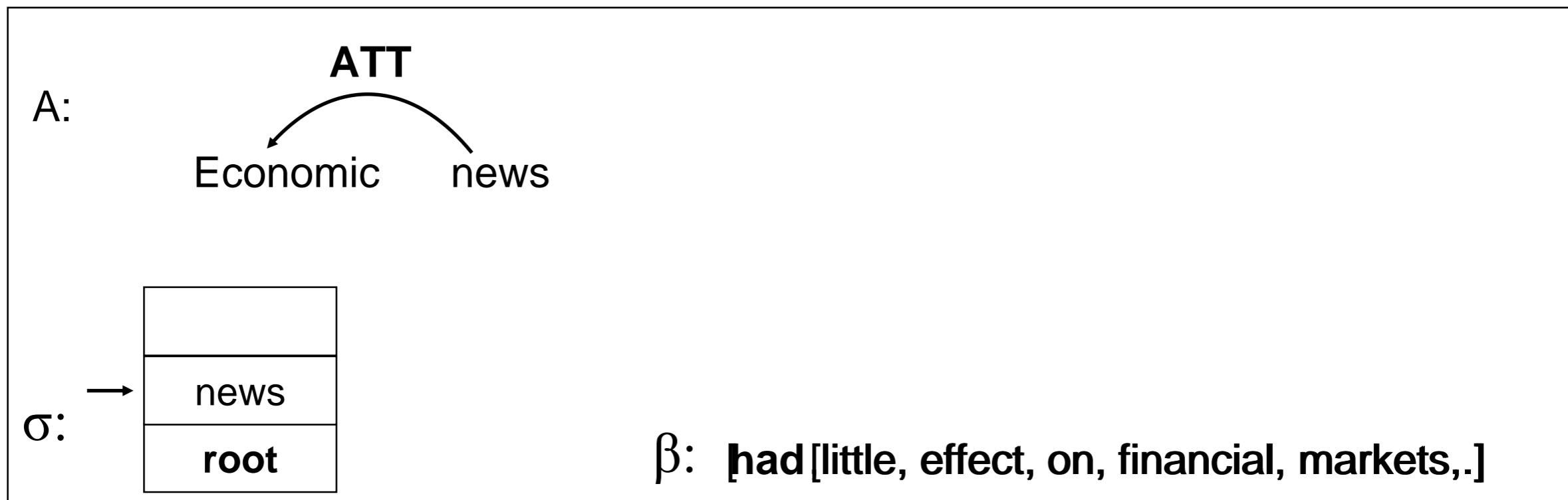
$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma, w_i | \beta, A \cup \{w_i, r, w_j\})$$

Transition-Based Parsing - Transitions

- **Shift**

Move next word from the buffer to the stack

$$(\sigma, w_i | \beta, A) \Rightarrow (\sigma | w_i, \beta, A)$$



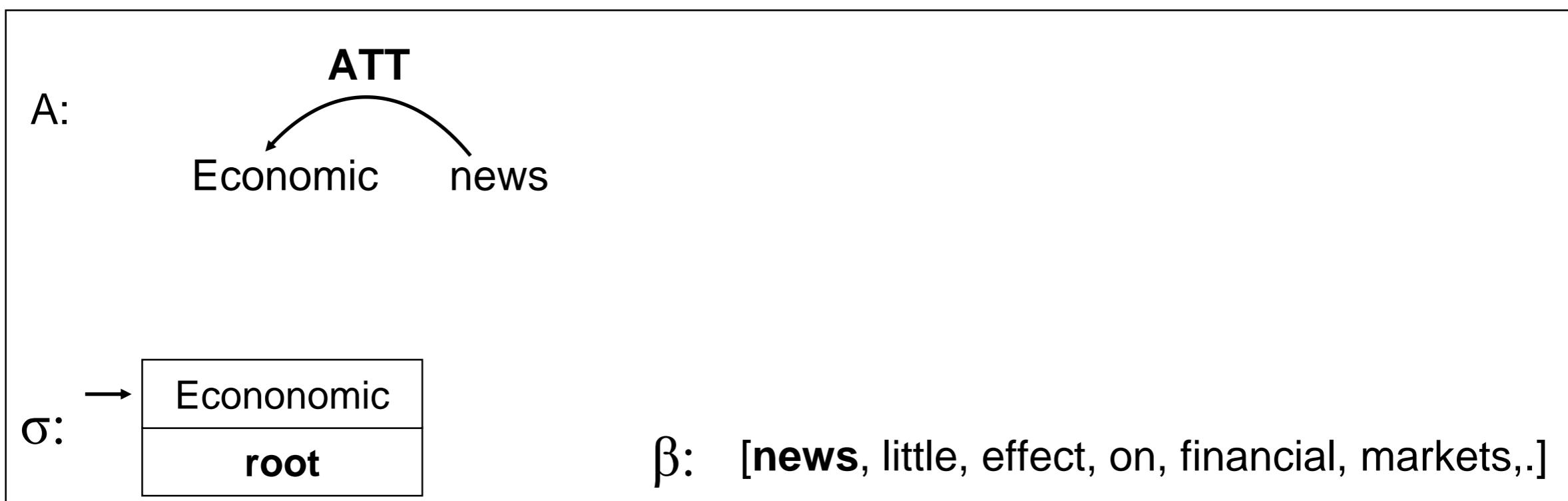
Transition-Based Parsing - Transitions

- **Arc-left_r**

Build an edge from the next word on the buffer to the top word on the stack.

$$(\sigma \mid w_i, w_j \mid \beta, A) \Rightarrow (\sigma, w_j \mid \beta, A \cup \{w_j, r, w_i\})$$

Not allowed if $i=0$ (root may not have a parent)

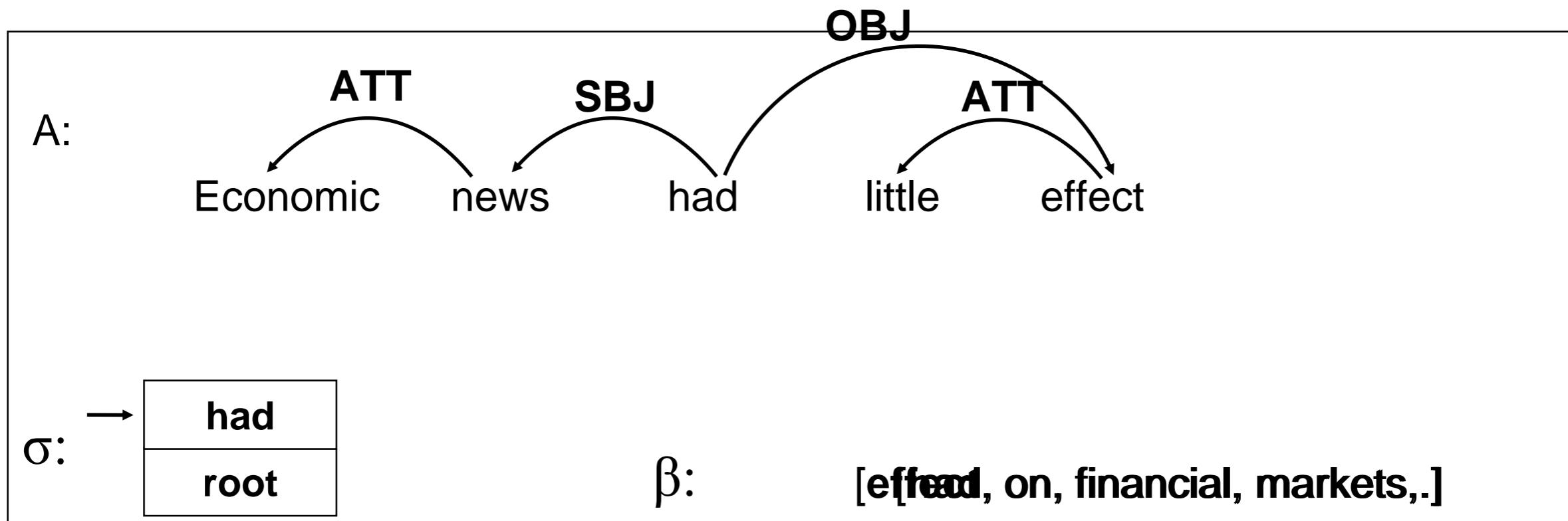


Transition-Based Parsing - Transitions

- **Arc-right**

Build an edge from the next word on the buffer to the top word on the stack.

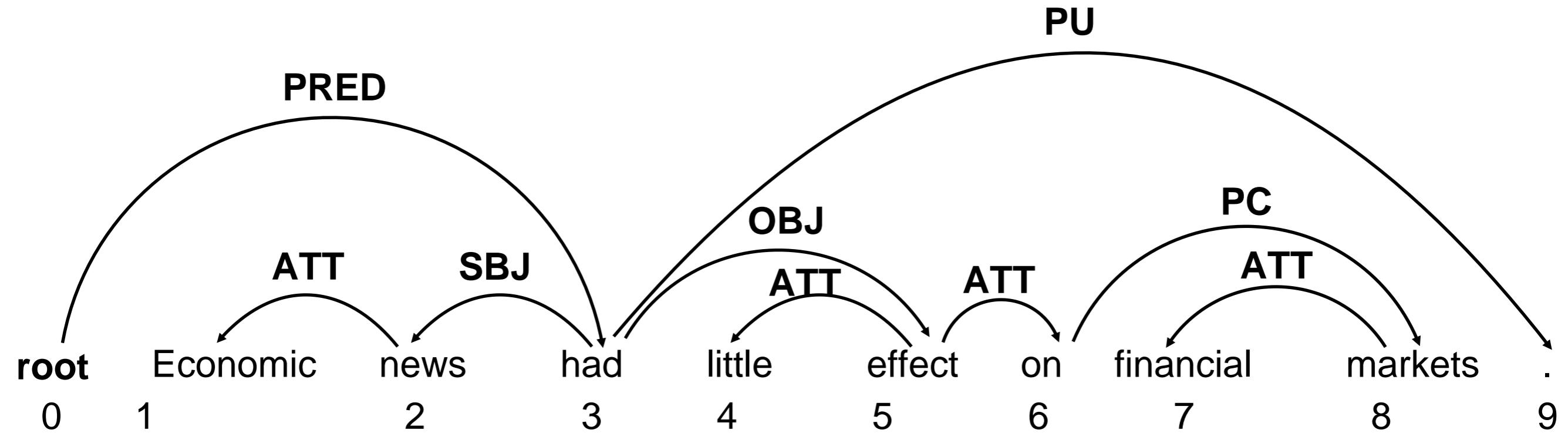
$$(\sigma \mid w_i, w_j \mid \beta, A) \Rightarrow (\sigma, w_i \mid \beta, A \cup \{w_i, r, w_j\})$$



Transition-Based Parsing - Some Observations

- Does the transition system contain dead ends? (states from which a terminal state cannot be reached)? No!
- What is the role of the buffer?
 - Contains words that can become dependents of a right-arc. Keep unseen words.
- What is the role of the stack?
 - Keep track of nodes that can become dependents of a left-arc.
- Once a word disappears from the buffer and the stack it cannot be part of any further edge!

Another Example



$$G = (V_s, A)$$

$V_s = \{\text{root}, \text{Economic}, \text{news}, \text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .\}$

$A = \{(\text{root}, \text{PRED}, \text{had}), (\text{had}, \text{SBJ}, \text{news}), (\text{had}, \text{OBJ}, \text{effect}), (\text{had}, \text{PU}, .),$
 $(\text{news}, \text{ATT}, \text{Economic}), (\text{effect}, \text{ATT}, \text{little}), (\text{effect}, \text{ATT}, \text{on}), (\text{on}, \text{PC}, \text{markets}),$
 $(\text{markets}, \text{ATT}, \text{financial})\}$

Transition-Based Parsing - Complete Example

initial state

next transition: shift (these are all predicted by discriminative ML classifier)

A:

σ :

root

β : [Economic, news, had, little, effect, on, financial, markets,.]

Transition-Based Parsing - Oracle Example

next-transition: Left-Arc_{ATT}

A:

σ :

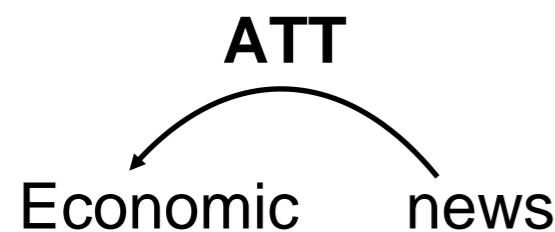
Economic
root

 β : [news, had, little, effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



σ :

root

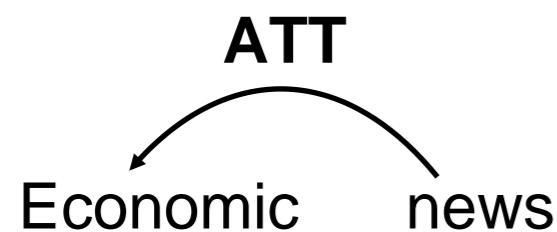
β :

[news, had, little, effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: Left-ArcsBJ

A:



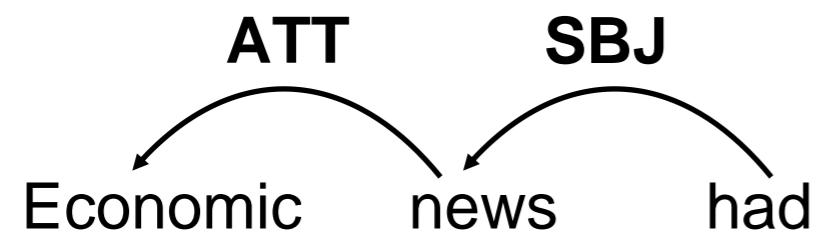
β :

[had, little, effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



σ :

root

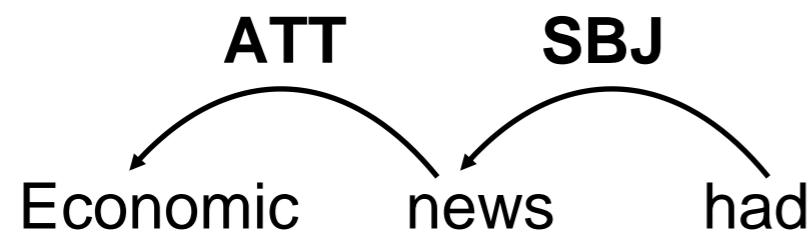
β :

[had, little, effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



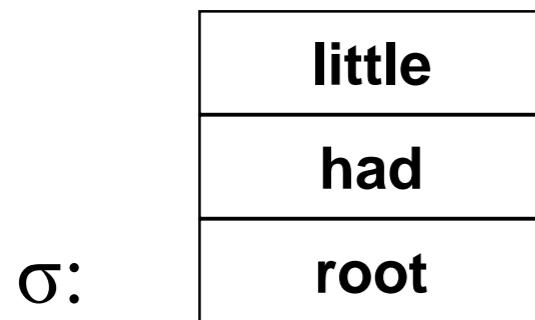
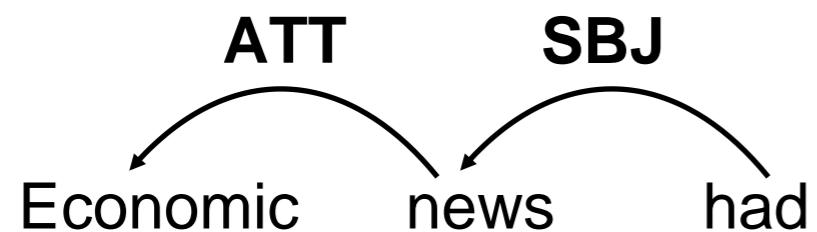
β :

[little, effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: Left-Arc_{SBJ}

A:



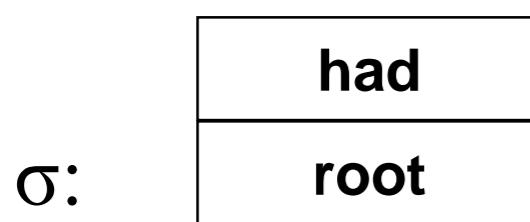
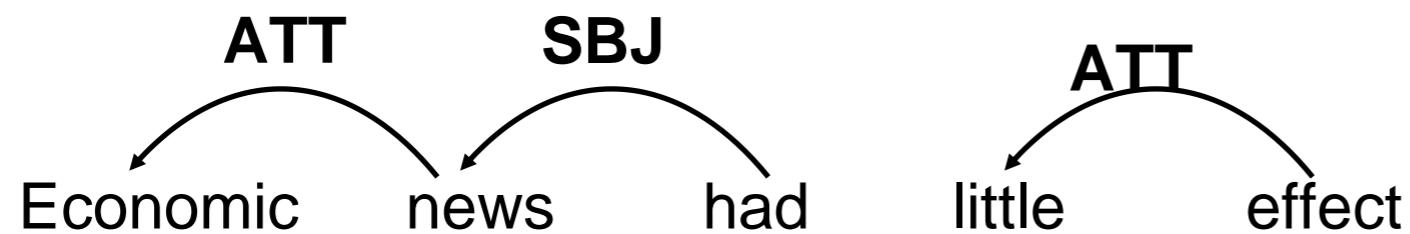
β :

[effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



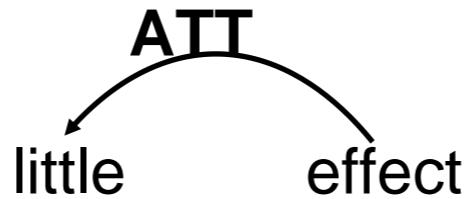
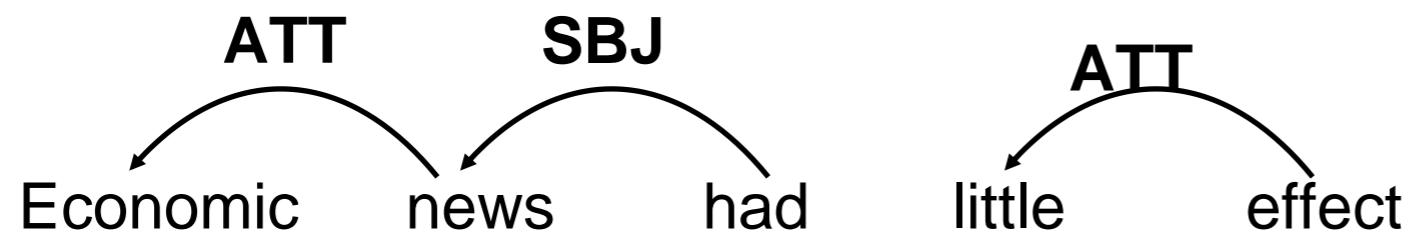
β :

[effect, on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



σ:	<table border="1"><tr><td>effect</td></tr><tr><td>had</td></tr><tr><td>root</td></tr></table>	effect	had	root
effect				
had				
root				

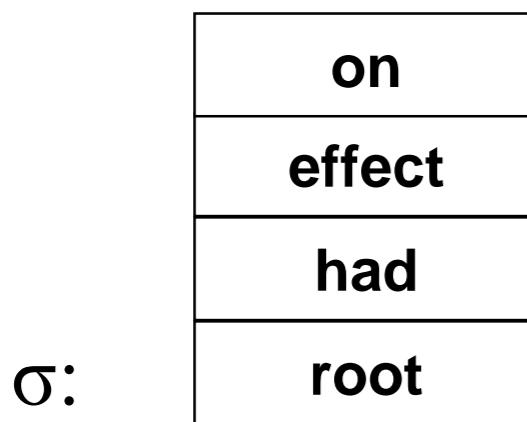
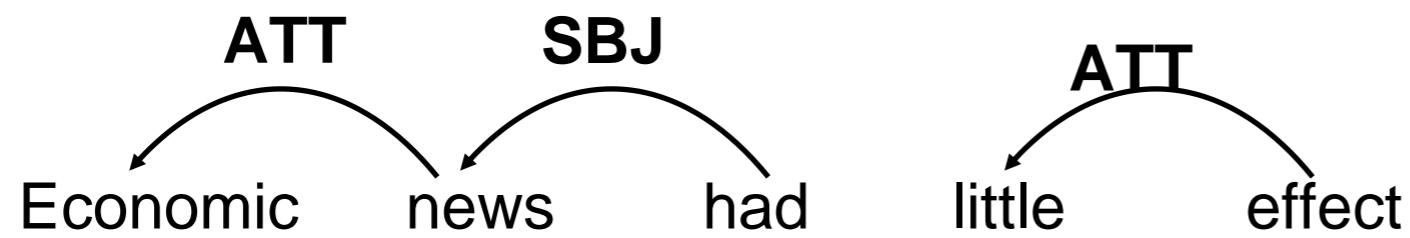
β:

[on, financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



σ :

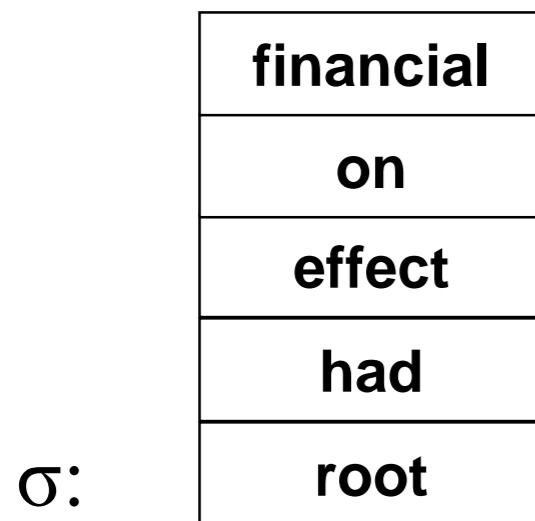
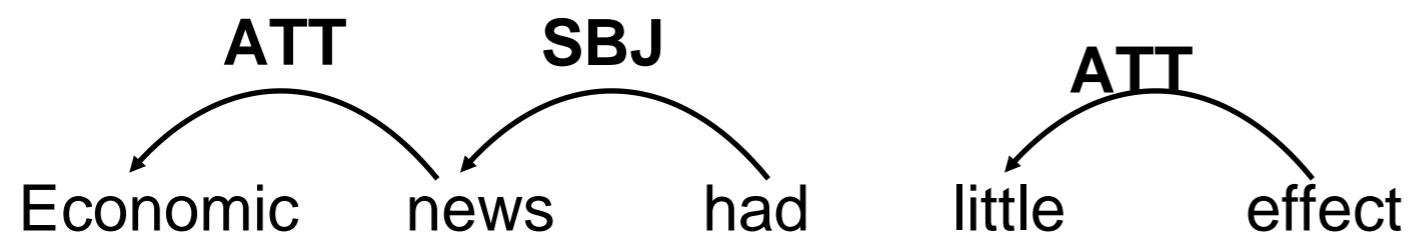
β :

[financial, markets,..]

Transition-Based Parsing - Oracle Example

next transition: Left-Arc_{ATT}

A:



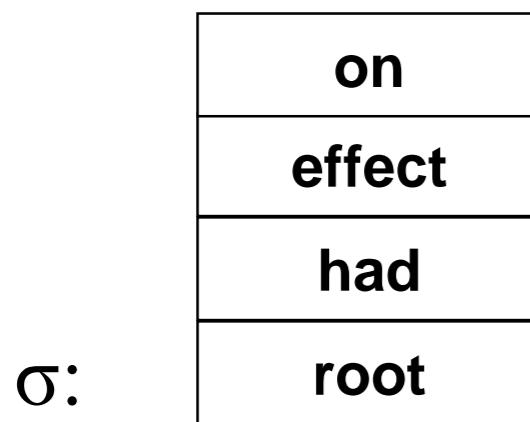
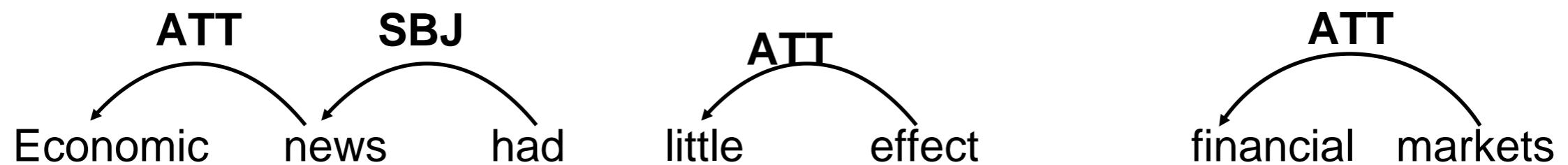
β :

[markets,..]

Transition-Based Parsing - Oracle Example

next transition: Right-Arc_{PC}

A:



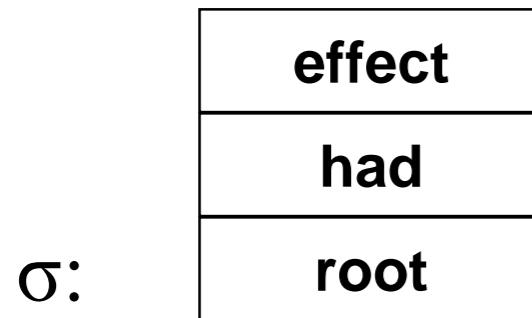
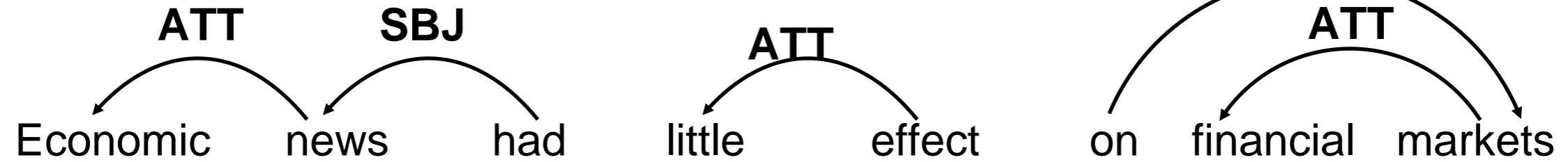
β :

[markets,..]

Transition-Based Parsing - Oracle Example

next transition: Right-Arc_{OBJ}

A:



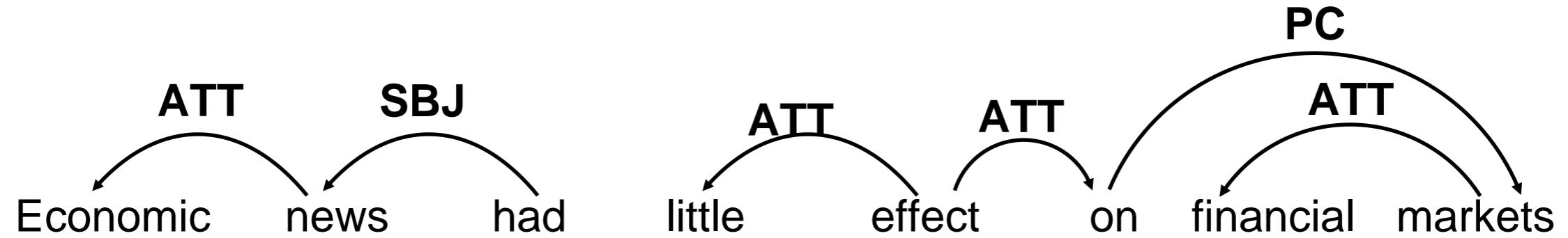
β :

[on,..]

Transition-Based Parsing - Oracle Example

next transition: Right-Arc_{OBJ}

A:



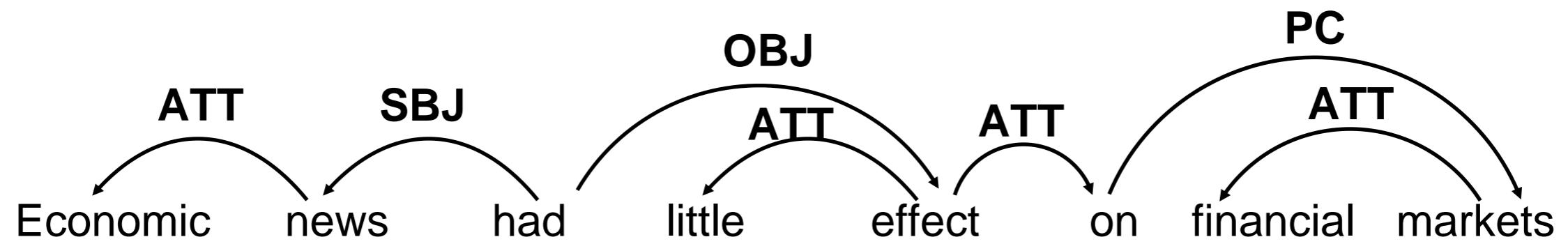
β :

[effect,..]

Transition-Based Parsing - Oracle Example

next transition: shift

A:



σ :

root

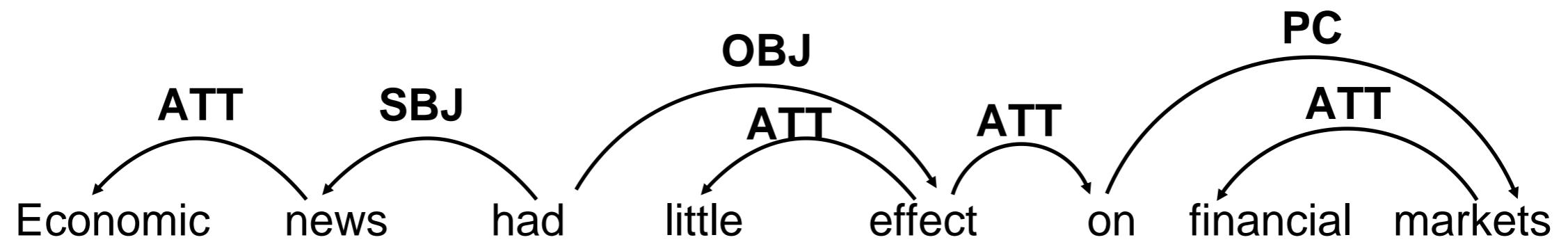
β :

[had,..]

Transition-Based Parsing - Oracle Example

next transition: Right-Arc_{PU}

A:



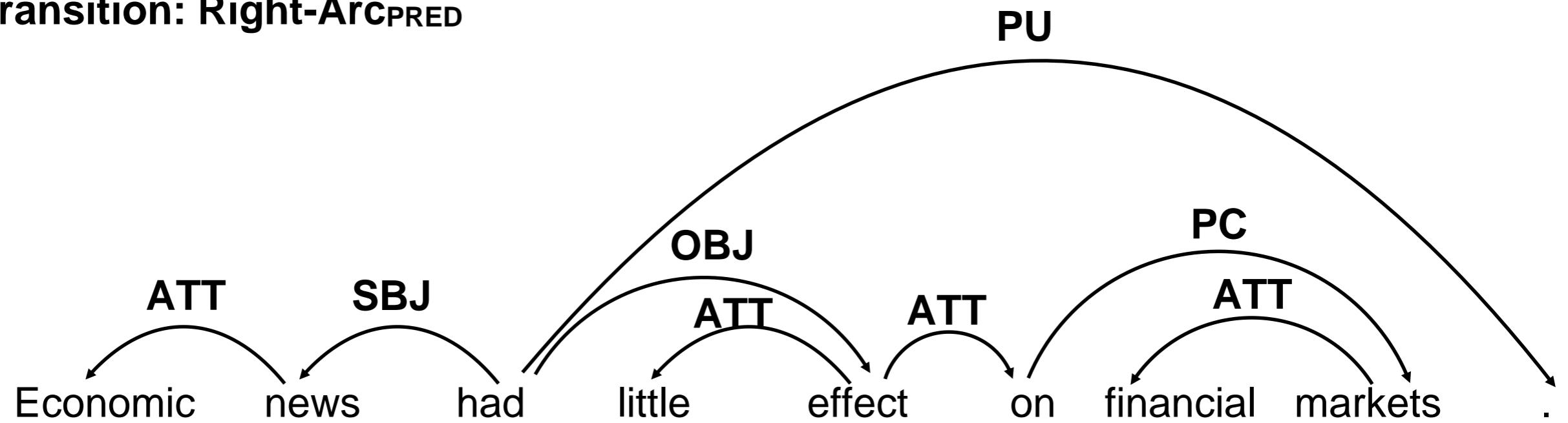
β:

[.]

Transition-Based Parsing - Oracle Example

next transition: Right-Arc_{PRED}

A:



σ :

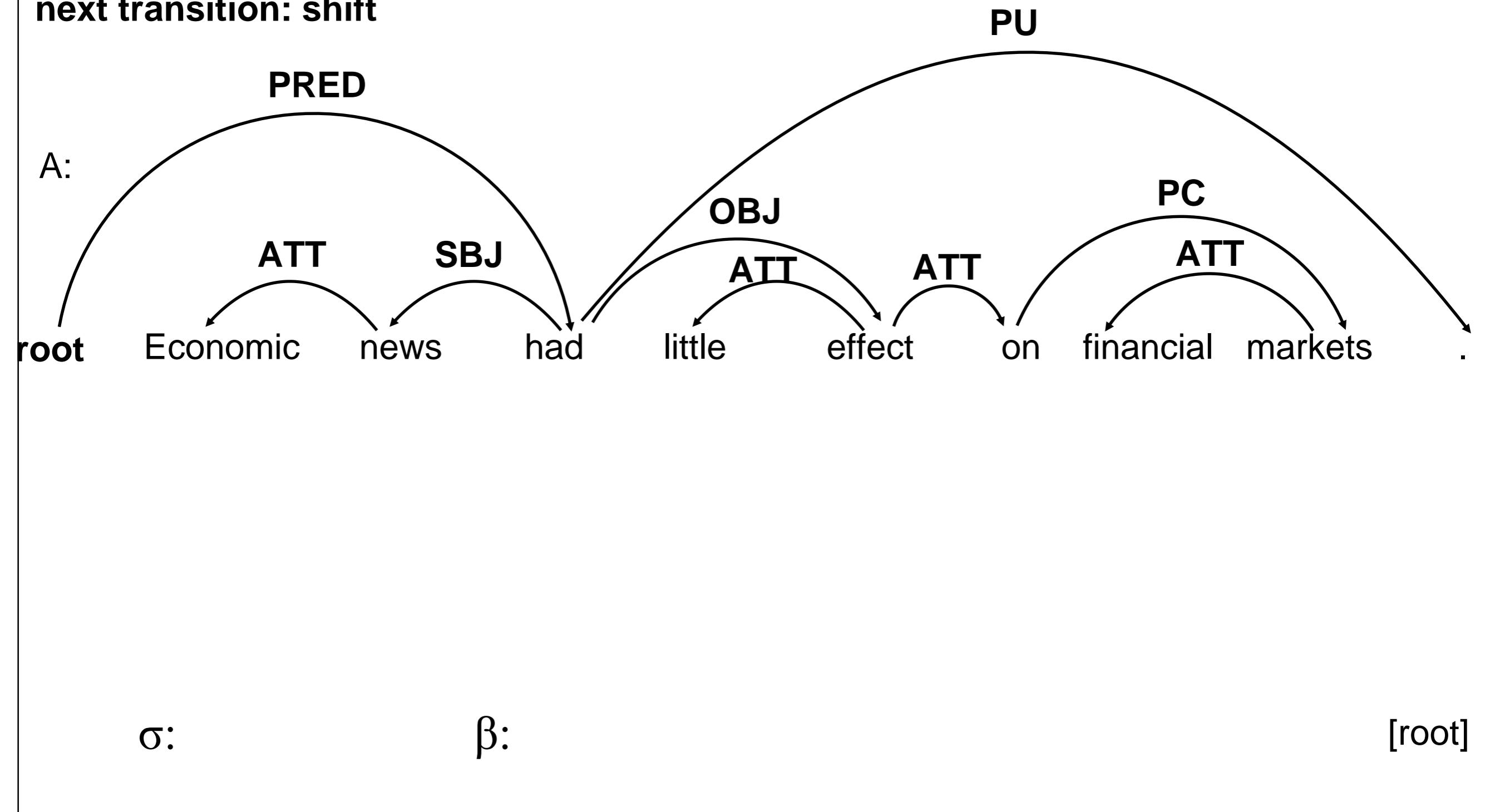
root

β :

[had]

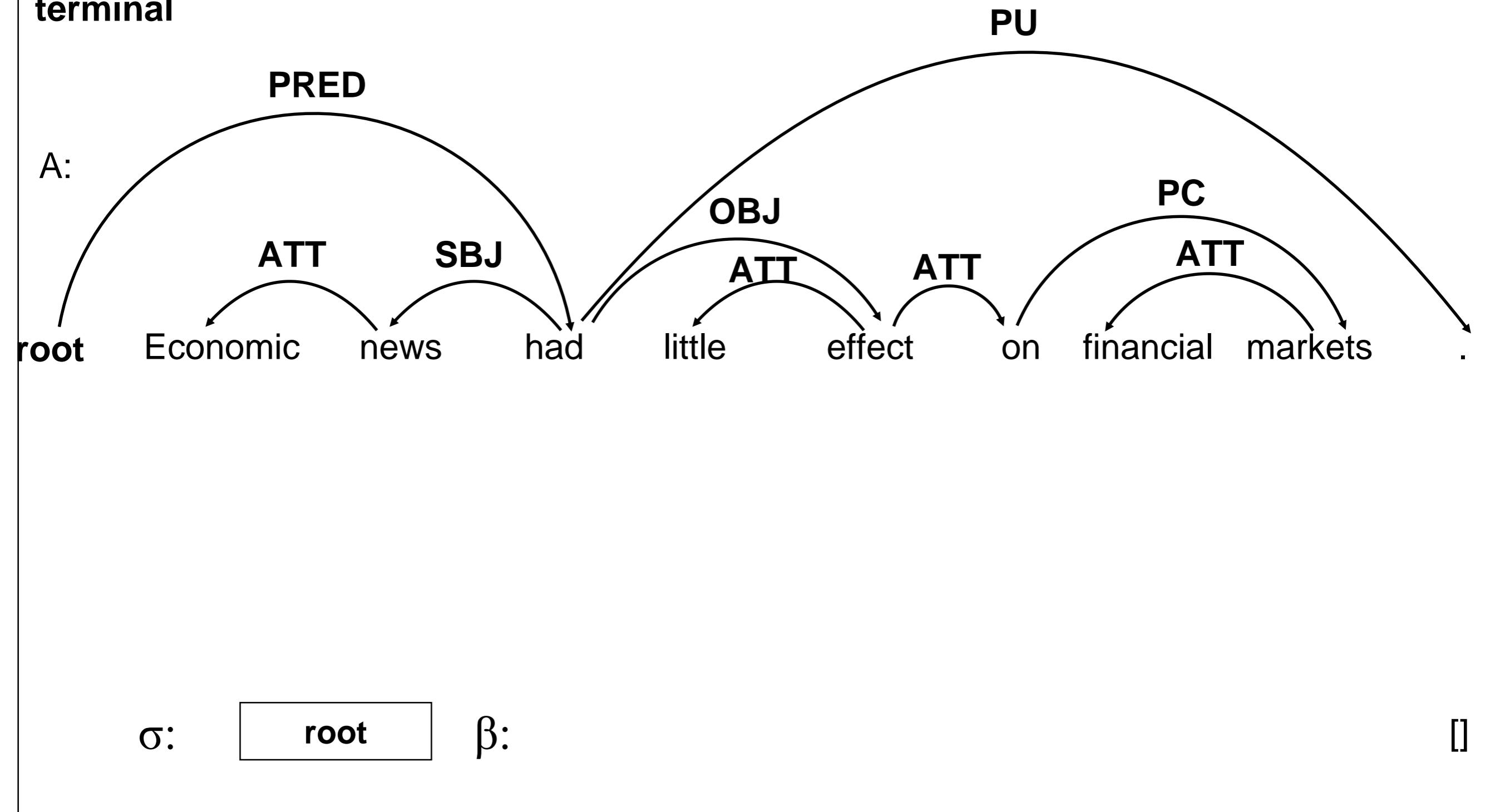
Transition-Based Parsing - Oracle Example

next transition: shift



Transition-Based Parsing - Oracle Example

terminal



Properties of the Transition System

- The time required to parse w_1, \dots, w_m with an oracle is $O(m)$. Why?
- Bottom-up approach: A node must *collect* all its children before its parent. Why?
- Can only produce projective trees. Why?
- This algorithm is complete (all projective trees over w_1, \dots, w_m can be produced by some sequence of transitions)
- Soundness: All terminal structures are projective forests (but not necessarily trees)

Deciding the Next Transition

- Instead of the unrealistic oracle, predict the next transition (and relation label) using a discriminative classifier.
 - Could use perceptron, log linear model, SVM, Neural Network, ...
 - This is a greedy approach (could use beam-search too).
 - If the classifier takes $O(1)$, the runtime for parsing is still $O(m)$ for m words.
- Questions:
 - What features should the classifier use?

*Local features from each state (buffer, stack, partial dependency structure)
... but ideally want to model entire history of transitions leading to the state.*
 - How to train the model?

Extracting Features

- Need to define a feature function that maps states to feature vectors.
- Each feature consists of:
 1. an address in the state description:
(identifies a specific word in the configuration, for example "top of stack").
 2. an attribute of the word in that address:
(for example POS, word form, lemma, word embedding, ...)

Example Features

Table 3.2: Typical feature model for transition-based parsing with rows representing address functions, columns representing attribute functions, and cells with + representing features.

Address	Attributes				
	FORM	LEMMA	POSTAG	FEATS	DEPREL
STK[0]	+	+	+	+	
STK[1]			+		
LDEP(STK[0])					+
RDEP(STK[0])					+
BUF[0]	+	+	+	+	
BUF[1]	+		+		
BUF[2]			+		
BUF[3]			+		
LDEP(BUF[0])					+
RDEP(BUF[0])					+

Training the Model

- Training data: Manually annotated (dependency) treebank
 - *Prague Dependency Treebank*
English/Czech parallel data, dependencies for full PTB WSJ.
 - *Universal Dependencies Treebank*
Treebanks for more than 80 languages (varying in size)
(<http://universaldependencies.org/>)
- **Problem: We have not actually seen the transition sequence, only the dependency trees!**
- Idea: Construct oracle transition sequences from the dependency tree.
Train the model on these transitions.

Constructing Oracle Transitions

- Start with initial state $([w_0]_\sigma, [w_1, w_2, \dots, w_m]_\beta, \{\}_A)$.
- Then predict the next transition using the annotated dependency tree A_d

$$o(c = (\sigma, \beta, A)) = \begin{cases} \text{LEFT-ARC}_r & \text{if } (\beta[0], r, \sigma[0]) \in A_d \\ \text{RIGHT-ARC}_r & \text{if } (\sigma[0], r, \beta[0]) \in A_d \text{ and, for all } w, r', \\ & \text{if } (\beta[0], r', w) \in A_d \text{ then } (\beta[0], r', w) \in A \\ \text{SHIFT}_r & \text{otherwise} \end{cases}$$

"Arc-Standard" Transitions

- **Shift:**

Move next word from the buffer to the stack

$$(\sigma, w_i | \beta, A) \Rightarrow (\sigma | w_i, \beta, A)$$

- **Left-Arc (for relation r):**

Build an edge from the next word on the buffer to the top word on the stack.

$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma, w_j | \beta, A \cup \{w_j, r, w_i\})$$

- **Right-Arc (for relation r)**

Build an edge from the top word on the stack to the next word on the buffer.

$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma, w_i | \beta, A \cup \{w_i, r, w_j\})$$

"Arc-Eager" Transitions

- **Shift:**

Move next word from the buffer to the stack

$$(\sigma, w_i | \beta, A) \Rightarrow (\sigma | w_i, \beta, A)$$

- **Left-Arc (for relation r):**

Build an edge from the next word on the buffer to the top word on the stack.

$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma, w_j | \beta, A \cup \{(w_j, r, w_i)\})$$

Precondition: $(w_j, *, w_i)$ is not yet in A.

- **Right-Arc (for relation r)**

Build an edge from the top word on the stack to the next word on the buffer.

$$(\sigma | w_i, w_j | \beta, A) \Rightarrow (\sigma | w_i | w_j, \beta, A \cup \{w_i, r, w_j\})$$

- **Reduce**

Remove a completed node from the stack.

$$(\sigma | w_i, \beta, A) \Rightarrow (\sigma, \beta, A)$$

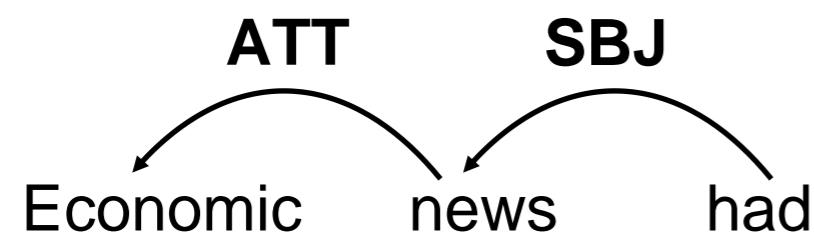
Precondition: there is some $(*, *, w_i)$ in A.

Arc-Eager Example

next transition: RightArc_{pred}

Can immediately attach *had* to root.

A:



σ :

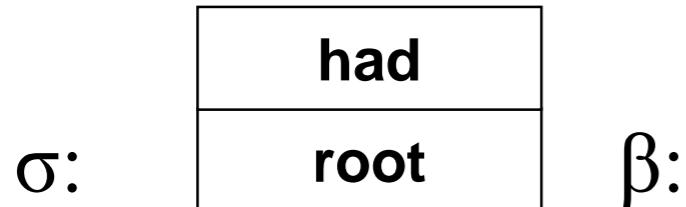
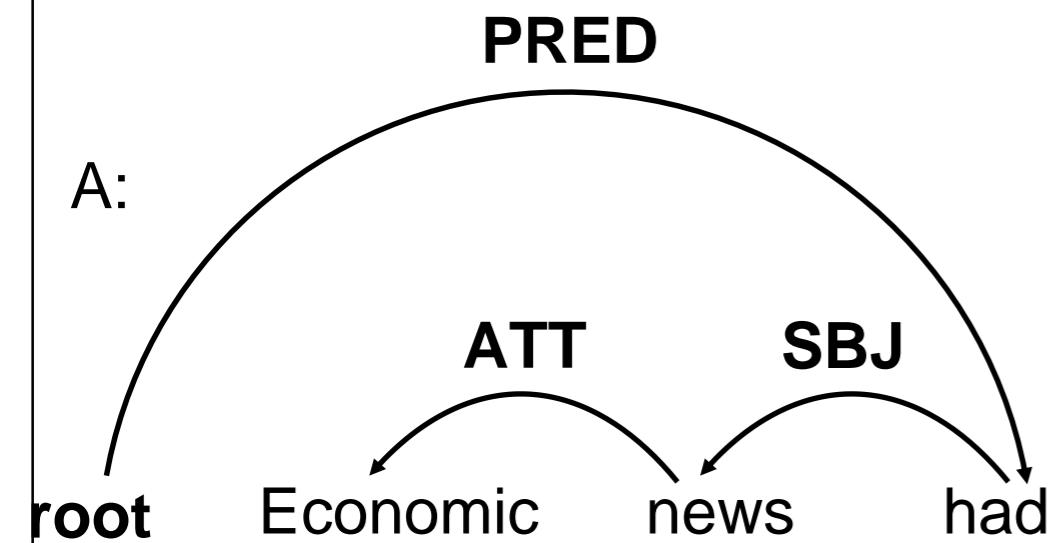
root

β :

[had, little, effect, on, financial, markets,..]

