13. Disambiguation

Ambiguous = capable of being understood in two or more senses or ways.

13.1 Ambiguity

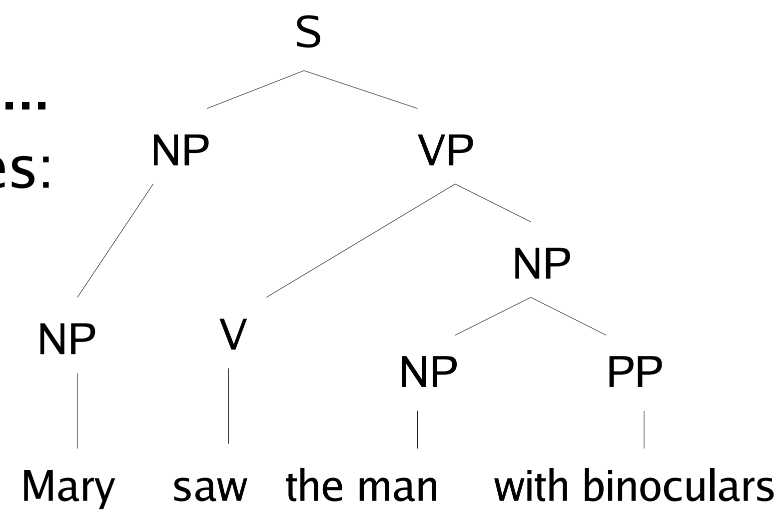
- Increases the number of possible interpretations $\rightarrow$ computational problem.
- Resolving ambiguities (disambiguation), especially word sense disambiguation, a major problem in NL understanding.
- Also important in tasks that do not aim at understanding: machine translation, information retrieval, text processing (spell check), speech processing.
- Requires
  - A good world model
  - Use of context
  - Use of probabilities (Prior probability, conditional probability)

13.2 Types of ambiguity

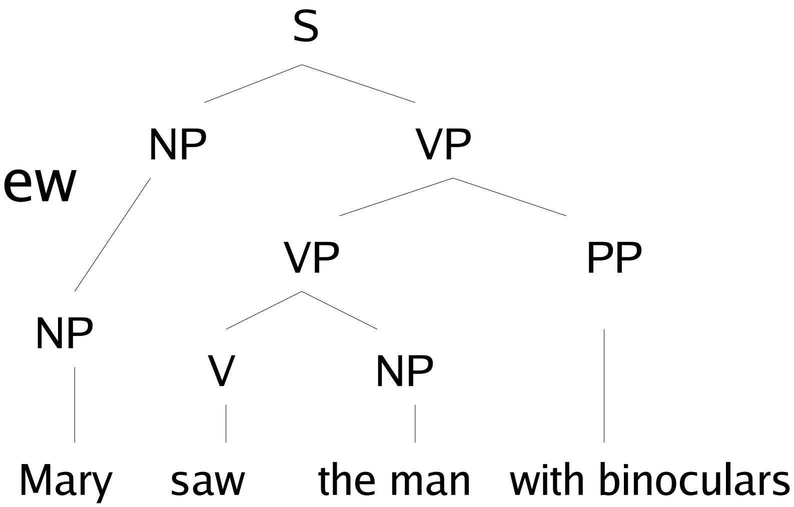
- Lexical
  - Word has several meanings: bat, bank, ...
  - Word can act in several syntactic classes: set, time, can, ...
  - In spoken language:

    'Ice cream' — 'I scream',

    'hear' — 'here'.



- Syntactic
  - Sentence has several parses:
    - “I saw the Grand Canyon flying to New York.”
    - “Squad helped dog bite victim.”
- Semantic
  - Often result of syntactic ambiguity.



- Pragmatic ambiguity
  - "I'll meet you next Friday." "I am here now."
- Referential ambiguity
  - Anaphoric expressions
    - "John gave Bob the sandwich. He smiled."
    - "John gave the dog the sandwich. It wagged its tail."
  - There are words for categories, not for individual objects.
    - "The apple I had today."
- Local ambiguity
  - The old train

    ... the young.

