Language Models Continued

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith

The Story So Far

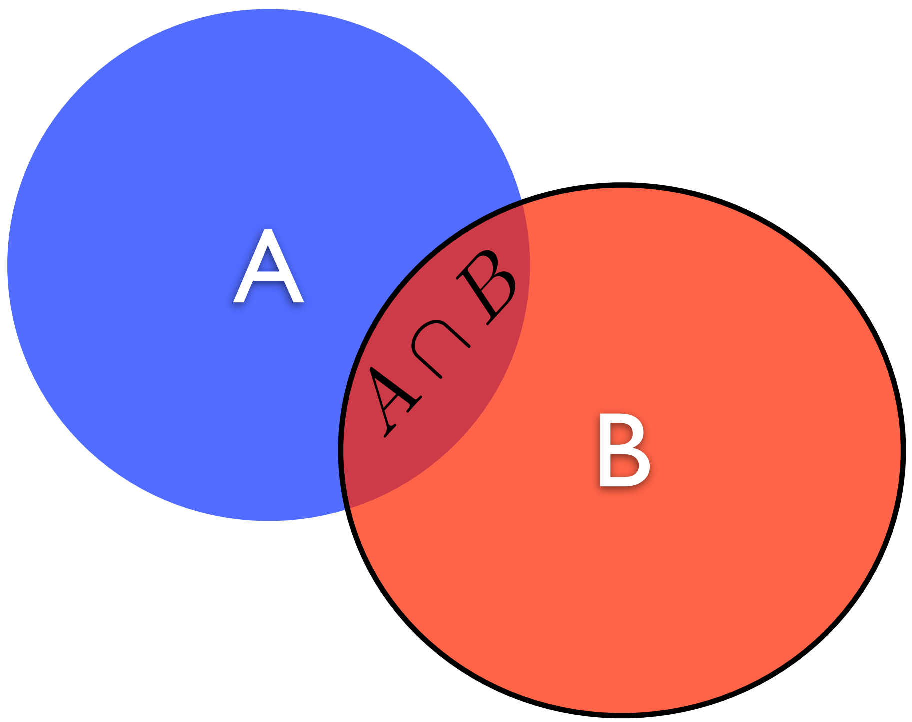
- Last time: simple LMs
  - Markov assumptions: bigrams, trigrams,...
  - Generating text from an n-gram model
- This time
  - More on probability
  - Bayes theorem and naive Bayes classifiers
  - Smoothing: expecting the unseen

Axioms of Probability

- Define event space $\bigcup_i \mathcal{F}_i = \Omega$
- Probability function, s.t. $P : \mathcal{F} \to [0, 1]$
  - Disjoint events sum $A \cap B = \emptyset \Leftrightarrow P(A \cup B) = P(A) + P(B)$
  - All events sum to one $P(\Omega) = 1$
- Show that: $P(\bar{A}) = 1 - P(A)$

Conditional Probability

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$



$$P(A,B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

$$\begin{aligned} P(A_1, A_2, \ldots, A_n) &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \\ &\quad \cdots P(A_n \mid A_1, \ldots, A_{n-1}) \end{aligned}$$

Chain rule

Independence

$$\begin{array}{rcl} P(A, B) & = & P(A)P(B) \\ & \Leftrightarrow & \\ P(A \mid B) = P(A) & \wedge & P(B \mid A) = P(B) \end{array}$$

In coding terms, knowing *B* doesn't help in decoding *A*, and vice versa.