Arc-Eager Example

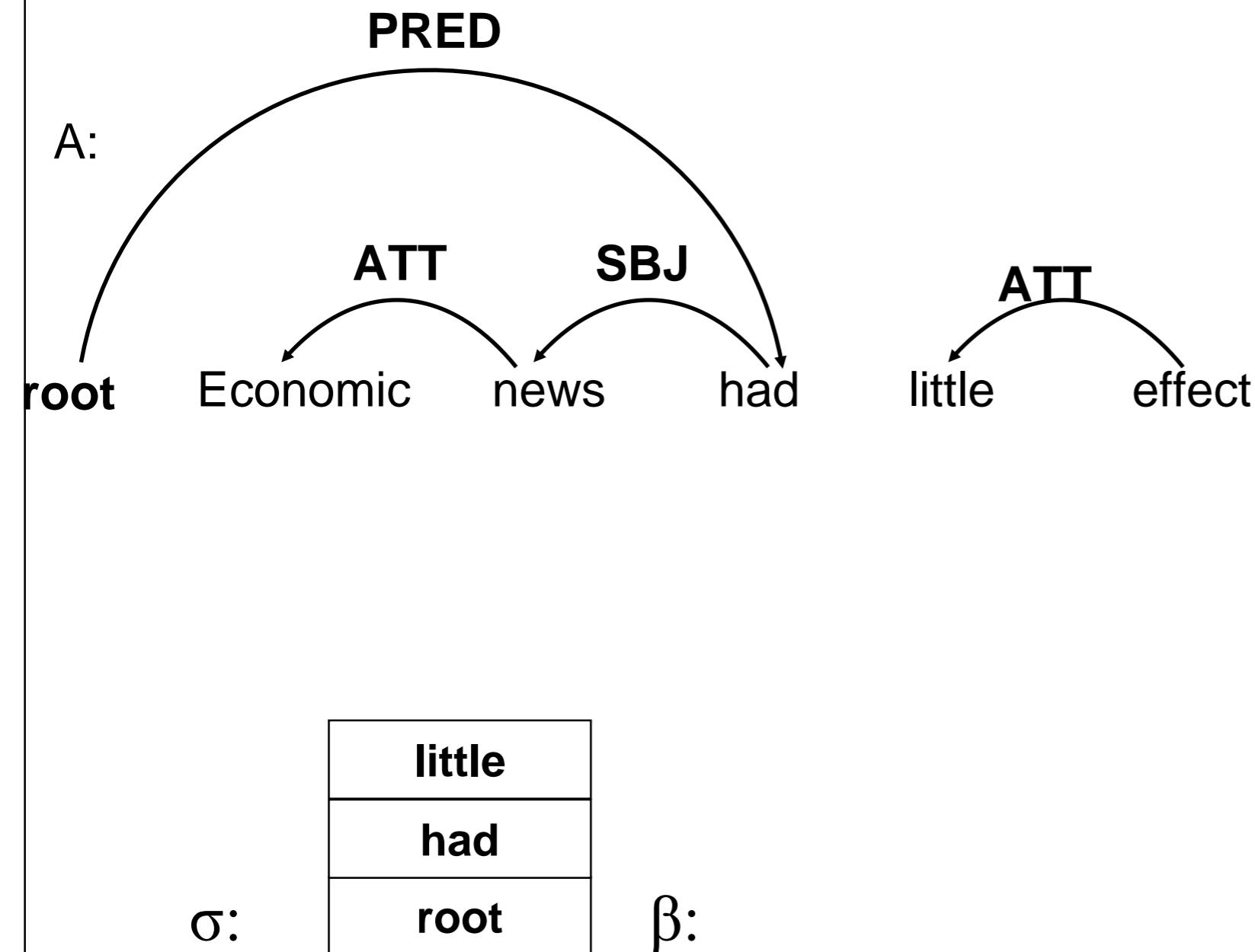
next transition: shift



[little, effect, on, financial, markets,..]

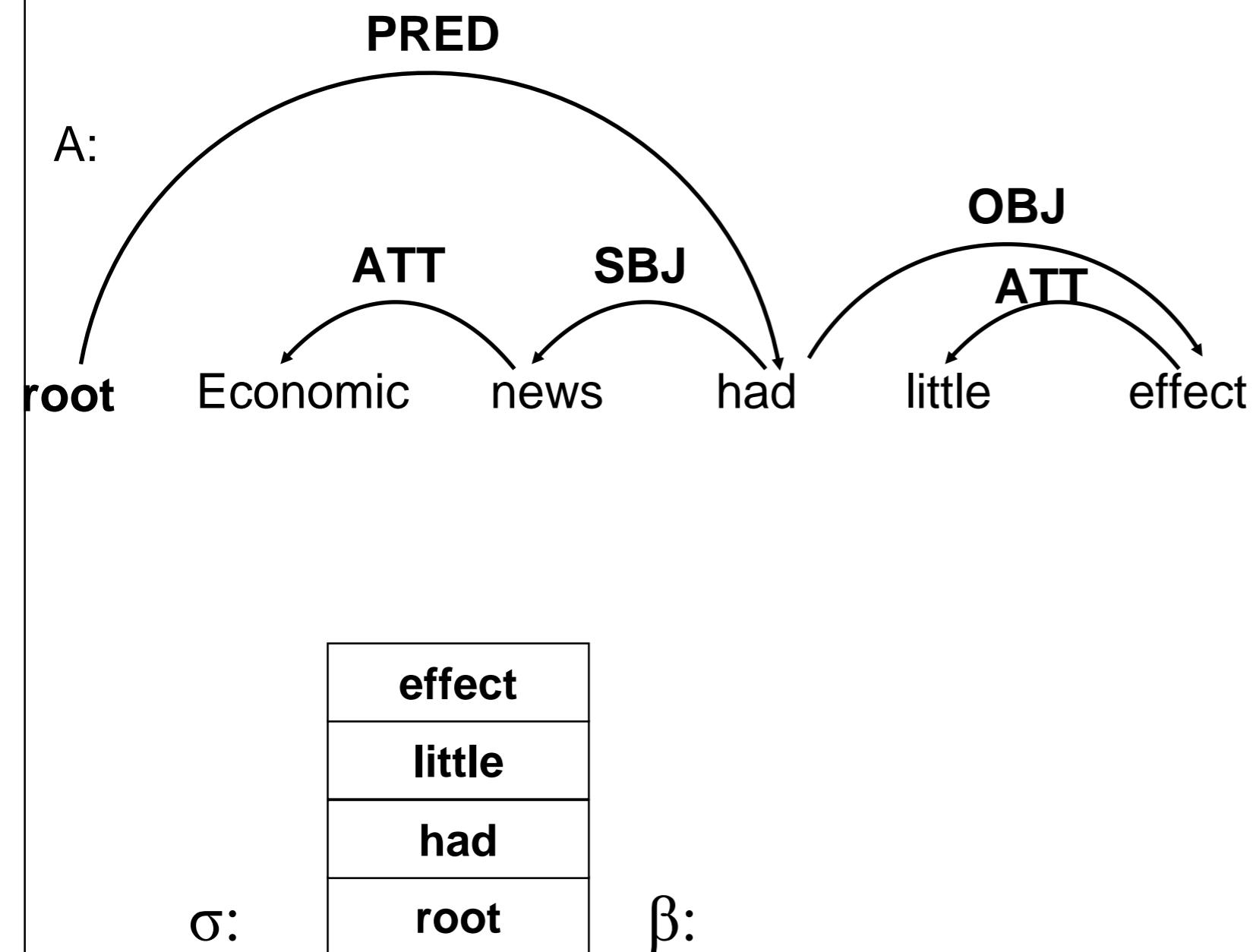
Arc-Eager Example

next transition: LeftArc_{ATT}



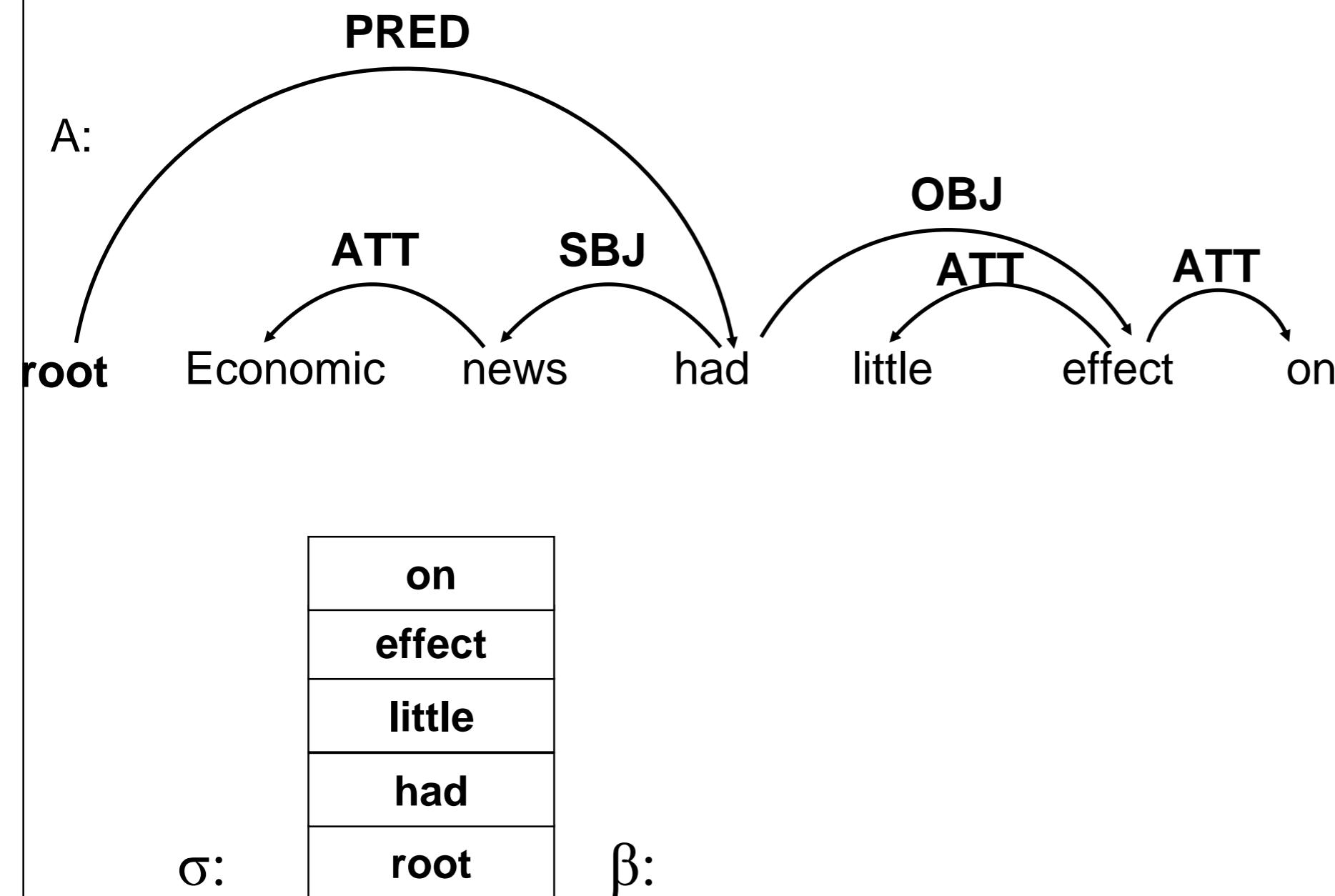
Arc-Eager Example

next transition: RightArcOBJ



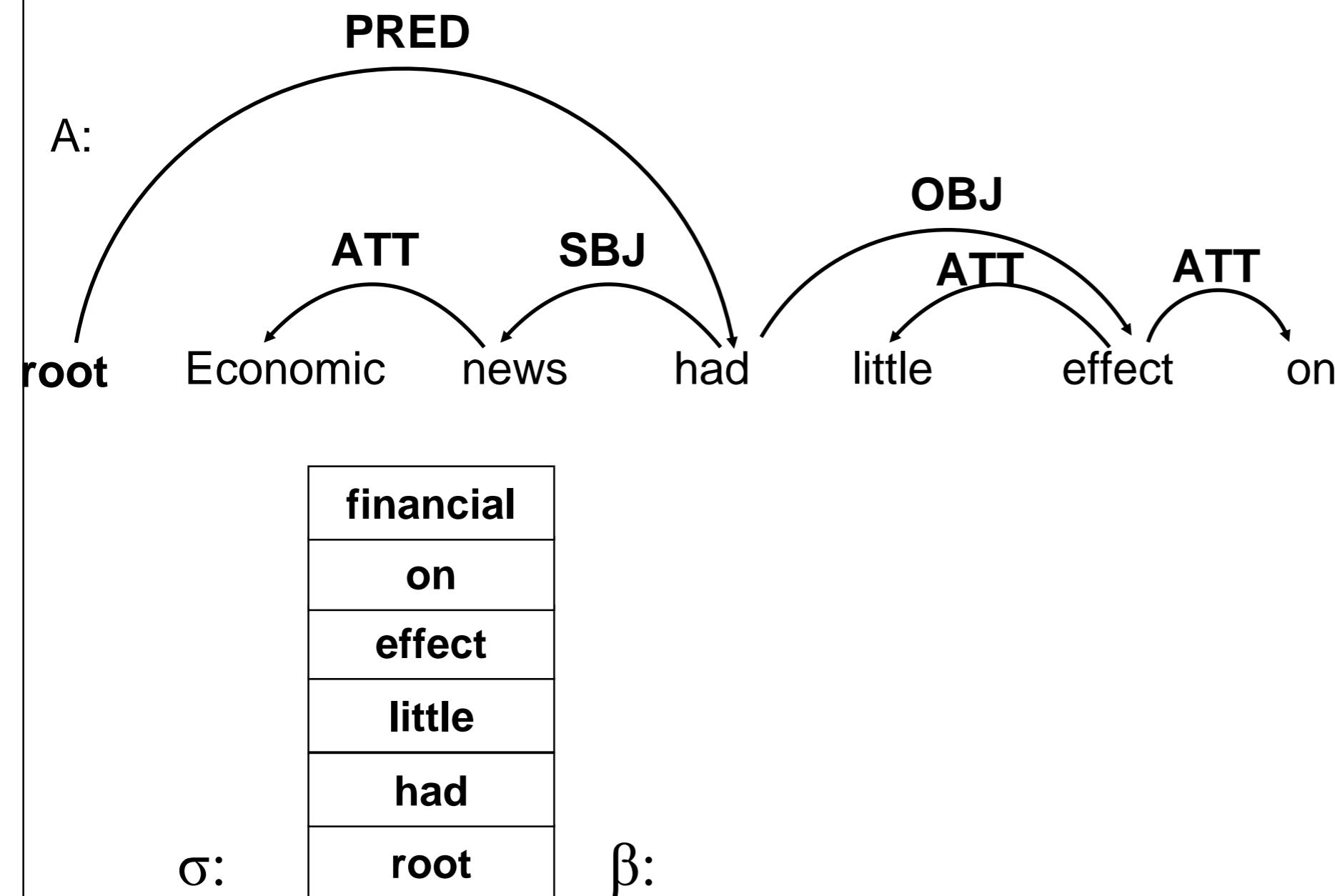
Arc-Eager Example

next transition: shift



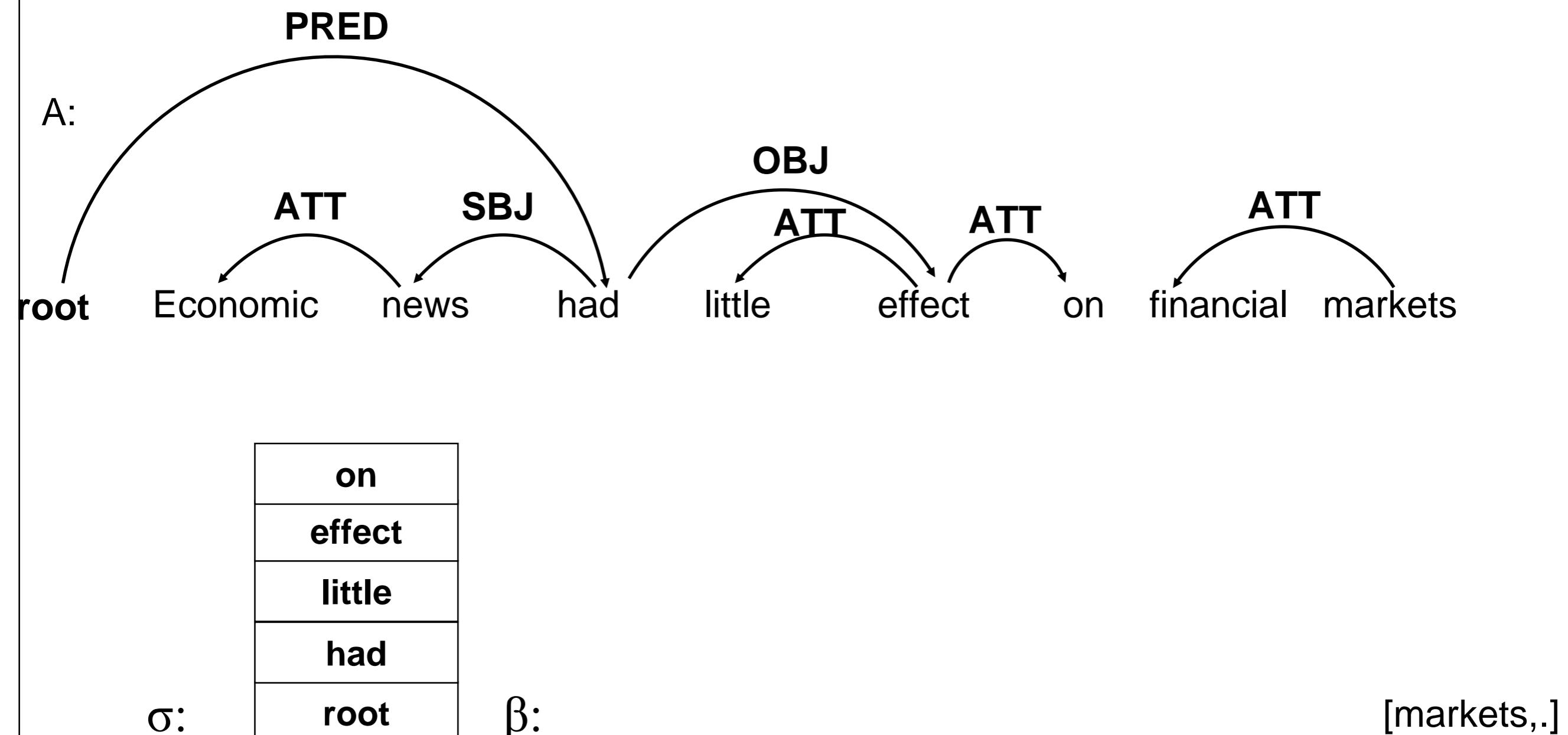
Arc-Eager Example

next transition: LeftArc_{ATT}



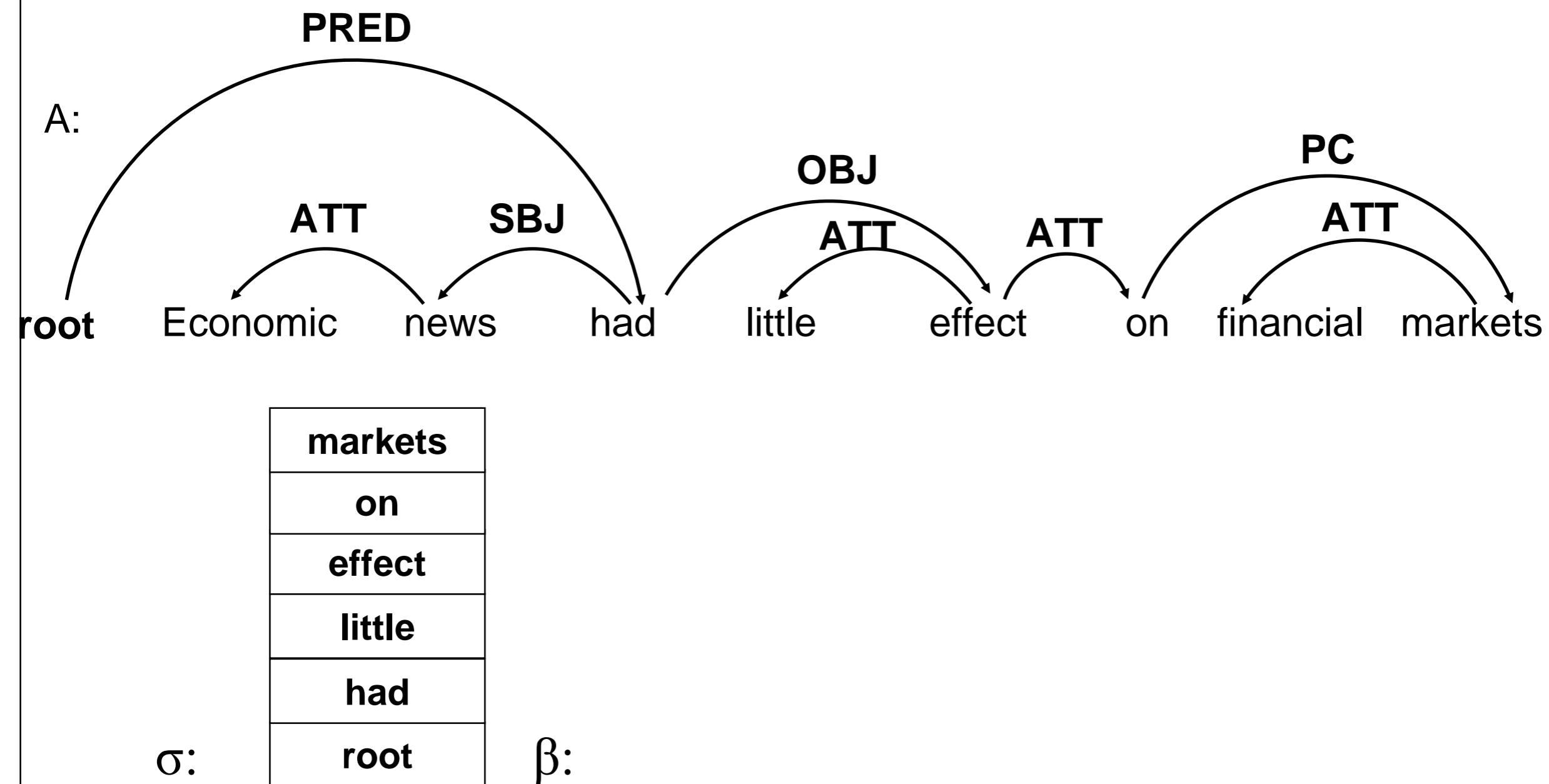
Arc-Eager Example

next transition: RightArc_{PC}



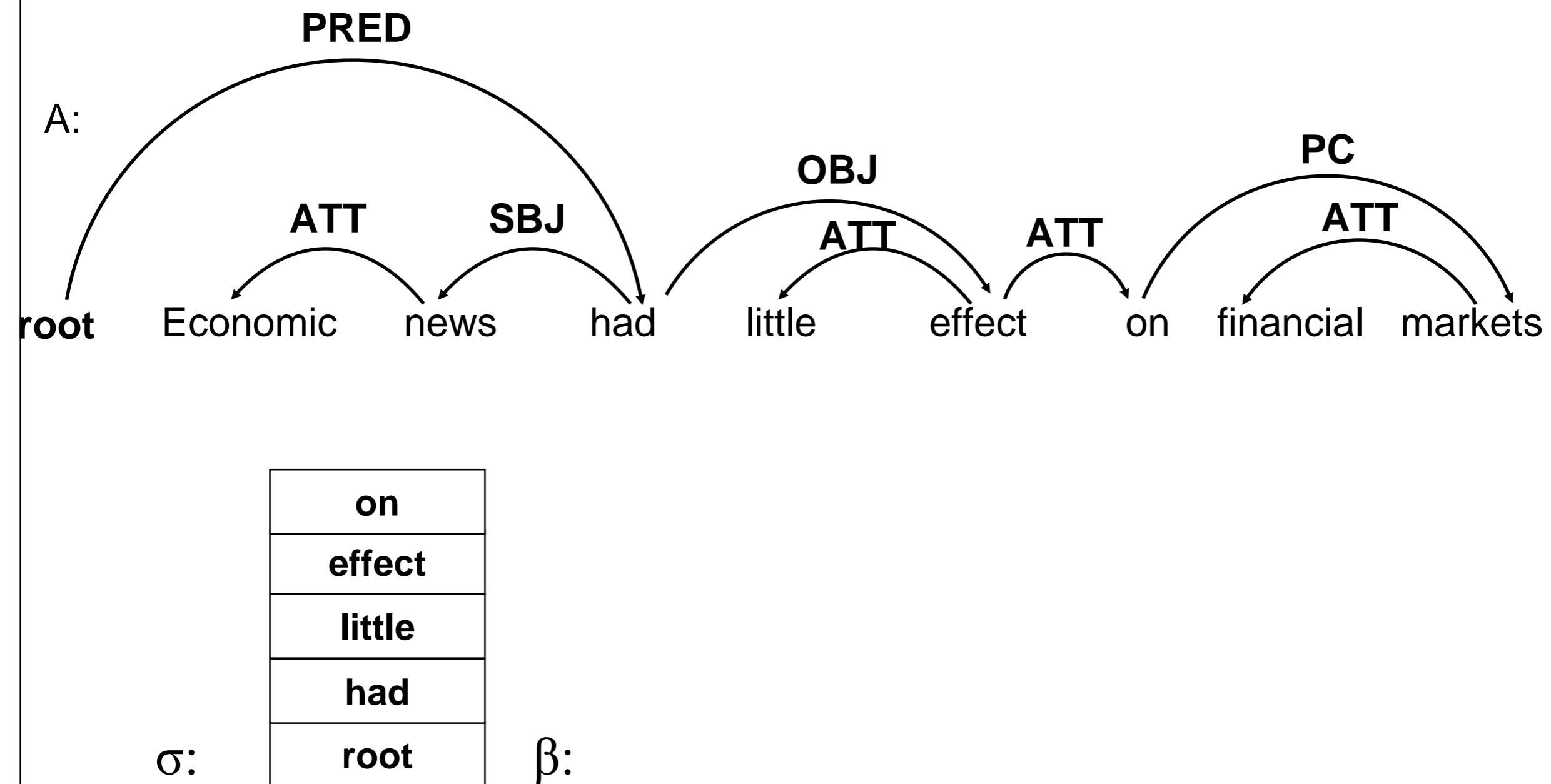
Arc-Eager Example

next transition: Reduce



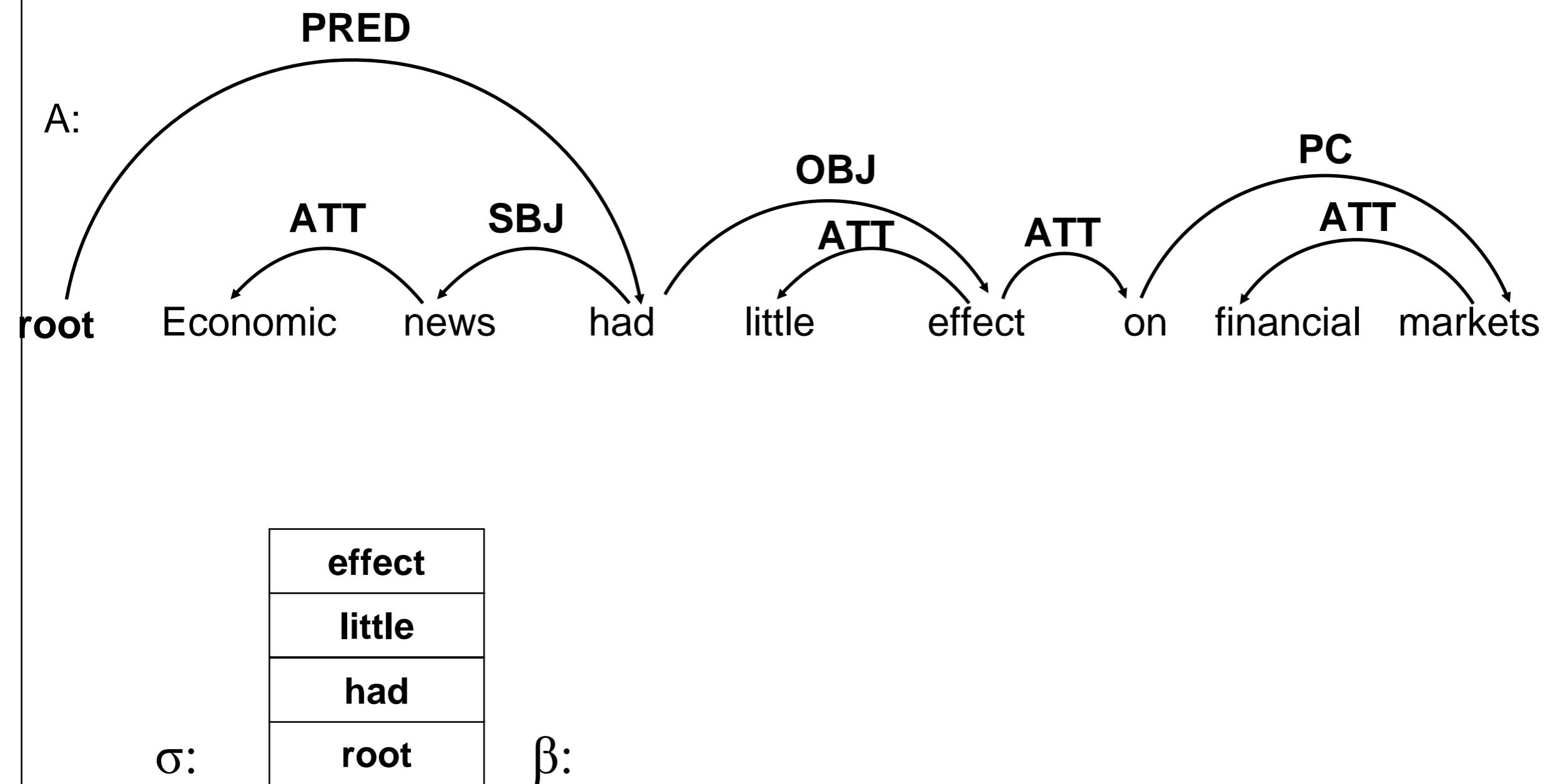
Arc-Eager Example

next transition: Reduce



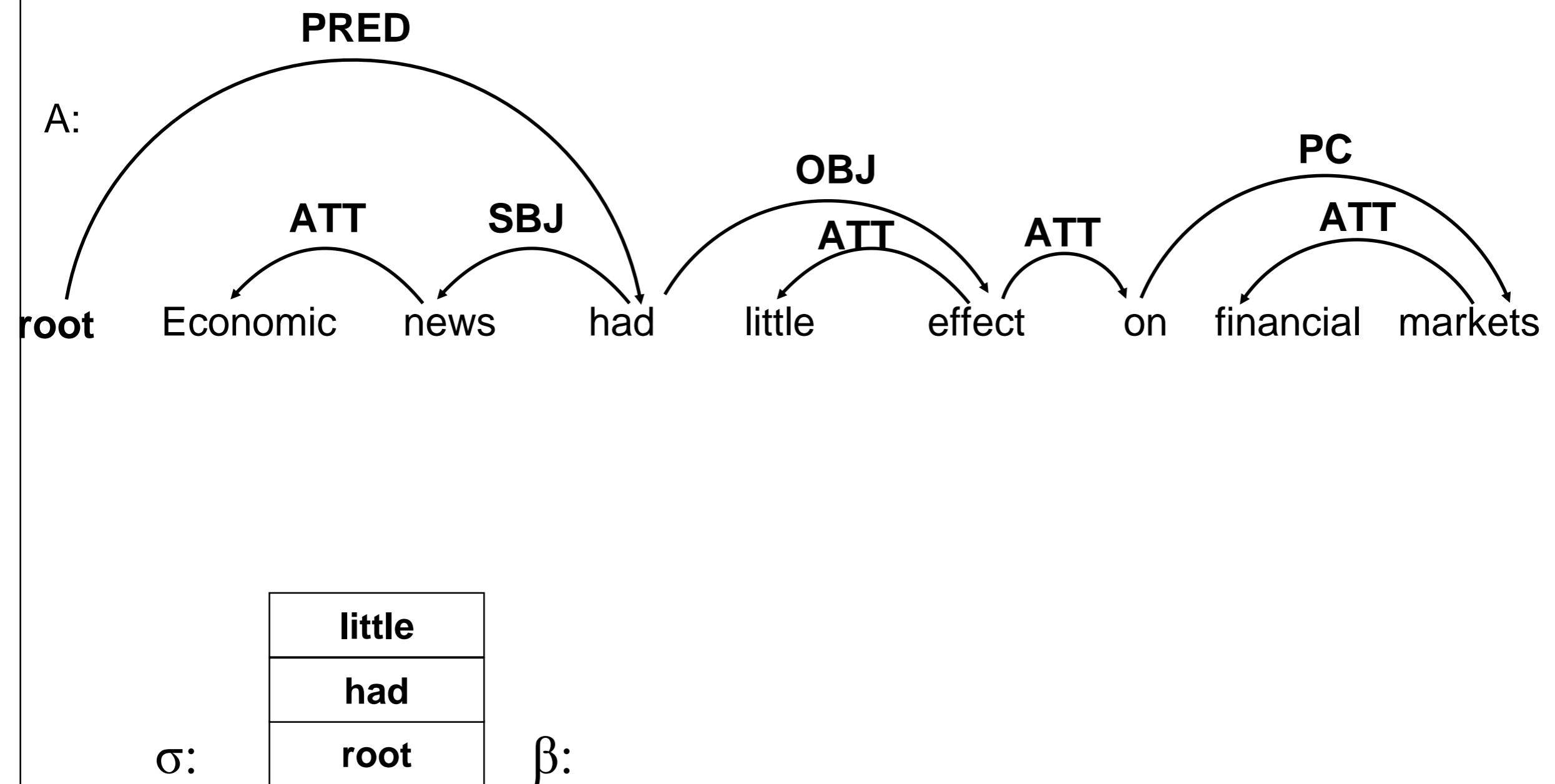
Arc-Eager Example

next transition: Reduce



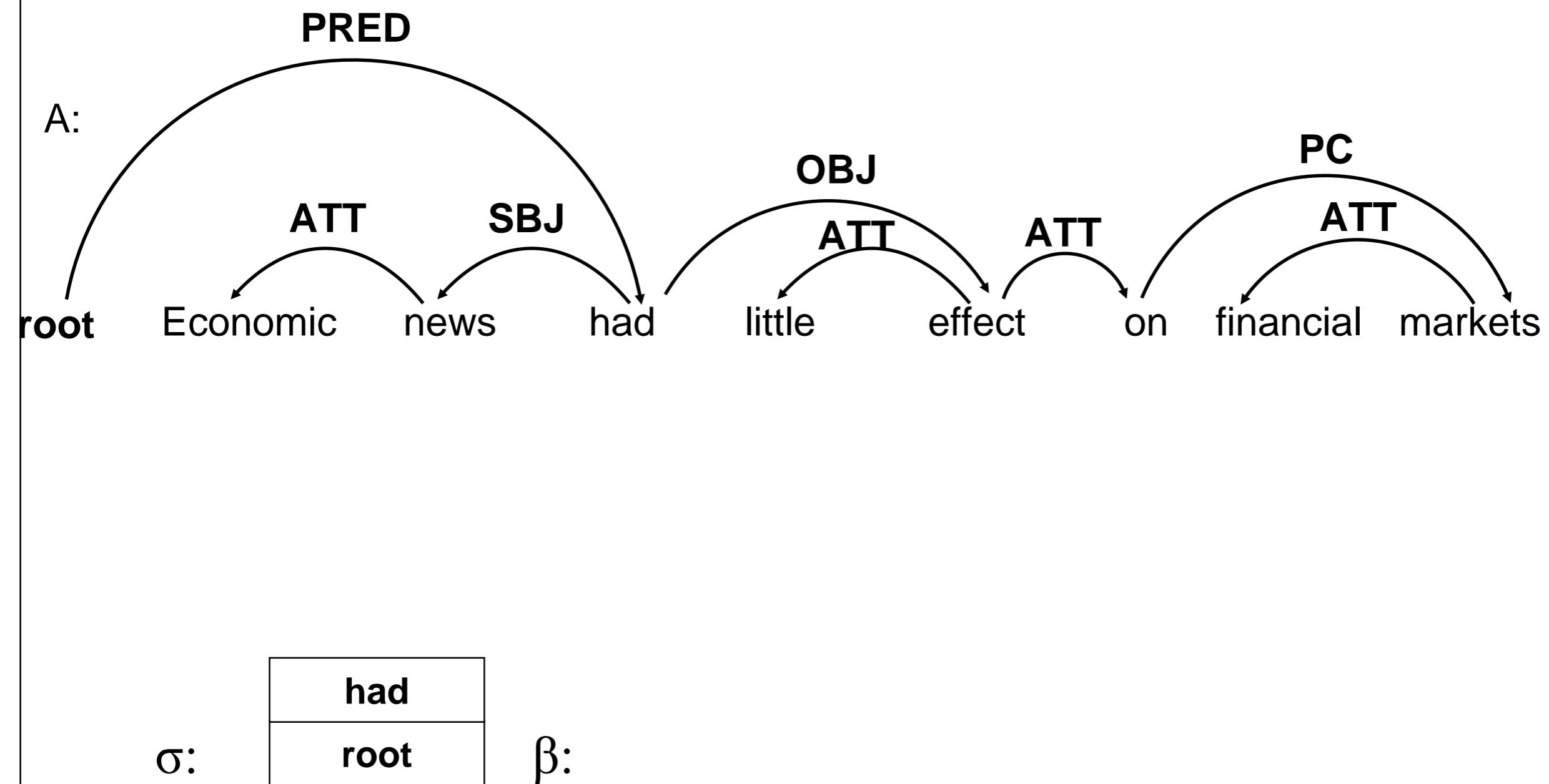
Arc-Eager Example

next transition: Reduce



Arc-Eager Example

next transition: Reduce

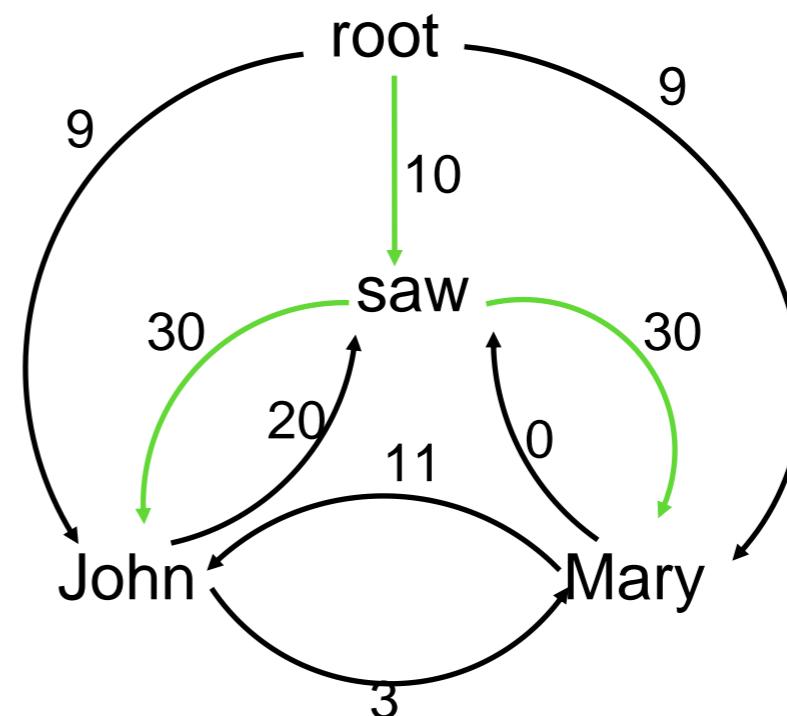
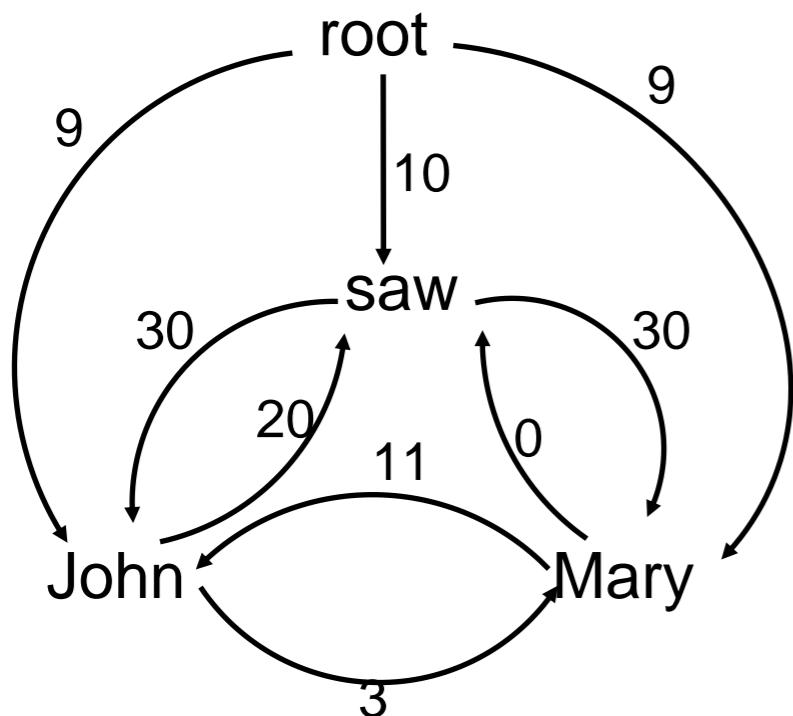


Graph-Based Approach

- Transition Based Parsing can only produce projective dependency structures? Why?
- Graph-based approaches do not have this restriction.
- Basic idea:
 - Each word is a vertex. Start with a completely connected graph.
 - Use standard graph algorithms to compute a Maximum Spanning Tree:
 - Need a model that assigns a score to each edge ("edge-factored model").

$$score(G) = \sum_{(w_i, r, w_j) \in A} \lambda(w_i, r, w_j)$$

MST Example



total score: 70

Computing the MST

- For undirected graphs, there are two common algorithms:
 - Kruskal's and Prim's, both run in $O(E \log V)$
- For dependency parsing we deal with directed graphs, so these algorithms are not guaranteed to find a tree.
 - Instead use Chu–Liu–Edmonds' algorithm, which runs in $O(EV)$ (naive implementation) or $O(E \log V)$ (with optimizations).

Earley Parser

Jurafsky's book example

Yassine Benajiba

Grammar (Same as the one in page 428 of the book with small changes)

$S \rightarrow NP\ VP$

$S \rightarrow Aux\ NP\ VP$

$S \rightarrow VP$

$NP \rightarrow \text{Pronoun}$

$NP \rightarrow \text{Proper-Noun}$

$NP \rightarrow \text{Det Nominal}$

$\text{Nominal} \rightarrow \text{Noun}$

$\text{Nominal} \rightarrow \text{Nominal Noun}$

$\text{Nominal} \rightarrow \text{Nominal PP}$

$VP \rightarrow \text{Verb}$

$VP \rightarrow \text{Verb NP}$

$VP \rightarrow \text{Verb NP PP}$

$VP \rightarrow \text{Verb PP}$

$VP \rightarrow VP\ PP$

$PP \rightarrow \text{Preposition NP}$

$\text{Det} \rightarrow \text{this}$

$\text{Noun} \rightarrow \text{flight}$

$\text{Verb} \rightarrow \text{book}$

$\text{Pronoun} \rightarrow \text{I}$

$\text{Proper-Noun} \rightarrow \text{Houston}$

$\text{Aux} \rightarrow \text{does}$

$\text{Preposition} \rightarrow \text{from}$

these are the
rules I've changed
for the sake of
brevity

Sentence to parse

book this flight

Earley Algorithm

- Keep track of parser states in a table (“chart”). $\text{Chart}[k]$ contains a set of all parser states that end in position k .

- **Input:** Grammar $G=(N, \Sigma, R, S)$, input string s of length n .

- **Initialization:** For each production $S \rightarrow \alpha \in R$ add a state $S \rightarrow \cdot \alpha [0,0]$ to $\text{Chart}[0]$.

let's
part do this
first

- for $i = 0$ to n :

- for each *state* in $\text{Chart}[i]$:

- if *state* is of form $A \rightarrow \alpha \cdot s[i] \beta [k,i]$:
 scan(state)

- elif *state* is of form $A \rightarrow \alpha \cdot B \beta [k,i]$:
 predict(state)

- else: // *state* is of form $A \rightarrow \alpha \cdot [k,i]$
 complete(state)



this is a mistake. it should be

• elif State is of form $A \rightarrow \alpha \cdot [k,i]$

Chart [0]

- S₀ S → • NP VP [0, 6]
S₁ S → • AUX NP VP [0, 6]
S₂ S → • VP [0, 0]

Earley Algorithm

- Keep track of parser states in a table (“chart”). $Chart[k]$ contains a set of all parser states that end in position k .
- **Input:** Grammar $G=(N, \Sigma, R, S)$, input string s of length n .

- **Initialization:** For each production $S \rightarrow \alpha \in R$ add a state $S \rightarrow \cdot \alpha [0,0]$ to $Chart[0]$.