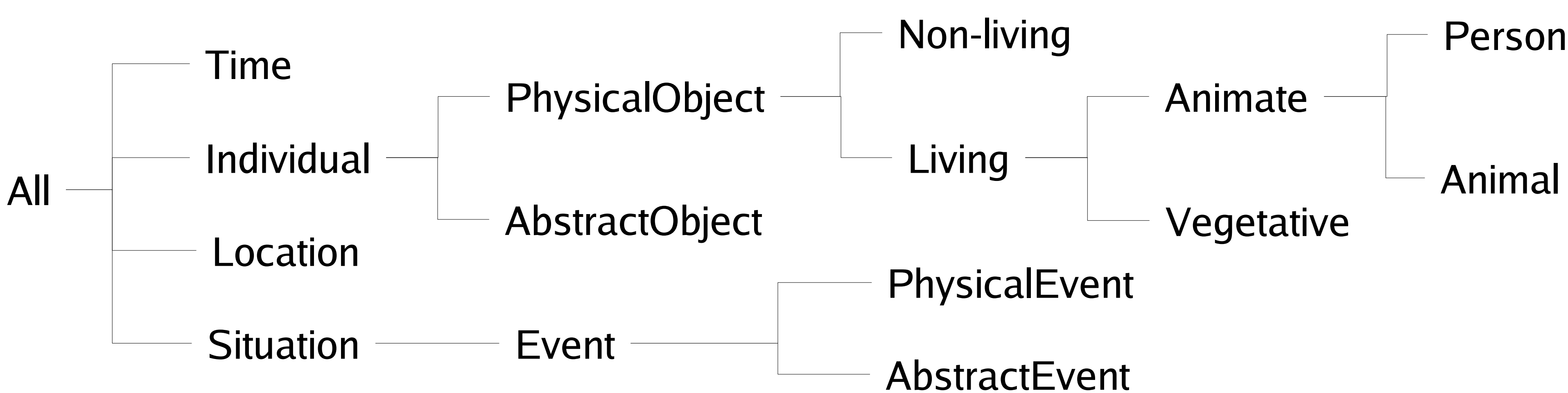
    ... left the station.
- Vagueness
  - "It's hot today."

13.3 Word sense disambiguation

- Association of a given word in a text or discourse with a meaning that is distinct from other potential meanings of that word.
  1. Determination of all different senses for every word that are relevant in the context.
  2. Assignment of words to senses. Relies in two sources of information:
     - Context
     - External knowledge