Another View of Markov Models

$$\begin{aligned} p(w_1, w_2, \ldots, w_n) &= p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \\ &\quad p(w_4 \mid w_1, w_2, w_3) \cdots p(w_n \mid p_1, \ldots, p_{n-1}) \end{aligned}$$

Another View of Markov Models

$$
\begin{aligned}
p(w_1, w_2, \ldots, w_n) \quad = \quad & p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \\
& p(w_4 \mid w_1, w_2, w_3) \cdots p(w_n \mid p_1, \ldots, p_{n-1})
\end{aligned}
$$

Markov independence assumption

$$p(w_i \mid w_1, \ldots w_{i-1}) \approx p(w_i \mid w_{i-1})$$

Another View of Markov Models

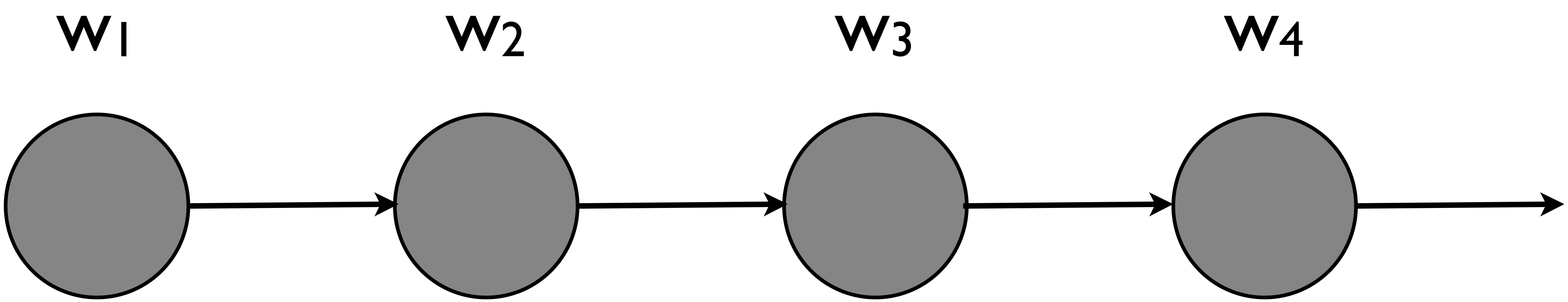
$$p(w_1, w_2, \ldots, w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \\ p(w_4 \mid w_1, w_2, w_3) \cdots p(w_n \mid p_1, \ldots, p_{n-1})$$

Markov independence assumption

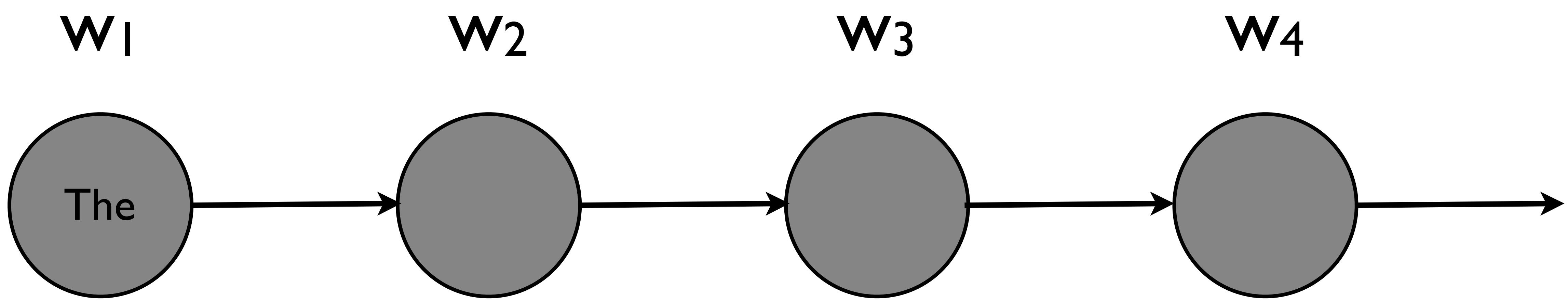
$$p(w_i \mid w_1, \ldots w_{i-1}) \approx p(w_i \mid w_{i-1})$$

$$p(w_1, w_2, \ldots, w_n) \approx p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_2) \\ p(w_4 \mid w_3) \cdots p(w_n \mid p_{n-1})$$