- for $i = 0$ to n :
 - for each *state* in $Chart[i]$:
 - if *state* is of form $A \rightarrow \alpha \cdot s[i] \beta [k,i]$:
 scan(state)
 - elif *state* is of form $A \rightarrow \alpha \cdot B \beta [k,i]$:
 predict(state)
 - else: // *state* is of form $A \rightarrow \alpha \cdot [k,i]$
 complete(state)

Now let's go inside
this loop

Chart [0]

S ₀	$S \rightarrow \bullet NP VP [0, 6]$
S ₁	$S \rightarrow \bullet AUX NP VP [0, 6]$
S ₂	$S \rightarrow \bullet VP [0, 6]$

When we process this we will call Predict since the element next to \bullet is a non-terminal.
 \Rightarrow new states with NP at their left hand side (LHS) will be added

Chart [0]

- S₀ S → • NP VP [0,6]
S₁ S → • AUX NP VP [0,6]
S₂ S → • VP [0,0]
S₃ NP → • Pronoun [0,0]
S₄ NP → • Proper-noun [0,0]
S₅ NP → • Det Nominal [0,0]

Chart [0]

S0	$S \rightarrow \bullet \text{NP VP } [0, 0]$
S1	$S \rightarrow \bullet \text{AUX NP VP } [0, 0]$
S2	$S \rightarrow \bullet \text{VP } [0, 0]$
S3	$\text{NP} \rightarrow \bullet \text{Pronoun } [0, 0]$
S4	$\text{NP} \rightarrow \bullet \text{Proper-noun } [0, 0]$
S5	$\text{NP} \rightarrow \bullet \text{Det Nominal } [0, 0]$

now we move to process the next state. Same as the one before except that now we're interested in 'Aux' at the LHS

Chart [0]

- S0 $S \rightarrow \bullet NP VP [0,0]$
- S1 $S \rightarrow \bullet AUX NP VP [0,0]$
- S2 $S \rightarrow \bullet VP [0,0]$
- S3 $NP \rightarrow \bullet Pronoun [0,0]$
- S4 $NP \rightarrow \bullet Proper\text{-}noun [0,0]$
- S5 $NP \rightarrow \bullet Det Nominal [0,0]$
- S6 $AUX \rightarrow \bullet does [0,0]$

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper-noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$

Same story for VP

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper-noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper\text{-}noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$

going through these one by one should do
the same thing for 'pronoun', 'proper-noun' and
'Det'.

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper\text{-}noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$

S14 Det $\rightarrow \bullet$ this $[0,0]$

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper-noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper-noun \rightarrow \bullet Houston [0,0]$

S14 Det $\rightarrow \bullet$ this [0,0]

now we're here. This will not call the scan() func
because even if the word next to \bullet is a terminal, it
is not $S[0]$, which is the word 'book'

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper-noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$\boxed{VP \rightarrow \bullet Verb [0,0]}$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$

this will

S14 Det $\rightarrow \bullet this [0,0]$

(all predict) for 'Verb' at the LHS

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper\text{-}noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$

|

S14 Det $\rightarrow \bullet$ this $[0,0]$
S15 Verb $\rightarrow \bullet$ book $[0,0]$

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper-noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$

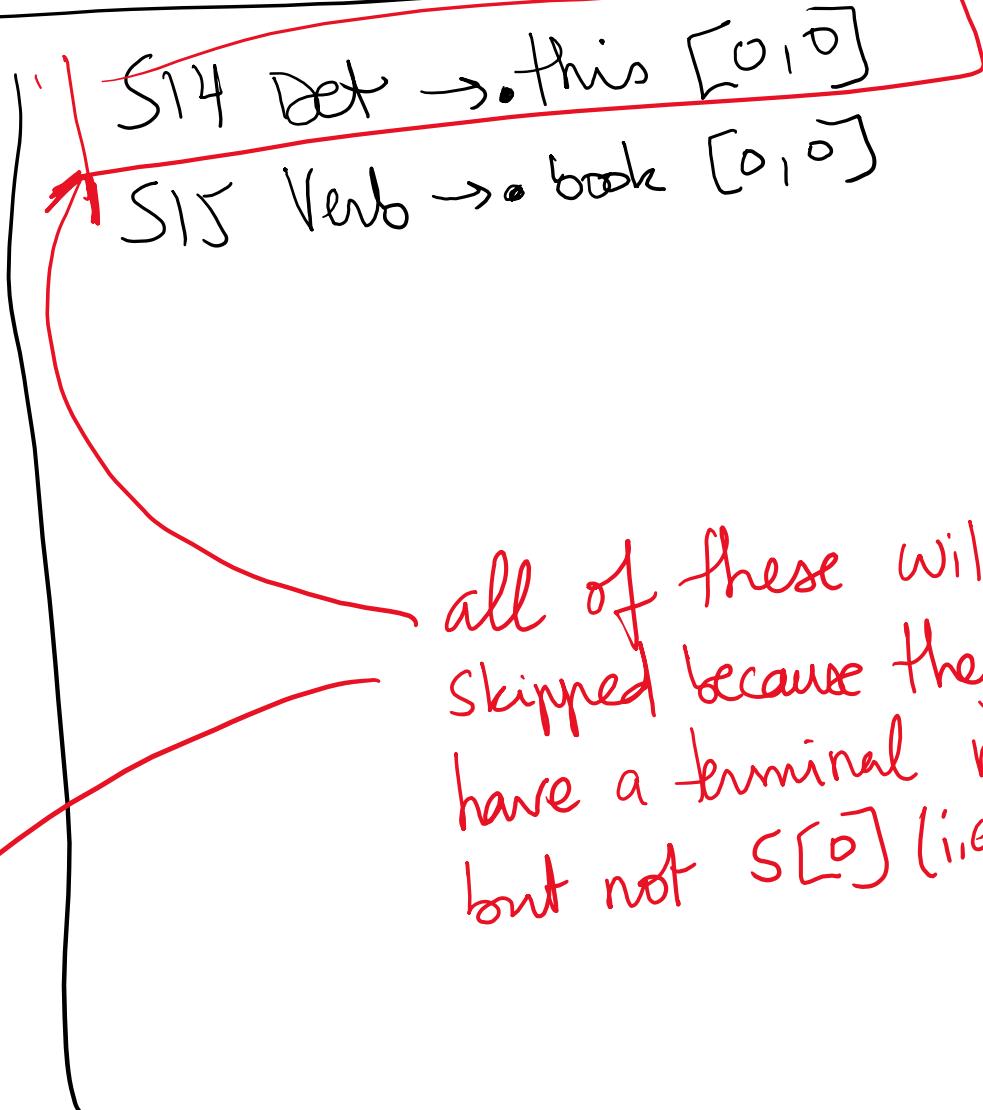
S14 Det $\rightarrow \bullet$ this $[0,0]$
 S15 Verb $\rightarrow \bullet$ book $[0,0]$

all of these should call Predict() for
 rules with 'Verb' at the LHS (but since
 we already did that nothing will be
 added to the chart)

Same for this (but for 'VP' at
 the LHS)

Chart [0]

S ₀	$S \rightarrow \bullet NP VP [0,0]$
S ₁	$S \rightarrow \bullet AUX NP VP [0,0]$
S ₂	$S \rightarrow \bullet VP [0,0]$
S ₃	$NP \rightarrow \bullet Pronoun [0,0]$
S ₄	$NP \rightarrow \bullet Proper-noun [0,0]$
S ₅	$NP \rightarrow \bullet Det Nominal [0,0]$
S ₆	$AUX \rightarrow \bullet does [0,0]$
S ₇	$VP \rightarrow \bullet Verb [0,0]$
S ₈	$VP \rightarrow \bullet Verb VP [0,0]$
S ₉	$VP \rightarrow \bullet Verb NP PP [0,0]$
S ₁₀	$VP \rightarrow \bullet Verb PP [0,0]$
S ₁₁	$VP \rightarrow \bullet VP PP [0,0]$
S ₁₂	$Pronoun \rightarrow \bullet I [0,0]$
S ₁₃	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$



all of these will be
 skipped because they all
 have a terminal next to \bullet ,
 but not $S[0]$ (i.e. 'book')

Chart [0]

S0	$S \rightarrow \bullet NP VP [0,0]$
S1	$S \rightarrow \bullet AUX NP VP [0,0]$
S2	$S \rightarrow \bullet VP [0,0]$
S3	$NP \rightarrow \bullet Pronoun [0,0]$
S4	$NP \rightarrow \bullet Proper\text{-}noun [0,0]$
S5	$NP \rightarrow \bullet Det Nominal [0,0]$
S6	$AUX \rightarrow \bullet does [0,0]$
S7	$VP \rightarrow \bullet Verb [0,0]$
S8	$VP \rightarrow \bullet Verb VP [0,0]$
S9	$VP \rightarrow \bullet Verb NP PP [0,0]$
S10	$VP \rightarrow \bullet Verb PP [0,0]$
S11	$VP \rightarrow \bullet VP PP [0,0]$
S12	$Pronoun \rightarrow \bullet I [0,0]$
S13	$Proper\text{-}noun \rightarrow \bullet Houston [0,0]$

S14 Det $\rightarrow \bullet this [0,0]$

S15 Verb $\rightarrow \bullet book [0,0]$

this will call Scan() and
new state will be added to
chart [1]

Also, at this point we're done with
chart [0] and we should move to
the next

Chart [1]

S/6 Verb → book • [0,1]

Chart [1]

S16 Verb → book • [0,1]

Since the • here is at the end we will call Complete(). Remember what Complete() looks like:

- function **complete(state)**: // state is of form $A \rightarrow \alpha \cdot [k, j]$

- for each state $B \rightarrow \beta \cdot A \gamma [i, k]$ add a new state $B \rightarrow \beta A \cdot \gamma [i, j]$ to Chart[j]

We need to create new states then in Chart[j], i.e. Chart[1]

in our case: $k = \emptyset, j = 1$

So basically we're looking for states that have 'Verb' next to • on the RHS and [0,0] as an index $\Rightarrow S7, S8, S9, S10$ in Chart[0]

Chart [1]

- S16 Verb → book • $[0, 1]$
- S17 VP → Verb • $[0, 1]$
- S18 VP → Verb • NP $[0, 1]$
- S19 VP → Verb • NP PP $[0, 1]$
- S20 VP → Verb • PP $[0, 1]$

Chart [1]

- S16 Verb → book • [0,1]
- S17 VP → Verb • [0,1]
- S18 VP → Verb • NP [0,1]
- S19 VP → Verb • NP PP [0,1]
- S20 VP → Verb • PP [0,1]

We're gonna call complete again but this time we'll look for states with 'VP' next to • on the RAS and [0,0] as an index \Rightarrow S2 and S11 in chart [0]

Chart [1]

- S16 Verb → book • $[0, 1]$
- S17 VP → Verb • $[0, 1]$
- S18 VP → Verb • NP $[0, 1]$
- S19 VP → Verb • NP PP $[0, 1]$
- S20 VP → Verb • PP $[0, 1]$
- S21 S → VP • $[0, 1]$
- S22 VP → VP • PP $[0, 1]$

Chart [1]

- S16 Verb → book • [0, 1]
- S17 VP → Verb • [0, 1]
- S18 VP → Verb • NP [0, 1]
- S19 VP → Verb • NP PP [0, 1]
- S20 VP → Verb • PP [0, 1]
- S21 S → VP • [0, 1]
- S22 VP → VP • PP [0, 1]
- this will call predict() for 'NP' on the LTS. The added states should have [1, 1] as the index*
- this will come right after but no states will be added*

Chart [1]

- S16 Verb → book • $[0, 1]$
- S17 VP → Verb • $[0, 1]$
- S18 VP → Verb • NP $[0, 1]$
- S19 VP → Verb • NP PP $[0, 1]$
- S20 VP → Verb • PP $[0, 1]$
- S21 S → VP • $[0, 1]$
- S22 VP → VP • PP $[0, 1]$
- S23 NP → • Pronoun $[1, 1]$
- S24 NP → • Proper-noun $[1, 1]$
- S25 NP → • Det Nominal $[1, 1]$

Chart [1]

-
- S16 Verb → book • [0,1]
 - S17 VP → Verb • [0,1]
 - S18 VP → Verb • NP [0,1]
 - S19 VP → Verb • NP PP [0,1]
 - S20 VP → Verb • PP [0,1] call predict()
 - S21 S → VP • [0,1]
 - S22 VP → VP • PP [0,1]
 - S23 NP → • Pronoun [1,1]
 - S24 NP → • Proper-noun [1,1]
 - S25 NP → • Det Nominal [1,1]

Chart [1]

-
- S16 Verb → book • $[0, 1]$
 - S17 VP → Verb • $[0, 1]$
 - S18 VP → Verb • NP $[0, 1]$
 - S19 VP → Verb • NP PP $[0, 1]$
 - S20 VP → Verb • PP $[0, 1]$
 - S21 S → VP • $[0, 1]$
 - S22 VP → VP • PP $[0, 1]$
 - S23 NP → • Pronoun $[1, 1]$
 - S24 NP → • Proper-noun $[1, 1]$
 - S25 NP → • tcf Nominal $[1, 1]$
 - S26 PP → • Preposition NP $[1, 1]$

Chart [1]

- S16 Verb → book • [0,1]
- S17 VP → Verb • [0,1]
- S18 VP → Verb • NP [0,1]
- S19 VP → Verb • NP PP [0,1]
- S20 VP → Verb • PP [0,1]
- S21 S → VP • [0,1]
- S22 VP → VP • PP [0,1]
- S23 NP → • Pronoun [1,1]
- S24 NP → • Proper-noun [1,1]
- S25 NP → • tcf Nominal [1,1]
- S26 PP → • Preposition NP [1,1]

This will call complete(), nothing will be added though because no states have 'S' in the RHS

Chart [1]

- S16 Verb → book • $[0,1]$
S17 VP → Verb • $[0,1]$
S18 VP → Verb • NP $[0,1]$
S19 VP → Verb • NP PP $[0,1]$
S20 VP → Verb • PP $[0,1]$
S21 S → VP • $[0,1]$
S22 VP → VP • PP $[0,1]$
S23 NP → • Pronoun $[1,1]$
S24 NP → • Proper-noun $[1,1]$
S25 NP → • Det Nominal $[1,1]$
S26 PP → • Preposition NP $[1,1]$

Call predict
on PP

these will all call predict()

Chart [1]

- S16 Verb → book • [0,1]
S17 VP → Verb • [0,1]
S18 VP → Verb • NP [0,1]
S19 VP → Verb • NP PP [0,1]
S20 VP → Verb • PP [0,1]
S21 S → VP • [0,1]
S22 VP → VP • PP [0,1]
S23 NP → • Pronoun [1,1]
S24 NP → • Proper-noun [1,1]
S25 NP → • Det Nominal [1,1]
S26 PP → • Preposition NP [1,1]

Call predict
on PP

these will all call predict()

Chart [1]

- S16 Verb → book • [0,1]
S17 VP → Verb • [0,1]
S18 VP → Verb • NP [0,1]
S19 VP → Verb • NP PP [0,1]
S20 VP → Verb • PP [0,1]
S21 S → VP • [0,1]
S22 VP → VP • PP [0,1]
S23 NP → • Pronoun [1,1]
S24 NP → • Proper-noun [1,1]
S25 NP → • Det Nominal [1,1]
S26 PP → • Preposition NP [1,1]

- S27 Pronoun → • I [1,1]
S28 Proper-noun → • Houston [1,1]
S29 Det → • this [1,1]
S30 Preposition → • from [1,1]

Chart [1]

- S16 Verb → book • $[0,1]$
- S17 VP → Verb • $[0,1]$
- S18 VP → Verb • NP $[0,1]$
- S19 VP → Verb • NP PP $[0,1]$
- S20 VP → Verb • PP $[0,1]$
- S21 S → VP • $[0,1]$
- S22 VP → VP • PP $[0,1]$
- S23 NP → • Pronoun $[1,1]$
- S24 NP → • Proper-noun $[1,1]$
- S25 NP → • Det Nominal $[1,1]$
- S26 PP → • Preposition NP $[1,1]$

- S27 Pronoun → • I $[1,1]$ Skip
- S28 Proper-noun → • Houston $[1,1]$ Skip
- S29 Det → • this $[1,1]$ Scan
- S30 Preposition → • from $[1,1]$ Skip

Done w/ chart [1]

Keep following the same algorithm until you finish chart [3]. Then you can tell whether this grammar recognizes this sentence if you have a state with the form: $S \rightarrow d \bullet [0,3]$

Natural Language Processing

Lecture 9: Lexical Semantics -
Word Senses and Lexical Relations

3/13/2019

COMS W4705
Yassine Benajiba

Natural Language Semantics

- Semantics is concerned with the meaning of language.
 - Lexical Semantics: What is the meaning of individual words?
 - Computational Semantics: How do we compute language meaning from word meaning? (next week)
 - How do we represent meaning?

Two Approaches to Language Meaning

- *Usage-based semantics:*
 - Distributional / Vector-space semantics. Word embeddings.
 - Core concept: Semantic similarity.
 - History in connectionist approaches (neural networks).
- Formal Semantics:
 - Use formal Meaning Representations for words /sentences / discourse.
 - Based on lexical resources.
 - History in cognitive science (mind as symbol manipulation device).

Why Natural Language Semantics?

- Meaning representations bridge between linguistic input and extra-linguistic knowledge.
 - Support inference and reasoning.
 - Examples:
 - Answering questions on an exam.
 - Deciding what to order at a restaurant by reading the menu.
 - Learning how to use a device by reading the manual.
 - Finding a joke funny.
 - Following a recipe.

Syntax and Semantics

- Can we do syntax without semantics?
 - "*He read the book with the blue cover*"
 - "*He saw the hill with the telescope*"
 - "*Last night I shot an elephant in my pajamas*"
- These examples require world knowledge (books have covers, telescopes are used for seeing, elephants are large, ...)

Semantics and Machine Translation

L'Avocat général _{fr} —————→ the general avocado _{en}
??

Word Senses

- Different word senses of the same word can denote different (more or less related) concepts.

bank (of a river) vs. *bank* (financial institution) vs. *bank* (storage facility)

mouse (animal) vs. *mouse* (computer accessory)

bright [light] vs. *bright* [idea] vs. *bright* [student] vs. *bright* [future]

- "Lexeme" - a pairing of a particular word form with its sense.

Homonymy

- Homonymy is a relation between concepts/senses:
 - Multiple **unrelated** concepts correspond to the same word form.
 - Examples:
 - *bank*
 - *check*
 - *kind*
 - *bass*

Polysemy

- Multiple semantically related concepts correspond to the same word form.
- Examples:
 - *wood* (material that trees are made of) vs. *wood* (a forested area)
 - *bank* (financial institution) vs. *bank* (building)

Metonymy

- A subtype of polysemy.
- Systematic and productive.
- One aspect of a concept is used to refer to other aspects of a concept (or the concept itself).
 - BUILDING <-> ORGANIZATION (*bank, school,...*)
 - ANIMAL <-> MEAT (the *chicken* was overcooked, the *chicken* eats a worm)

Zeugma

- when a single word is used with two other parts of a sentence but must be understood differently (word sense) in relation to each.
- *Does United serve breakfast and JFK?*
- *He lost his gloves and his temper.*

Semantic Relations

- Synonym / Antonym
- Hypernym / Hyponym (IS-A)
- Meronym / Holonym (part-of relationship)

Synonyms

- Two lexemes that share a sense.
 - couch/sofa
 - vomit/throw up
 - car/automobile
 - hazelnut/filbert
 - water/ H_2O

Note that even though the sense is the same, there may be differences in politeness, slang, register, genre, etc.

Lexical Substitution

- Two lexemes are synonyms if they can be substituted for each other in a sentence, such that the sentence retains its meaning (truth conditions).
- Note that synonymy is not a relationship between words, but between lexemes.

large
How big is that plane?

large (?)
She was like a big sister to him.

Antonyms

- Senses are opposites with respect to one specific feature of their meaning.
- Otherwise, they are very similar!
 - *dark / light* (level of luminosity)
 - *short / long* (length)
 - *hot / cold* (temperature)
 - *rise / fall* (direction)
 - *front / back* (relative position)

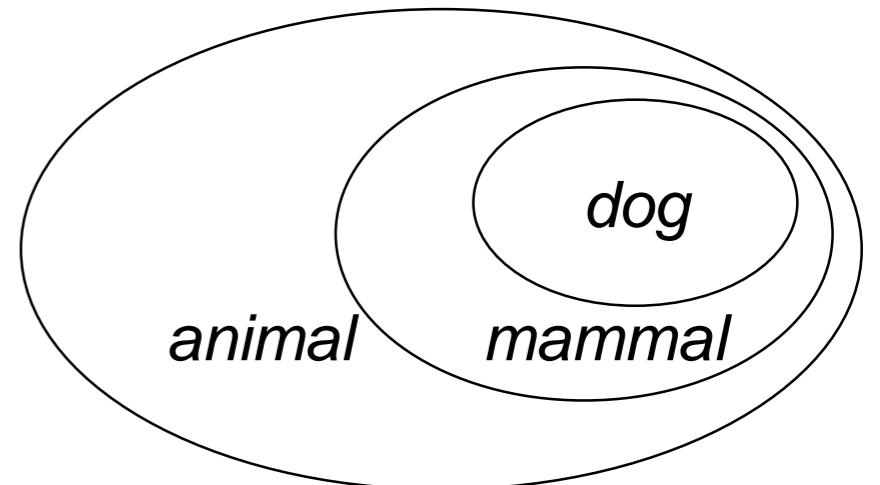
Antonyms typically describe opposite ends of a scale, or opposite direction/position with respect to some landmark.

Hyponymy

- One sense is a hyponym (or subordinate) of another sense if the first sense is more specific, denoting a subclass of the other. (IS-A relationship).
 - *dog* is a hyponym of *mammal*.
 - *mammal* is a hyponym of *animal*.
 - *desk* is a hyponym of *furniture*.
 - *sprint* is a hyponym of *run*.
- The inverse relation is called **hyponymy**, so *furniture* is a hypernym (or superordinate) of *desk*.

Hyponymy

- Can think of Hyponymy as a set relationship.
 - *for all x, if x is a mammal, then x is an animal.*
- Related to **entailment**:



The woman ate spaghetti with cheese.



The person consumed pasta with dairy product.

Meronymy

- Part-whole relationship.
- A meronym is a part of another concept.
 - *leg* is a meronym of *chair*.
 - *wheel* is a meronym of *car*.
 - *cellulose* is a meronym of *paper*. (substance meronymy)
- The inverse relation is **holonymy**.
Car is a holonym of *wheel*.

WordNet

- WordNet is a lexical database containing English word senses and their relations.
- Represents word sense as **synsets**, sets of lemmas that have synonymous lexemes in one context.

{*composition, paper, report, theme*}

{*newspaper, paper*}

{*paper*}

- Version 3.1 Contains synonyms, antonyms, hyponyms, (some) meronyms, and frequency information for about 117.000 nouns, 11.500 verbs, 22.000 adjectives, and 4.500 adverbs.

WordNet Senses

- each sense comes with a *gloss* (dictionary definition).

```
$ wn mouse -over
```

Overview of noun mouse

The noun mouse has 4 senses (first 1 from tagged texts)

1. (14) mouse -- (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
2. shiner, black eye, mouse -- (a swollen bruise caused by a blow to the eye)
3. mouse -- (person who is quiet or timid)
4. mouse, computer mouse -- (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad; "a mouse takes much more room than a trackball")

WordNet Hypernyms

```
$ wn mouse -hyphen
```

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun mouse

4 senses of mouse

Sense 1

mouse

=> rodent, gnawer

=> placental, placental mammal, eutherian, eutherian mammal

=> mammal, mammalian

=> vertebrate, craniate

=> chordate

=> animal, animate being, beast, brute, creature, fauna

=> organism, being

=> living thing, animate thing

=> whole, unit

=> object, physical object

=> physical entity

=> entity

WordNet Meronyms

```
$ wn mouse -meron
```

Meronyms of noun mouse

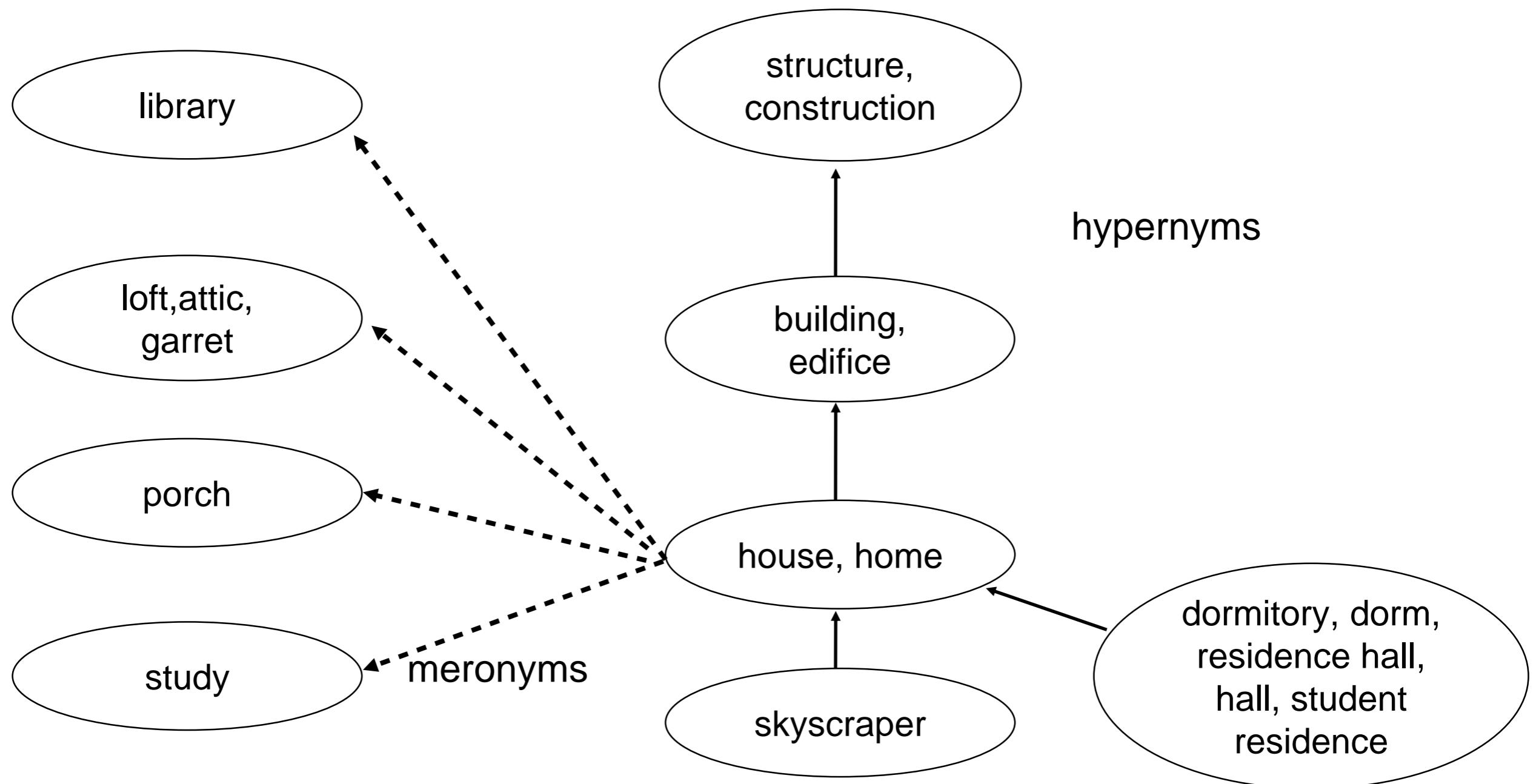
1 of 4 senses of mouse

Sense 4

mouse, computer mouse

 HAS PART: mouse button

Semantic Relations as a Graph



Word Sense Disambiguation (WSD)

- Given a word token in context, identify its correct word sense (from a list of possible word-senses).
- Why? Machine translation. Question answering. Information Retrieval, Speech Synthesis...
- What set of senses:
 - For MT: Possible translations of a word (e.g. English to Spanish)
 - For Speech Synthesis: Homographs (*bass*, *bread*,...)
 - Senses from a dictionary like WordNet

Two types of WSD tasks

- Lexical Sample task
 - Small pre-selected set of target words.
 - Inventory of senses for each word.

*Largemouth **bass** are willing to bite a variety of baits.*
- All-words task
 - Every word in an entire text.
 - A lexicon with senses for each word.
 - Similar to POS tagging (but specific set of tags for each word).

WSD Methods

- Supervised Machine Learning
 - Train a classifier for each word.
 - Requires hand-labeled training data (sense annotations).
- Dictionary Methods
 - No training data. Instead use a sense dictionary (like WordNet).
 - Or exploit sense relations (WordNet graph).
- Semi-supervised learning
 - Use a small hand-annotated data set and generalize ("bootstrapping").

Supervised WSD

- Given:
 - A sense inventory (for example, WordNet senses).
 - An annotated training corpus.
 - A set of features extracted from the training corpus.
 - A classifier.
- How would you solve this?

Annotated Training Corpus

- Lexical sample task:
 - SENSEVAL 1 to 3 (1998,2001). Small number of target words (~50), 12.000 to 20.000 instances.
SemEval 2007.
 - "line, hard, serve" dataset. 4.000 instances each for the words line, hard, and serve.
- All-words task:
 - SENSEVAL 3, SemEval 2007.
 - SemCor ("Semantic concordance"). 234,000 words from Brown Corpus, manually tagged with WordNet senses.

Feature Definition - Intuition

"If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words...

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word...

The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?''

Feature Definition

- Can you identify the word sense?

bass

Feature Definition

- Approaches for representing context:
 - Collocational features:
 - Take the position of each context word into account.
 - Bag-of-word features:
 - Simply keep the set of all context words.
 - Using word embeddings (or phrase embeddings)

Bag-of-Words features

*He learned how to play **guitar** and **bass** watching **Youtube** videos.*



guitar,
watching,
Youtube

fish	0
:	:
guitar	1
:	:
watching	1
:	:
Youtube	1

Collocational Features

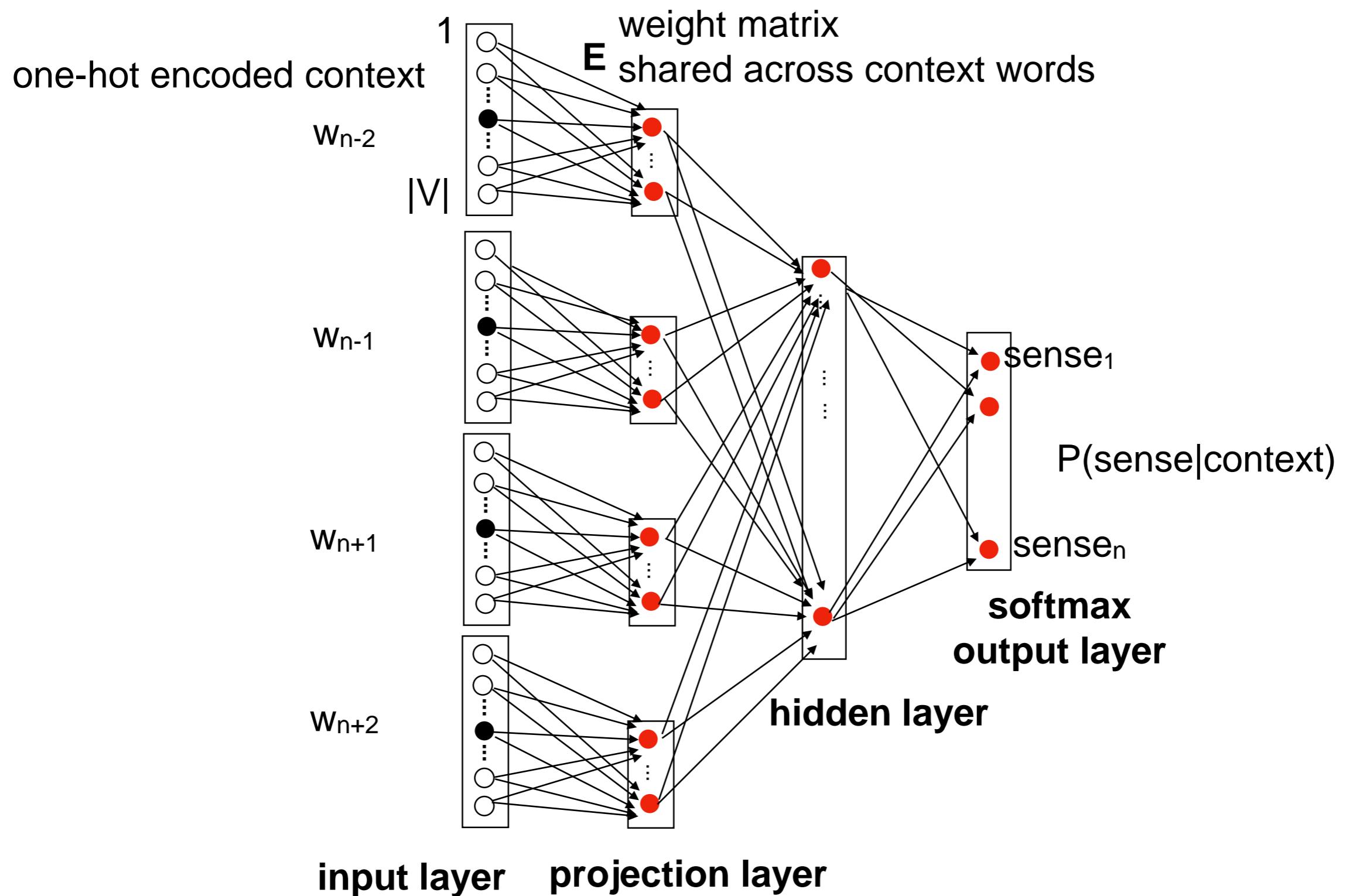
*He learned how to play **guitar** and **bass** watching Youtube videos.*

- Assume feature window: +/- 2 words.
- Position specific information about the words in the window

[guitar, NN, and, CC, watching, VBG, Youtube, NNP]

[word_{n-2}, POS_{n-2}, word_{n-1}, POS_{n-1}, word_{n+1}, POS_{n+1}]

Feed-forward NN for WSD



Dictionary-Based Methods

- Supervised WSD requires a lot of annotated training data.
- Instead, we can use a separate dictionary, such as WordNet:
 - to obtain candidate senses.
 - to obtain information that allows us to identify which of the candidate senses is correct.

1. (25) **bank**

sloping land (especially the slope beside a body of water)

"they pulled the canoe up on the bank"

"he sat on the bank of the river and watched the currents"

2. (20) **depository financial institution, bank, banking company**

a financial institution that accepts deposits and channels the moneyin:

"he cashed a check at the bank"

"that bank holds the mortgage on my home"

Simplified Lesk Algorithm

Lesk 1986

- Use dictionary glosses for each sense.
- Choose the sense that has the highest word overlap between gloss and context (ignore function words).

*The **bank** can guarantee **deposits** will eventually cover future tuition costs because it invests in adjustable-rate **mortgage** securities.*

1. (25) **bank**

sloping land (especially the slope beside a body of water)

"they pulled the canoe up on the bank"

"he sat on the bank of the river and watched the currents"

2. (20) **depository financial institution, bank, banking company**

a financial institution that accepts **deposits** and channels the money i:

"he cashed a check at the bank"

"that bank holds the **mortgage** on my home"

Extensions to Lesk Algorithm

- Often the available definitions and examples do not provide enough information. Overlap is 0.
- Different approaches to extending definitions:
 - "Corpus-Lesk": Use a sense-tagged-example corpus, add context from example sentences.
 - Extended Gloss Overlap (Banerjee & Pedersen 2003): Add glosses from related words (hyponyms, meronyms, ...)
- Use embedded representations for each word in the definition. Choose the sense with highest average vector similarity to the context.

WSD evaluation

- Ideally, we would like to use an **extrinsic** (task-based) evaluation.
 - Use WSD as a component of some task (for example, MT) and see if the results improve.
- For convenience we often use **intrinsic** evaluation.
 - Measure sense accuracy (% of correct sense tags).
- What are some good baselines for this task?
 - Most frequent sense (according to some tagged corpus).
 - Lesk often used as a baseline for more elaborate approaches.

WSD Performance

- Varies widely depending on how difficult the disambiguation task is.
- Accuracies of over 90% are commonly reported on some of the classic WSD tasks.
- Senseval brought careful evaluation of difficult WSD (many senses, co-occurrence)
- Senseval 1: more fine grained senses, wider range of types:
 - Overall: about 75% accuracy
 - Nouns: about 80% accuracy
 - Verbs: about 70% accuracy

Upper Bound

- How well do people do on this task?
- Inter-annotator agreement:
 - Compare multiple human annotations on the same data, given the same annotations guidelines.
- Human agreement on all-words task with WordNet senses:
 - 75%-80%

Semi-Supervised WSD

- What if we only have a few labeled training examples.
- Idea: Bootstrapping. Generalize from a small hand-labeled dataset.
Yarowsky, 1995
- "One sense per collocation" rule
 - A word reoccurring in collocation with the same word will almost surely have the same sense.
 - Example:
 - the word *play* occurs with the music sense of **bass**
 - the word *fish* occurs with the fish sense of **bass**

Extracting New Sentences

- Using a few keywords and the "One sense per collocation" rule we can label new sentences for the known senses.

We need more good teachers – right now, there are only a half a dozen who can play the free **bass** with ease.

An electric guitar and **bass** player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps

The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper.

And it all started when **fishermen** decided the striped **bass** in Lake Mead were too skinny.

A Lexical Substitution Task

- Instead of identifying word sense
 - find a substitute for the target word, such that the meaning is preserved.

*He was **bright** and independent and proud .*

*Snow covered areas appear **bright** blue in the image which was taken in early spring.*

*... an institution that nurtures the best and **brightest** young musicians.*

- How would you solve this task?

Acknowledgments

- Some slides by Kathy McKeown, Bob Coyne, Dan Jurafsky.

Natural Language Processing

Lecture 10: Semantic Role Labeling.

3/13/2019

COMS W4705
Yassine Benajiba

Word Meaning and Sentence Meaning

- So far we have discussed the meaning of individual words.
- Now: meaning of entire predicate-argument structures and sentences.
- What should the representations be?
- How do we compute predicate or sentence-level representations from word representations?
 - What is the role of syntax?

Approaches to Sentence Level Semantics

- Semantic Role Labeling (SRL) / Frame Semantic Parsing.
 - Target representation: PropBank predicate argument structures, FrameNet-style annotations.
- Full-sentence semantics (next week)
 - Target representations: Predicate-logic, Abstract Meaning Representation

Frame Semantics

(Fillmore, 1992)

- Long history in cognitive science, AI, ... (Minsky 1974, Barsalou 1992)
- A frame represents a situation, object, event providing background needed to understand a word ('cognitive schemata').
- Different words (of different part-of-speech) can evoke the same frame

Giving → {*donate.v*, *gift.n*, *give.v*, *hand over.v*, *treat.v*, ... }

- A pair of a word and a frame is called a lexical unit (LU).

Frame Elements

- Frames describe the interaction/relation between a set of frame-specific semantic roles called *Frame Elements* (FEs).

Giving: A Donor transfers a Theme from a Donor to a Recipient.

Core:

Donor

The person that begins in possession of the Theme and causes it to be in the possession of the Recipient
The entity that ends up in possession of the Theme.

Recipient

Theme

The object that changes ownership.

Non-core:

The Means by which the Donor gives the Theme to the Recipient.

The Purpose for which the Donor gives the Theme to the Recipient.

:

FrameNet

(Baker et al, 1998)

- Lexical resource based on Frame Semantics: 13640 lexical units in 1087 frames.
- Example **annotations** illustrate how frame elements are realized linguistically.
 - Frames evoked by frame evoking elements (FEE).
 - Central interest: mapping from Grammatical Function (Subj, Obj, ...) to Frame Elements.

POS	Apple NNP	wanted to VVD	TO	donate VB	a computer DT NN	to every school in the country PRP DT NN IN DT NN .
FE	Donor			FEE	Theme	Recipient
GF	Subj				Obj NP	Dep-to PPto
PT	NP					

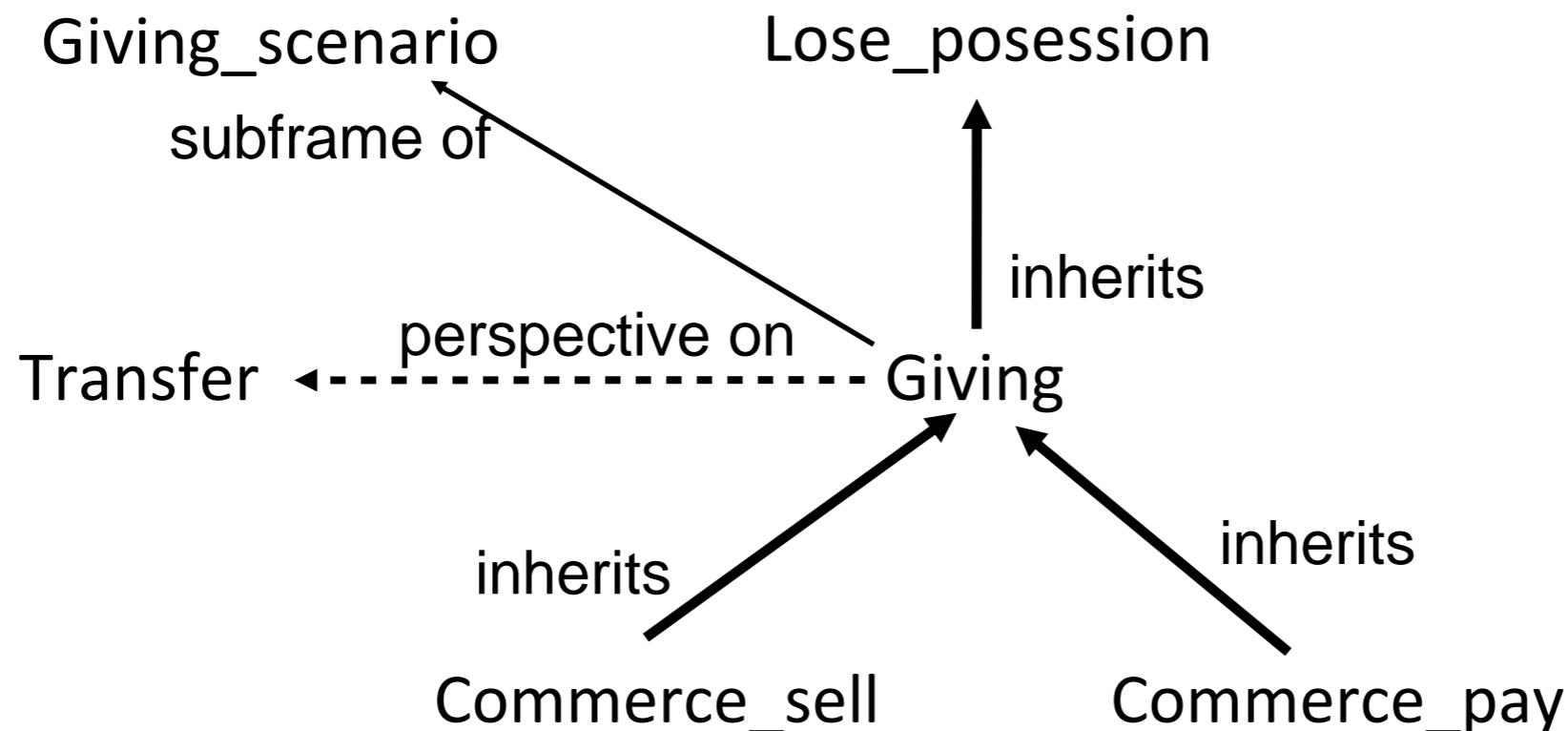
Valence Pattern

- Valence patterns (derived from annotated sentences) specify different ways grammatical roles (subject, object, ...) can be mapped to frame elements for a given lexical unit.

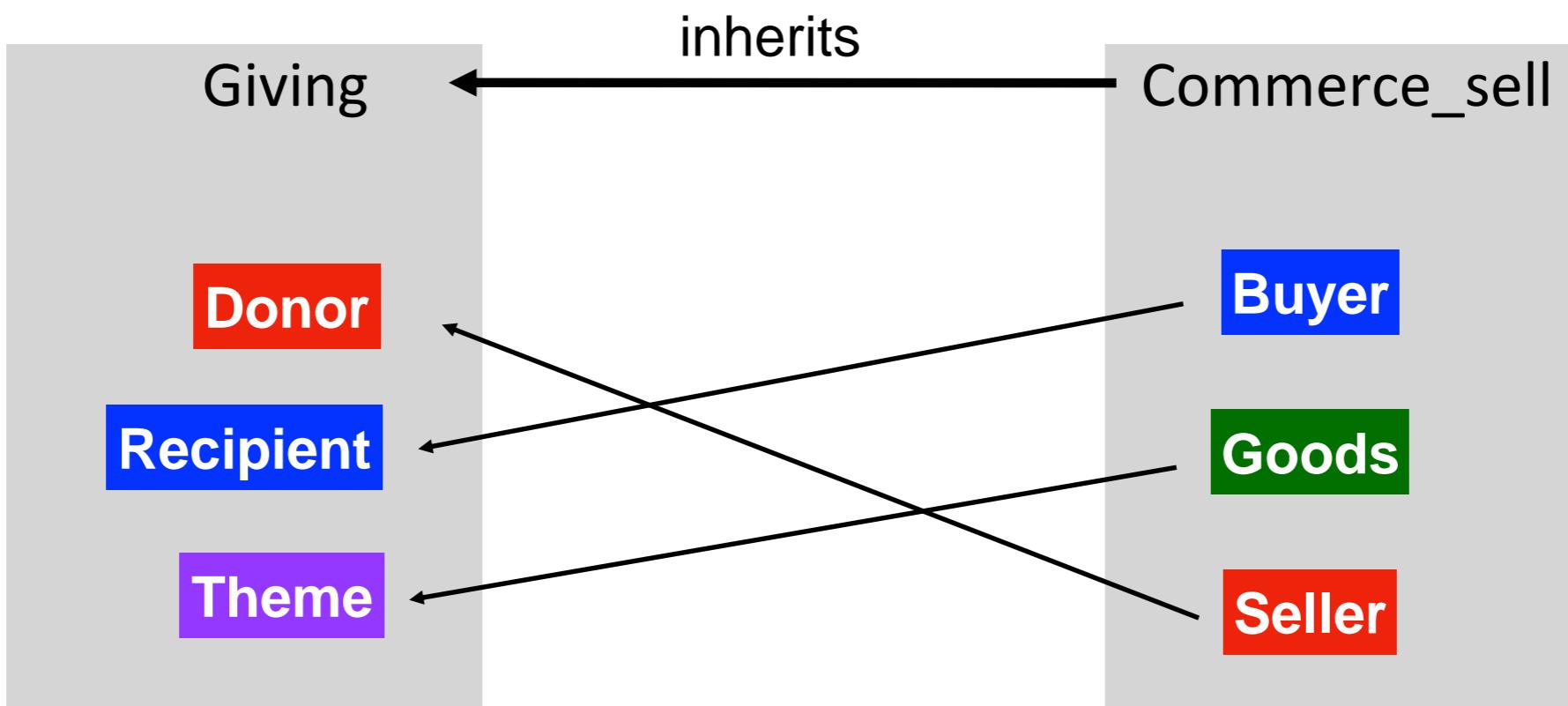
Valence pattern	Example sentence
(subj/ DONOR) V (obj/ RECIPIENT) (obj2/ THEME)	<i>John gave Mary the book</i>
(subj/ DONOR) V (obj/ THEME) (dep-to/ RECIPIENT)	<i>John gave the book to Mary</i>
(subj/ DONOR) V (dep-of/ THEME) (dep-to/ RECIPIENT)	<i>John gave of his time to people like M.</i>
(subj/ DONOR) V (dep-to/ RECIPIENT)	<i>John gave to charity</i>

Frame-to-Frame Relations

- Frames are related via frame-to-frame relations.