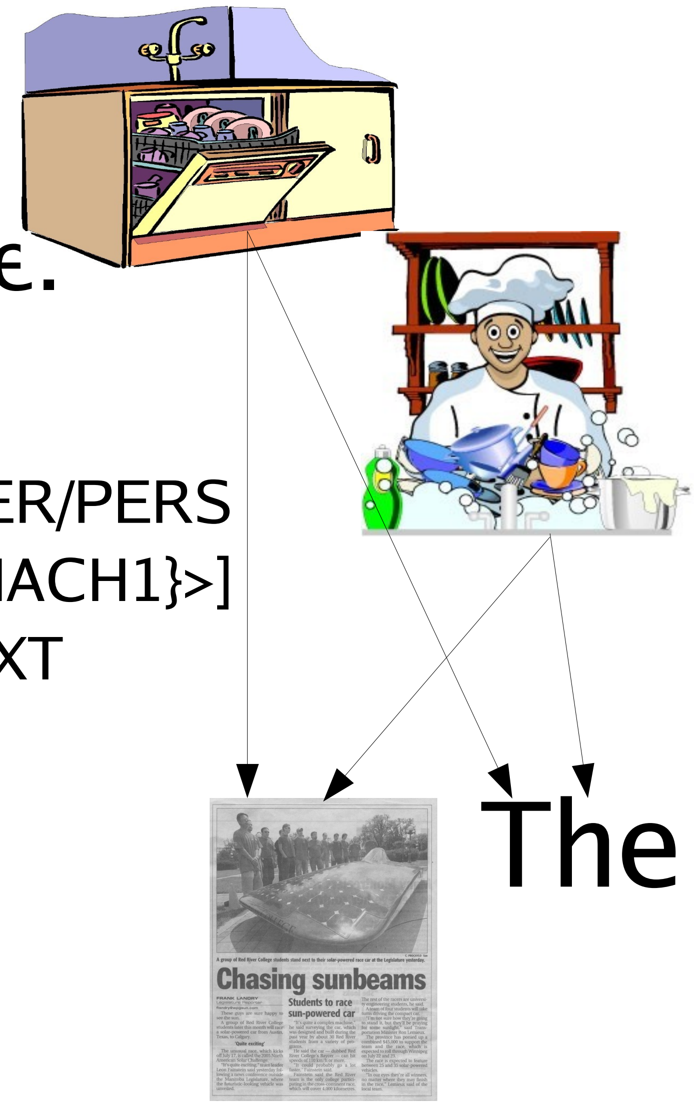
13.4 Selectional restrictions

- Specification of legal combinations of senses that can
  - Co-occur
  - Be applied to discard incoherent parses.
- Word senses can be disjoint, overlapping or subclasses of each other, for instance:
  - CAT1 — DOG1 , PET1 — MAMMAL1, MAMMAL1 Đ OG1
- Abstraction hierarchy:



- Example: the verb 'read' takes two arguments:
  - Agent, capable of reading
  - Theme, object containing text
- “The dishwasher read the article.”
- Logical form:

(READS1 *r1* [AGENT <THE *d1* {DISHWASHER/PERS
DISHWAHSER/MACH1}>]
[THEME <THE *p1* {ARTICLE/TEXT
ARTICLE1}>])



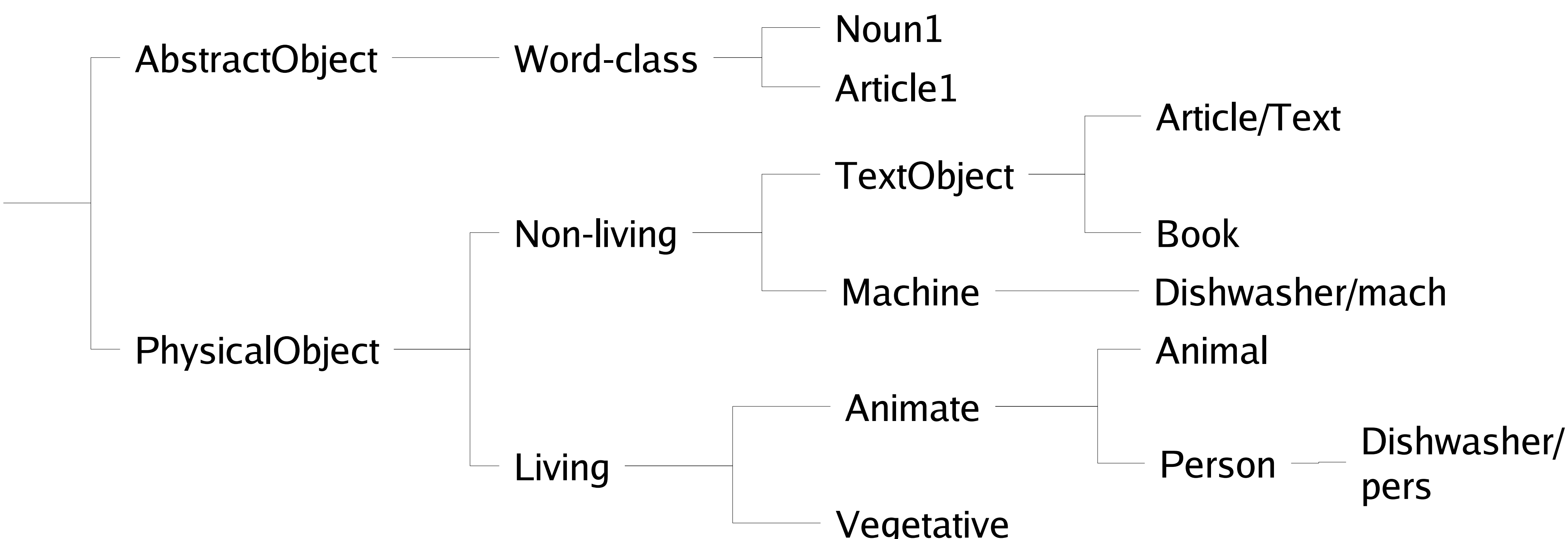
- Finding the allowable combinations can be viewed as constraint satisfaction.
- Apply selectional restrictions for READS1:

  (AGENT READS1 Person)

  (THEME READS1 TextObject)
- One reading makes sense:

  (READS1 [AGENT <THE d1 DISHWASHER/PERS>]

  [THEME <THE p1 ARTICLE/TEXT>])



13.5 Main problems of the approach

- Proved extremely useful and popular; works well in limited domains, however ...
- Semantic well-formedness not binary condition:
  - Absolute conditions: Eliminate possible interpretations
  - Abstract constraints: Allow almost anything.
    - "I ate a pizza."
    - "I ate the box."
    - "I ate some chips."

(THEME EATS1 FoodObject)
vs.
(THEME EATS1 PhysicalObject)

- Sensitive to context:
  - "I could not eat the car."
  - "My car eats gasoline."

OK in negated context

Or in metaphors

13.6 Other approaches

- Selectional restrictions give only very coarse classification of acceptable senses $\rightarrow$ need more predictive method.
- Distinguish common interpretations of senses from uncommon. Use *prior probabilities*:
  - Given all data, if a word occurs in **SENSE1** 95% of the time as opposed in **SENSE2** 5% of the time, we would predict **SENSE1** 100% of the time.
- Use *context*
  - Other words in the document give a good indication of the word's actual sense.

13.7 Word collocations

- Words that tend to appear together within a window: bigrams, trigrams, N-grams, whole sentences or entire texts.
- A sentence S of length n:

Window of size m centered on word $\mathbf{W}_t$

$$S = W_1\ W_2 \ldots\ W_{t-\lfloor m/2 \rfloor} \ldots W_{t-2}\ W_{t-1}\ \mathbf{W}_{\mathbf{t}}\ W_{t+1}\ W_{t+2} \ldots W_{t+\lfloor m/2 \rfloor} \ldots W_n$$

- Predict the sense of word $\mathbf{W}_{\mathbf{t}}$ from the words within the window, i.e., find σ that maximizes:
  - $\mathrm{ArgMax}_S\ \Pr(\sigma(\mathbf{W}_{\mathbf{t}}) \mid W_{t-m/2} \ldots \mathbf{W}_{\mathbf{t}} \ldots W_{t+m/2})$ (1)

- Use Bayes' rule to estimate (1):

These all are the same

$$\Pr(W_{t-m/2} \ldots W_{t+m/2} \mid \sigma(\mathbf{W_t})) * \Pr(\sigma(\mathbf{W_t})) / \Pr(W_{t-m/2} \ldots W_{t+m/2})$$

This can be approximated by:

$$\prod_{i=1,m} \Pr_m(W_i \mid \sigma(\mathbf{W_t}))$$

With sense σ

$$\frac{\text{\# of times } W_i \text{ occurs in a window centered at } \mathbf{W_t}}{\text{\# of times } \mathbf{W_t} \text{ is a center of a window}}$$

13.8 Semantics vs. statistics

- Semantic and statistical approaches are complementary.
- Given enough data, statistical method eventually captures all senses of words, and semantic relations between them.
- Semantic definition can be used even if the word is seen for the first time:

  "I sailed the lake in my ∗preef yesterday."

13.9 Part-of-speech (POS) tagging

- Assign lexical tags (syntactic categories) to each word in the corpus:
  - Noun, verb, adjective, preposition, and so on
  - Verbs: active, passive, transitive, intransitive
  - Nouns: singular, plural
- Some words may appear in several categories: e.g., can, swing, time.
- Example (Picasso):

Noun	Verb	Article	Noun	Pronoun	Verb	Pronoun	Verb	Article	Noun
Art	is	a	lie	that	lets	us	see	the	truth.