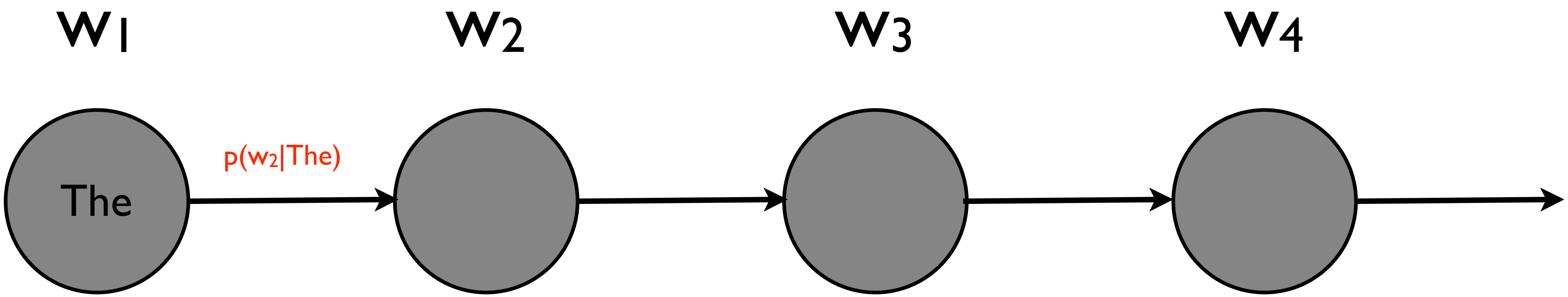
Yet Another View



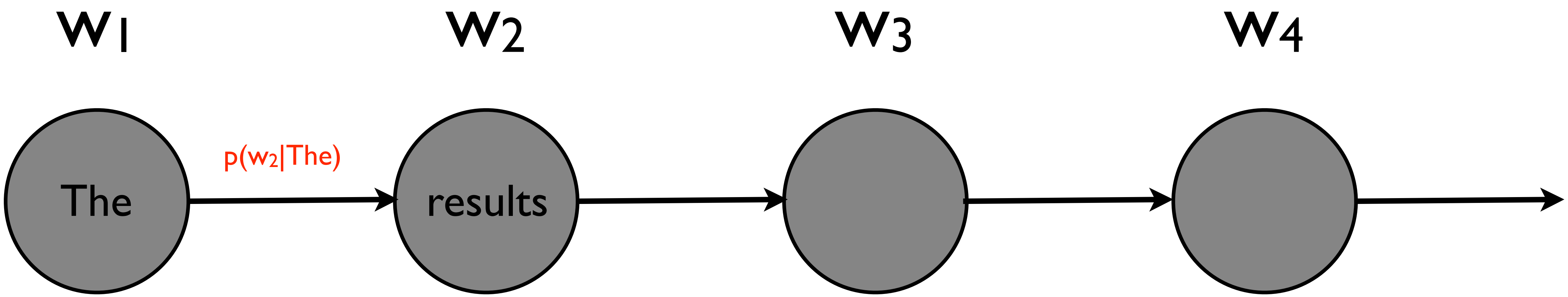
Yet Another View



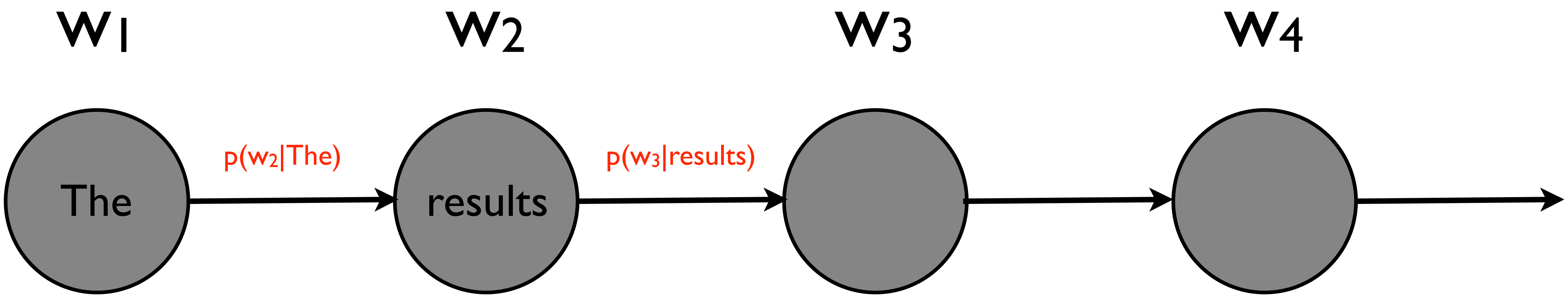