Frame-Element Relations



PropBank

(Baker et al, 2005)

- Another corpus annotated with semantic roles, based on English Penn Treebank & OntoNotes 5.0. (~2m Words)
- Also available: Chinese, Hindi/Urdu, Arabic.
- Full-text annotation (only verbs).
- Numbered arguments (semantic roles).
 - Interpretation is specific to each verb.

Frameset for *donate*.01

Arg0: *giver*

Arg1: *thing given*

Arg2: *entity given to*

the company	donate d rel	over \$35,000	to residents
Arg0	Arg1	Arg2	

Proto Roles

(Dowty 1991)

- Proto-Agent
 - Volitional involvement in event or state.
 - Sentience (and/or perception)
 - Causes an event or change of state in another participant
 - Movement (relative to position of another participant)
- Proto-Patient
 - Undergoes change of state
 - Causally affected by another participant
 - Stationary relative to movement of another participant

PropBank Roles

- Each frameset has numbered argument: Arg0, Arg1, Arg2, ...
 - Arg0:PROTO-AGENT
 - Arg1:PROTO-PATIENT
 - Arg2: usually: benefactive, instrument, attribute, or end state
 - Arg3: usually: start point, benefactive, instrument, or attribute
 - Arg4 the end point (Arg2-Arg5 are not really that consistent, causes a problem for labeling)

PropBank FrameSets

- Different framesets correspond to different senses.

Frameset for tend.01, care for

Arg0: tender

Arg1: thing tended (to)

John Arg0	tends rel	to the needs of his patrons Arg1
--------------	--------------	-------------------------------------

Frameset for tend.02, have a tendency

Arg0: theme

Arg2: attribute

The cost, or premium Arg0	tends rel	to get fat in times of crisis Arg2
------------------------------	--------------	---------------------------------------

Another Example

Frameset for increase.01, go up incrementally

Arg0: causer of increase

Arg1: thing increasing

Arg2: amount increased by

Arg3: start point

Arg4: end point

[Arg₀ Big Fruit Co.] **increased** [Arg₁ the price of bananas]

[Arg₁ The price of bananas] was **increased** again [Arg₀ by Big Fruit Co.]

[Arg₁ The price of bananas] **increased** [Arg₂ 5%]

Observations:

Syntax and semantics do not map 1:1. Generalize away from syntactic variations.

Semantic Role Labeling (SRL)

- Input: raw sentence.
- Goal: automatically produce PropBank or FrameNet-style annotations ("frame-semantic parsing").
- Applications:
 - Question Answering (Shen and Lapata 2007, Surdeanu et al. 2011)
 - Machine Translation (Liu and Gildea 2010, Lo et al. 2013)
 - Stock prediction, spoken dialog segmentation, ...
- How would you approach this problem?

Generic SRL Algorithm

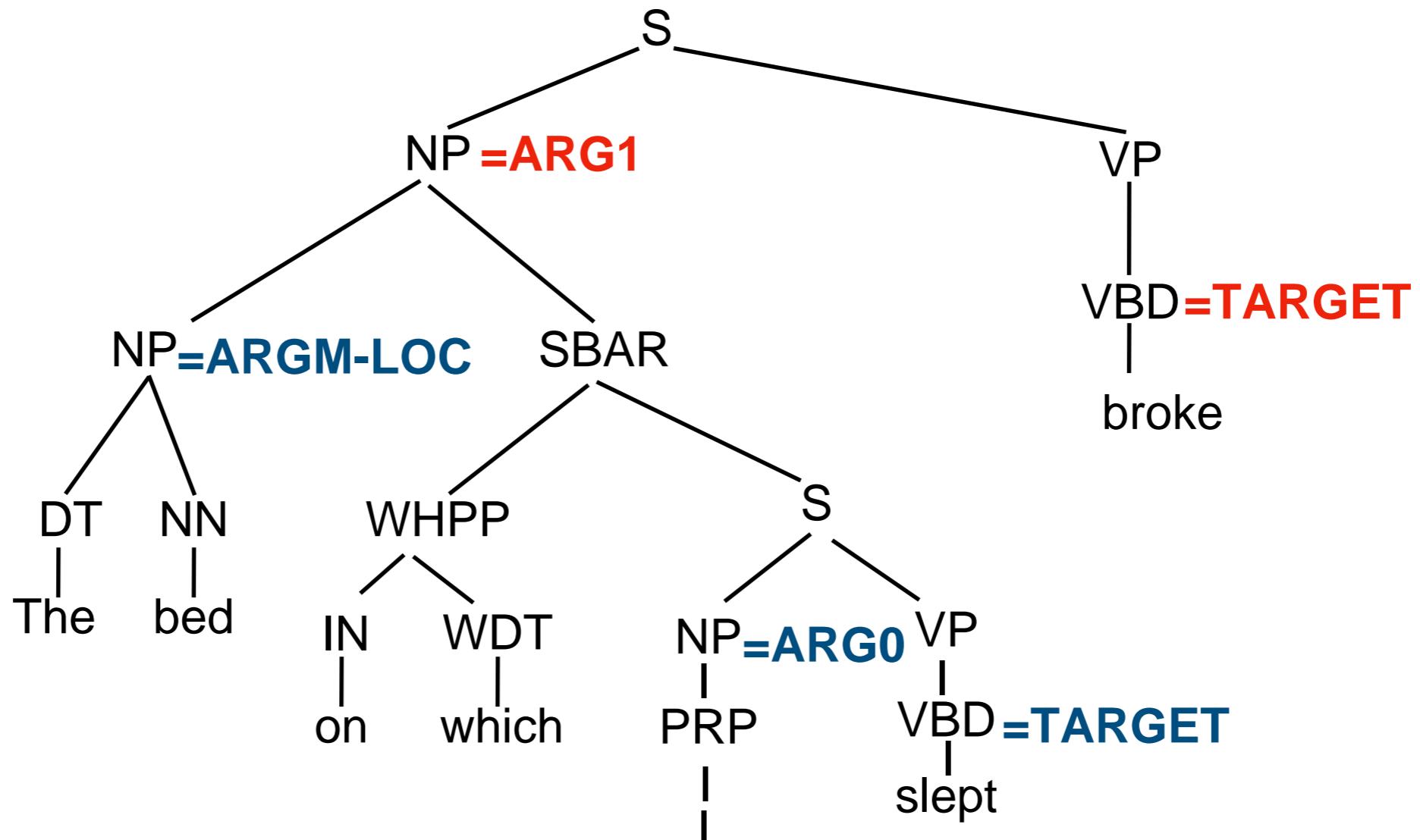
Algorithm outline:

- Parse the sentence (dependence or constituency parse)
- Detect all potential targets (predicates / frame evoking elements)
- For each predicate:
 - For each node in the parse tree use supervised ML classifiers to:
 1. identify if it is an argument.
 2. label the argument with a role.

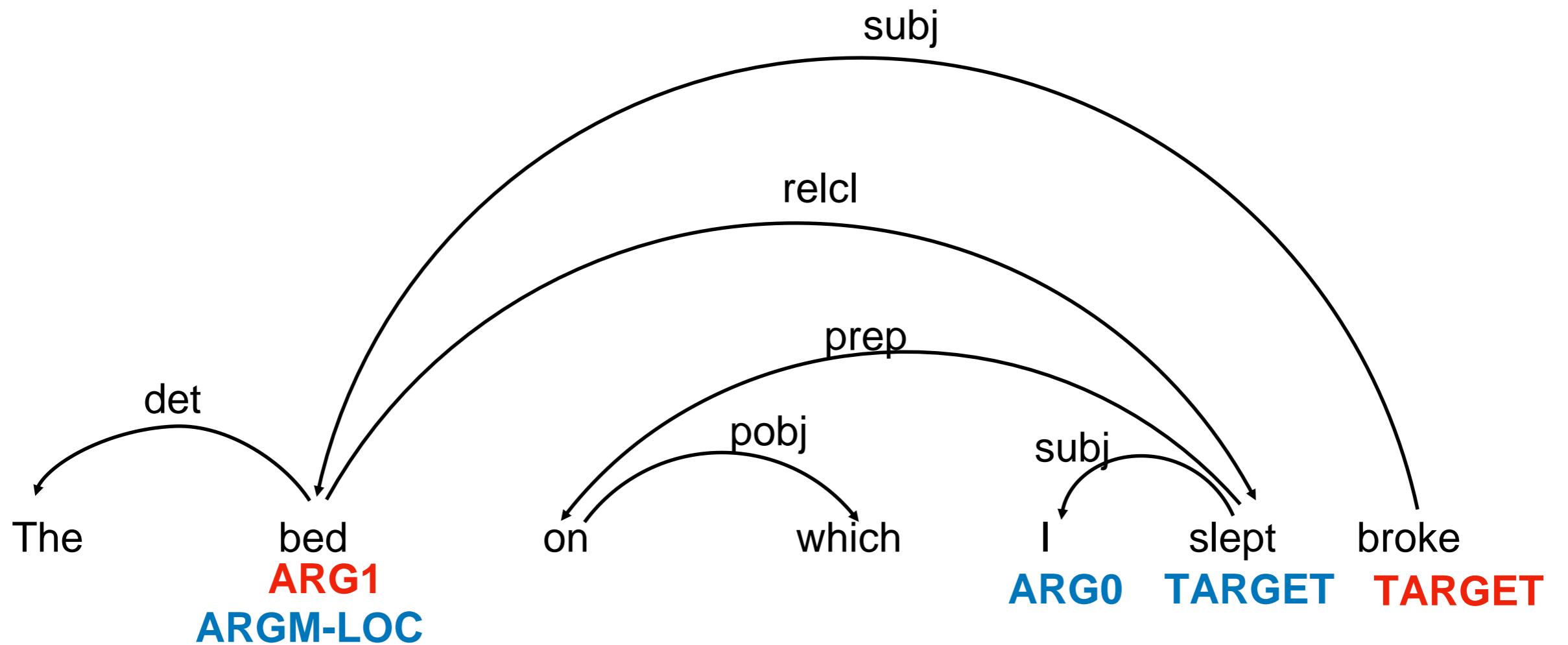
Choosing Targets

- For PropBank:
 - Choose all verbs.
 - Possibly remove light verbs.
- For FrameNet:
 - Choose all lexical items (verbs, nouns, adjectives) that are in the annotated FrameNet training data.

SRL Example



SRL Example



Selectional Restrictions and Preferences

- Different semantic roles might have restrictions on the semantic type of arguments they can take.

I want to eat someplace nearby

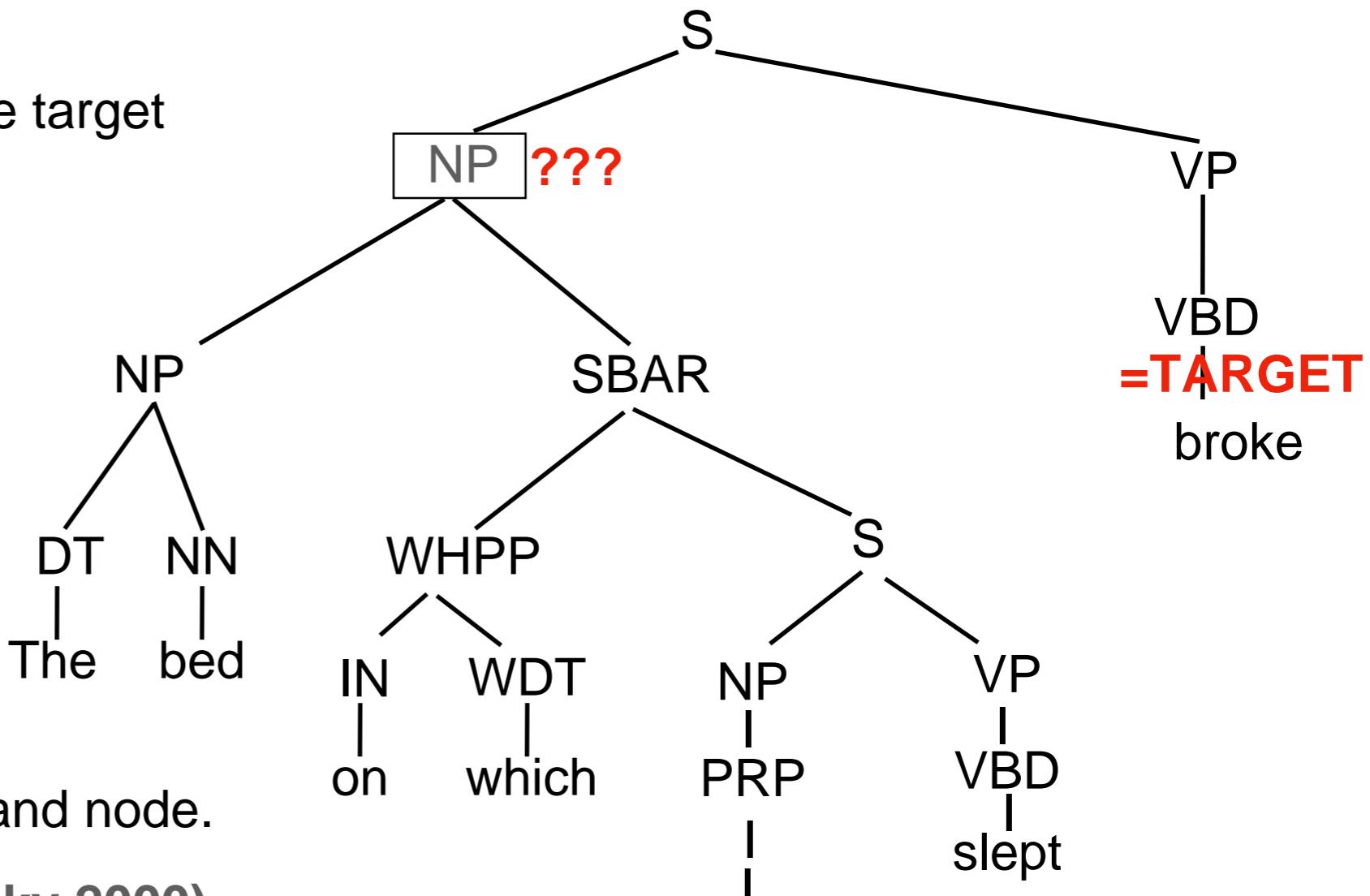
I want to eat Korean for lunch

- Food FE (or ARG1) needs to be *edible*.
- But what about:
...people realized you can't eat gold for lunch if you're hungry
- How could you model these?

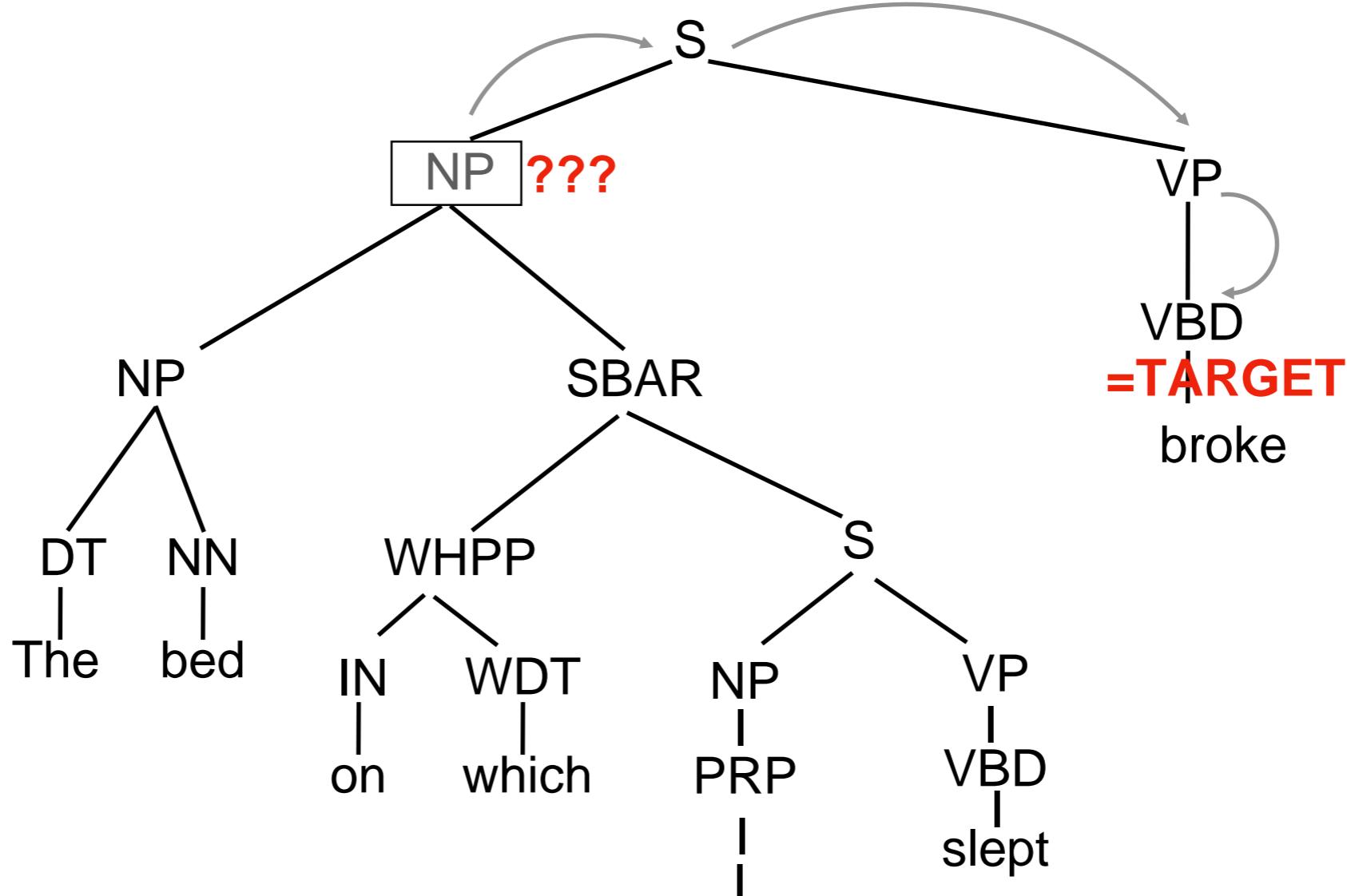
Features

What features should we use for argument detection and labeling?

- target predicate: broke
- headword (+POS): *bed NN*
- phrase type: NP
- linear position: before or after the target
- argument structure of the verb.
"NP broke"
- target voice: active
- possibly semantic features
(named entity class,
WordNet synsets of head word,
...)
- first and last word of constituent and their POS.
- Parse tree path between target and node.

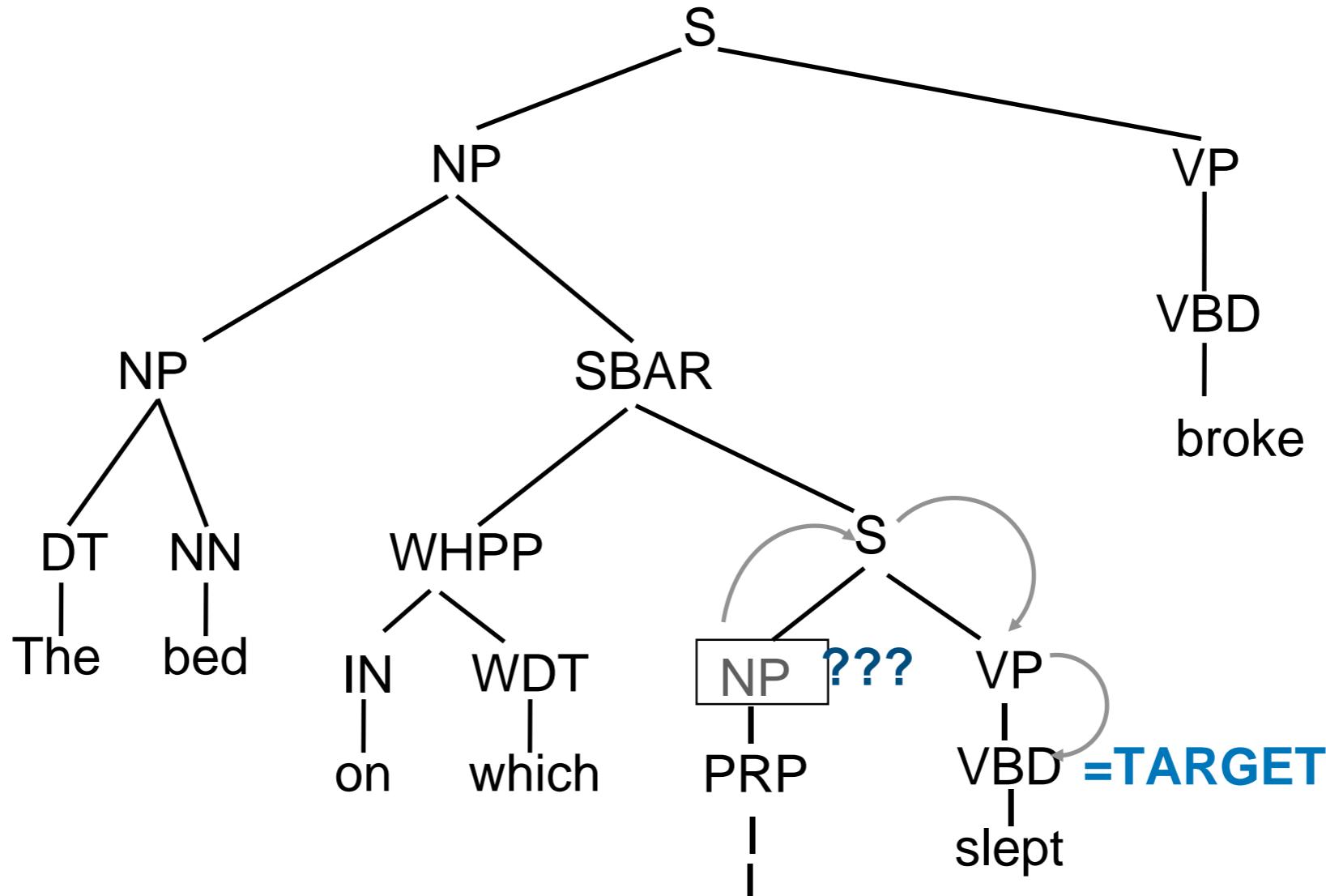


Parse Tree Path



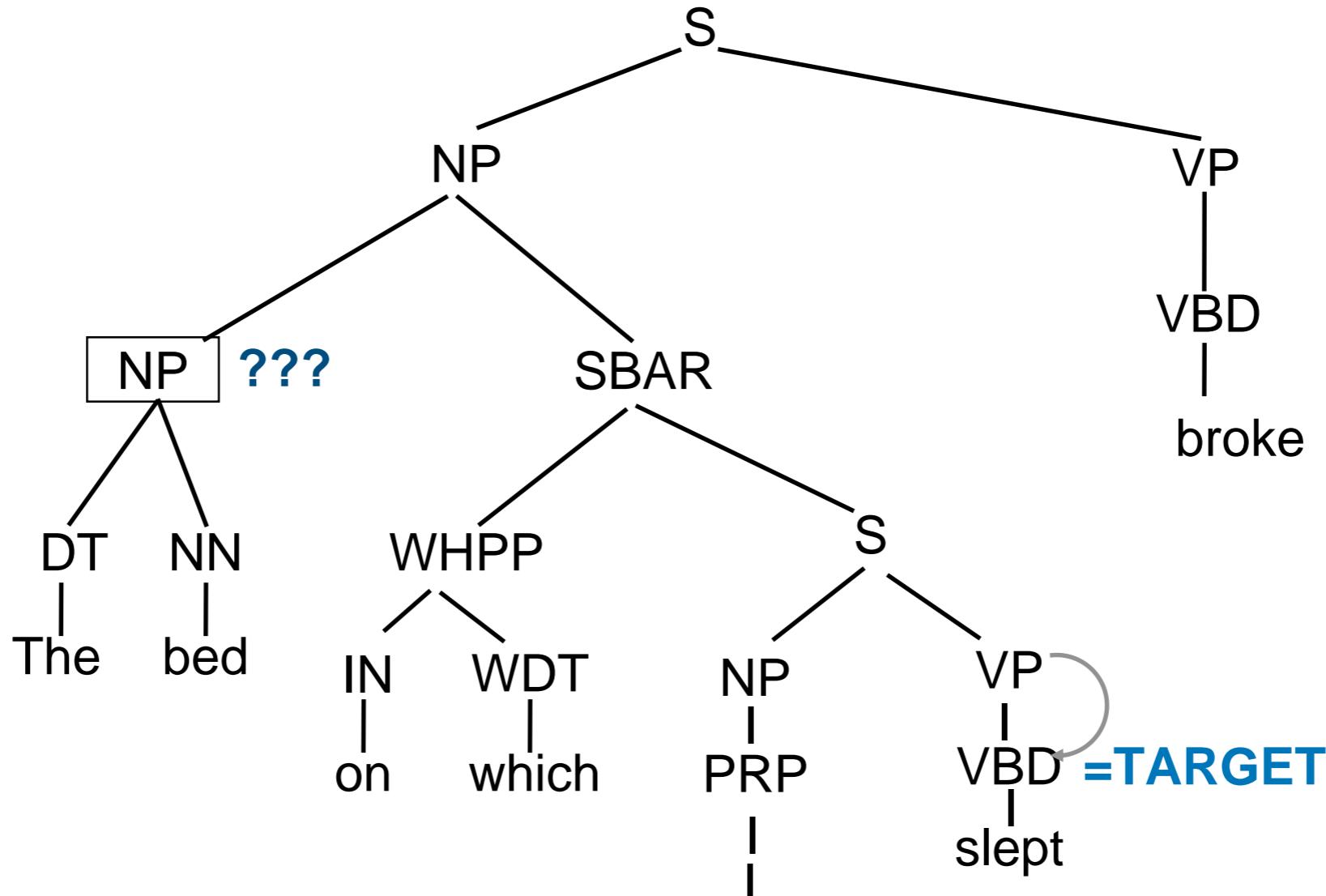
NP↑S↓VP↓VBD

Parse Tree Path



NP↑S↓VP↓VBD

Parse Tree Path



NP↑NP↓SBAR↓S↓VP↓VBD

Frequent Path Features

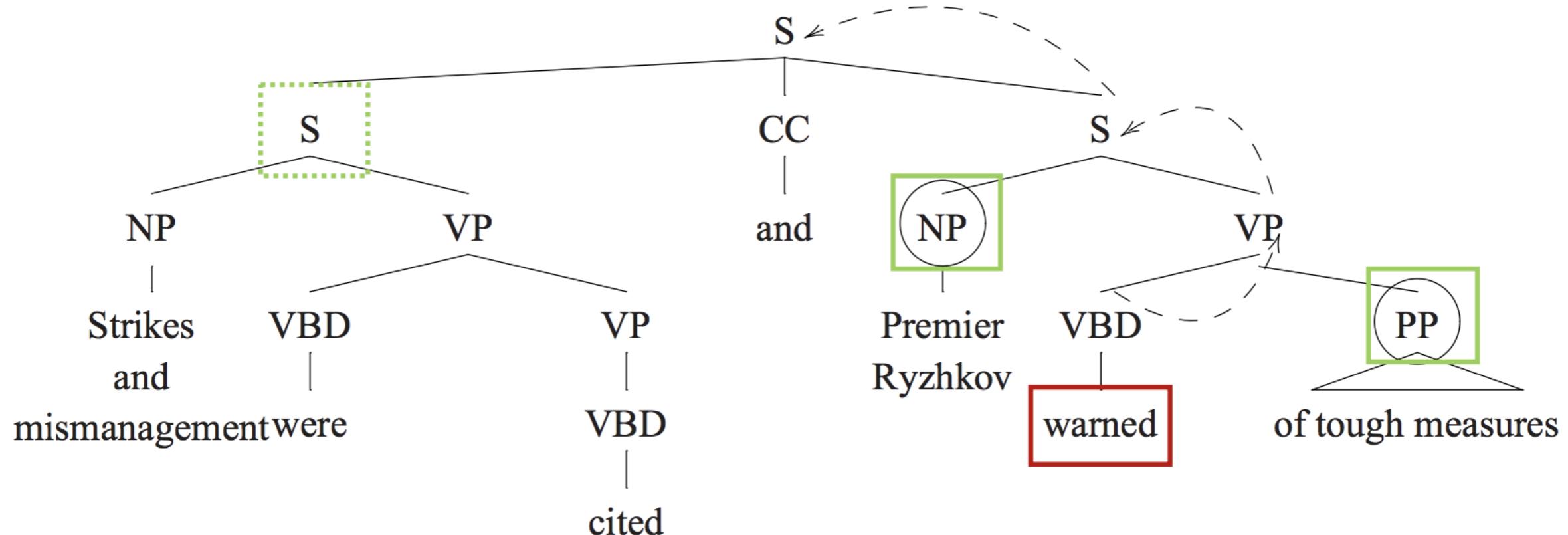
Frequency	Path	Description
14.2%	VB↑VP↓PP	PP argument/adjunct
11.8	VB↑VP↑S↓NP	subject
10.1	VB↑VP↓NP	object
7.9	VB↑VP↑VP↑S↓NP	subject (embedded VP)
4.1	VB↑VP↓ADVP	adverbial adjunct
3.0	NN↑NP↑NP↓PP	prepositional complement of noun
1.7	VB↑VP↓PRT	adverbial particle
1.6	VB↑VP↑VP↑VP↑S↓NP	subject (embedded VP)
14.2		no matching parse constituent
31.4	Other	

(from Palmer, Gildea, Xiu, 2010, SRL book)

Candidate Pruning

(Xue and Palmer 2004)

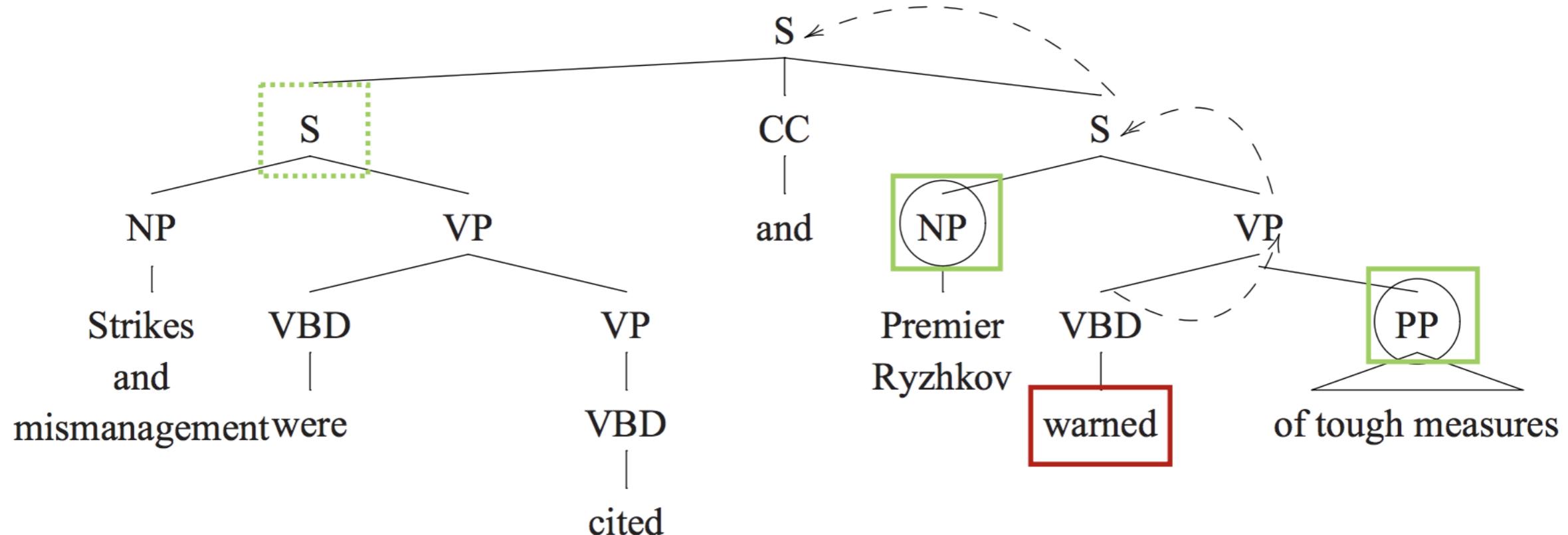
- Algorithm looks at one target at a time. Very few phrases can possibly be arguments.
- Difficult for classifiers to learn: Few positive samples (phrases that are arguments), few negative samples.
- Syntax should tell us *something* about possible arguments.



Pruning Heuristic

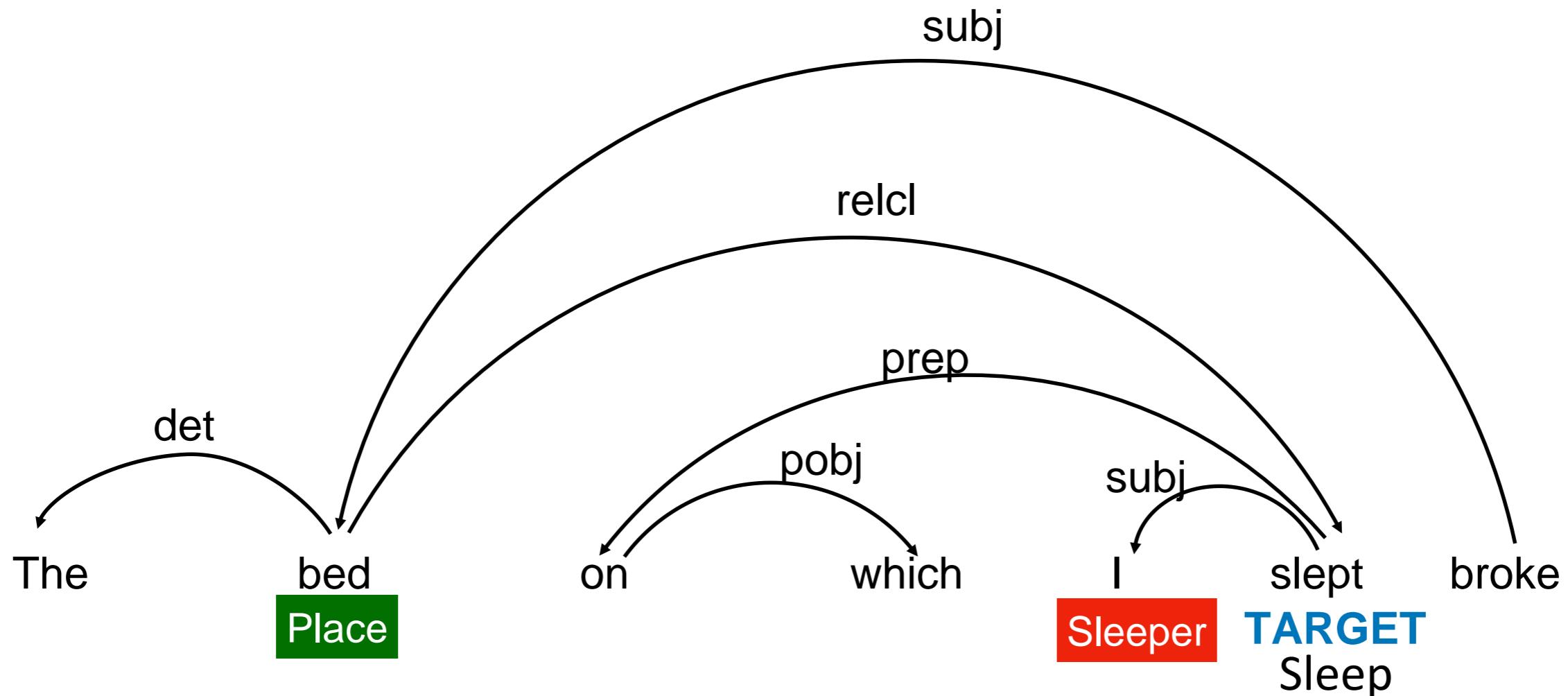
(Xue and Palmer 2004)

- Add sisters of the predicate, then aunts, then great-aunts, etc.
 - Ignore nodes in the subtrees of the selected nodes.
 - Ignore anything in coordinated structures.



FrameNet Parsing

- Slightly more complex: Need to decide on the frame first, then use frame-specific classifiers for the semantic roles.



Frame Semantic Parsing Systems

FrameNet I		SemEval 2007; FrameNet 1.3	
Gildea & Jurafsky 2002	Thompson et al. 2003	Johansson & Nugues 2007	"SEMAFOR" Das et al. 2010, 2012
Argument Classification	$P(\text{fe} \mid \text{features})$	generative prob. model	SVM
Argument Identification	$P(\text{arg} \mid \text{features})$		heuristics+ SVM
Frame Selection	✗	SVM	log-linear
Target Identification	✗		heuristics
Input Syntactic representation	Constituency	Dependency	

More recent work uses Neural Networks (e.g. Swayamdipta et al. 2017)

Features used in FrameNet Parsing

	G&J	J&N	SEMAFOR
Syntactic Representation	PS Collins	DepMST	DepMST
Target Dependency Labels and Words		✓	✓
Target parent word / POS		✓	✓
Target word/ POS	✓	✓	✓
Voice (for verb targets)	✓	✓	✓
Relative Position (before/after/on)	✓	✓	✓

Global Inference

- So far, classifier just decided on one argument at a time.
 - But there are interactions between arguments!
 - FEs may not overlap.
 - Labeling one constituent as ARG0 should increase the probability of another constituent to be ARG1.
 - Some argument combinations are impossible.
- Solutions: Beam Search (Das et al. 2010/2014), Dual Decomposition (Des et al. 2010/2014), DP algorithm (Täckström et al. 2015)

Acknowledgments

- Some slides by Martha Palmer, Shumin Wu, Dan Jurafsky, Nathan Schneider.

Natural Language Processing

Lecture 11: Semantic Parsing -
Abstract Meaning Representation (AMR)

3/27/2019

COMS W4705
Yassine Benajiba

Logical Forms

- Logical form satisfies many goals for meaning representations (unambiguous, canonical form, supports inference, expressiveness)
- But difficult to annotate on a large scale.

We skipped this, so let's briefly talk about it

Abstract Meaning Representation (AMR)

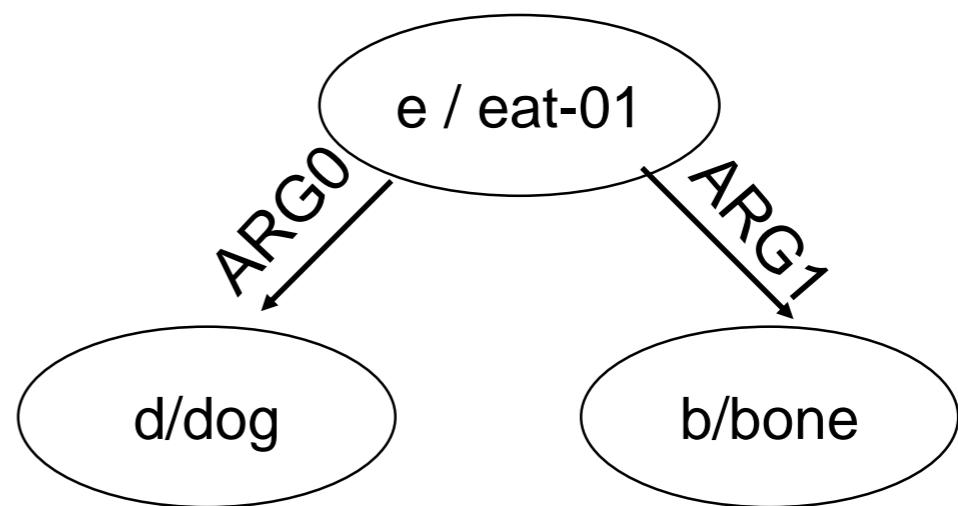
(Banarescu et al., 2013)

- Uses a single, simple data structure (feature structures / directed graphs) to represent many aspects of meaning.
- Focus on "who does what to whom" but leave out details (tense, quantifiers, etc.)
- This level of abstraction facilitates **consistent, large-scale human annotation**.
Goal: build a giant "semantics bank" (comparable to treebanks for syntax).

AMR Example

The dog is eating a bone.

```
(e / eat-01  
  :ARG0 (d / dog)  
  :ARG1 (b / bone) )
```

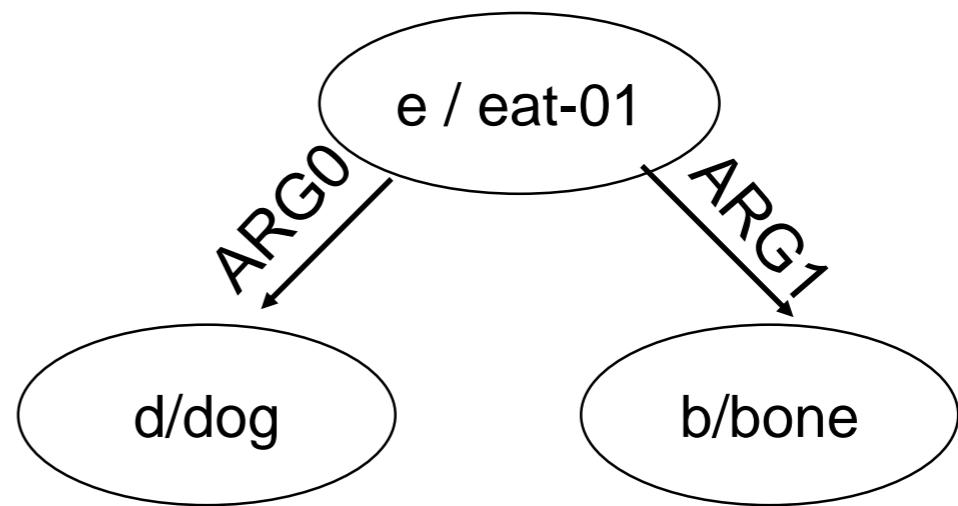


- Edges are labeled with **relations (including semantic roles)**
- Each node has a **variable**.
- Nodes are labeled with **concepts**.
- PropBank framesets used wherever possible.

AMR Example

The dog is eating a bone.

```
(e / eat-01  
  :ARG0 (d / dog)  
  :ARG1 (b / bone) )
```

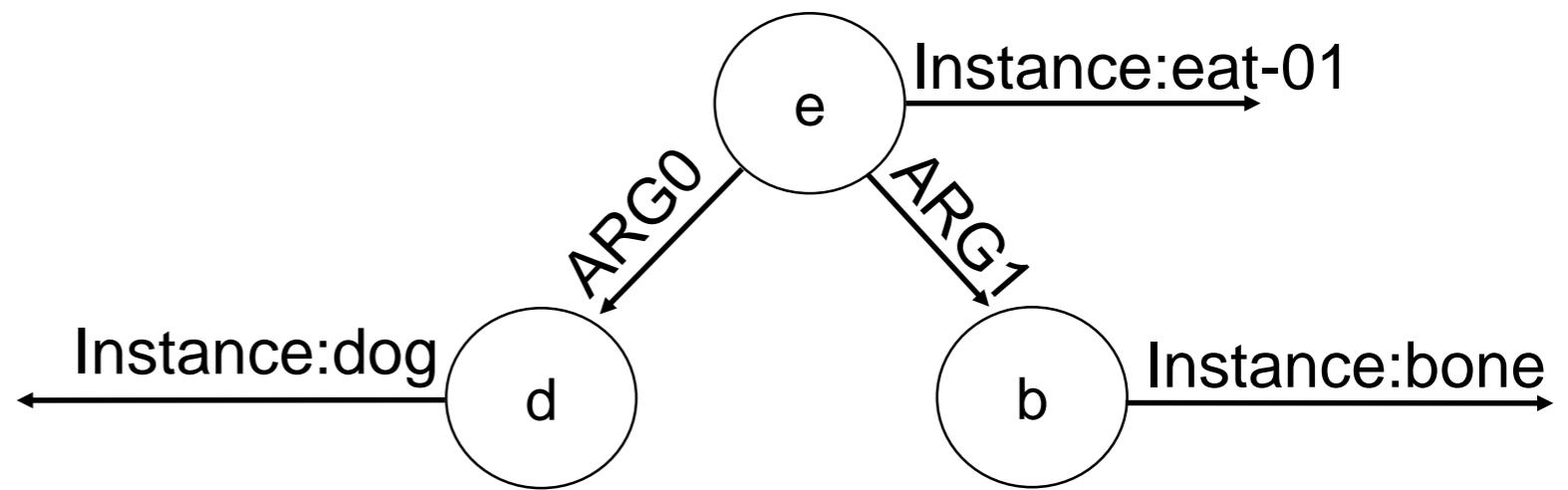


- Edges are labeled with **relations (including semantic roles)**
- Each node has a **variable**.
- Nodes are labeled with **concepts**.
- PropBank framesets used wherever possible.

AMR Example

The dog is eating a bone.

```
(e / eat-01  
  :ARG0 (d / dog)  
  :ARG1 (b / bone) )
```

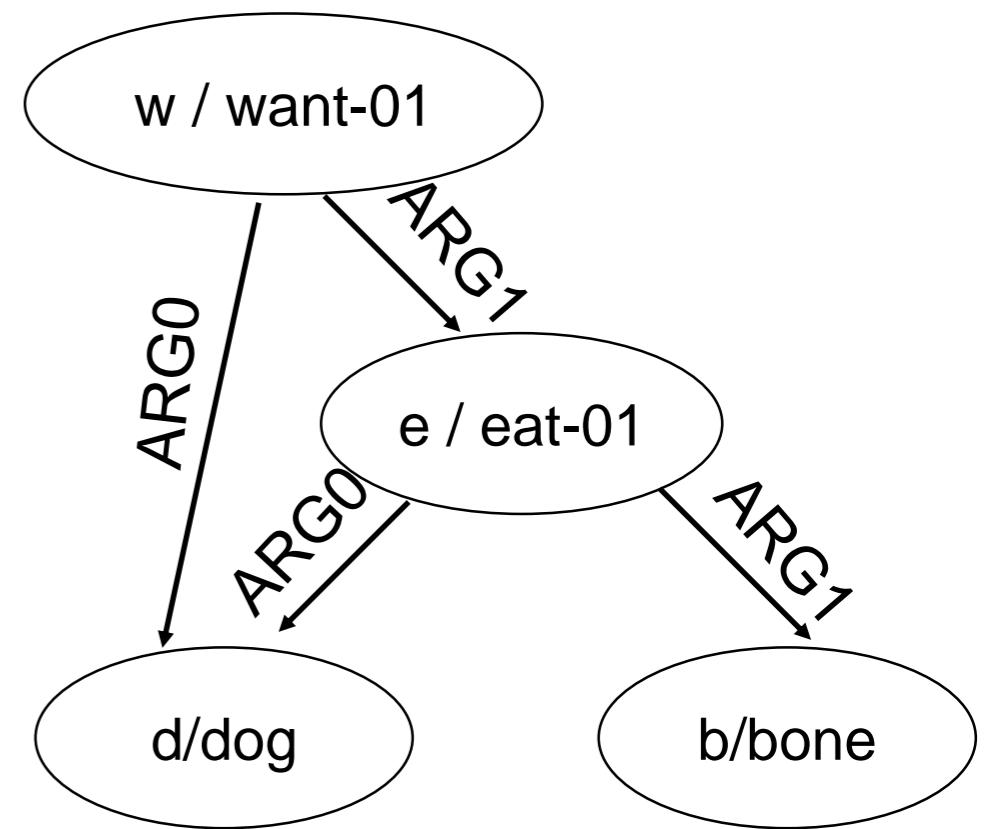


- Edges are labeled with **relations (including semantic roles)**
- Each node has a **variable**.
- Nodes are labeled with **concepts**.
 - Concepts can also be represented as edges.
- PropBank framesets used wherever possible.