13.10 Language as random process

- Randomness implies lack of structure, definition or understanding.
- Stochastic methods generalize the deterministic aspect.
- Given
  - words of language: $S_w = \{w_1, \ldots w_N\}$
  - set of tags: $S_t = \{t_1, \ldots, t_M\}$.
  - Sentence is a sequence of random variables $(W_1\ W_2 \ldots W_n)$, which can take any of the values in S_w.
  - Tags are random variables $T_1, T_2, \ldots, T_m$, which can take values in S_t.
- We want to find $t_1, \ldots, t_n$ that maximize:
  - $\Pr(T_1=t_1, \ldots, T_n=t_n \mid W_1=w_1, \ldots, W_n=w_n) = \Pr(t_1, \ldots t_n \mid w_1, \ldots, w_n)$

13.11 Hidden Markov Model

- Due to potential ambiguity, and lack of space and time, we try to maximize $Pr(t_1, \ldots, t_n \mid w_1, \ldots, w_n)$

  $\propto Pr(t_1)Pr(w_1 \mid t_1)\, Pr(t_2 \mid t_1, w_1) \ldots Pr(t_n \mid w_1, \ldots, w_n, t_1, \ldots t_{n-1})$

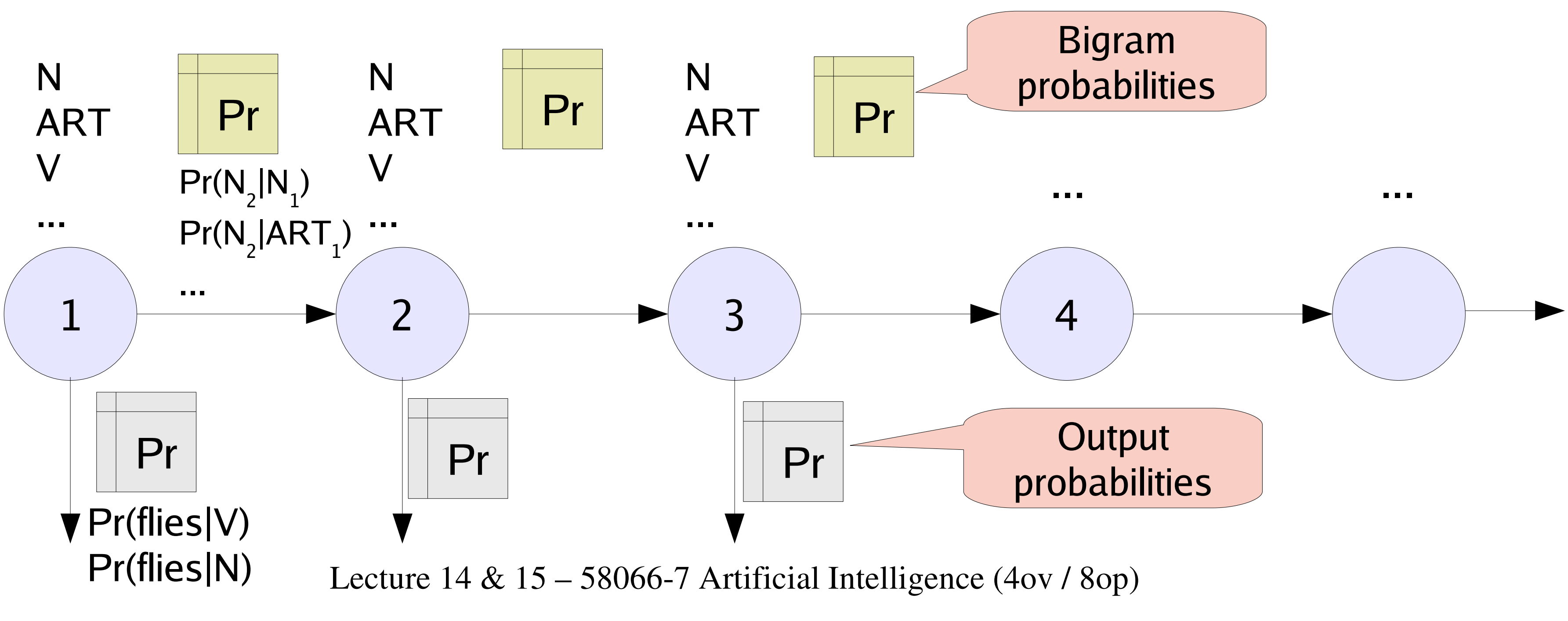
  $= \prod_{i=1,n} Pr(t_i \mid t_1, \ldots, t_{i-1}, w_1, \ldots, w_{i-1})\, Pr(w_i \mid t_1, \ldots, t_i, w_1, \ldots w_{i-1})$
- Markov assumptions:
  - $Pr(t_i \mid t_1, \ldots, t_{i-1}, w_1, \ldots, w_{i-1})$ approaches $Pr(t_i \mid t_{i-1})$
  - $Pr(w_i \mid t_1, \ldots, t_i, w_1, \ldots, w_{i-1})$ approaches $Pr(w_i \mid t_i)$
- So we approximate $\prod_{i=1,n} Pr(t_i \mid t_{i-1})\, Pr(w_i|t_i)$