Yet Another View



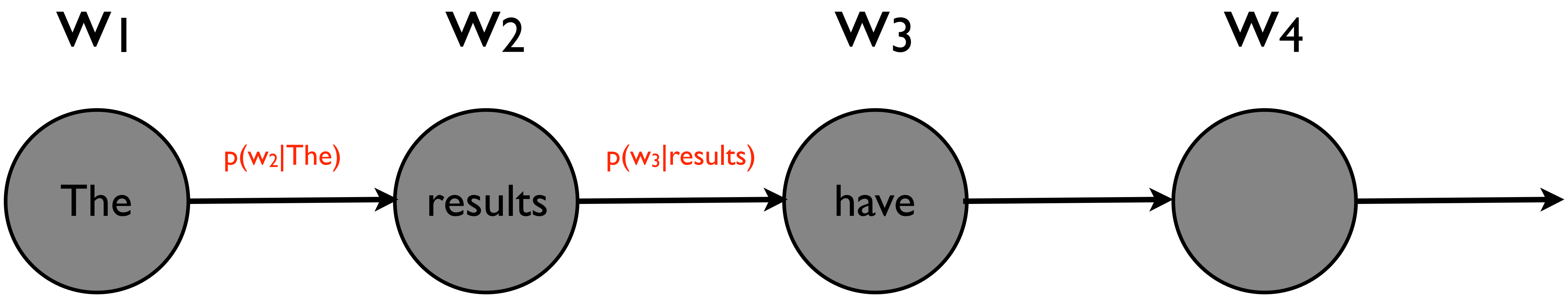
Yet Another View



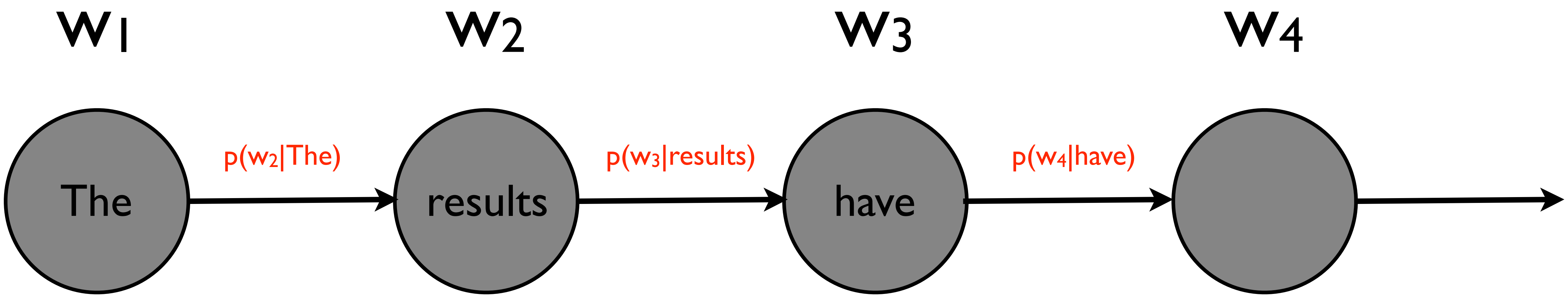
Yet Another View