Reentrancy

The dog wants to eat a bone.

```
(w / want-01
  :ARG0 (d / dog)
  :ARG1 (e / eat-01
    :ARG0 d
    :ARG1 (b / bone))
```

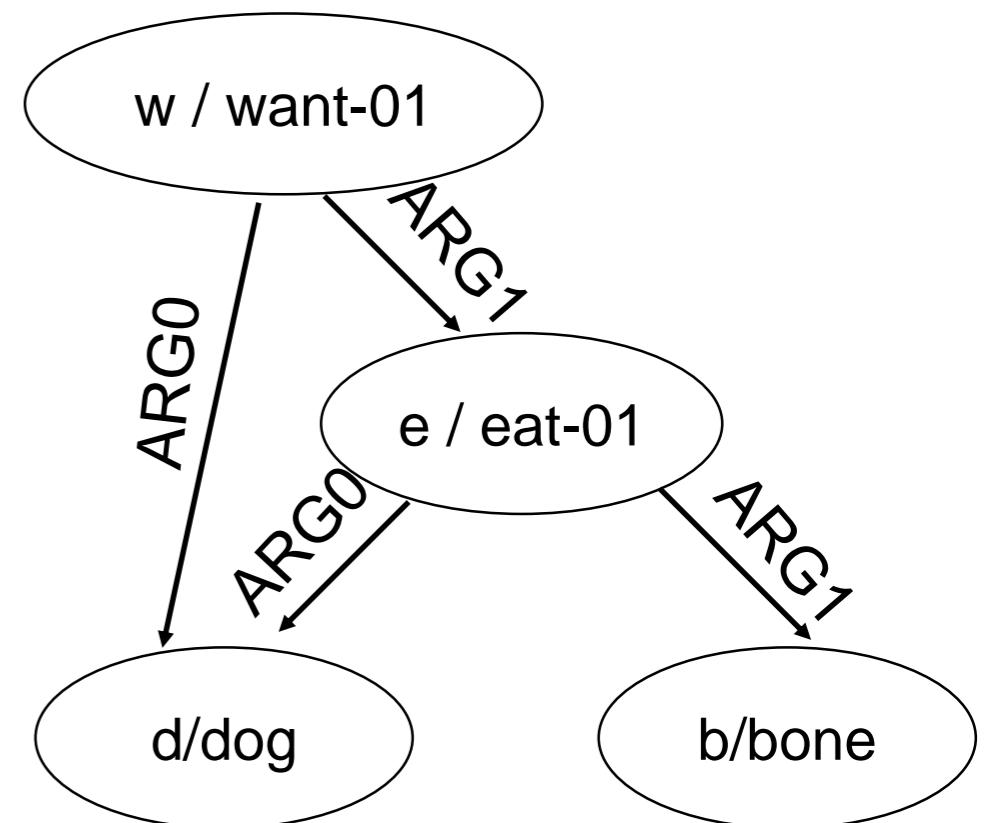


- Why the graph representation? Entities can play multiple roles.
- Two incoming edges in the graph, re-used variable in string notation.

AMR and Event Logic

The dog wants to eat a bone.

(w / want-01
 :ARG0 (d / **dog**)
 :ARG1 (e / eat-01
 :ARG0 **d**
 :ARG1 (b / bone))



- AMR is related to event logic:
 - All concepts are existentially quantified.
 - Relations and concept labels are predicates.

$$\exists w \exists d \exists e \exists b \text{Want}(w) \wedge \text{Dog}(d) \wedge \text{Eat}(e) \wedge \text{Bone}(b) \wedge \\ \text{ARG0}(w,d) \wedge \text{ARG1}(w,e) \wedge \text{ARG0}(e,d) \wedge \text{ARG1}(e,b)$$

Canonical Representaiton

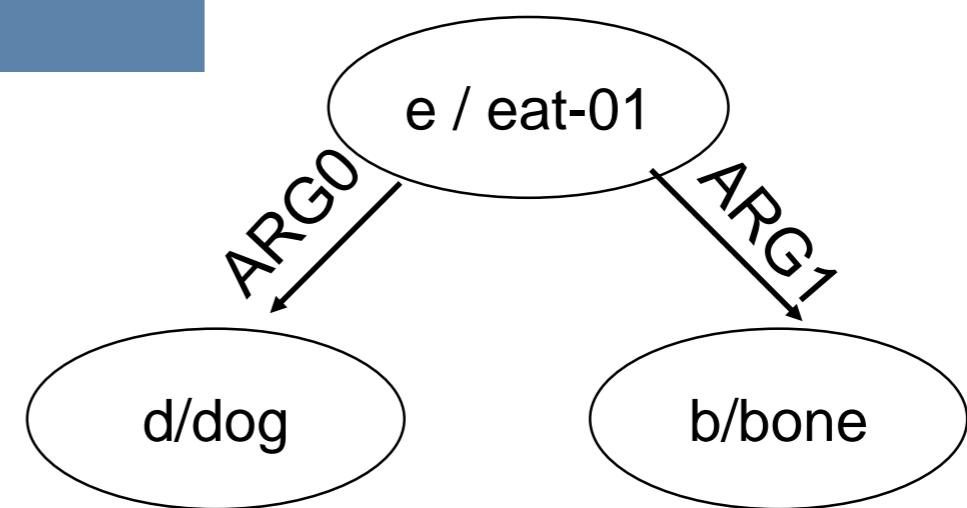
The dog is eating a bone.

The bone was eaten by the dog.

The dog's eating of the bone.

...

```
(e / eat-01  
:ARG0 (d / dog)  
:ARG1 (b / bone) )
```



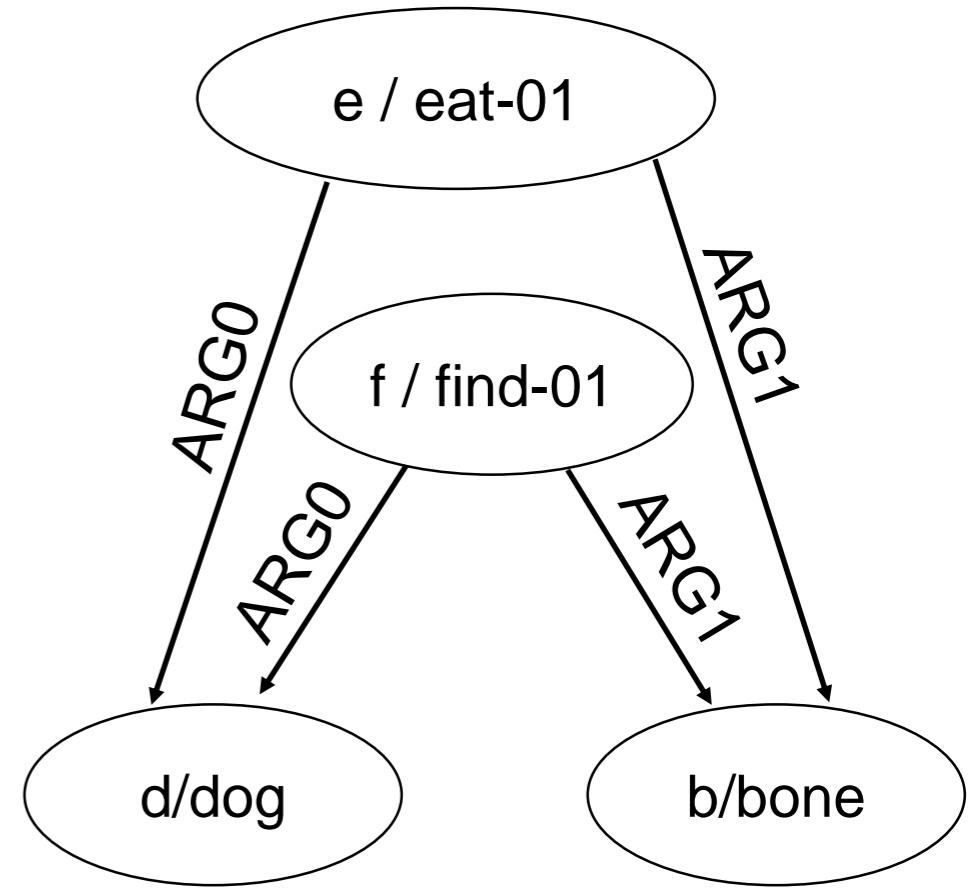
- Many different sentences can have the same AMR representation.
- Nouns can describe events too.

Inverse relations

The dog ate a bone that he found.

(e / eat-01
 :ARG0 (d / dog)
 :ARG1 (b / bone))

(f / find-01
 :ARG0 d
 :ARG1 b)

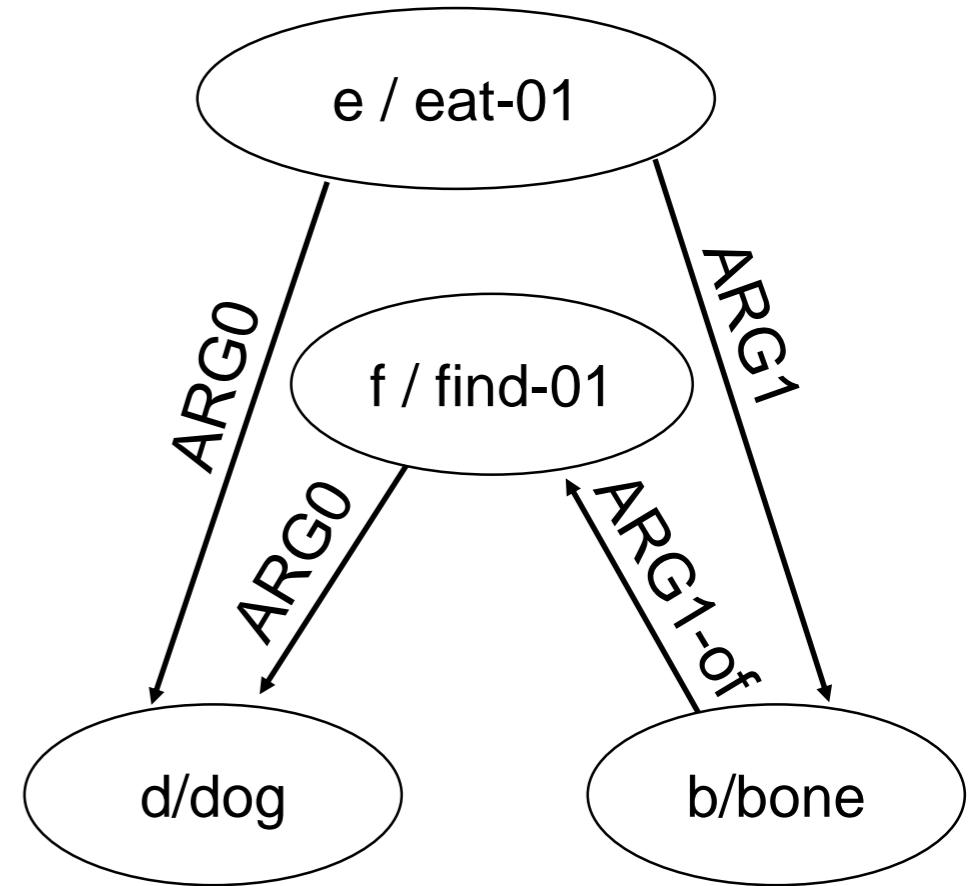


- AMR annotations are typically single-rooted (tree plus reentrancy)
- The single root is the "focus" of the sentence.

Inverse relations

The dog ate a bone that he found.

```
(e/ eat-01
  :ARG0 (d / dog)
  :ARG1 (b / bone
    :ARG1-of (f / find
      :ARG0 d) )
```

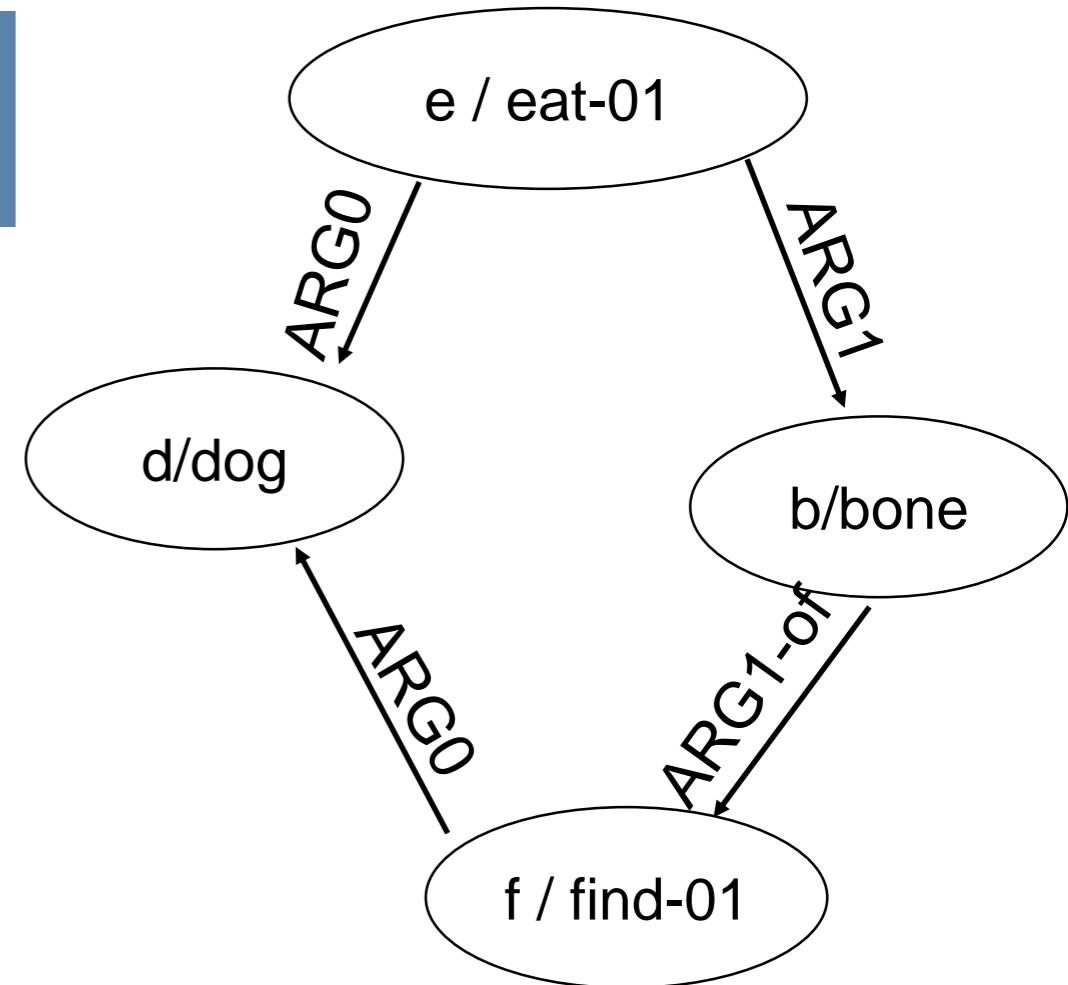


- AMR annotations are typically single-rooted (tree plus reentrancy)
- The single root is the "focus" of the sentence.

Inverse relations

The dog ate a bone that he found.

```
(e/ eat-01
  :ARG0 (d / dog)
  :ARG1 (b / bone
    :ARG1-of (f / find
      :ARG0 d) )
```

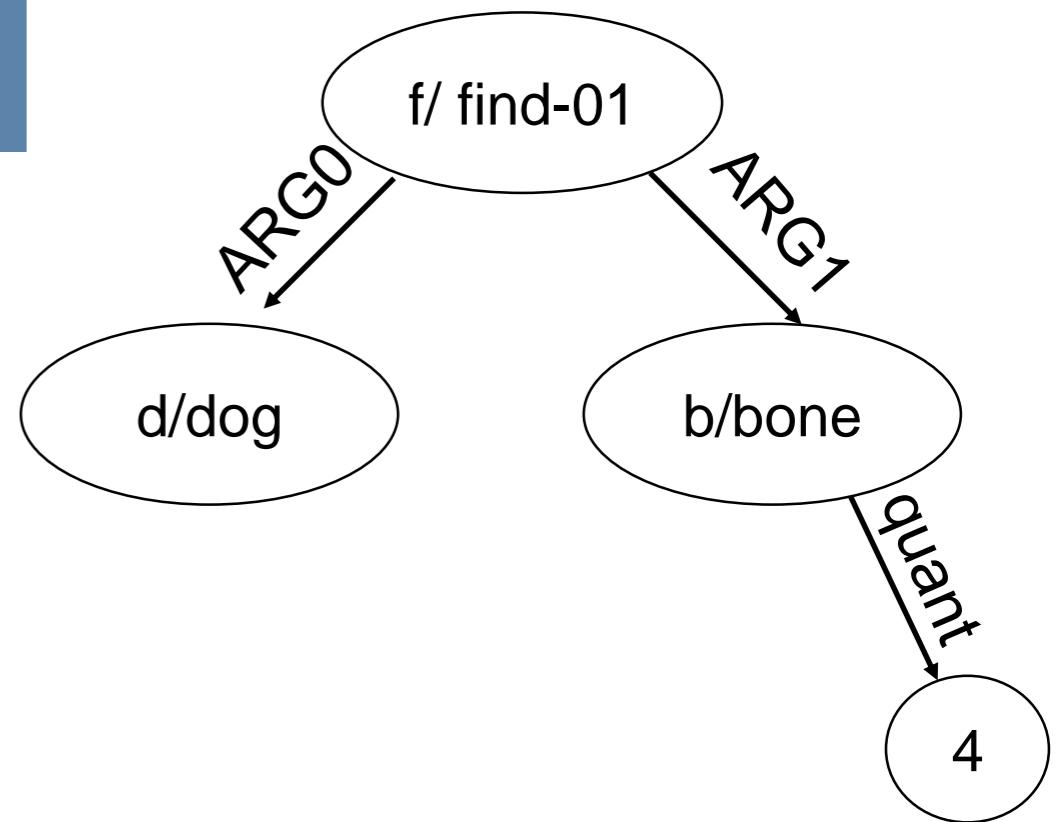


- AMR annotations are typically single-rooted (tree plus reentrancy)
- The single root is the "focus" of the sentence.

Constants

*The dog found **four** bones.*

```
(f/ find-01  
  :ARG0 (d / dog)  
  :ARG1 (b / bone  
         :quant 4) )
```

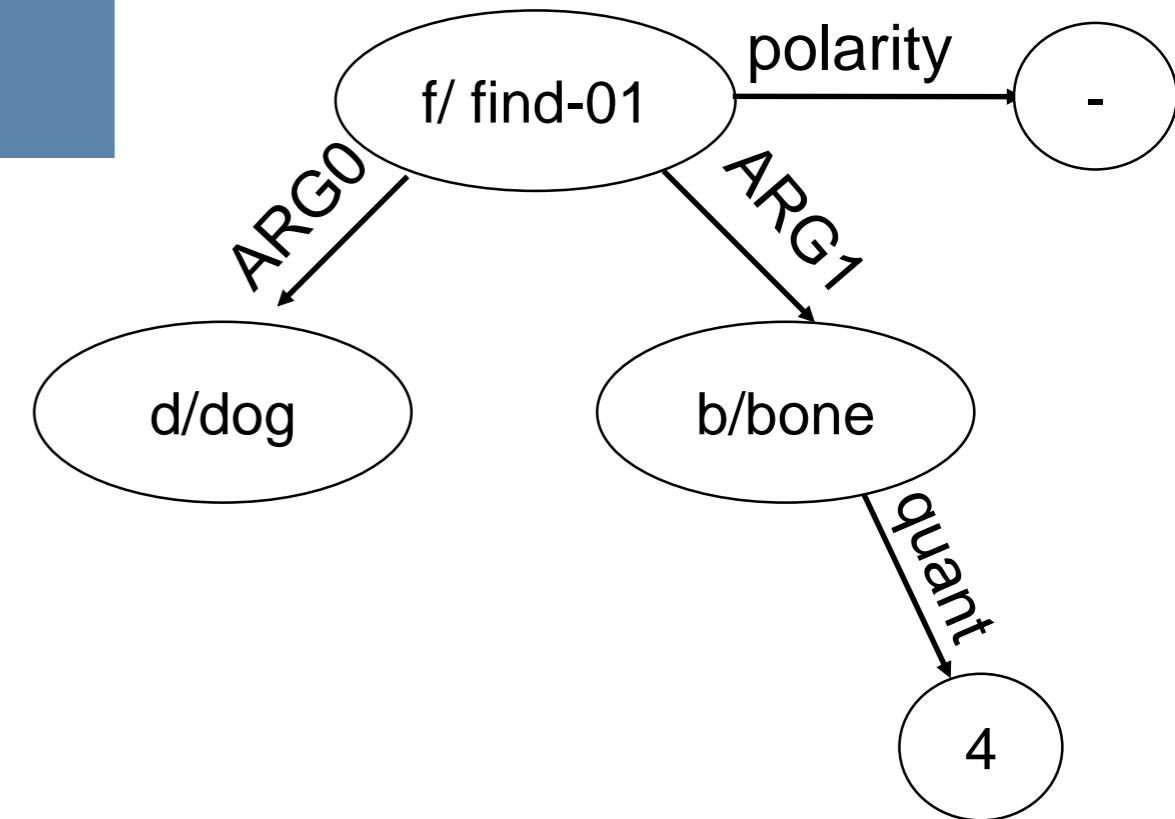


- Constants are used to represent quantities (node gets no variable).
- Also used for negation.

Constants

*The dog did not find **four** bones.*

```
(f/ find-01  
  :ARG0 (d / dog)  
  :ARG1 (b / bone  
         :quant 4)  
  :polarity -)
```



- Constants are used to represent quantities (node gets no variable).
- Also used for negation.

Non-Core Roles

- AMR annotations use some built-in relations (not in PropBank)
:time, :location, :manner, :part, :frequency
- :mod and :domain for attributes
- :op1, op2, ...for lists of arguments (for example in conjunctions).

```
(t/ truck  
  :mod (m / monster) )
```

a monster truck.

```
(s/see-01  
  (y / yummy  
    :domain (f / food) )
```

seeing that the food is yummy.

```
(a / and  
  :op1 (a / apple)  
  :op2 (o / orange) )
```

apples and oranges.

Names and Dates

```
(j / join-01
  :ARG0 (p / person :wiki -
    :name (p2 / name :op1 "Pierre" :op2 "Vinken")
    :age (t / temporal-quantity :quant 61
      :unit (y / year)))
  :ARG1 (b / board
    :ARG1-of (h / have-org-role-91
      :ARG0 p
      :ARG2 (d2 / director
        :mod (e / executive :polarity -))))))
  :time (d / date-entity :month 11 :day 29)))
```

AMR to English

(r / read-01
 :arg0 (j / judge)
 :arg1 (t / thing
 :arg1-of (p /propose-01))

(p / picture-01
 :ARG0 (i / it)
 :ARG1 (b2 / boa
 :mod (c / constrictor)
 :ARG0-of (d / digest-01
 :ARG1 (e / elephant))))

English to AMR

- "*The girl wants the boy to like her*"
- "*The girl wants the boy to believe that she likes him*"

AMR Data

- The Little Prince
(publicly available, <http://amr.isi.edu/download.html>):
 - English and Chinese
 - Biomedical Data
- "AMRBank", 14k sentence, PTB and other corpora
(including online discussion forums)

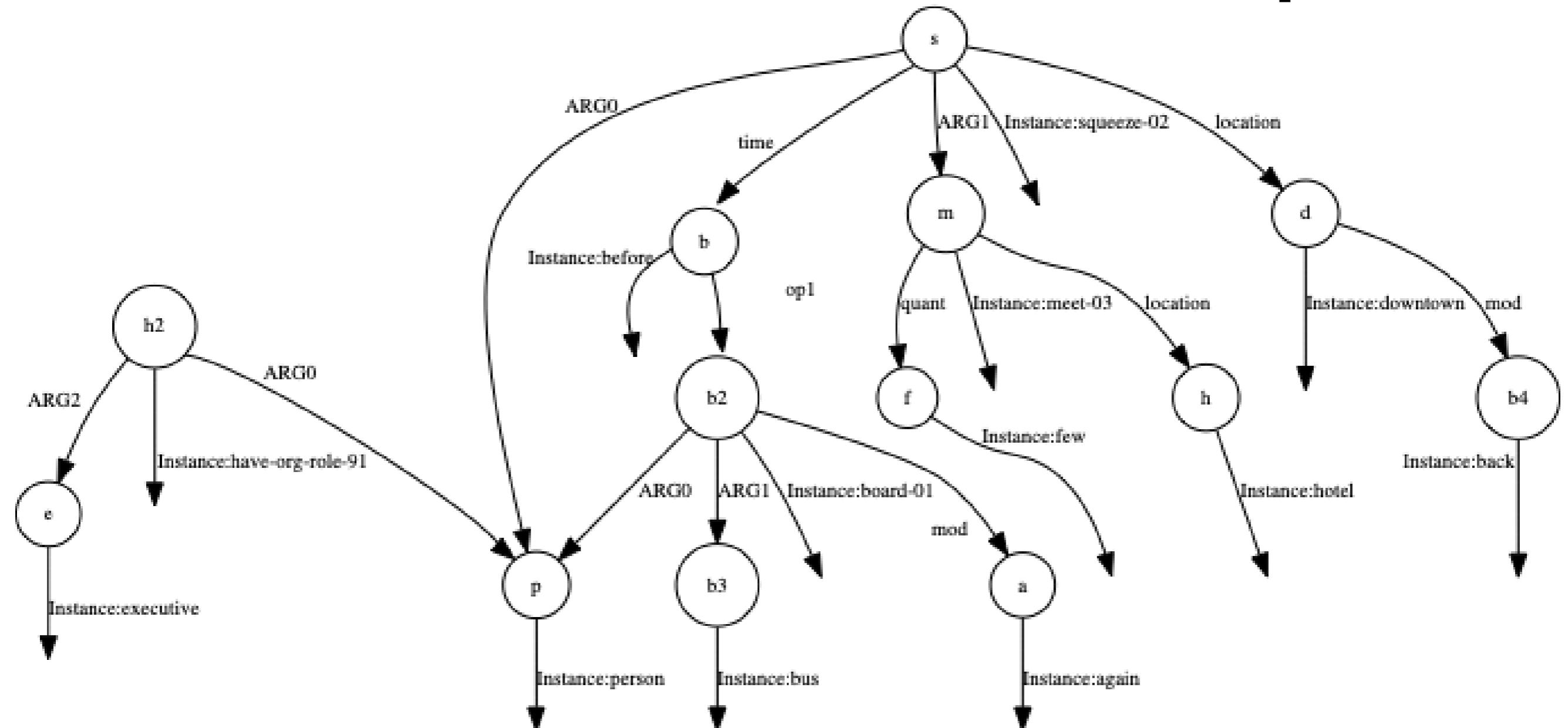
Another AMR Example

```
(s / squeeze-02
  :ARG0 (p / person
           :ARG0-of (h2 / have-org-role-91
                      :ARG2 (e / executive) ))
  :ARG1 (m / meet-03
           :location (h / hotel)
           :quant (f / few) )

  :location (d / downtown
              :mod (b4 / back) )
  :time (b / before
           :op1 (b2 / board-01
                     :ARG0 p
                     :ARG1 (b3 / bus)
                     :mod (a / again) )))
```

Back downtown, the execs squeezed in a few meetings at the hotel before boarding the buses again.

Another AMR Example



Back downtown, the execs squeezed in a few meetings at the hotel before boarding the buses again.

Applications of AMR

- Semantics-Based Machine Translation
(Jones, Andreas, Bauer, Hermann & Knight, 2012)
- Summarization:
 - Abstractive Summarization
(Liu, Flanigan, Thomson, Sadeh & Smith, 2015)
 - Text Compression (text-to-text generation)
(Thadani, 2015)
- Predicting stock price movement from financial news
(Xie, 2015)

Acknowledgments

- Some slides from Nathan Schneider & Jeff Flanigan's AMR tutorial at NAACL 2015.

Natural Language Processing

Lecture 12: Lexical Semantics (part I) -
Word Representations and Word Embeddings.

3/27/2019

COMS W4705
Yassine Benajiba

Jabberwocky

- Can you identify what the words in this poem mean?



Beware the jabberwock, my son
the jaws that bite, the claws that catch!

Beware the jubjub bird, and
the frumious bandersnatch!

"Jabberwocky", Lewis Carroll, 1871

Semantic Similarity and Relatedness

- We can often tell that two words are similar or related, even if they aren't exact synonyms.
- "**fast**" is similar to "**rapid**" and "**speed**"
- "**tall**" is similar to "**high**" and "**height**"
- Question answering:
 - Q: "*How tall is Mt. Everest?*"
 - Candidate A: "*The official height of Mount Everest is 29029 feet*"

Relatedness

- "**cat**" is more similar to "**dog**" than to "**table**"
- "**table**" is more similar to "**chair**" than to "**dog**"
- "**run**" is more similar to "**fly**" than to "**think**".
- "**cat**" is more similar to "**meow**" than to "**bark**".

Single Word Representation: One-Hot Vector

$$\begin{matrix} 0 & a \\ \vdots & \vdots \\ \text{fish} & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \text{|V|} & \text{zythum} \end{matrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

What about unseen words?

Unknown Words

*A bottle of **tesgüino** is on the table.*

*Everybody likes **tesgüino**.*

***Tesgüino** makes you drunk.*

*We make **tesgüino** out of corn.*

Example from Nida, 1975.

- Can you figure out from context what **tesgüino** means?
 - Some kind of alcoholic beverage, maybe beer or whisky.
- Intuition: Two words should be similar if they have similar typical word **contexts**.

How would you represent context?

Distributional Hypothesis

- Wittgenstein ("Philosophical Investigations):
"the meaning of a word is in its use in the language"
- Zelig Harris (1954):
"oculist and eye-doctor ... occur in almost the same environments"
"If A and B have almost identical environments we say that they
are synonyms."
- J.R. Firth (1957)
"you shall know a word by the company it keeps!"

Co-occurrence Matrix

	⌚	חלון	*	₪	±	₩
⌘	51	20	84	0	3	0
▷	52	58	4	4	6	26
⊗	115	89	10	42	33	17
◎	59	39	23	4	0	0
✖	98	14	6	2	1	0
✳	12	17	3	2	9	27
⬡	11	2	2	0	18	0

$\text{sim}(\boxtimes, ⌘) = 0.770$
 $\text{sim}(\boxtimes, *) = 0.939$
 $\text{sim}(\boxtimes, \triangleright) = 0.961$

- Numbers are co-occurrence counts (how often the symbols appear together in context).
- Which symbol is most similar to \boxtimes ?

What it really looks like

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	89	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
berry	11	2	2	0	18	0

Verb-Object counts

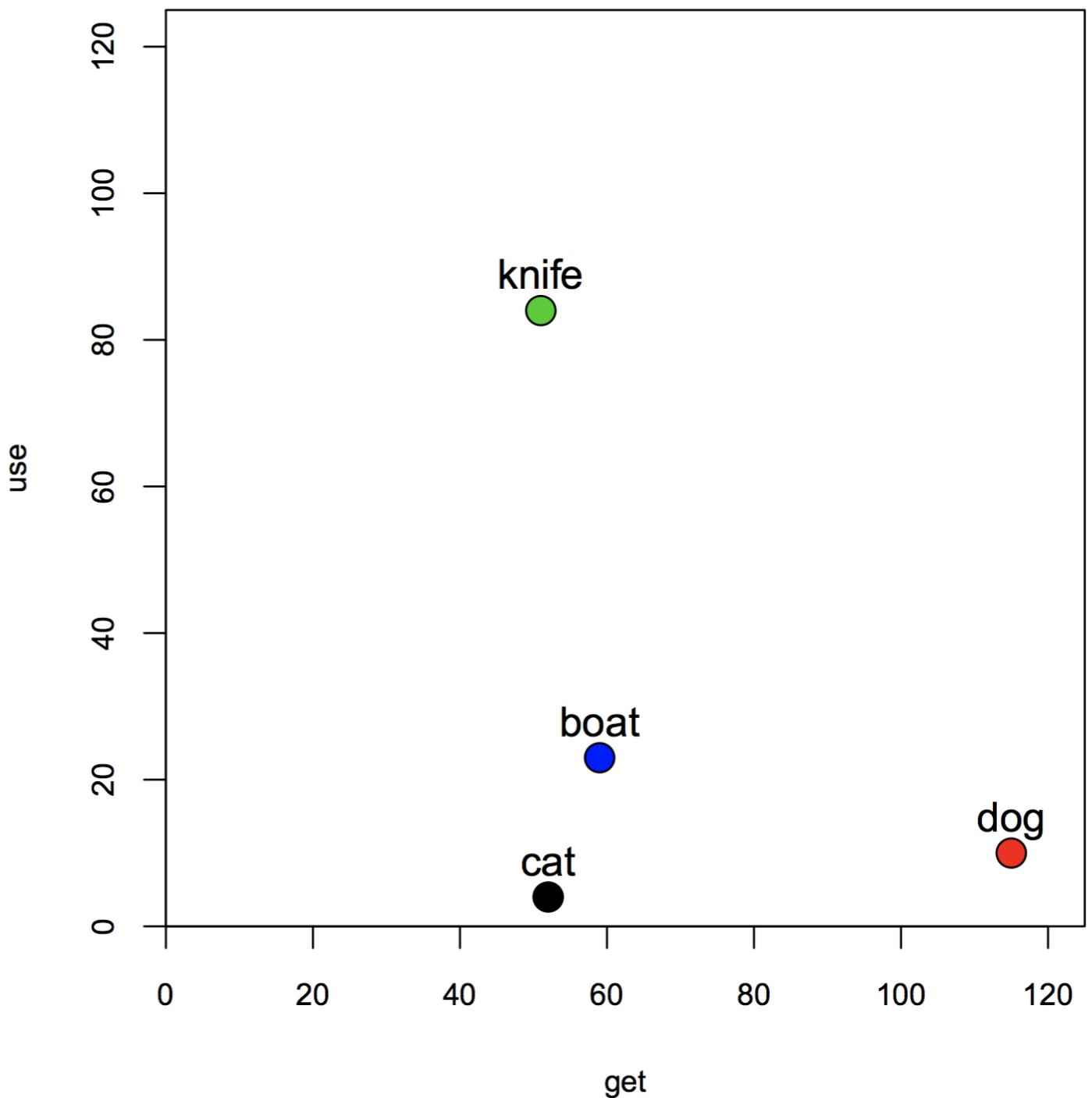
$\text{sim}(\text{dog}, \text{knife}) = 0.770$
 $\text{sim}(\text{dog}, \text{boat}) = 0.939$
 $\text{sim}(\text{dog}, \text{cat}) = 0.961$

- Row vector \mathbf{x}_{dog} describes usage of *dog* as a grammatical object in the corpus.
- Can be seen as coordinates in n-dimensional Euclidean space.

Geometric Interpretation

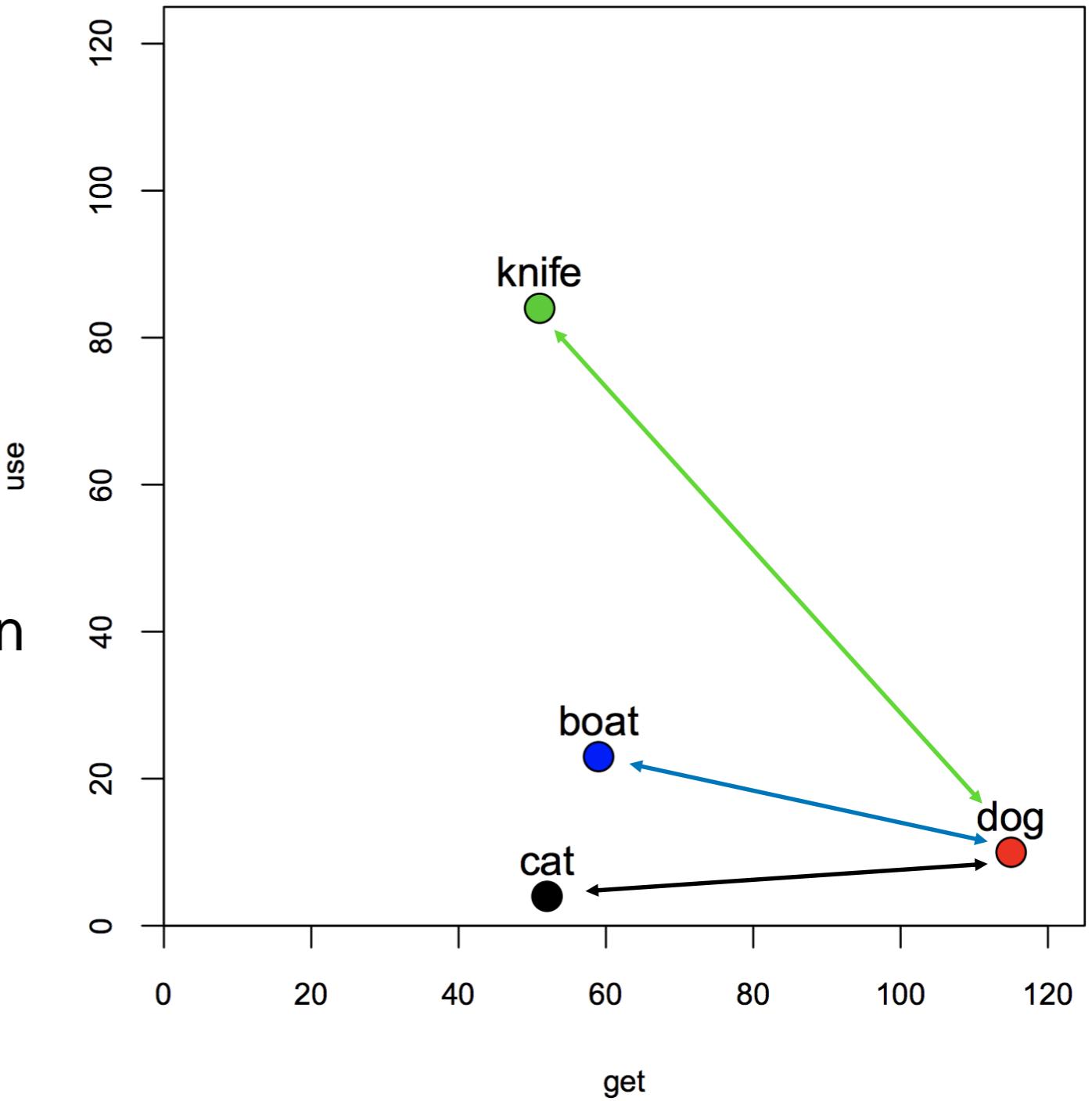
- Row vector \mathbf{x}_{dog} describes usage of *dog* in the corpus.
- Can be seen as coordinates in n -dimensional Euclidean space.
- Illustrated for two dimensions "get" and "use".

$$\mathbf{x}_{\text{dog}} = (115, 10)$$



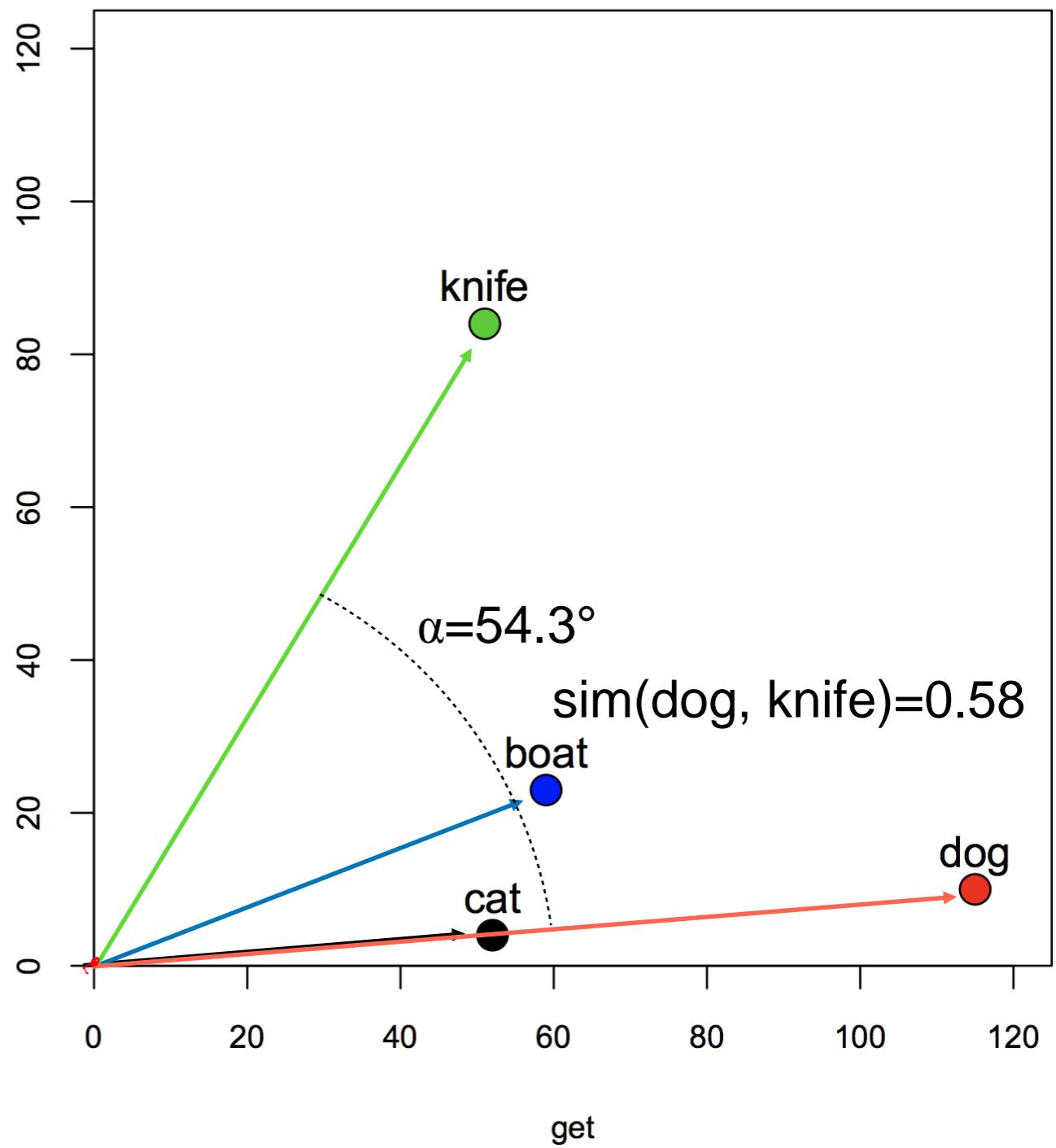
Geometric Interpretation

- How should we compute similarity?
 - First approach: Spatial distance between words.
 - (lower distance = higher similarity)
 - Potential problem: location depends on frequency of noun
 $\text{count(dog)} \approx 2.7 \text{ count(cat)}$



Geometric Interpretation

- How should we compute similarity?
 - Second approach:
 - Direction is more important than location.
 - Normalize "length" $\|x_{\text{dog}}\|$ of vector.
 - or use angle α as distance measure (or cos of these angles).