Output probability

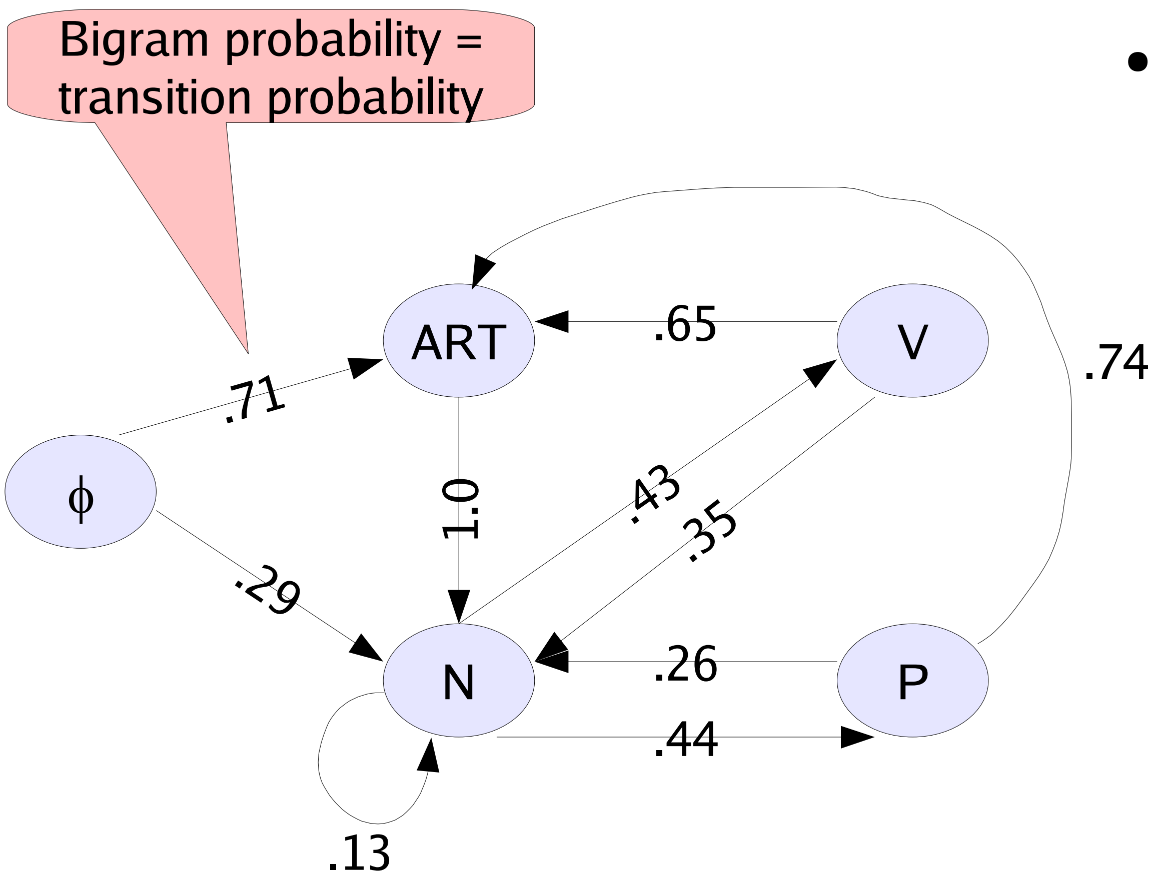
Transition probability

13.12 Transition and output probabilities

- *Output probabilities* gives the probability to each possible output of the node.
- *Transition probability* gives the probability of moving from one state to another.
- For instance, the node N will have a probability table that indicates for each word how likely it'll be selected, if we randomly select a noun.