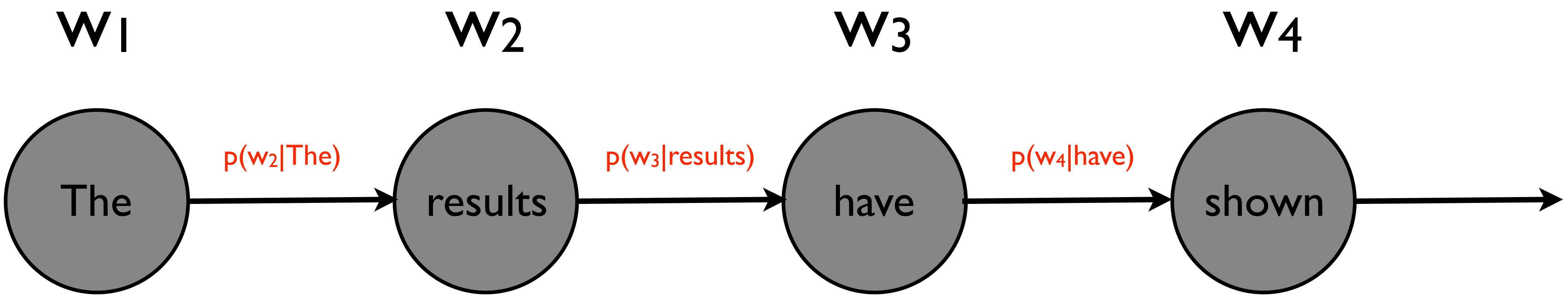
Yet Another View



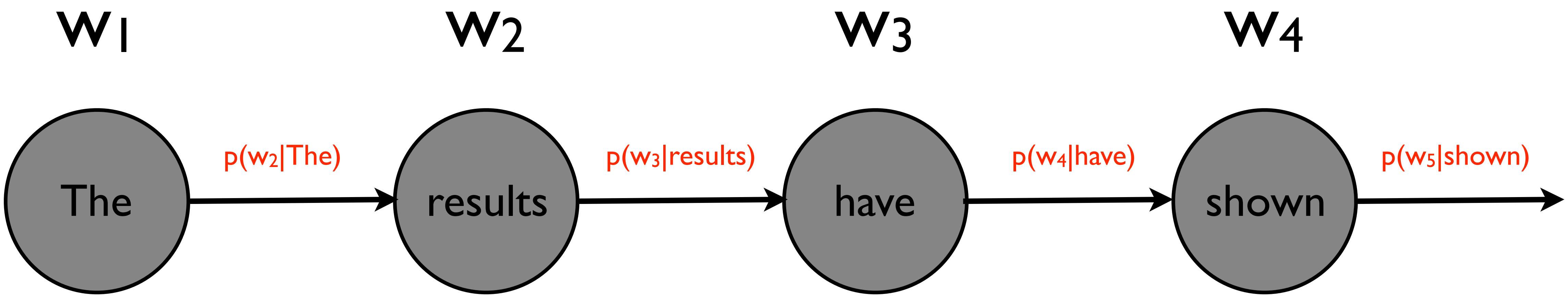
Yet Another View



Yet Another View

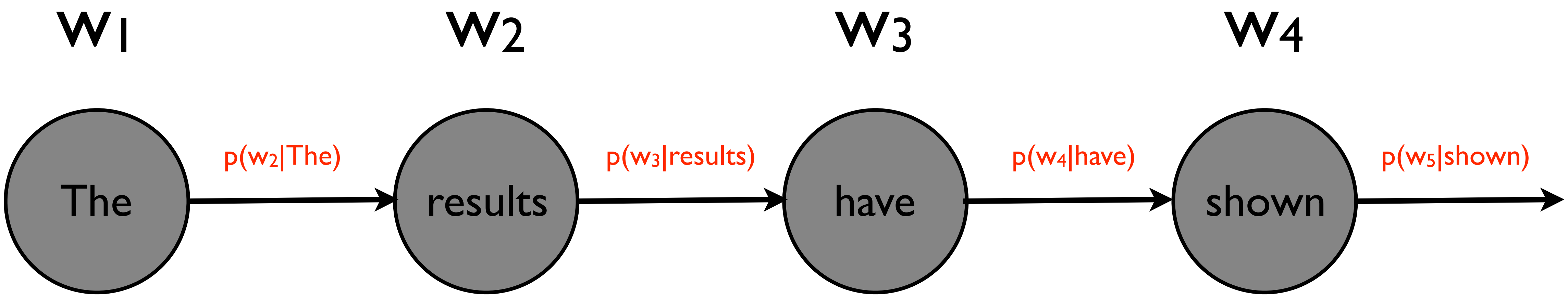