Cosine Similarity

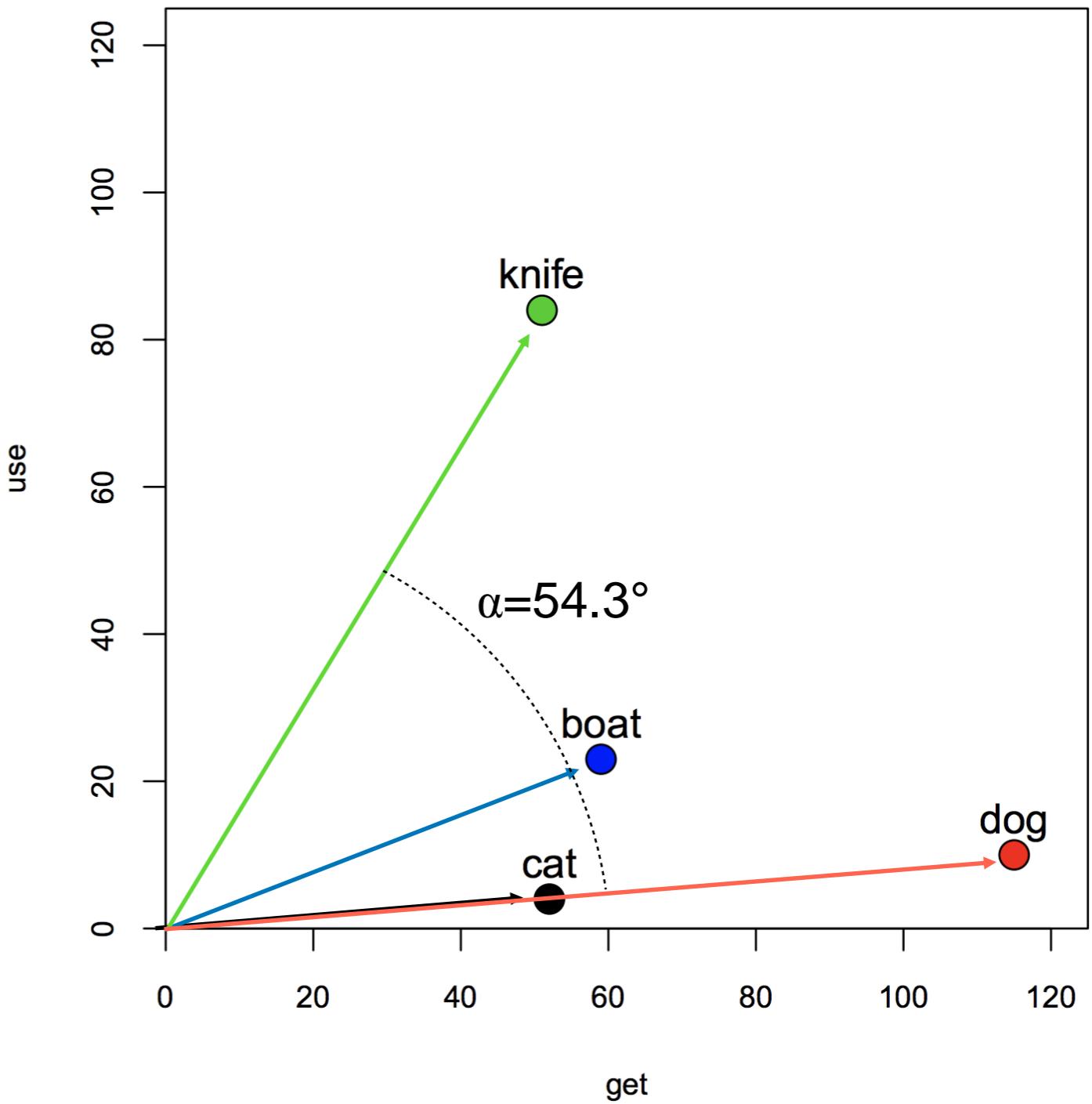
$$sim_{cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2} = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}$$

Colinear vectors (same direction):

$$sim_{cos}(\mathbf{x}, \mathbf{y}) = 1$$

Orthogonal vectors
(90° angle, no shared attributes):

$$sim_{cos}(\mathbf{x}, \mathbf{y}) = 0$$



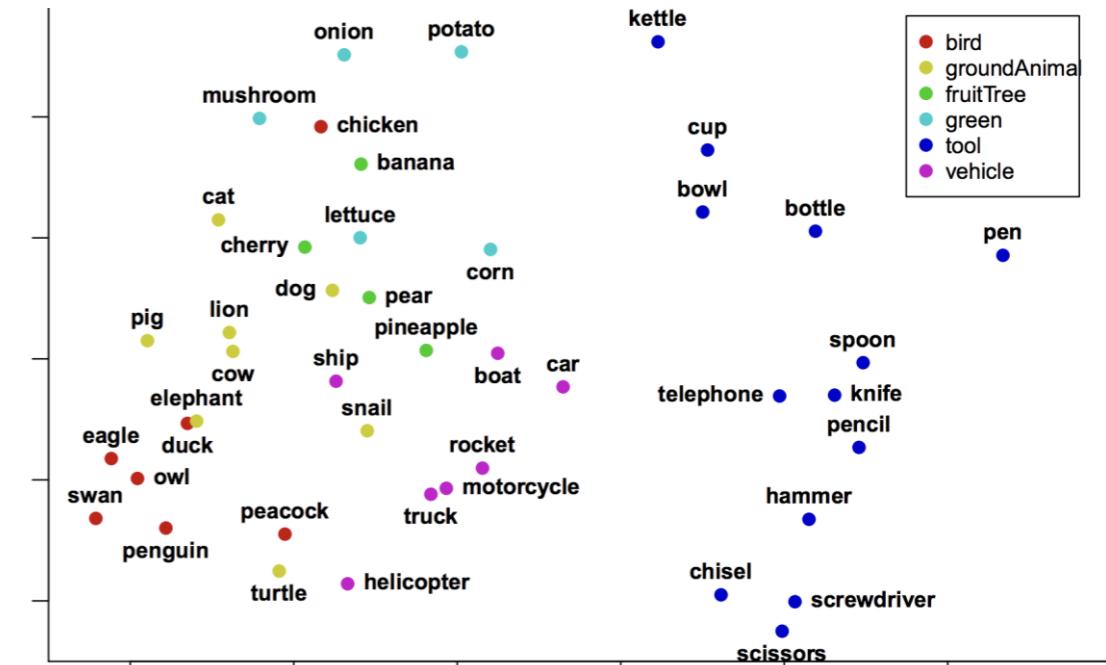
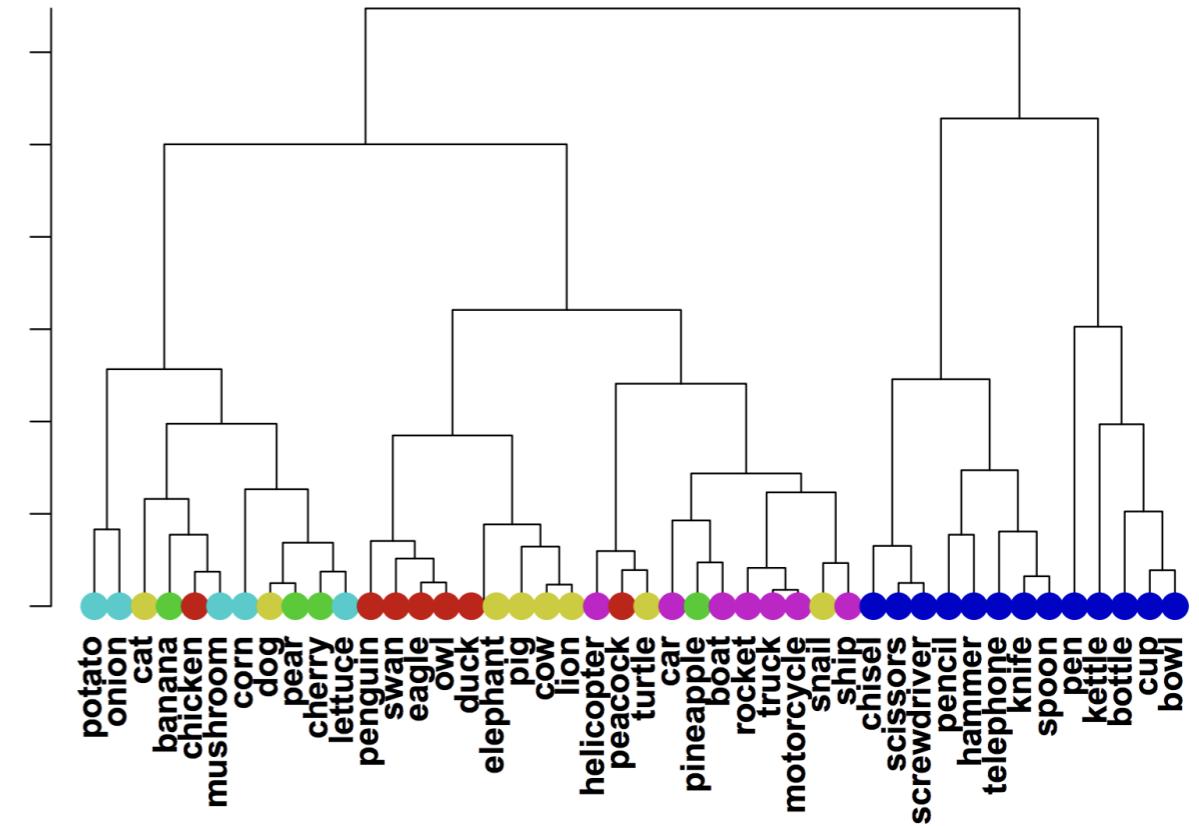
What to do with DSM similarities

- Most similar to **school**:
country (49.3), church (52.1), hospital (53.1), house (54.4), hotel (55.1), industry (57.0), company (57.0), home (57.7), family (58.4), university (59.0), party (59.4), group (59.5), building (59.8), market (60.3), bank (60.4), business (60.9), area (61.4), department (61.6), club (62.7), town (63.3), library (63.3), room (63.6), service (64.4), police (64.7),...

Clustering and Semantic Maps

- Distributional Similarity/Distance can be used to
 - find nearest neighbors (similar words)
 - cluster related words into hierarchical categories.
 - construct semantic maps.

Word space clustering of concrete nouns (V-Obj from BNC)



Variations of Distributional Semantic Models

- A Distributional Semantic Model (DSM) is any matrix M such that each row represents the distribution of a term x across contexts, together with a similarity measurement.
- The previous example shows one particular semantic space (frequency counts of Verb-object co-occurrences).
- There are many different models we could choose.
- Different models might capture different "types" of similarity.

Dimensions of Distributional Semantic Models

1. Preprocessing, definition of "terms" (word form, lemmas, POS, ...).
2. Context definition:
 - Type of context (word, syntactic dependents (with or without relation labels), remove stop-words, etc.)
 - Size of context window.
3. Feature scaling / term weighting (association measures).
4. Normalization of rows / columns.
5. Dimensionality reduction.
6. Similarity measure.

Effect of context size

Nearest neighbors of *dog*

2-word window:

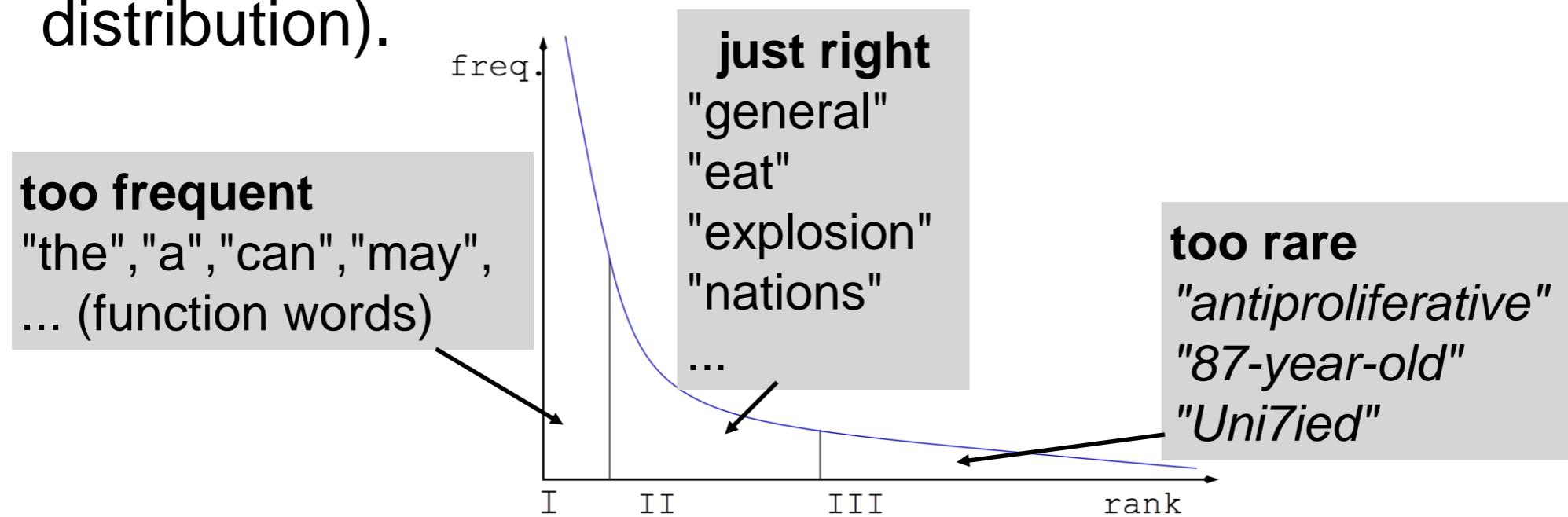
cat, horse, fox, pet, rabbit, pig, animal, mongrel, sheep, pigeon

30-word window:

kennel, puppy, pet, terrier, rottweiler, canine, cat, to bark, Alsatian

Term Weighting

- Problem: Not all context terms are equally relevant to characterize the meaning of a word.
 - Some appear too often, some are too rare (Zipfian distribution).



- One solution: TF*IDF (term frequency * inverse-document frequency)

TF*IDF

- Originates in document retrieval (find document relevant to a keyword). For DSM: 'document' = target word d .
- Term frequency: How often does the term t appear in the context window of the target word?

$$tf_{t,d} = \text{count}(d, t)$$

- Inverse document frequency: For how many words does t appear in the context window

$$idf_{t,D} = \log \frac{|D|}{|\{d \in D, \text{count}(d, t) > 0\}|}$$

- TF*IDF:

$$tf_{t,d} \cdot idf_{t,D}$$

Sparse vs. Dense Vectors

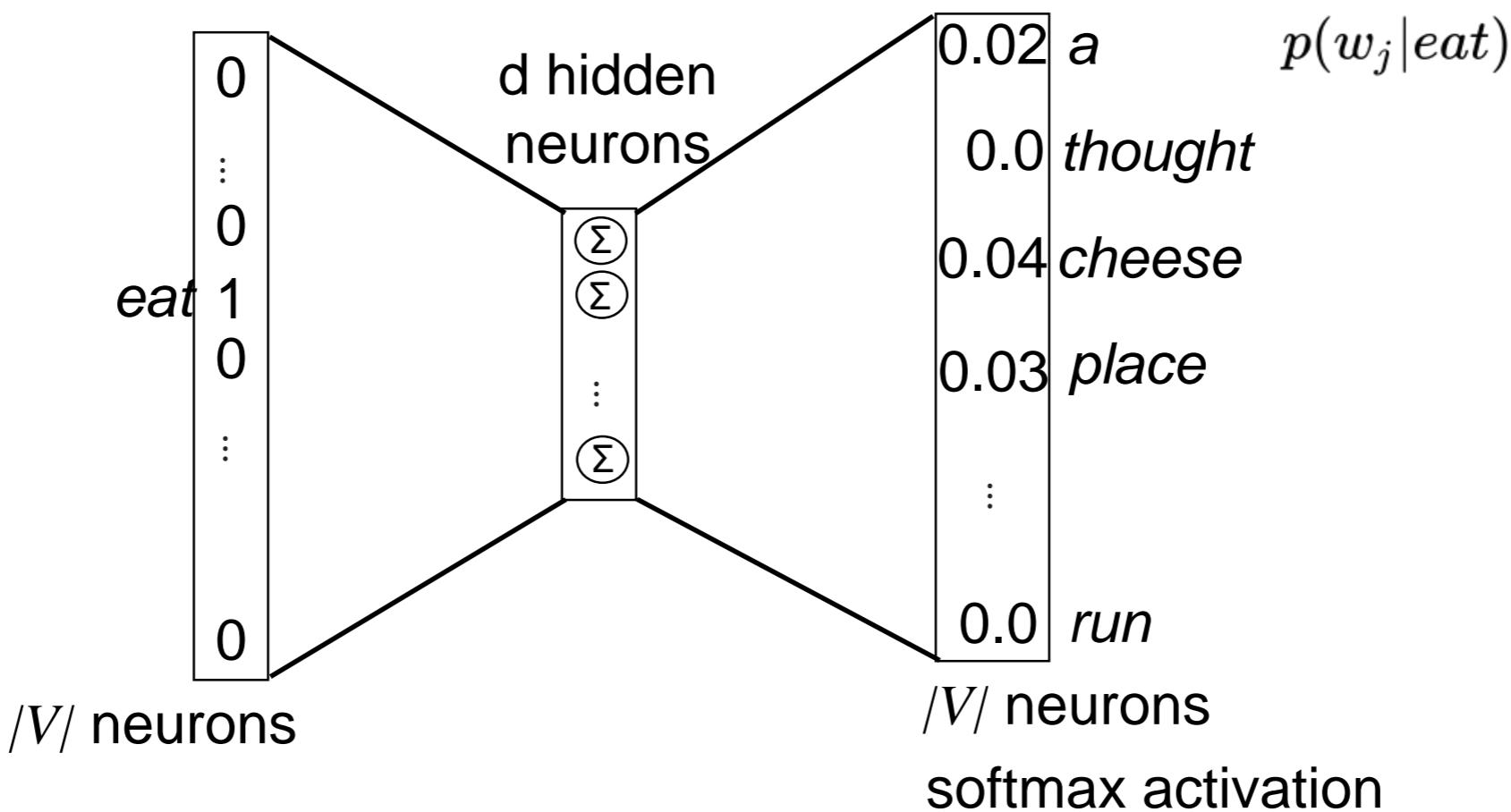
- Full co-occurrence matrix is very big and contains a lot of 0 entries.
 - Potentially inconvenient to store. Slow computation.
 - Synonyms may still contain orthogonal dimensions, which are irrelevant.
- **Word embeddings** are representations of words in a low-dimensional, dense vector space. There are two main approaches:
 - Use matrix decomposition on co-occurrence matrix, for example Singular Value Decomposition (SVD).
 - **Learn embeddings using neural networks. Minimal feature-engineering required.**

Learning Word Embeddings with Neural Networks

- The neural network should capture the relationship between a word and its context.
- Two models:
(Word2Vec, Mikolov et al. 2013)
 - **Skip-Gram model:** Input is a single word.
Predict a probability for each context word.
 - **Continuous bag-of-words (BOW):**
Input is a representation of the context window.
Predict a probability for each target word.
- Inspired by Neural Language Models *(Bengio et al. 2003)*

Skip-Gram Model

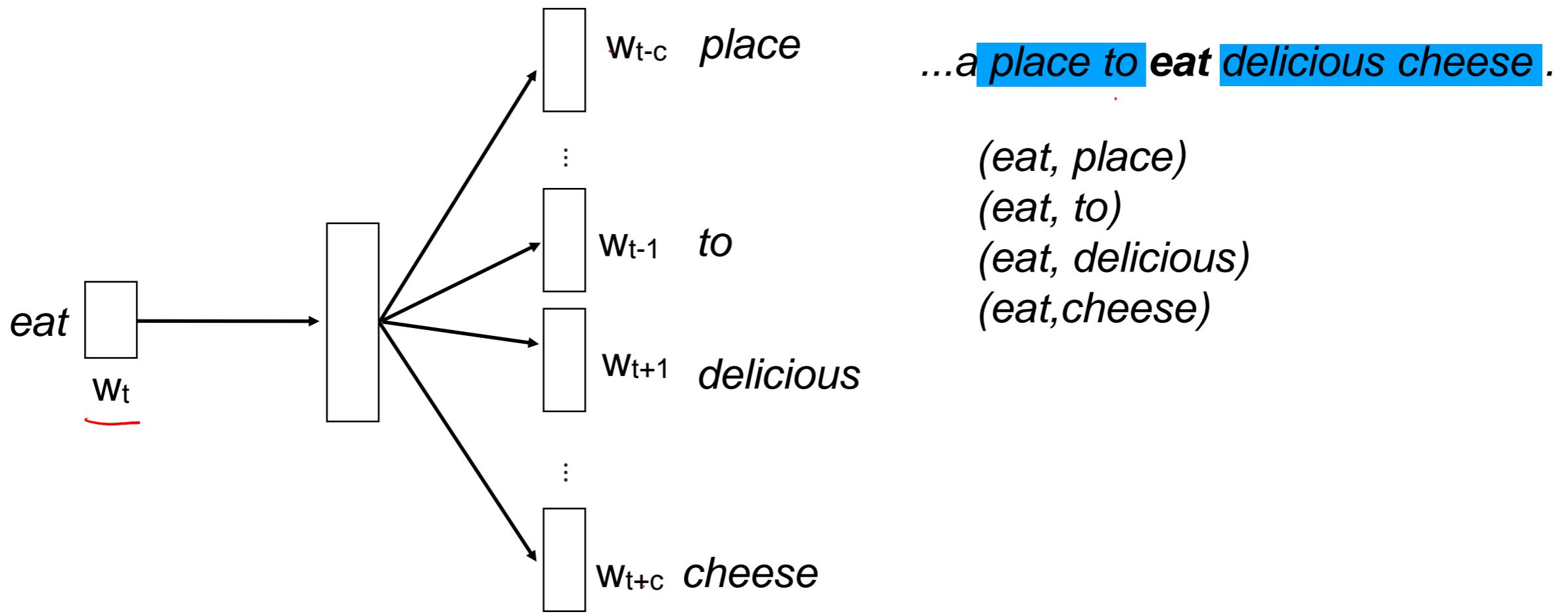
- Input:
A single word in one-hot representation.
- Output: probability to see any single word as a context word.



- Softmax function normalizes the activation of the output neurons to sum up to 1.0.

Skip-Gram Model

- Compute error with respect to each context word.



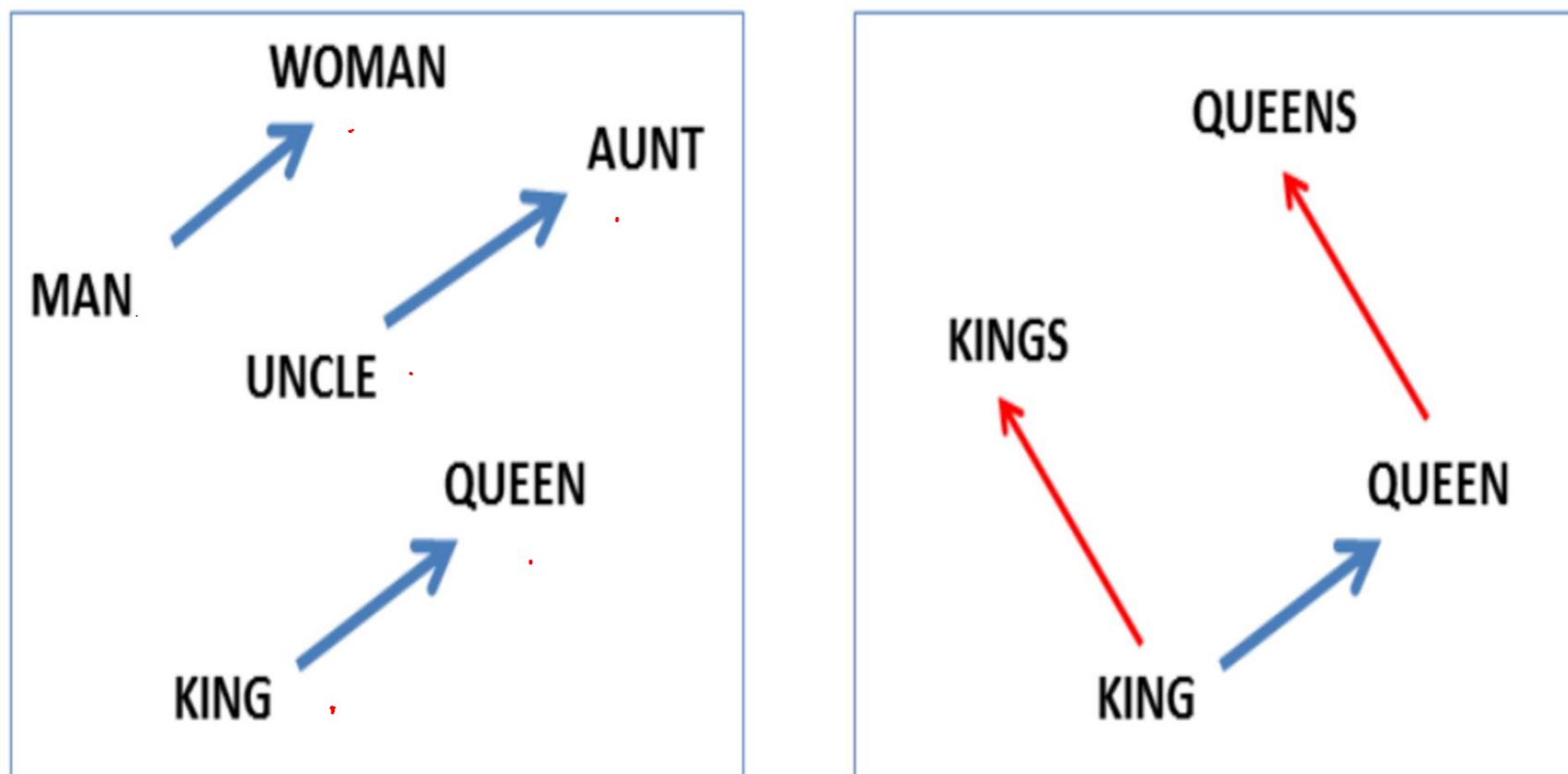
- Combine errors for each word, then use combined error to update weights using back-propagation.

$$error = - \sum_{i=-c, i \neq 0}^c \log p(w_{t+i} | w_t)$$

Embeddings are Magic

(Mikolov 2016)

$$\text{vector('king')} - \text{vector('man')} + \text{vector('woman')} \approx \text{vector('queen')}$$



Application: Word Pair Relationships

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Using Word Embeddings

- Word2Vec:
 - <https://code.google.com/archive/p/word2vec/>
- GloVe: Global Vectors for Word Representation
 - <https://nlp.stanford.edu/projects/glove/>
- Can either use pre-trained word embeddings or train them on a large corpus.

Acknowledgments

- Some content adapted from slides by Kathy McKeown, Dan Jurafsky, Stefan Evert, Marco Baroni

Natural Language Processing

Lecture 13:
Machine Learning: Linear and Log-Linear Models

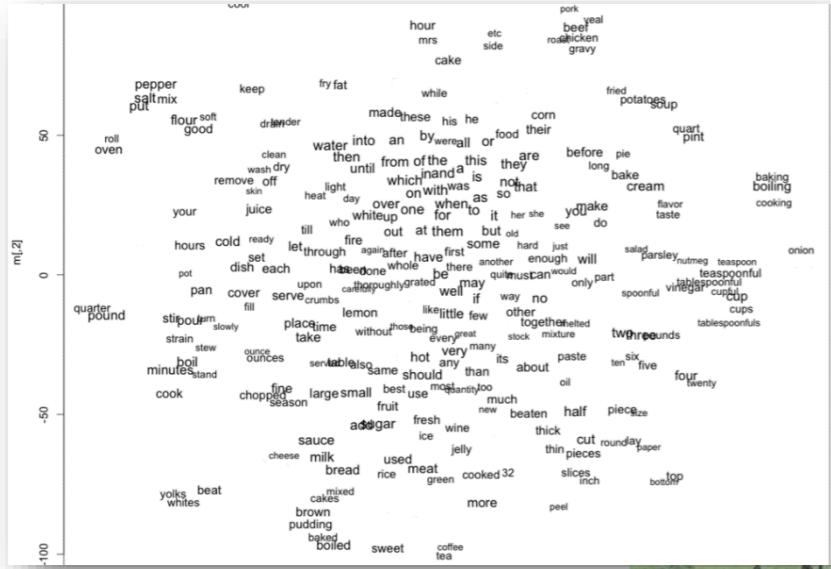
4/3/2019

COMS W4705
Yassine Benajiba

Intro

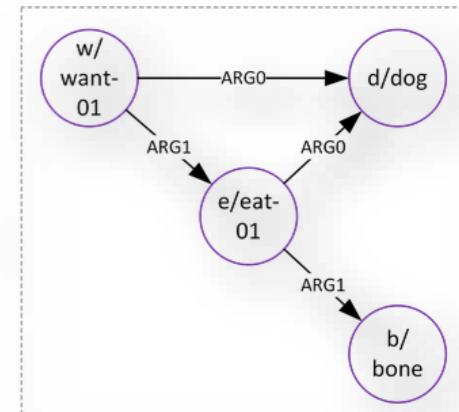


Intro



$\exists w, e, d, b$
 $\text{instance}(w, \text{want-01}) \wedge \text{instance}(d, \text{dog})$
 $\wedge \text{instance}(e, \text{eat}) \wedge \text{instance}(b, \text{bone})$
 $\wedge \text{arg0}(w, d) \wedge \text{arg1}(w, e)$
 $\wedge \text{arg0}(e, d) \wedge \text{arg1}(e, b)$

$(w / \text{want-01}$
 $: \text{ARG0 } (d / \text{dog})$
 $: \text{ARG1 } (e / \text{eat-01})$
 $: \text{ARG0 } d$
 $: \text{ARG1 } (b / \text{bone}))$



Machine Learning and NLP

- We have encountered many different situations where we had to make a prediction:
 - Text classification, language modeling, POS tagging, constituency/dependency parsing,
 - These are all classification problems of some form.
- Today: Some machine learning background. Linear/log-linear models. Basic neural networks.

Generative Algorithms

- Assume the observed data is being “generated” by a “hidden” class label.
- Build a different model for each class.
- To predict a new example, check it under each of the models and see which one matches best.
- Model $P(x|y)$ and $P(y)$. Then use bases rule

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

Discriminative Algorithms

- Model conditional distribution of the label given the data
 $P(y|x)$
- Learns decision boundaries that separate instances of the different classes.
- To predict a new example, check on which side of the decision boundary it falls.

Machine Learning Definition

- “Creating systems that improve from experience.”
- “*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*”
(Tom Mitchel, Machine Learning 1997)

Inductive Learning (a.k.a. *Science*)

- **Goal:** given a set of input/output pairs (training data), find the function $f(x)$ that maps inputs to outputs.
Problem: We did not see all possible inputs!
- Learn an approximate function $h(x)$ from the training data and hope that this function *generalize well* to unseen inputs.
- **Ockham's razor:** Choose the *simplest* hypothesis that is consistent with the training data.

Classification and Regression

- Recall: In **supervised learning**, training data consisting of training examples $(x_1, y_1), \dots, (x_n, y_n)$, where x_j is an input example (a d -dimensional vector of attribute values) and y_j is the label.
- Two types of supervised learning problems:

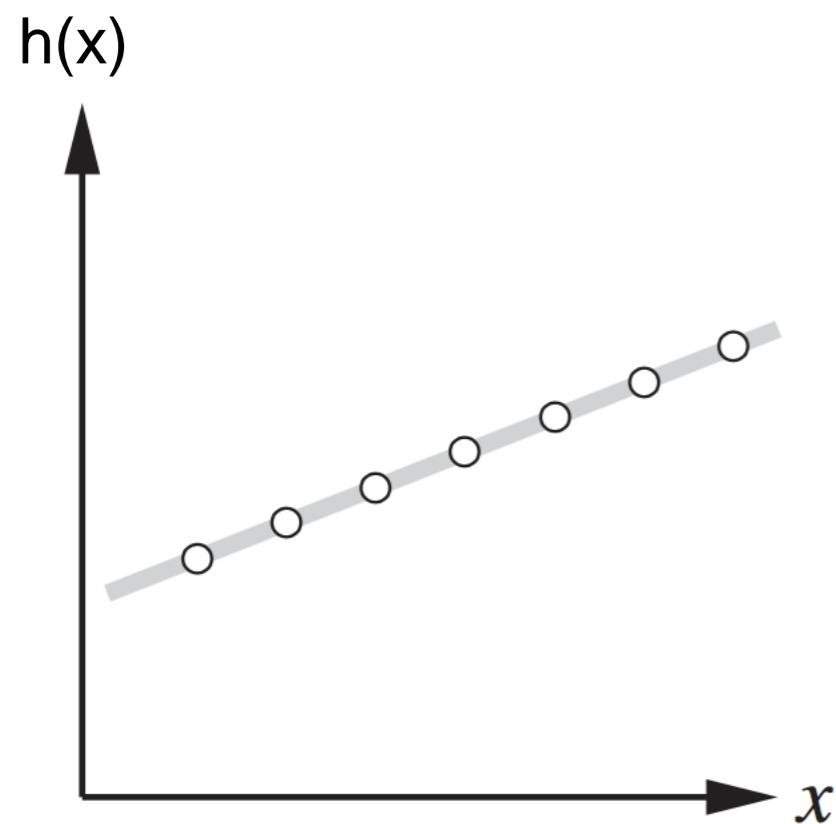
- In classification: y_j is a finite, discrete set.
Typically $y_j \in \{-1, +1\}$. i.e. predict a label from a set of labels.
Learn a **classifier** function:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

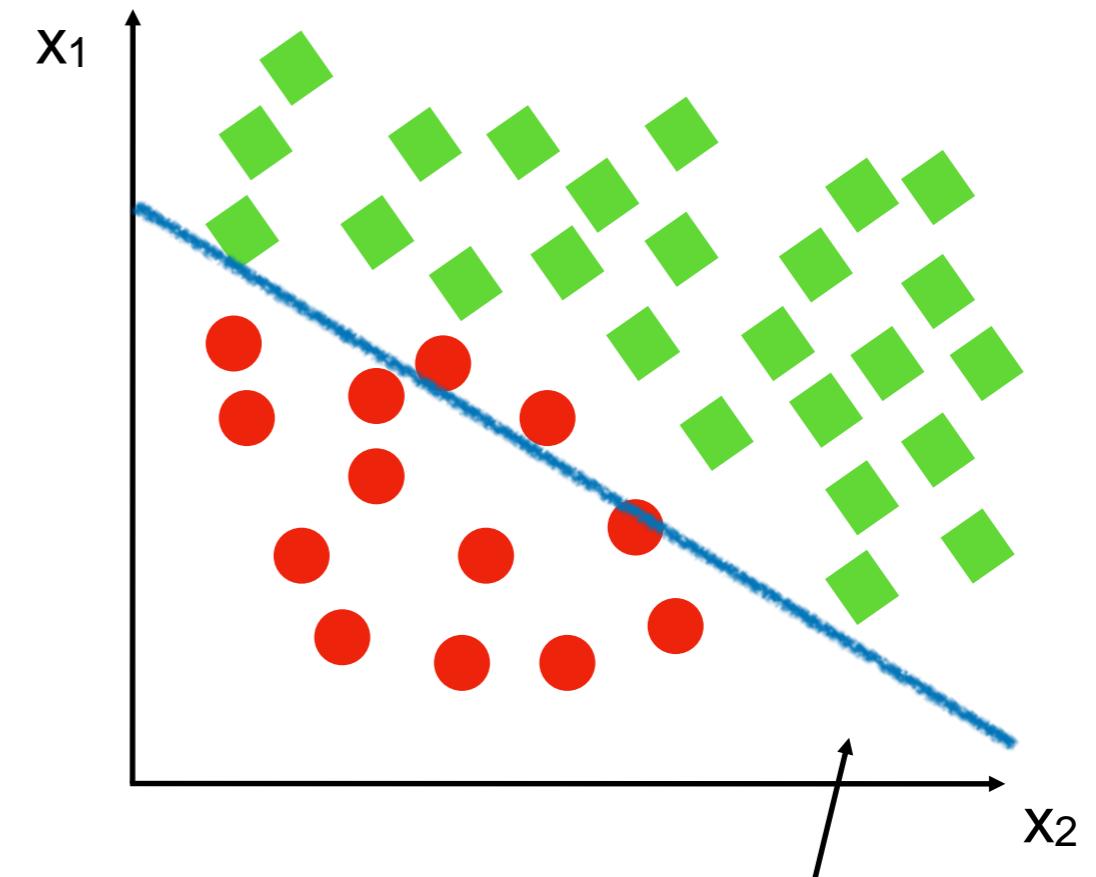
- In regression: $x_j \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. i.e. predict a numeric value.
Learn a **regressor** function:

$$h : \mathbb{R}^d \rightarrow \mathbb{R}$$

Linear Classification and Regression

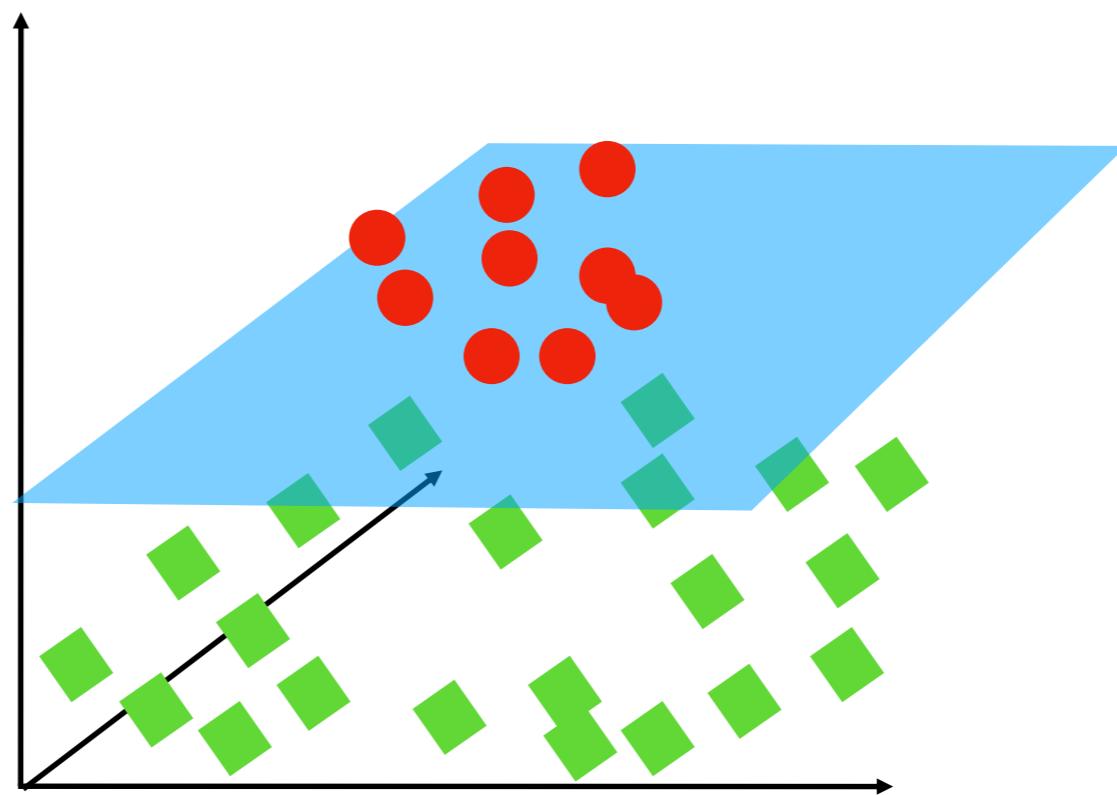


Regression

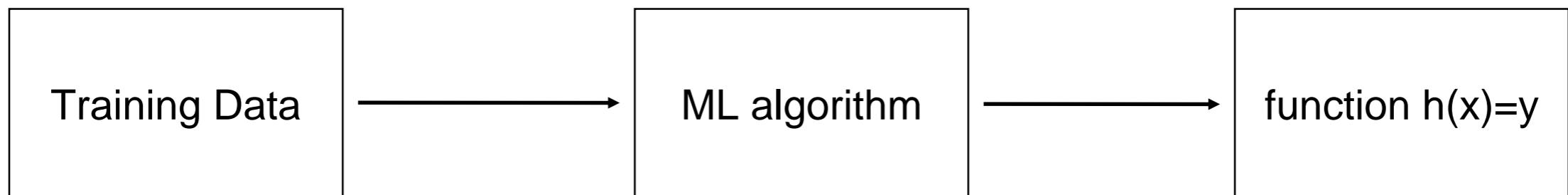


Classification

Linear Classification



Training ML models

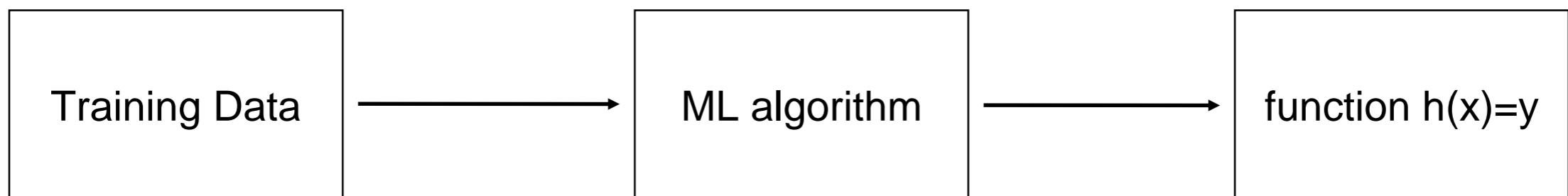


- How can we be confident about the learned function?
- Can compute empirical error/risk on the training set:

$$E_{train}(h) = \sum_{i=1}^n loss(y_i, h(x_i))$$

- Typical loss functions:
 - Least square loss (L2): $loss(y_i, h(x_i)) = (y_i - h(x_i))^2$
 - Classification error: $loss(y_i, h(x_i)) = \begin{cases} 1 & \text{if } sign(h(x_i)) \neq sign(y_i) \\ 0 & \text{otherwise.} \end{cases}$

Training ML models



- Empirical error/risk:

$$E_{train}(h) = \sum_{i=1}^n loss(y_i, h(x_i))$$

- Training aims to minimize E_{train} .
- We hope that this also minimizes E_{test} , the test error.

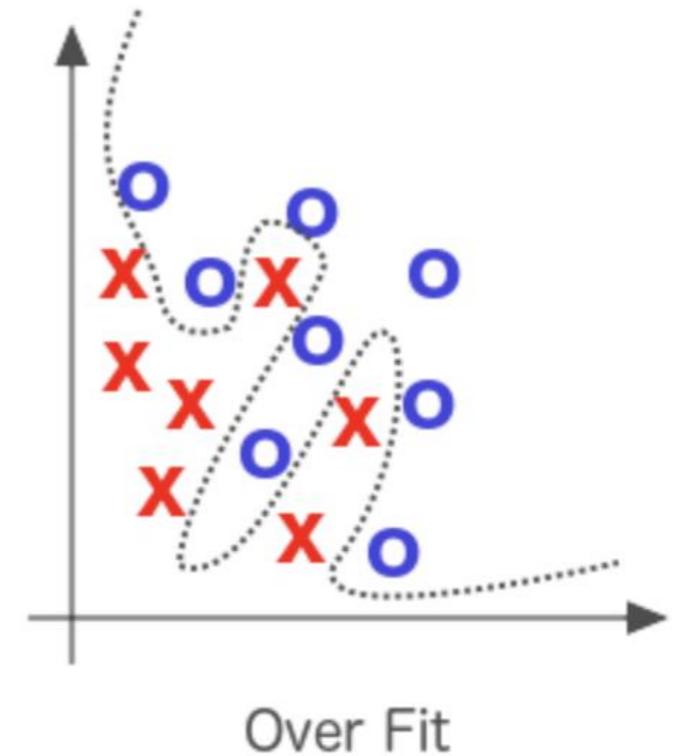
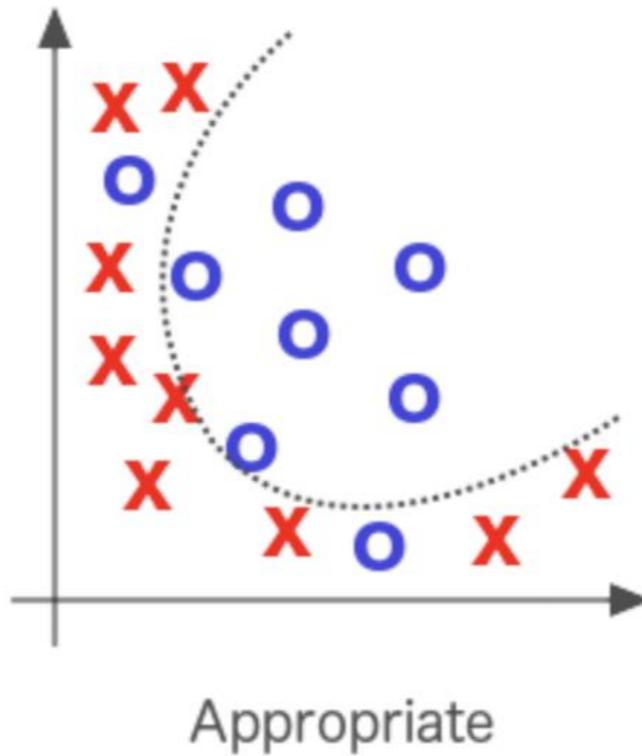
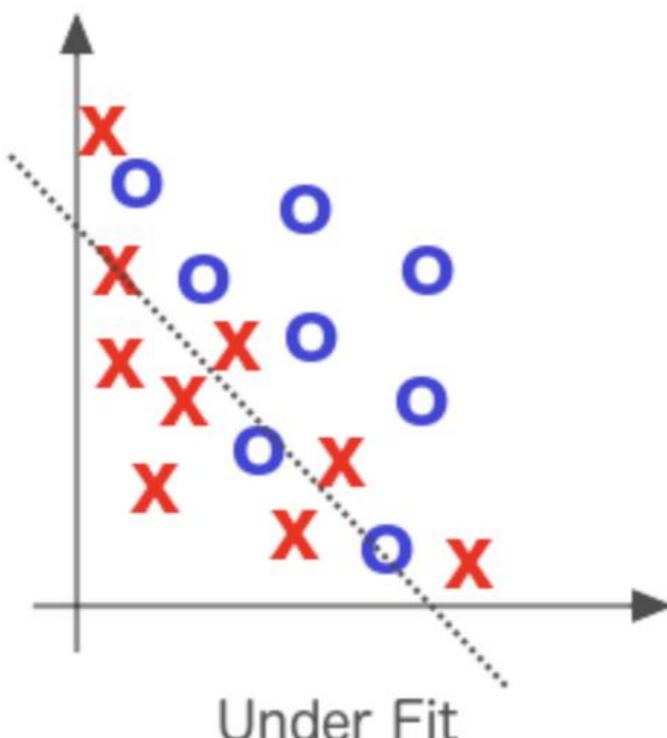
Overfitting

- Problem: Minimizing empirical risk can lead to **overfitting**.
 - This happens when a model works well on the training data, but it does not **generalize** to testing data.
 - Data sets can be noisy. Overfitting can model the noise in the data.