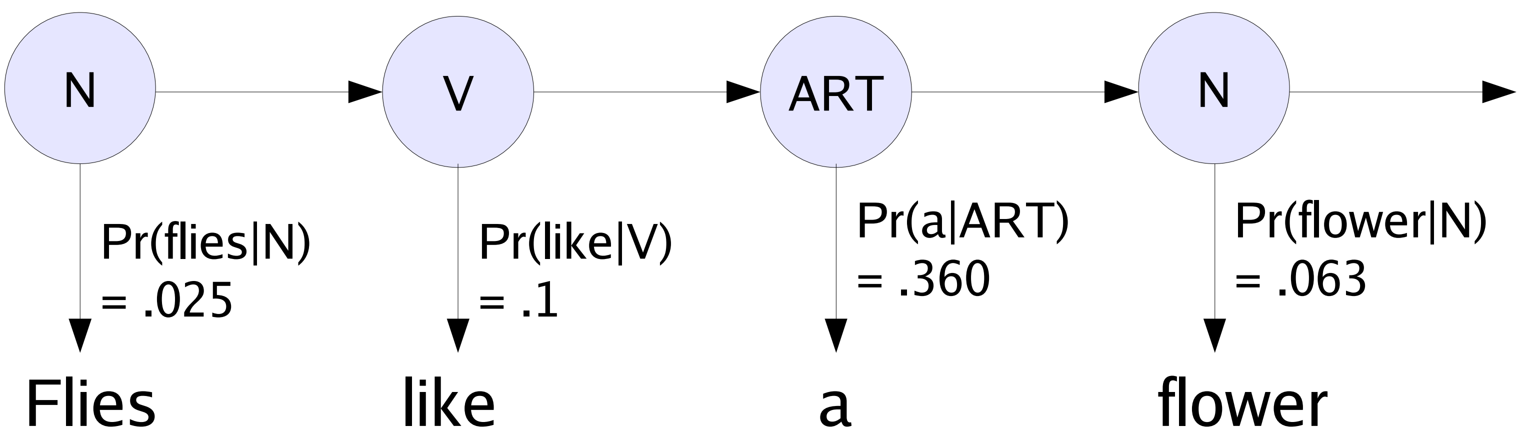


13.13 Example



- Output probabilities:
  - Pr(the|ART) = .54
  - Pr(flies|N) = .025
  - Pr(flies|V) = .076
  - Pr(like|V) = .1
  - Pr(like|P) = .068
  - Pr(like|N) = .012
  - Pr(a|ART) = .360
  - Pr(a|N) = .001
  - Pr(flower|N) = .063
  - Pr(flower|V) = .05
  - Pr(birds|N) = .076

- Calculate the probability that the sequence N V ART N generates the sentence “Flies like a flower.”
  - Probability of sequence N V ART N, given by the Markov chain, is $.29 \times .43 \times .65 \times 1 = .081$ $\prod_{i=1,n} \Pr(t_i \mid t_{i-1})$
  - Probability of “Flies like a flower”, given by the output probabilities, is $.025 \times .1 \times .360 \times .063 = 5.4 \times 10^{-5}$ $\prod_{i=1,n} \Pr(w_i|t_i)$
  - Likelihood that HMM generates the sentence is 4.37×10^{-6}



13.15 Early work

- Use of window (Weaver 1949)
  - In context of machine translation
  - Window size: with a couple of words on either side of the word to be disambiguated vs. the whole sentence.
  - Weaver also introduced the statistical aspect
- AI-methods in 50's and 60's
  - Semantic networks (Quillian): spreading activation
  - Frames (Hayes)
- Connectionist methods 60's and 70's
  - Inhibition (Rumelhart and McClelland)