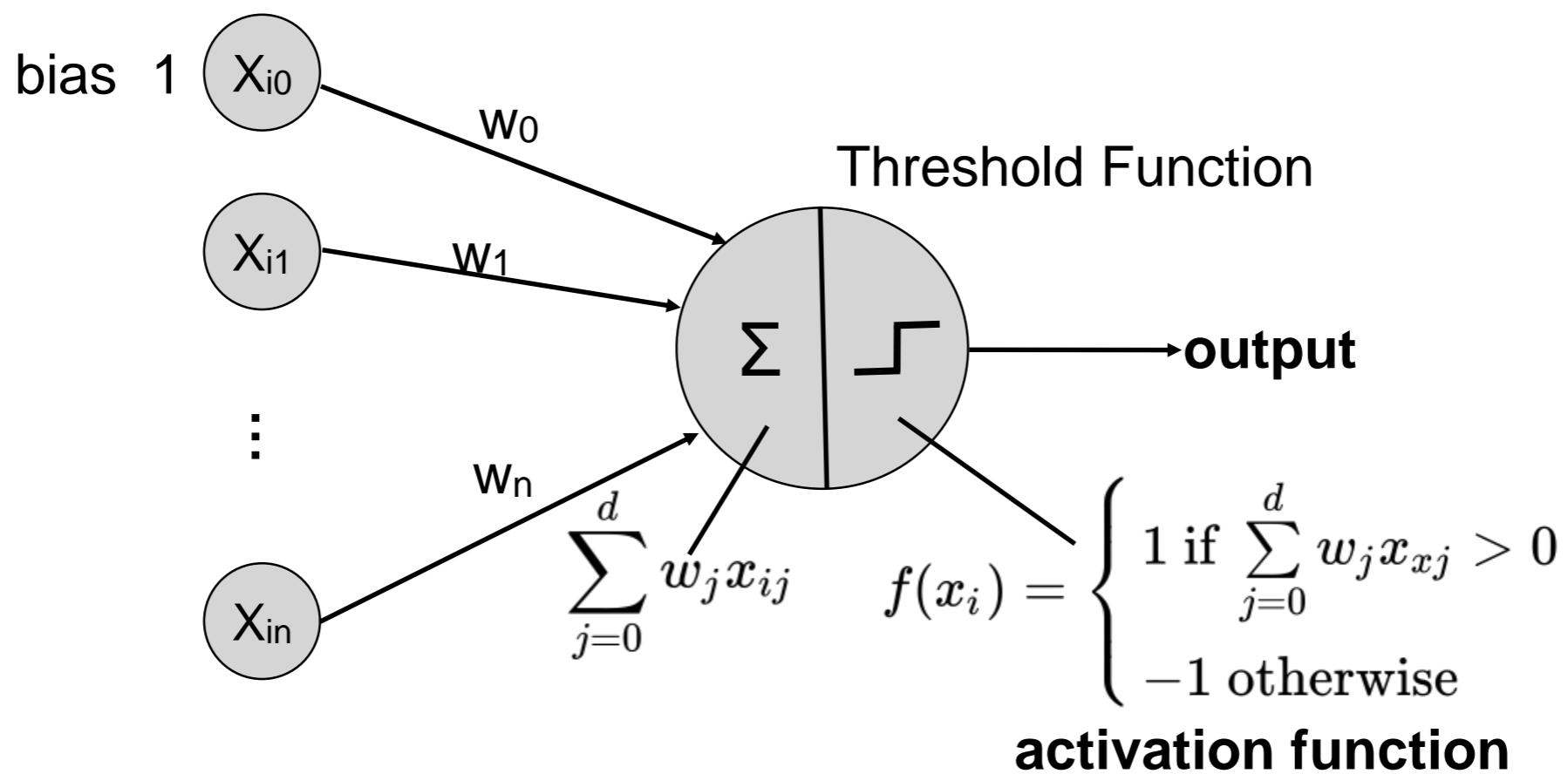
Preventing Overfitting

- Solutions: Simpler models.
 - Reduce the number of features (feature selection).
 - Model selection.
 - Regularization.
 - Cross validation.
- However: Adding wrong assumptions (bias) to the training algorithm can lead to **underfitting!**

Goodness of Fit



Linear Model



$$f(x_i) = sign\left(\sum_{j=0}^d w_j x_{ij}\right)$$

Linear Models

- We have chosen a function class (linear separators).
 - Specified by parameter \mathbf{w} .
- Need to estimate \mathbf{w} on the basis of the training set.
- What loss should we use? One option: minimize classification error:

$$\text{loss}(y_i, h(x_i)) = \begin{cases} 1 & \text{if } \text{sign}(h(x_i)) \neq \text{sign}(y_i) \\ 0 & \text{otherwise.} \end{cases}$$

Perceptron Learning

- Problem: Threshold function is not differentiable, so we cannot find a closed-form solution or apply gradient descent.
- Instead use iterative perceptron learning algorithm:
 - Start with arbitrary hyperplane.
 - Adjust it using the training data.
 - Update rule: $w_j \leftarrow w_j + (y - h_{\mathbf{w}}(\mathbf{x})) \times x_j$
- **Perceptron Convergence Theorem** states that any linear function can be learned using this algorithm in a finite number of iterations.

Perceptron Learning Algorithm

Input: Training examples $(x_1, y_1), \dots, (x_n, y_n)$

Output: A perceptron defined by (w_0, w_1, \dots, w_d)

Initialize $w_j \leftarrow 0$, for $j=0 \dots d$

while not converged:

shuffle training examples.

for each training example (x_i, y_i) :

if output - target $\neq 0$: #(output and prediction do not match)

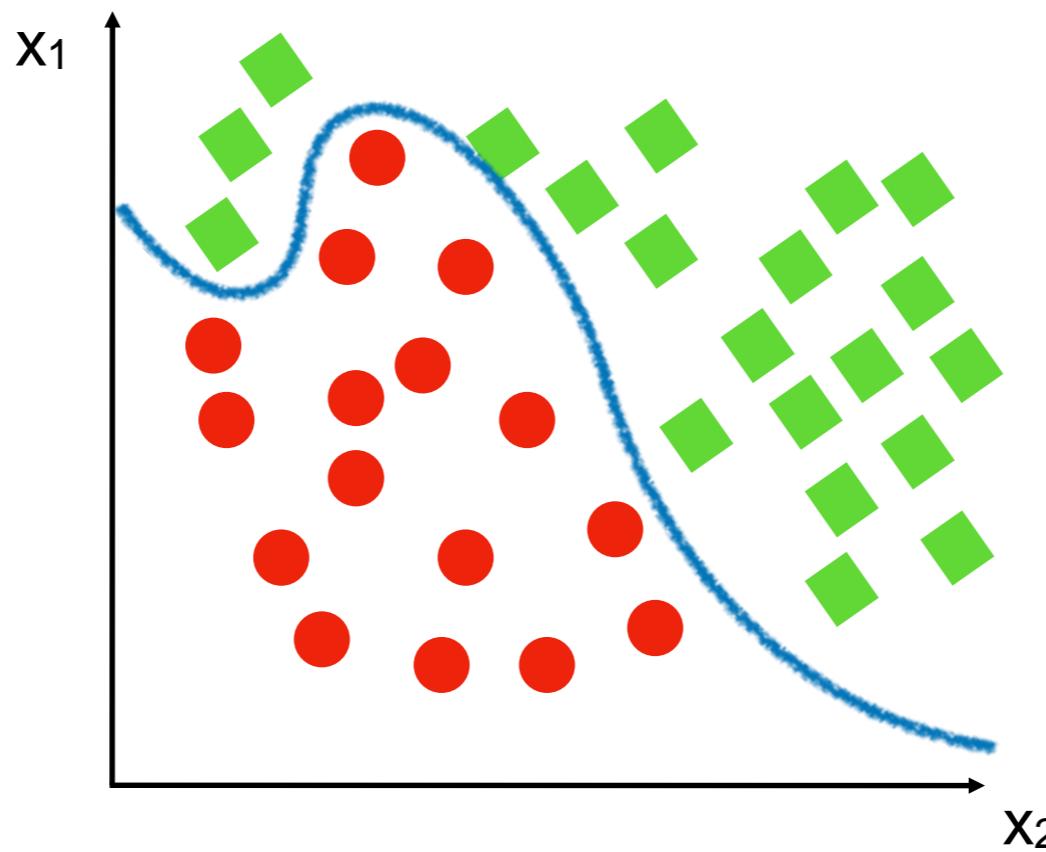
for each weight w_j :

$$w_j \leftarrow w_j + (y - h_{\mathbf{w}}(\mathbf{x})) \times x_j$$

"convergence" means that the weights don't change for one entire iteration through the training data.

Perceptron

- Simple learning algorithm. Guaranteed to converge after a finite number of steps.
 - But **only** if the data is linearly separable.



perceptron cannot learn this

Feature Functions

- In NLP we often need to make multi-class decisions.
Linear models provide only binary decisions.
- Use a feature function $\phi(x, y)$ where x is an input object
and y is a possible output.
- The values of ϕ are d -dimensional vectors.

$$\phi(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$$

Log-Linear Model

(a.k.a. "Maximum Entropy Models")

- Define conditional probability $P(y|x)$

$$P(y|x; \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot \phi(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{w} \cdot \phi(x, y'))}$$

- $\exp(z) = e^z$ is positive for any z .
- $\sum_y P(y|x; \mathbf{w}) = 1$
- But how should we estimate \mathbf{w} ?

Log-Likelihood

- Define the log-likelihood of some model \mathbf{w} on the training data $(x_1, y_1), \dots, (x_n, y_n)$ as

$$LL(\mathbf{w}) = \sum_{i=1}^n \log P(y_i | x_i; \mathbf{w})$$

- We want to compute the maximum likelihood

$$LL^*(\mathbf{w}) = \arg \max_{\mathbf{w}} \sum_{i=1}^n \log P(y_i | x_i; \mathbf{w})$$

- Unfortunately, there is no general analytical solution. Can use gradient-based optimization.

Simple Gradient Ascent

Initialize $\mathbf{w} \leftarrow$ any setting in the parameter (weight) space
for a set number of iterations T:

for each w_i in \mathbf{w} :

$$w'_i \leftarrow w_i + \alpha \frac{\partial}{\partial w_i} LL(\mathbf{w})$$

update each w_i to w'_i

- Follow the gradients (partial derivatives) to find a parameter setting that maximizes $LL(\mathbf{w})$
- $\alpha > 0$ is the **learning rate** or **step size**.

Partial Derivative of the Log Likelihood

$$\frac{\partial}{\partial_i} LL(\mathbf{w}) = \frac{\partial}{\partial_i} \sum_{i=1}^n \log P(y_i | x_i; \mathbf{w})$$

$$= \sum_i \phi_j(x_i, y_i) - \sum_i \sum_y P(y|x_i; \mathbf{w}) \phi_j(x_i, y)$$

Regularization

- Problem: Parameter estimation can overfit the training data.
- Can include a regularization term. For example L₂ regularizer:

$$LL(\mathbf{w}) = \sum_{i=1}^n \log P(y_i | x_i; \mathbf{w}) - \frac{\lambda}{2} |\mathbf{w}|^2$$

$$LL(\mathbf{w}) = \sum_{i=1}^n \log P(y_i | x_i; \mathbf{w}) - \frac{\lambda}{2} |\mathbf{w}|^2$$

- $\lambda > 0$ controls the strength of the regularization.
- Since we are maximizing $\mathbf{w}^* = \arg \max_{\mathbf{w}} LL(\mathbf{w})$, there is now a trade-off between fit and model 'complexity'.

POS Tagging with Log-Linear Models

- Previously we used a generative model (HMM) for POS tagging.

- Now we want to use a discriminative model for

$$P(t_1, t_2, \dots, t_n | w_1, w_2, \dots, w_n)$$

$$= \prod_{i=1}^m P(t_i | t_1, \dots, t_{i-1}, x_1, \dots, x_m)$$

- Next tag is conditioned on previous tag sequence and all observed words.

Maximum Entropy Markov Models (MEMM)

- Make an independence assumption (similar to HMM):

$$\begin{aligned} & P(t_1, t_2, \dots, t_n | w_1, w_2, \dots, w_n) \\ &= \prod_{i=1}^m P(t_i | t_1, \dots, t_{i-1}, w_1, \dots, w_m) \\ &= \prod_{i=1}^m P(t_i | t_{i-1}, w_1, \dots, w_m) \end{aligned}$$

- Probability only depends on the previous tag.

MEMMs

$$\prod_{i=1}^m P(t_i | t_{i-1}, w_1, \dots, w_m)$$

- Model each term using a log-linear model

$$P(t_i | t_{i-1}, w_1, \dots, w_m) = \frac{\exp(\mathbf{w} \cdot \phi(w_1, \dots, w_m, i, t_{i-1}, t_i))}{\sum_{t'} \exp(\mathbf{w} \cdot \phi(w_1, \dots, w_m, i, t_{i-1}, t'))}$$

- ϕ is a feature function defined over:
 - the observed words w_1, \dots, w_m
 - the position of the current word
 - the previous tag t_{i-1}
 - the suggested tag for the current word t_i
- t' is a variable ranging over all possible tags.

MEMMs

$$P(t_i | t_{i-1}, w_1, \dots, w_m) = \frac{\exp(\mathbf{w} \cdot \phi(w_1, \dots, w_m, i, t_{i-1}, t_i))}{\sum_{t'} \exp(\mathbf{w} \cdot \phi(w_1, \dots, w_m, i, t_{i-1}, t'))}$$

- Training: same as any log-linear model.
- Decoding: Need to find $\arg \max_{t_1, \dots, t_m} P(t_i, \dots, t_m | t_{i-1}, w_1, \dots, w_m)$
 - Can use Viterbi algorithm!

Feature Function

(Ratnaparkhi, 1996)

- $\phi(w_1, \dots, w_m, i, t_{i-1}, t_i)$ is a feature vector of length d.
- $(w_i, t_i), (w_{i-1}, t_i), (w_{i-2}, t_i), (w_{i+1}, t_i), (w_{i+2}, t_i)$
- (t_{i-1}, t_i)
- $(w_i \text{ contains numbers}, t_i),$
 $(w_i \text{ contains uppercase characters}, t_i)$
 $(w_i \text{ contains a hyphen}, t_i)$
- $(\text{prefix}_1 \text{ of } w_i, t_i), (\text{prefix}_2 \text{ of } w_i, t_i), (\text{prefix}_3 \text{ of } w_i, t_i), (\text{prefix}_4 \text{ of } w_i, t_i)$
 $(\text{suffix}_1 \text{ of } w_i, t_i), (\text{suffix}_2 \text{ of } w_i, t_i), (\text{suffix}_3 \text{ of } w_i, t_i), (\text{suffix}_4 \text{ of } w_i, t_i)$

Feature Example

The stories about well-heeled communities and developers ...

DT NNS IN ??

- (*well-heeled*,JJ), (*about*,JJ), (*stories*,JJ), (*communities*, JJ), (*and*,JJ)
- (IN,JJ)
- (w_i contains a hyphen, JJ)
- (*w*,JJ), (*we*,JJ), (*weI*,JJ), (*well*, JJ)
 (*d*,JJ), (*ed*,JJ), (*Ied*,JJ), (*eIed*, JJ)

Natural Language Processing

Lecture 14:
Machine Learning: Feed-forward Neural Networks,
Autoencoders/embeddings, Dense networks

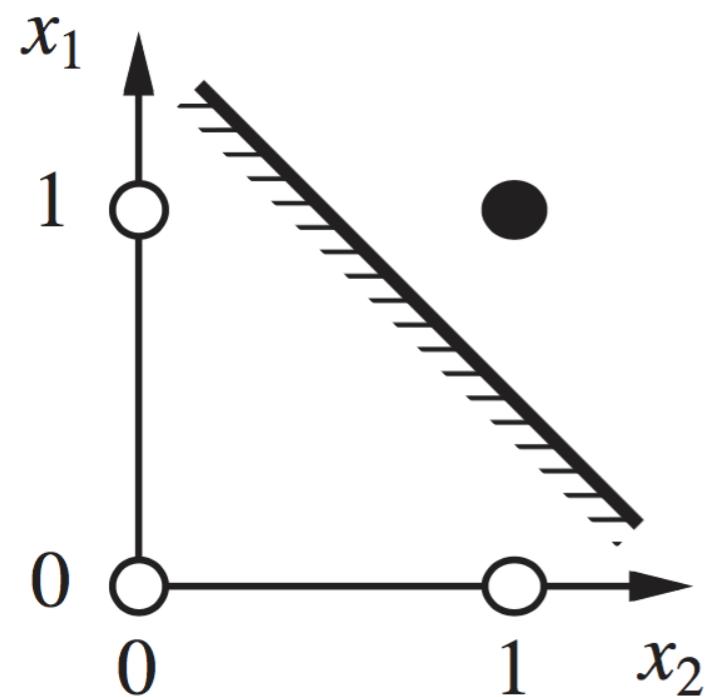
4/3/2019

COMS W4705
Yassine Benajiba

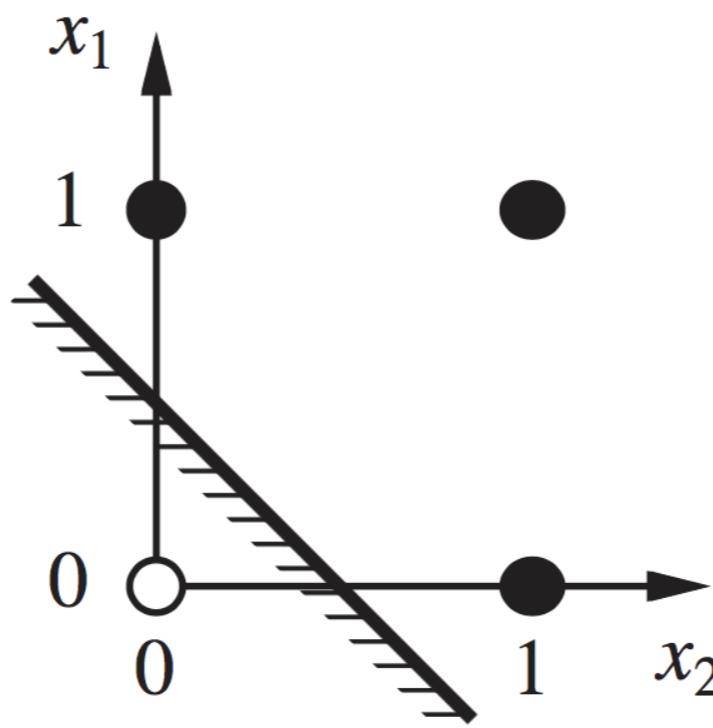
Perceptron Expressiveness

- Simple perceptron learning algorithm, starts with an arbitrary hyperplane and adjusts it using the training data.
 - Step function is not differentiable, so no closed-form solution.
- Perceptron produces a linear separator.
 - Can only learn linearly separable patterns.
- Can represent boolean functions like **and**, **or**, **not** but not the **xor** function.

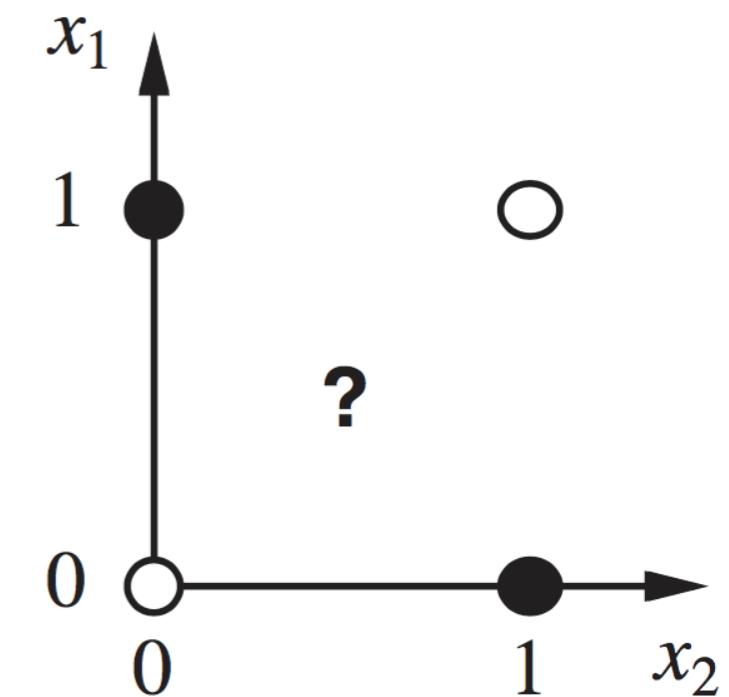
The problem with xor



(a) x_1 **and** x_2

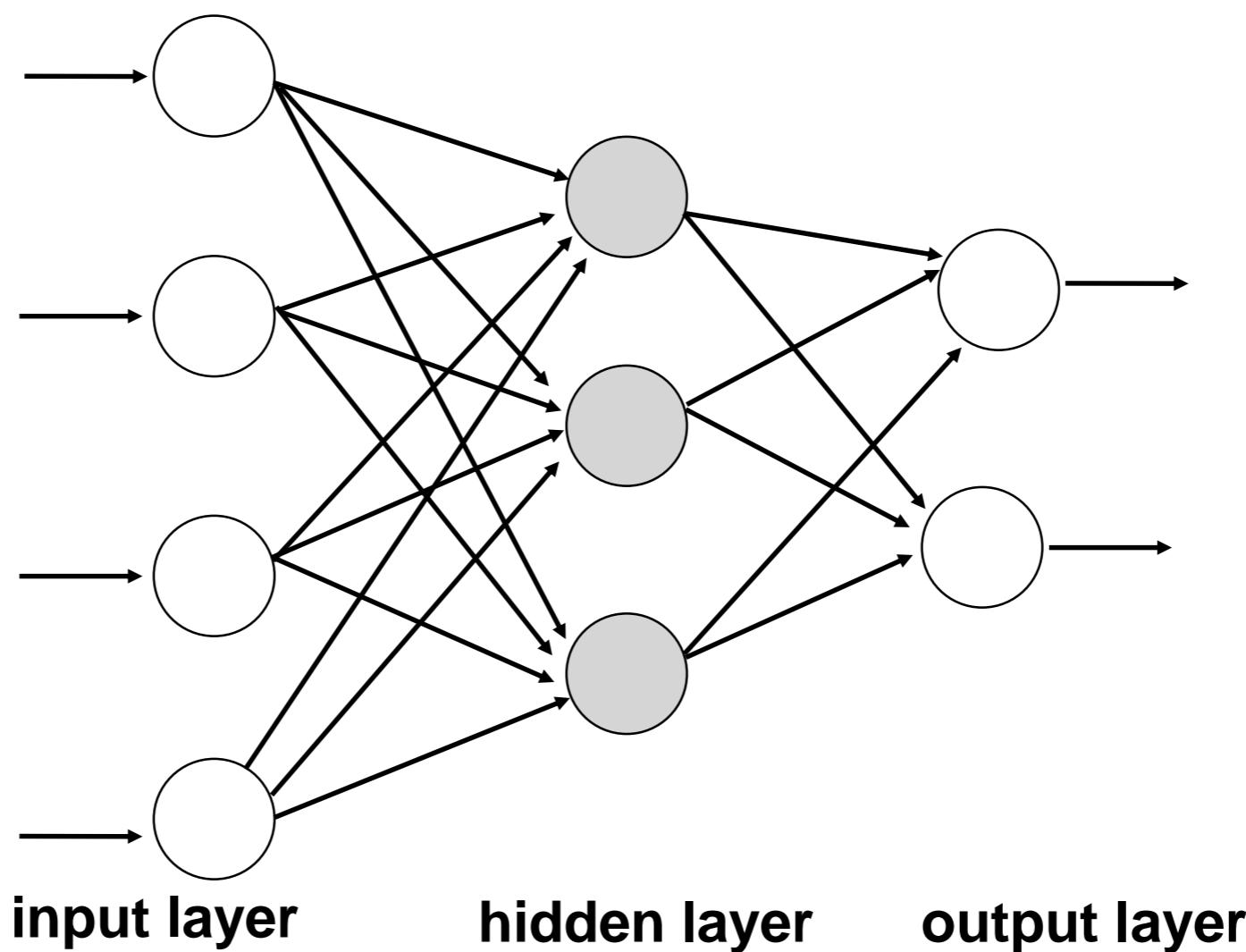


(b) x_1 **or** x_2



(c) x_1 **xor** x_2

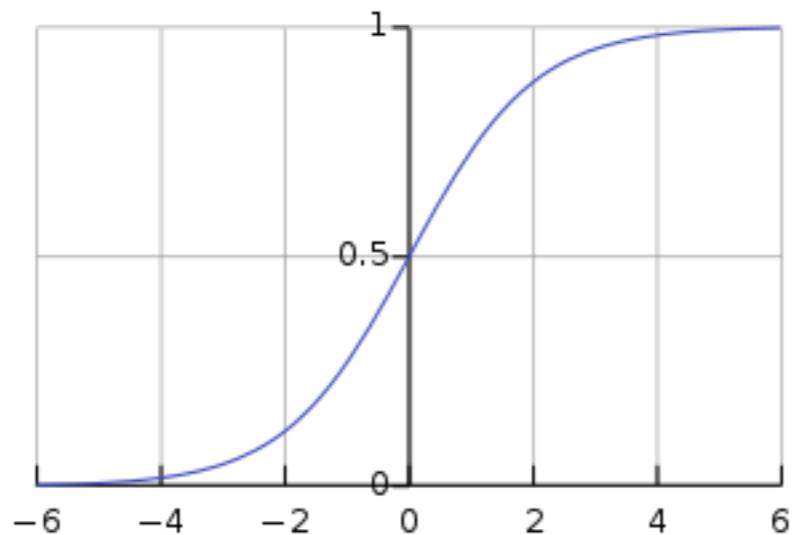
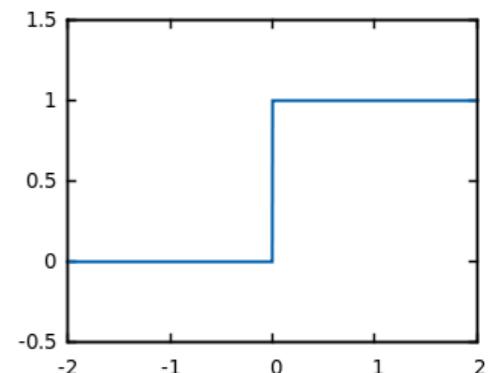
Multi-Layer Neural Networks



- Basic idea: represent any (non-linear) function as a composition of soft-threshold functions. This is a form of non-linear regression.
- Lippmann 1987: Two hidden layers suffice to represent any arbitrary region (provided enough neurons), even discontinuous functions!

Activation Functions

- One problem with perceptrons is that the **threshold function (step function)** is undifferentiable.
- It is therefore unsuitable for gradient descent.
- One alternative is the **sigmoid (logistic) function**:

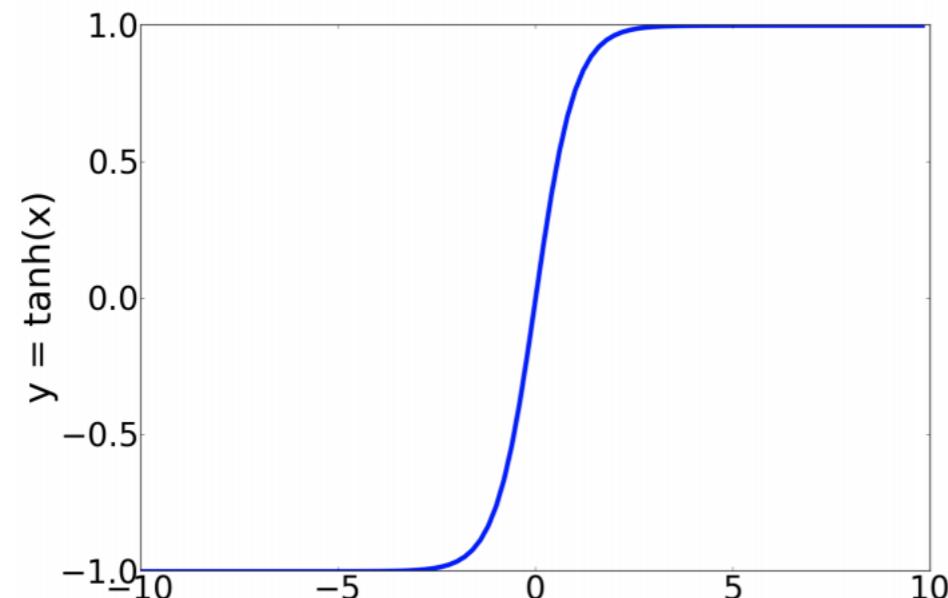


$$g(z) = \frac{1}{1 + e^{-z}}$$

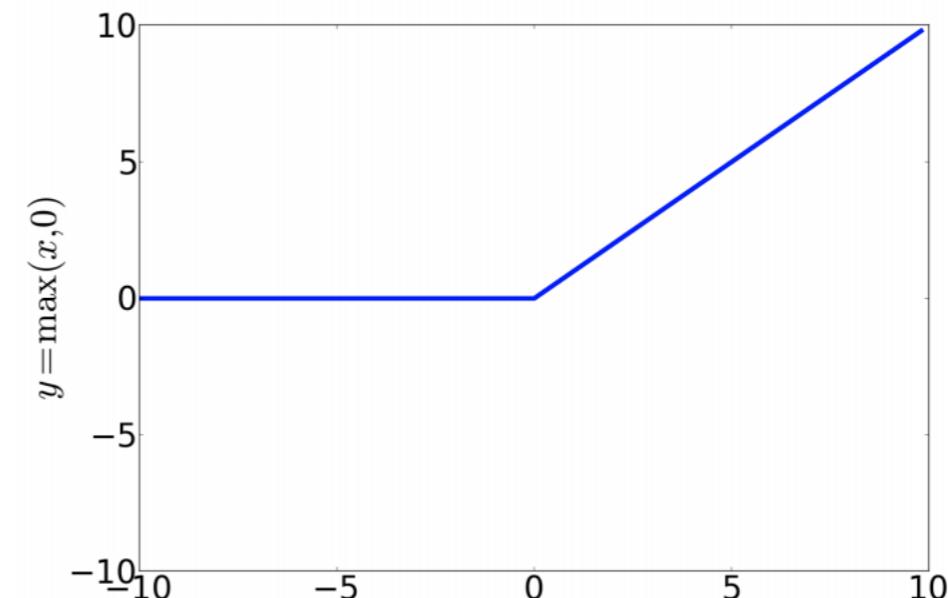
$$g(z) = 0 \text{ if } z \rightarrow -\infty$$
$$g(z) = 1 \text{ if } z \rightarrow \infty$$

Activation Functions

- Two other popular activation functions:



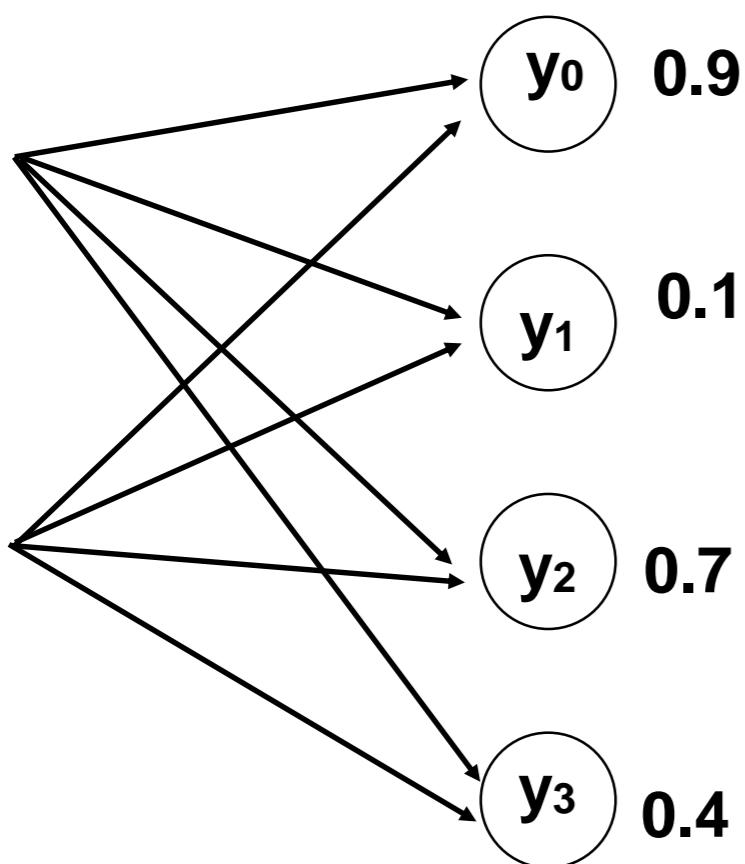
$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$



$$relu(z) = \max(z, 0)$$

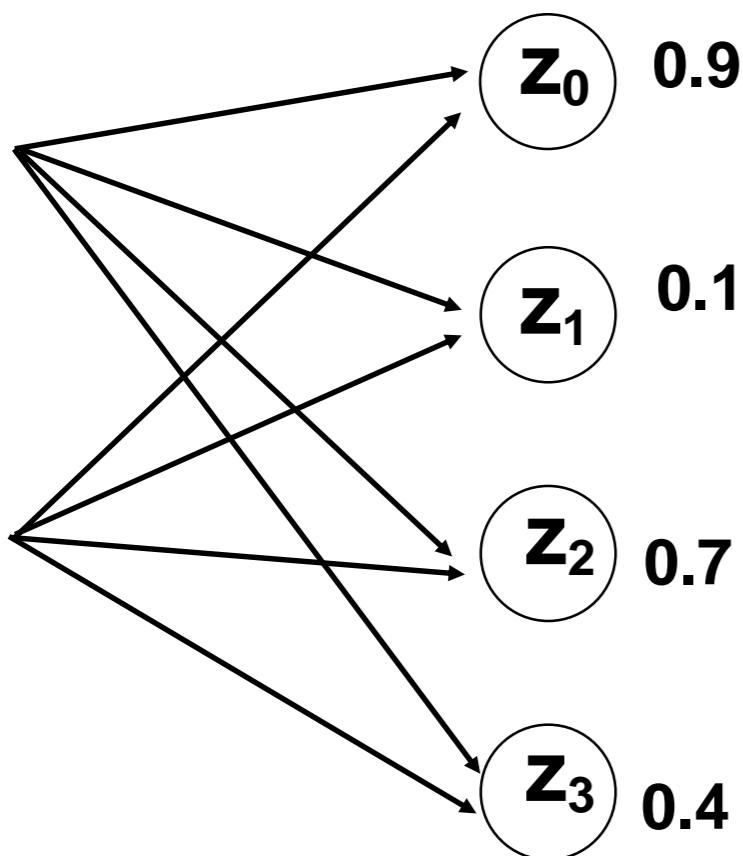
Output Representation

- Many NLP Problems are multi-class classification problems.
- Each output neuron represents one class. Predict the class with the highest activation.



Softmax

- We often want the activation at the output layer to represent probabilities.
- Normalize activation of each output unit by the sum of all output activations (as in log-linear models).



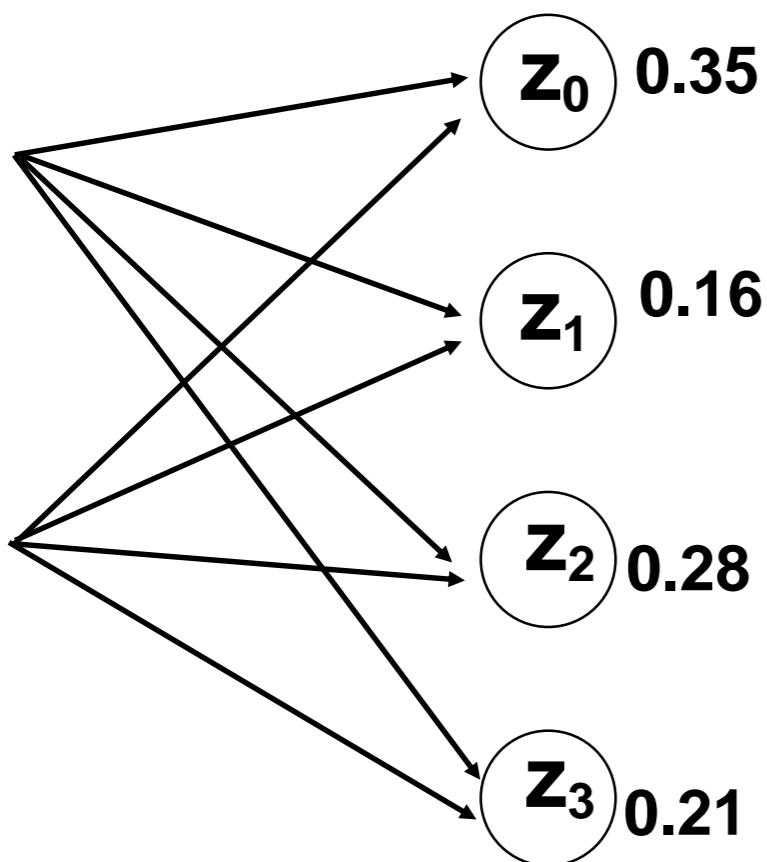
$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^k \exp(z_j)}$$

The network computes a probability

$$P(c_i | \mathbf{x}; \mathbf{w})$$

Softmax

- We often want the activation at the output layer to represent probabilities.
- Normalize activation of each output unit by the sum of all output activations (as in log-linear models).



$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^k \exp(z_j)}$$

The network computes a probability

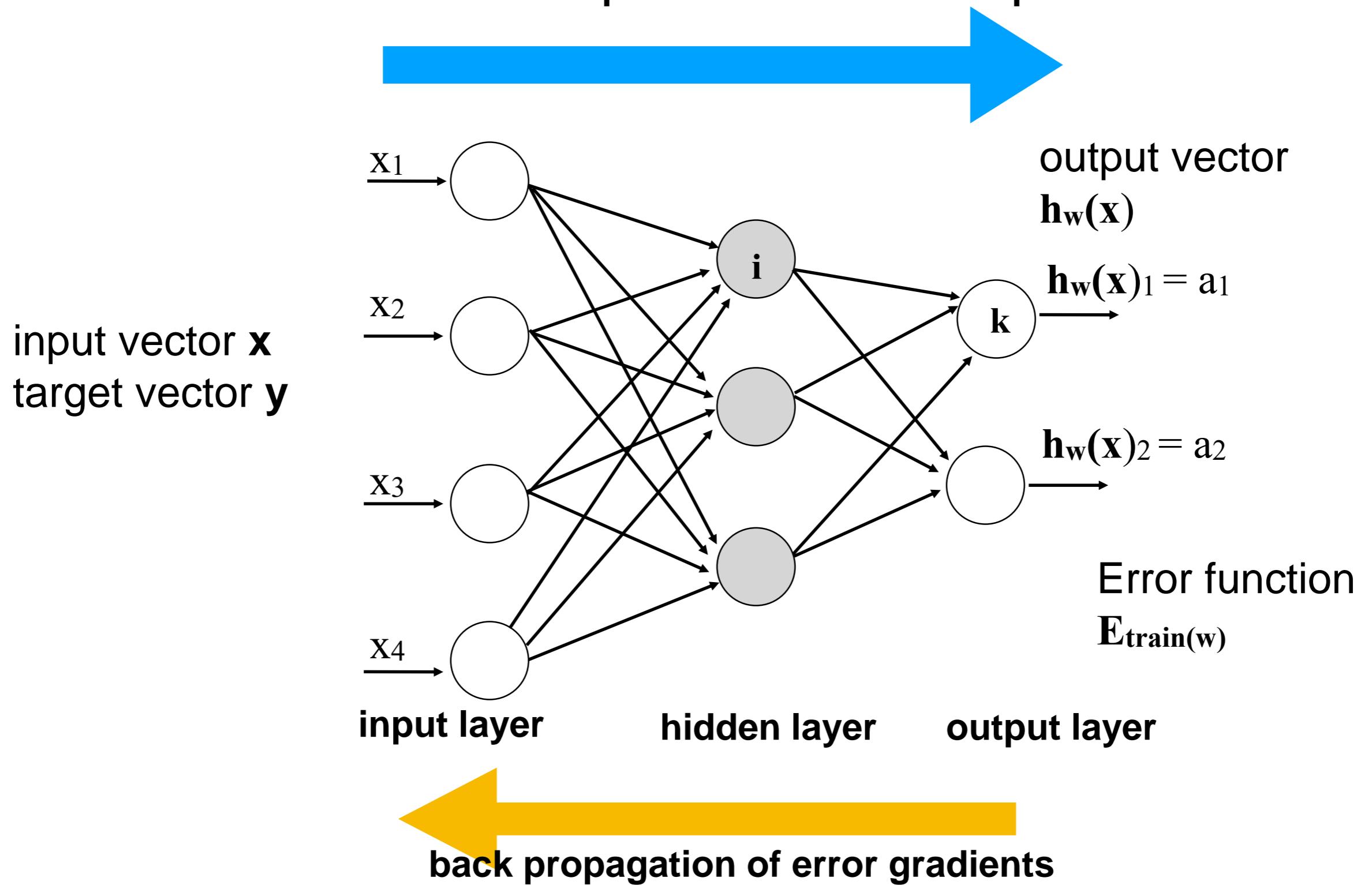
$$P(c_i | \mathbf{x}; \mathbf{w})$$

Learning in Multi-Layer Neural Networks

- Network structure is fixed, but we want to train the weights. Assume **feed-forward** neural networks: no connections that are loops.
- **Backpropagation Algorithm:**
 - Given current weights, get network output and compute loss function (assume multiple outputs / a vector of outputs).
 - Can use gradient descent to update weights and minimize loss.
 - Problem: We only know how to do this for the last layer!
 - Idea: Propagate error backwards through the network.

Backpropagation

feed-forward computation of network outputs



Negative Log-Likelihood

(also known as cross-entropy)

- Assume target output is a one-hot vector and $c(y)$ is the target class for target \mathbf{y} .
- Compute the negative log-likelihood for a single example

$$Loss(\mathbf{y}, h_{\mathbf{w}}(x)) = -\log P(c(\mathbf{y})|\mathbf{x}; \mathbf{w})$$

- Empirical error for the entire training data:

$$E_{train}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N -\log P(c(\mathbf{y}^{(i)})|\mathbf{x}^i; \mathbf{w})$$

Stochastic Gradient Descent (for a single unit)

- Goal: Learn parameters that minimize the empirical error.

Randomly initialize w

for a set number of iterations T:

shuffle training data $\mathcal{D} = (x^{(j)}, y^{(j)})|_{j=1}^n$

for $j = 1 \dots N$:

for each w_i (all weights in the network):

$$w_i \leftarrow w_i - \eta \frac{\partial}{\partial w_i} Loss(y^{(j)}, h_w(x^{(j)}))$$

- η is the learning rate.
- It often makes sense to compute the gradient over batches of examples, instead of just one ("mini-batch").

Backpropgation

- Simplified multi-layer case (a single unit per layer):



- Stochastic Gradient Descent should perform the following update:

$$w_2 \leftarrow w_2 - \eta \frac{\partial Loss(y, f(g(x)))}{\partial w_2}$$

$$w_1 \leftarrow w_1 - \eta \frac{\partial Loss(y, f(g(x)))}{\partial w_1}$$

- Problem: How do we compute the gradient for parameters w_1 and w_2 ?

Chain Rule of Calculus

- To compute gradients for hidden units, we need to apply the chain rule of calculus:

The derivative of $f(g(x))$ is

$$\frac{df(g(x))}{dx} = \frac{df(g(x))}{dg(x)} \cdot \frac{dg(x)}{dx}$$

Backpropagation



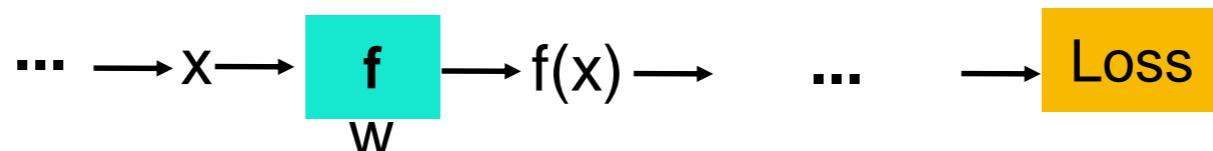
$$\frac{\partial Loss}{w_2} = \left(\frac{\partial Loss}{g(f(x))} \right) \left(\frac{\partial g(f(x))}{\partial w_2} \right)$$

$$\frac{\partial Loss}{w_1} = \left(\frac{\partial Loss}{\partial f(x)} \right) \left(\frac{\partial f(x)}{\partial w_1} \right)$$

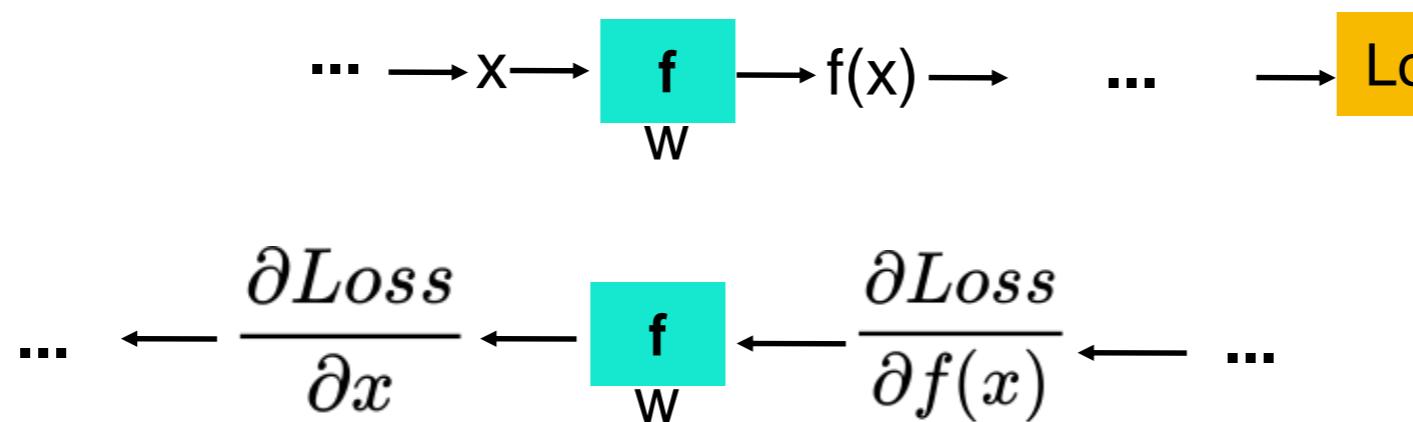
$$= \left(\frac{\partial Loss}{g(f(x)))} \right) \left(\frac{\partial g(f(x))}{\partial f(x)} \right) \left(\frac{\partial f(x)}{\partial w_1} \right)$$

Backpropagation

forward



backward



Assume we know

$$\frac{\partial \text{Loss}}{\partial f(x)}$$

We want to compute

$$\frac{\partial \text{Loss}}{\partial x}$$

to propagate it back.

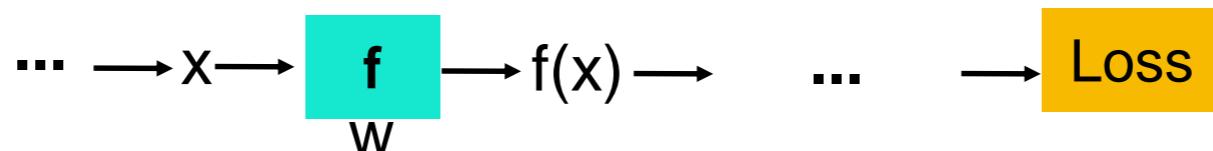
and

$$\frac{\partial \text{Loss}}{\partial w}$$

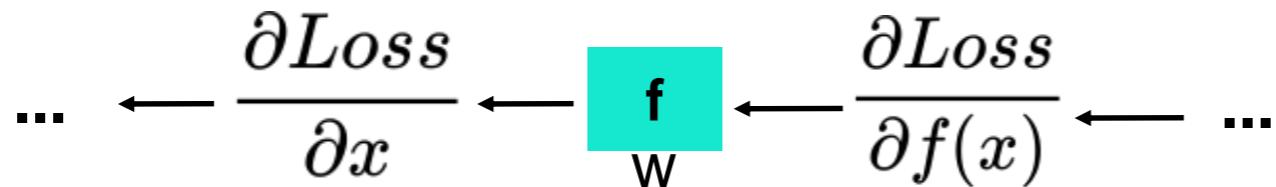
(for the weight update)

Backpropagation

forward



backward



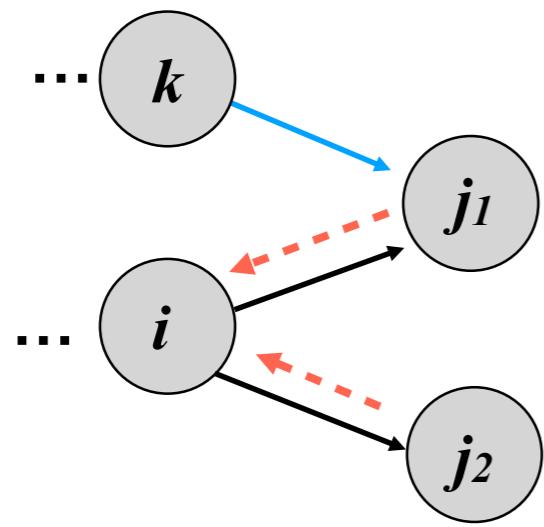
$$\frac{\partial Loss}{\partial x} = \left(\frac{\partial Loss}{\partial f(x)} \right) \left(\frac{\partial f(x)}{\partial x} \right)$$

$$\frac{\partial Loss}{\partial w} = \left(\frac{\partial Loss}{\partial f(x)} \right) \left(\frac{\partial f(x)}{\partial w} \right)$$

**to compute these
we have to know
the derivate of the
function f**

Backpropagation with Multiple Neurons

- Let $\Delta_j = \frac{\partial Loss}{\partial j}$ be the derivative of the loss w.r.t to the output of unit j .



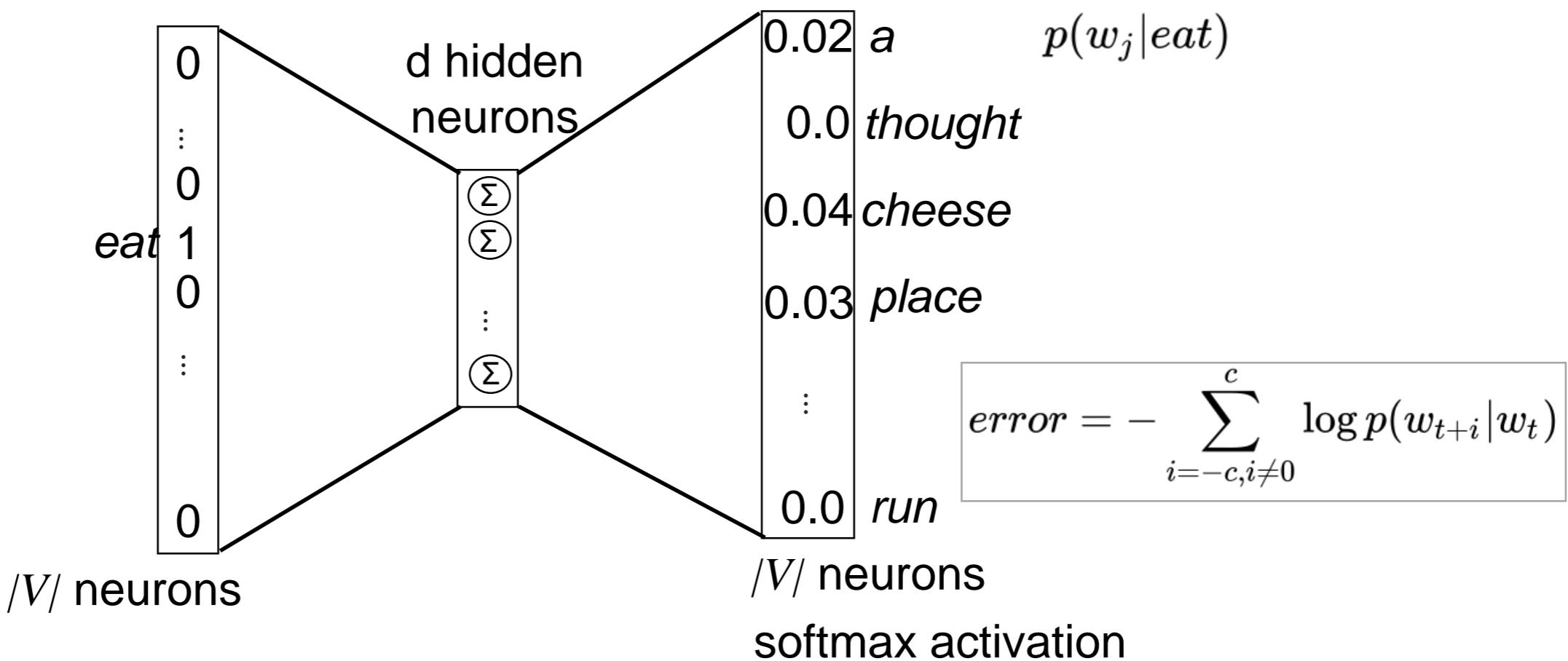
$$\begin{aligned}\Delta_i &= \frac{\partial Loss}{\partial i} = \sum_j \left(\frac{\partial Loss}{\partial j} \right) \left(\frac{\partial j}{\partial i} \right) \\ &= \sum_j \Delta_j \left(\frac{\partial j}{\partial i} \right)\end{aligned}$$

- The output of j is computed during the forward pass.

Autoencoders Embeddings

Skip-Gram Model

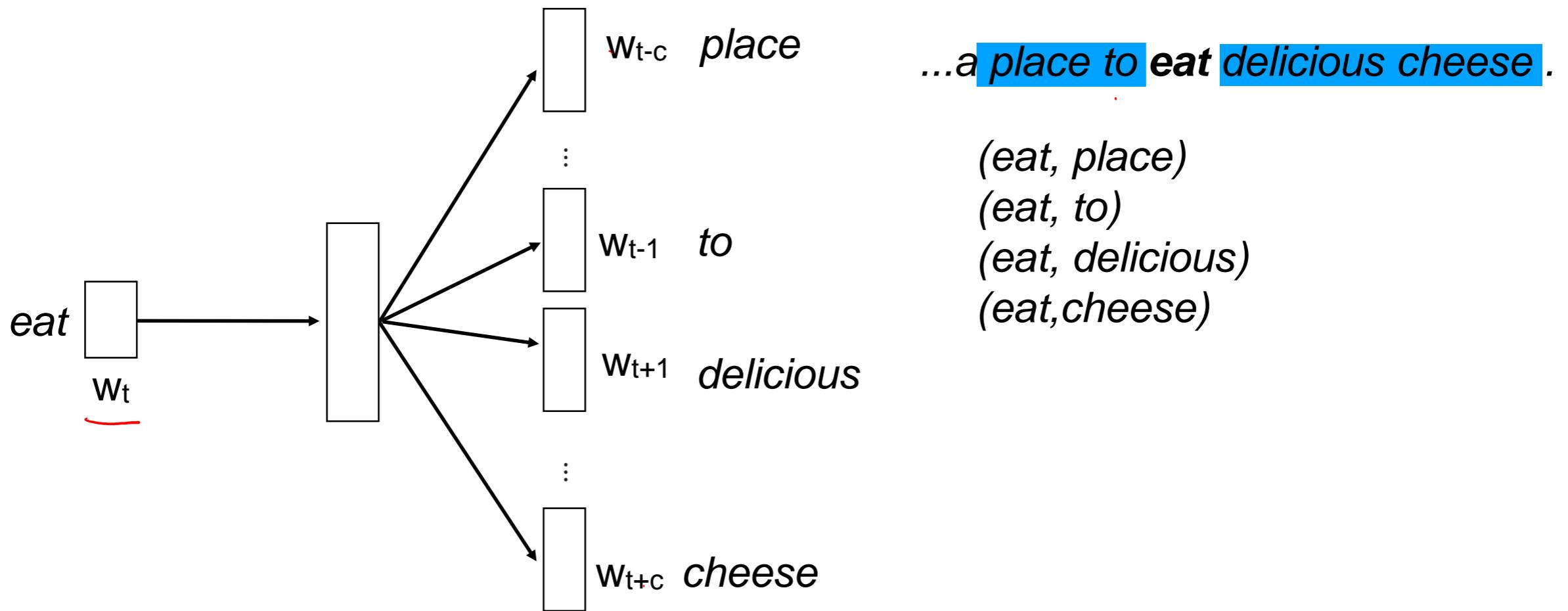
- Input:
A single word in one-hot representation.
- Output: probability to see any single word as a context word.



- Softmax function normalizes the activation of the output neurons to sum up to 1.0.

Skip-Gram Model

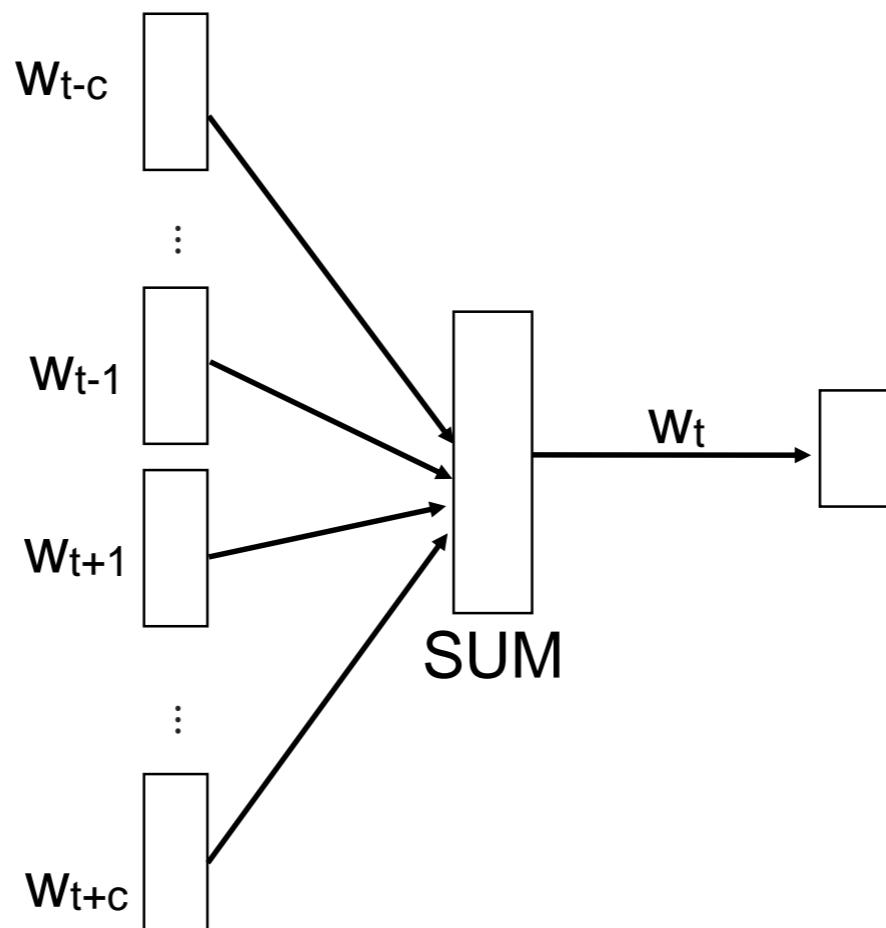
- Compute error with respect to each context word.



- Combine errors for each word, then use combined error to update weights using back-propagation.

$$\text{error} = - \sum_{i=-c, i \neq 0}^c \log p(w_{t+i} | w_t)$$

Continuous Bag-of-Words Model (CBOW)

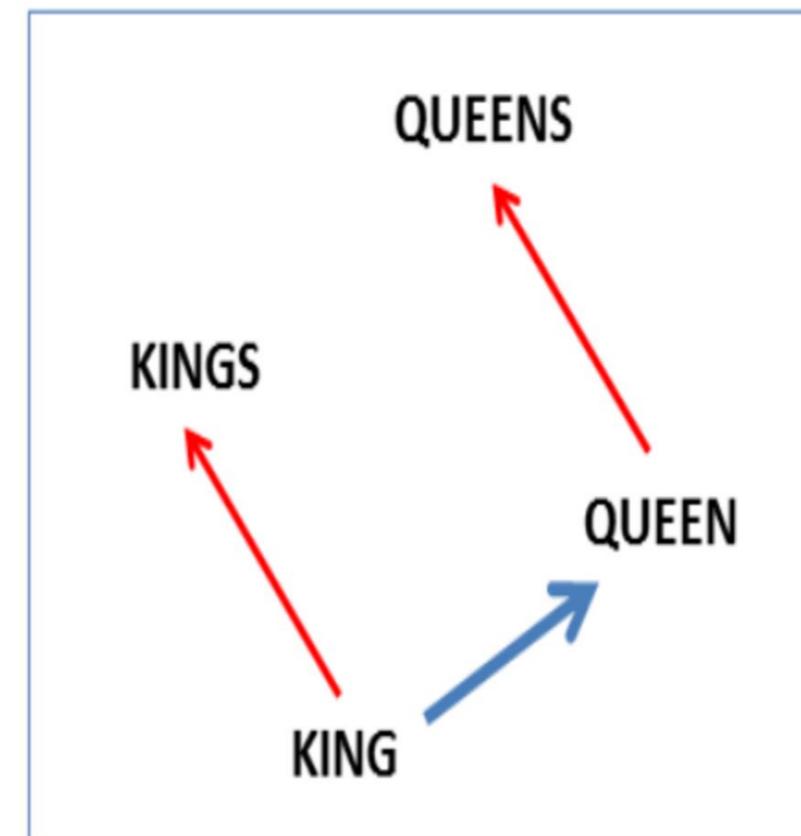
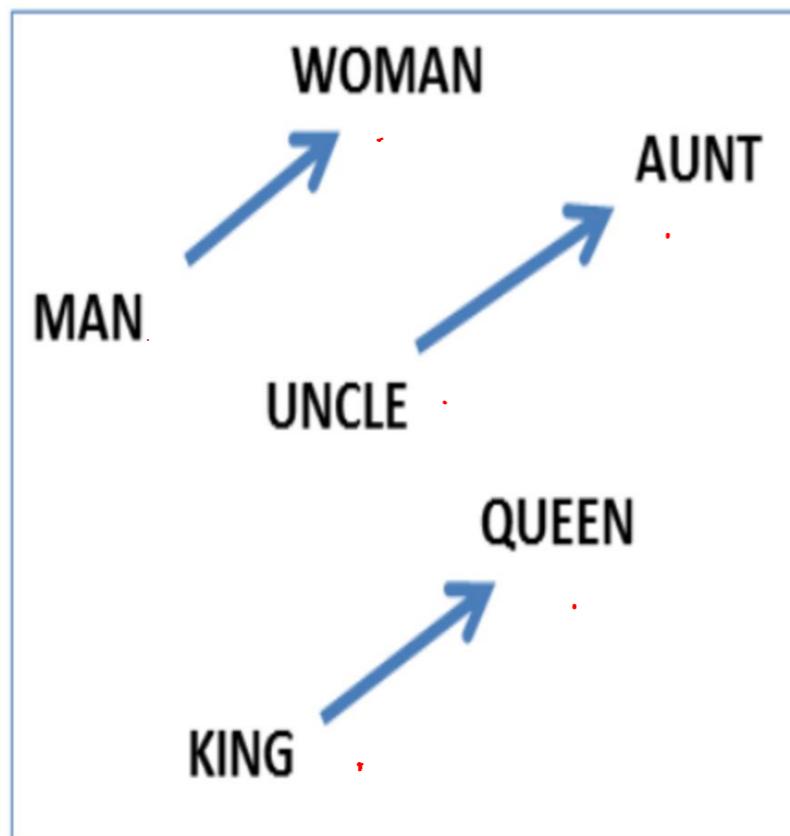


- Input: Context words. Averaged in the hidden layer.
- Output: Probability that each word is the target word.

Embeddings are Magic

(Mikolov 2016)

$$\text{vector('king')} - \text{vector('man')} + \text{vector('woman')} \approx \text{vector('queen')}$$



Application: Word Pair Relationships

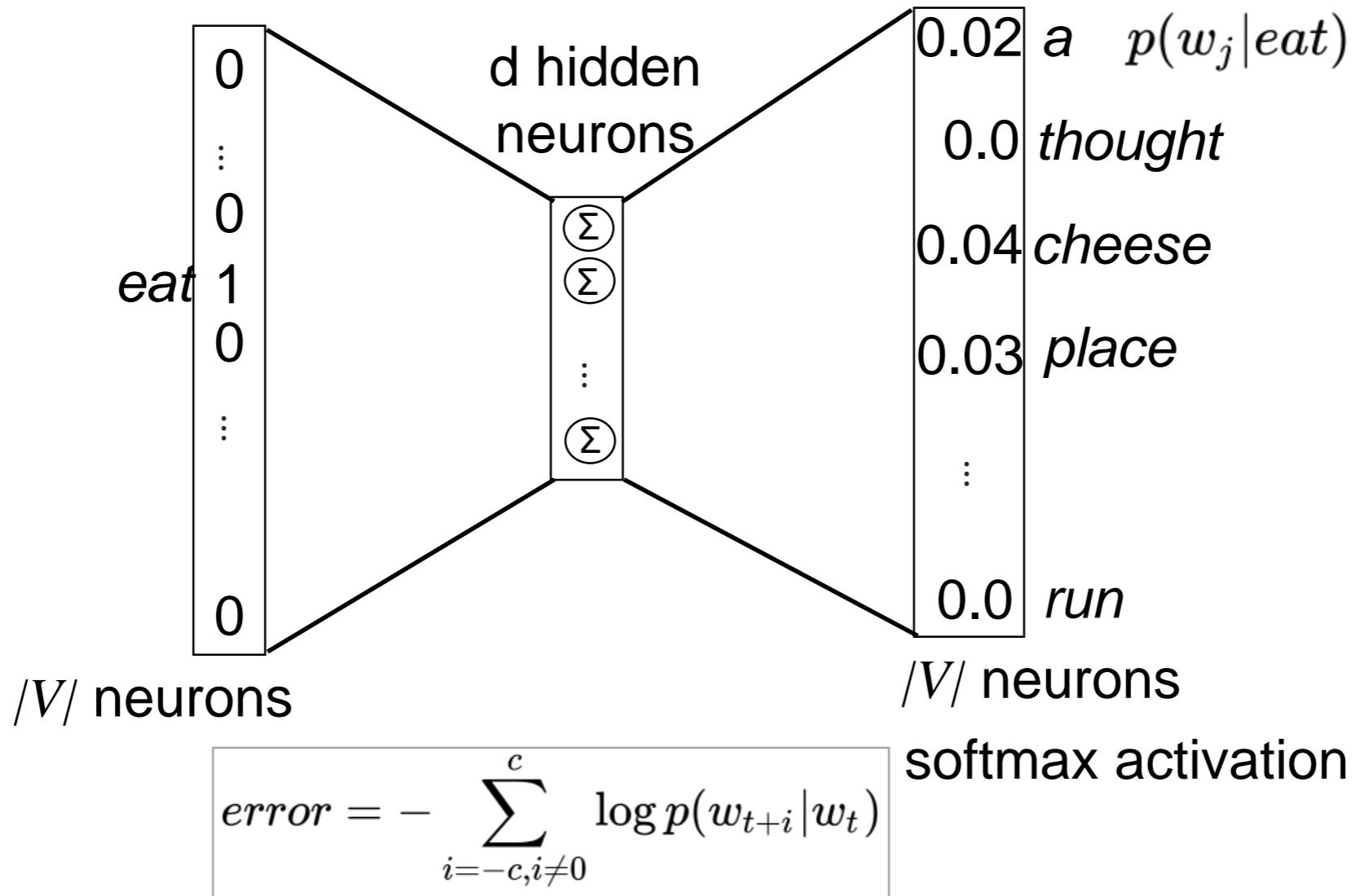
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Using Word Embeddings

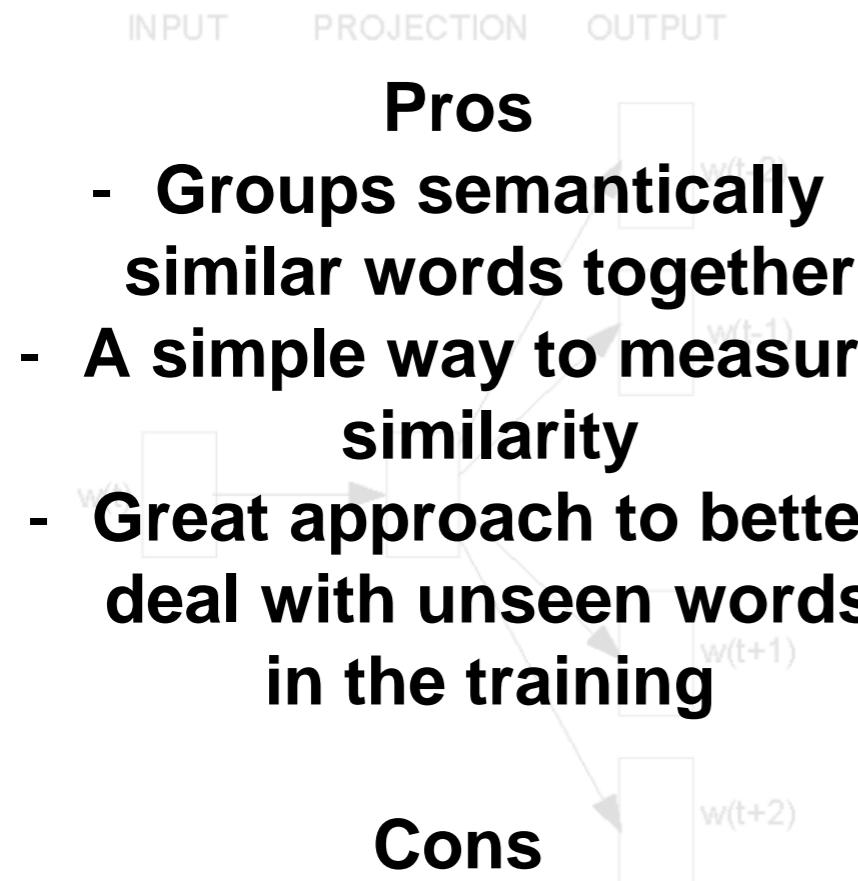
- Word2Vec:
 - <https://code.google.com/archive/p/word2vec/>
- GloVe: Global Vectors for Word Representation
 - <https://nlp.stanford.edu/projects/glove/>
- Can either use pre-trained word embeddings or train them on a large corpus.

Word embeddings



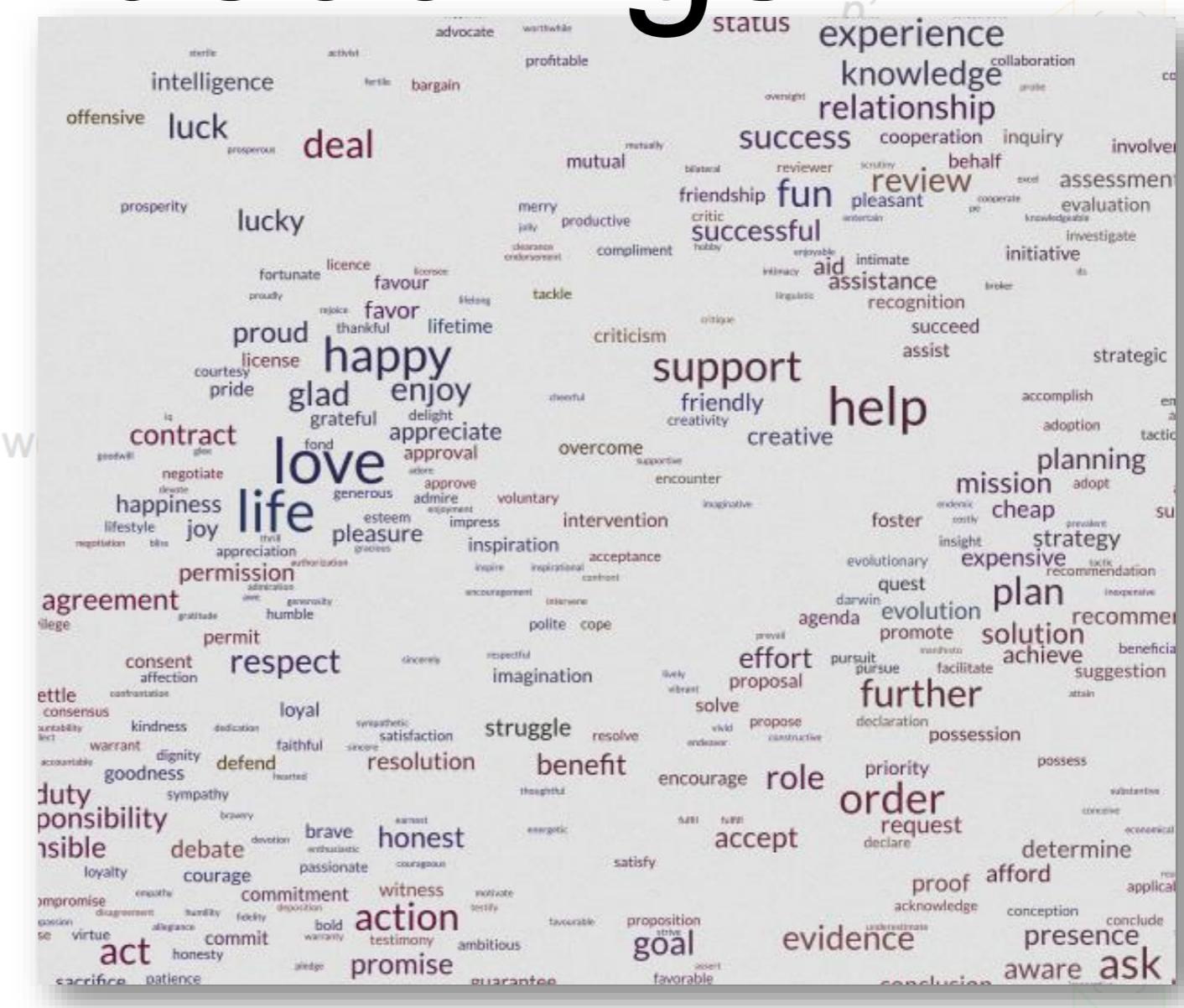
Word embeddings

w(t+j)
Where j is a
context



Cons

- Doesn't make a difference between function and content words
 - Only one representation for polysemous words
 - Non interpretable semantic dimensions



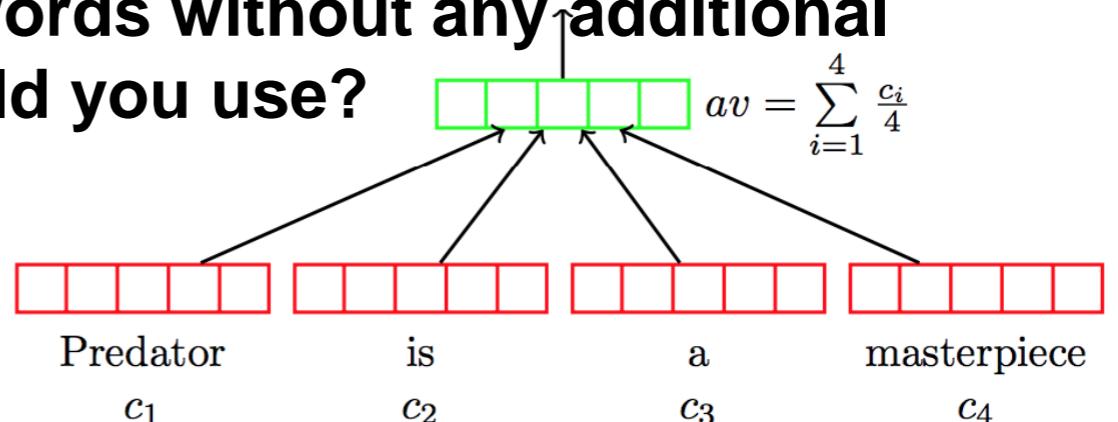
How can we build a sentence representation using word-level distributional representations?

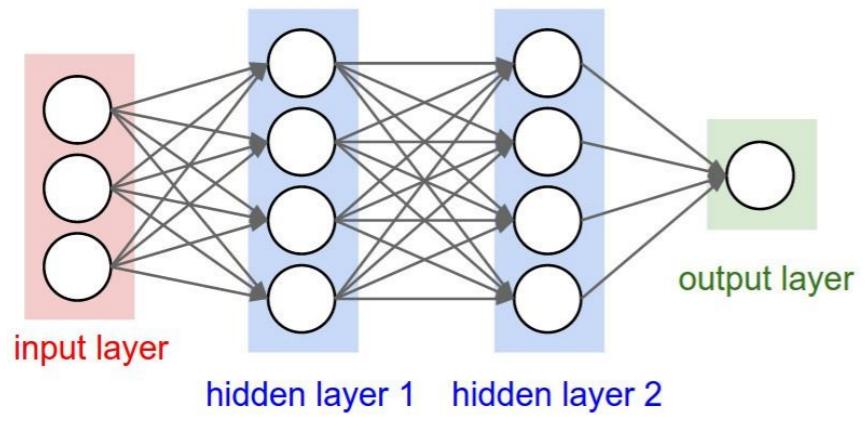
No learning

No learning

If you don't want to use any additional learning, you could just stop here and take the average of the embeddings as the representation of the sentence. What are the pros and cons of such an approach?

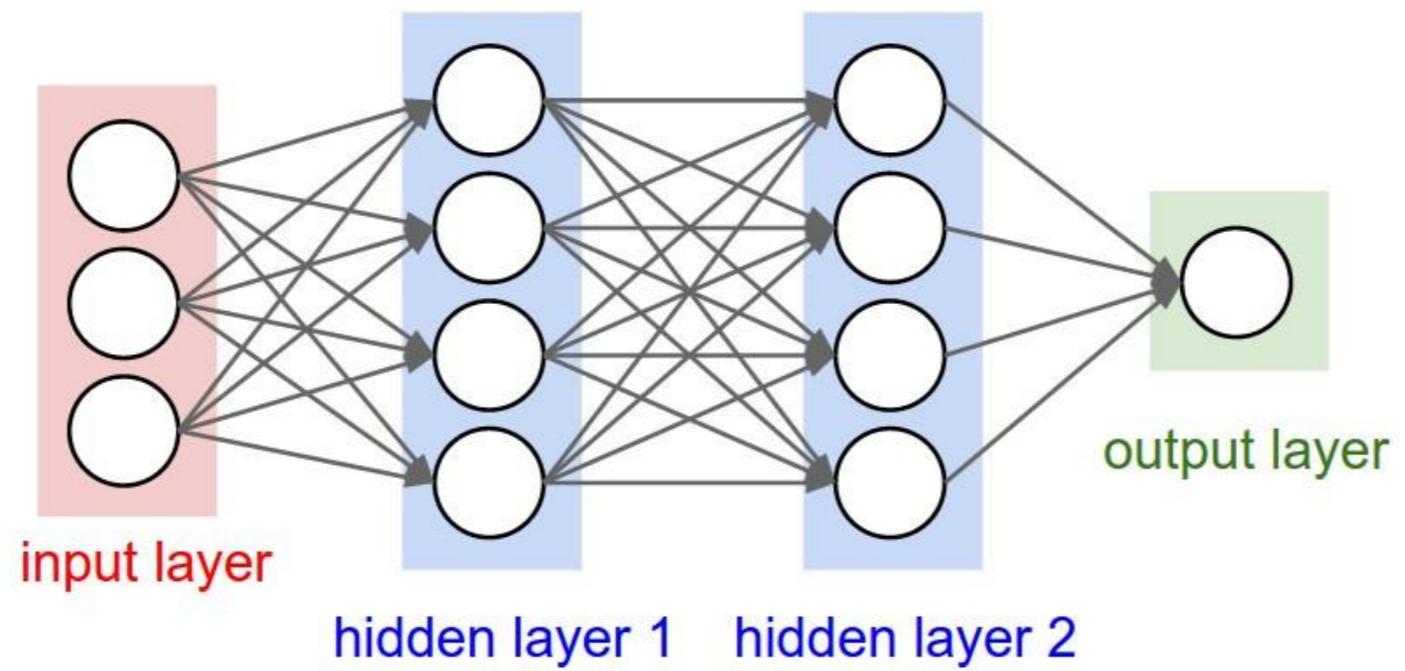
What if you wanted to weight the words without any additional learning, what could you use?





dense nets

dense nets



dense nets

Deep Averaging Networks

(DANs)

**Deep Unordered Composition Rivals Syntactic Methods
for Text Classification**

Mohit Iyyer,¹ Varun Manjunatha,¹ Jordan Boyd-Graber,² Hal Daumé III¹

¹University of Maryland, Department of Computer Science and UMIACS

²University of Colorado, Department of Computer Science

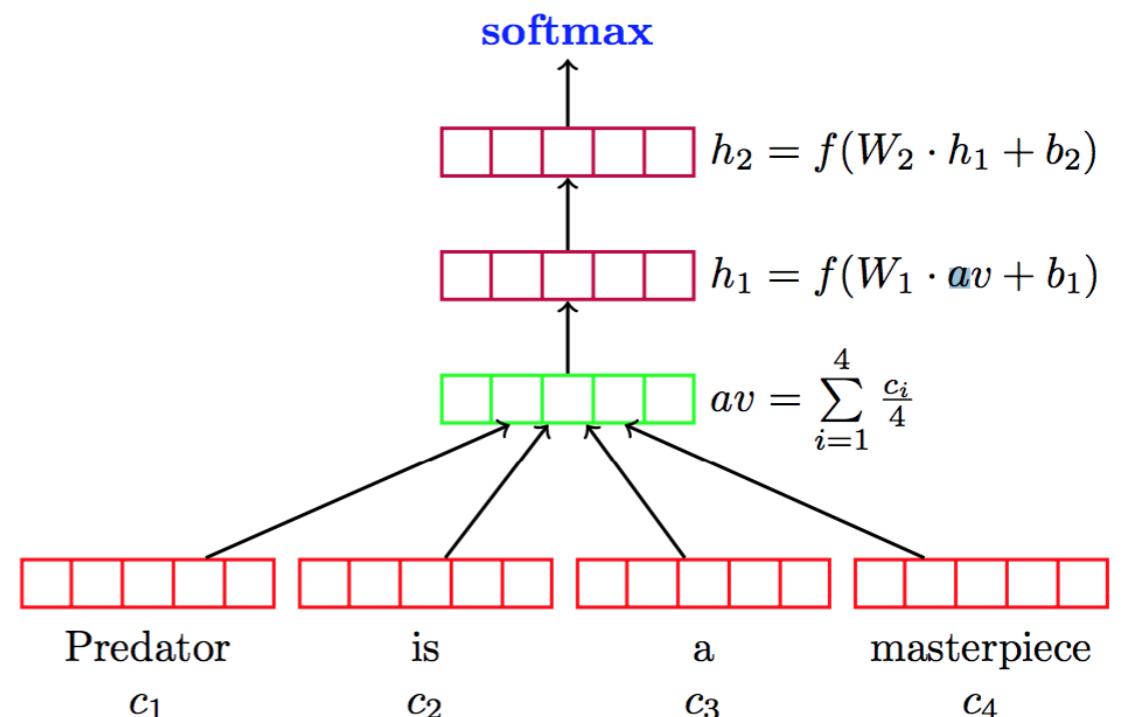
{miyyer, varunm, hal}@umiacs.umd.edu, Jordan.Boyd.Grabber@colorado.edu

https://cs.umd.edu/~miyyer/pubs/2015_acl_dan.pdf

dense nets

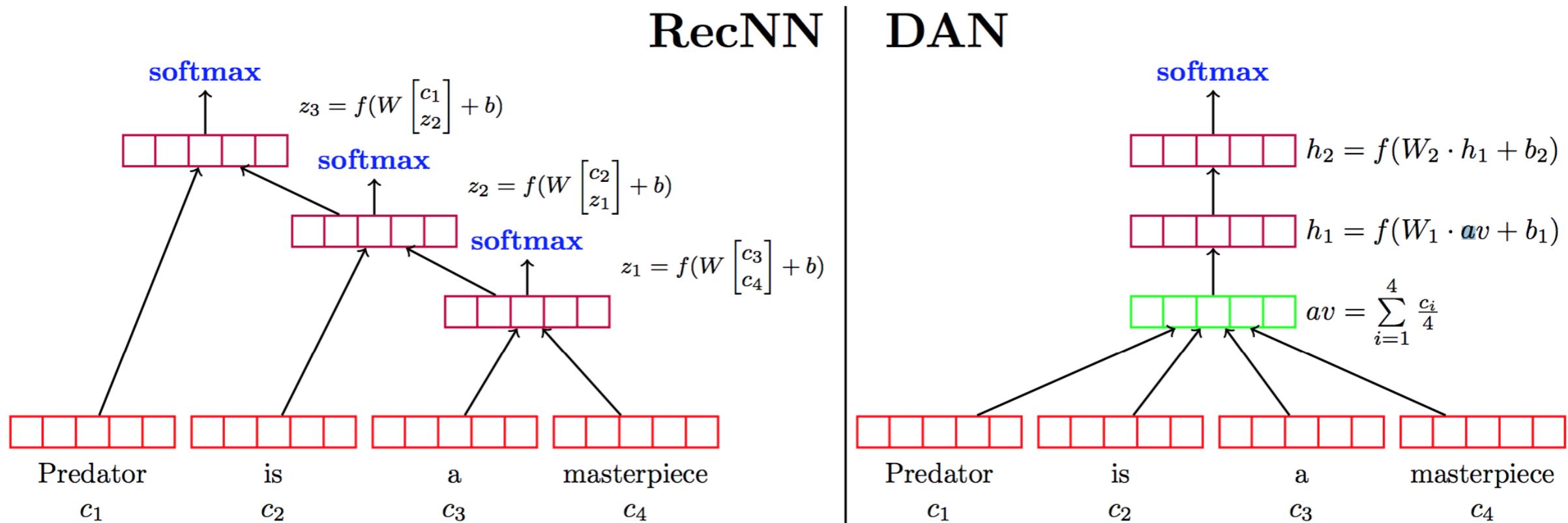
Deep Averaging Networks (DANs)

DAN



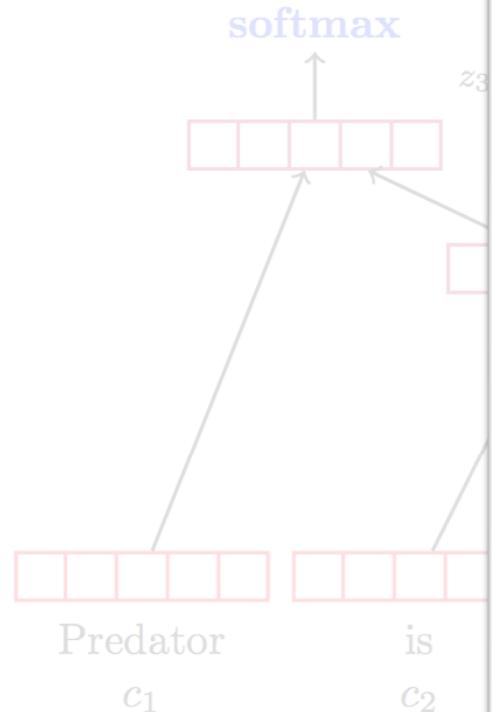
dense nets

Deep Averaging Networks (DANs)



dense nets

Deep Averaging Networks (DANs)



ties? Socher et al. (2013b) report that removing the nonlinearities from their **RecNN** models drops performance on the Stanford Sentiment Treebank by over 5% absolute accuracy. Most unordered functions are linear mappings between bag-of-words features and output labels, so might they suffer from the same issue? To isolate the effects of syntactic composition from the nonlinear transformations that are crucial to **RecNN** performance, we investigate how well a deep version of the **NBOW** model performs on tasks that have recently been dominated by syntactically-aware models.

$$h_2 = f(W_2 \cdot h_1 + b_2)$$

$$h_1 = f(W_1 \cdot av + b_1)$$

$$av = \sum_{i=1}^4 \frac{c_i}{4}$$

masterpiece
 c_3 c_4

dense nets

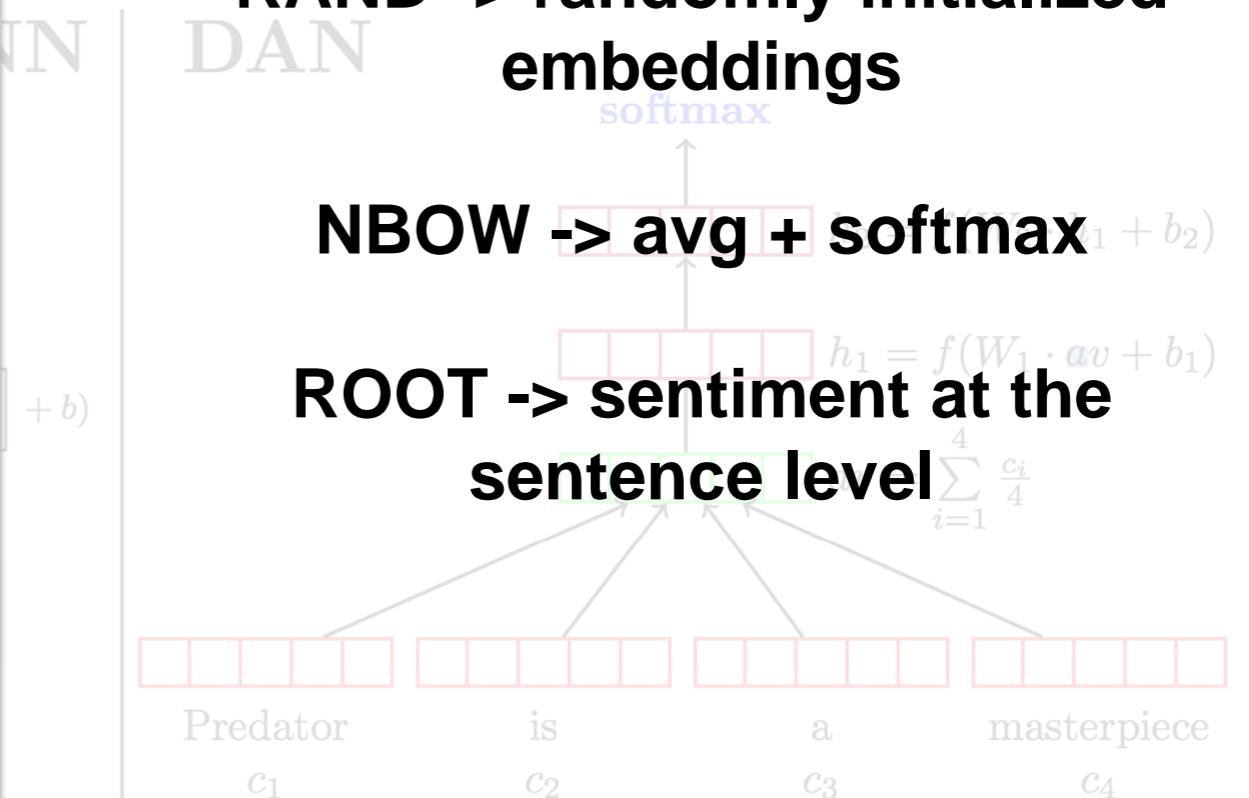
Deep Averaging Networks (DANs)

Model	RT	SST fine	SST bin	IMDB	Time (s)
DAN-ROOT	—	46.9	85.7	—	31
DAN-RAND	77.3	45.4	83.2	88.8	136
DAN	80.3	47.7	86.3	89.4	136
NBOW-RAND	76.2	42.3	81.4	88.9	91
NBOW	79.0	43.6	83.6	89.0	91
BiNB	—	41.9	83.1	—	—
NBSVM-bi	79.4	—	—	91.2	—
RecNN*	77.7	43.2	82.4	—	—
RecNTN*	—	45.7	85.4	—	—
DRecNN	—	49.8	86.6	—	431
TreeLSTM	—	50.6	86.9	—	—
DCNN*	—	48.5	86.9	89.4	—
PVEC*	—	48.7	87.8	92.6	—
CNN-MC	81.1	47.4	88.1	—	2,452
WRRBM*	—	—	—	89.2	—

RAND \rightarrow randomly initialized embeddings

NBOW \rightarrow avg + softmax

ROOT \rightarrow sentiment at the sentence level

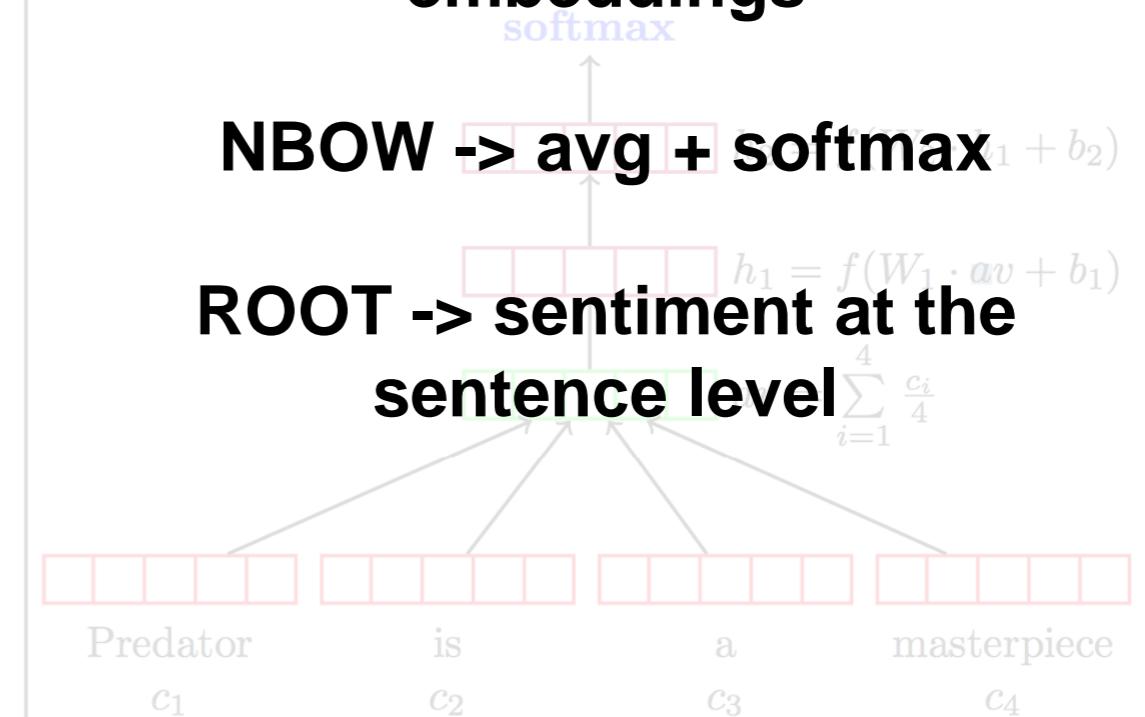


dense nets

Deep Averaging Networks (DANs)

Model	RT	SST fine	SST bin	IMDB	Time (s)
DAN-ROOT	—	46.9	85.7	—	31
DAN-RAND	77.3	45.4	83.2	88.8	136
DAN	80.3	47.7	86.3	89.4	136
NBOW-RAND	76.2	42.3	81.4	88.9	91
NBOW	79.0	43.6	83.6	89.0	91
BiNB	—	41.9	83.1	—	—
NBSVM-bi	79.4	—	—	91.2	—
RecNN*	77.7	43.2	82.4	—	—
RecNTN*	—	45.7	85.4	—	—
DRecNN	—	49.8	86.6	—	431
TreeLSTM	—	50.6	86.9	—	—
DCNN*	—	48.5	86.9	89.4	—
PVEC*	—	48.7	87.8	92.6	—
CNN-MC	81.1	47.4	88.1	—	2,452
WRRBM*	—	—	—	89.2	—

RAND \rightarrow randomly initialized embeddings



Acknowledgments

- Some slides by Chris Kedzie