Yet Another View



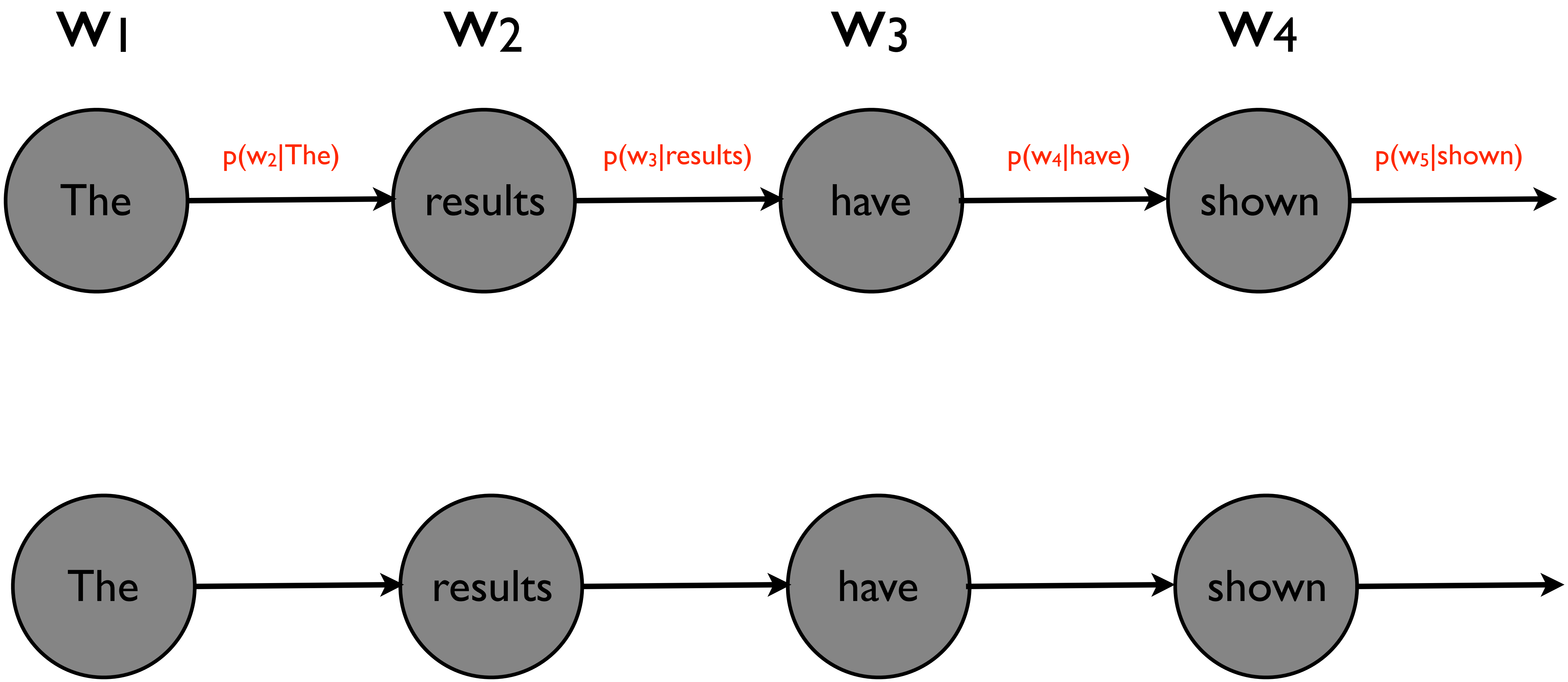
Yet Another View

Directed graphical models: *lack* of edge means conditional independence



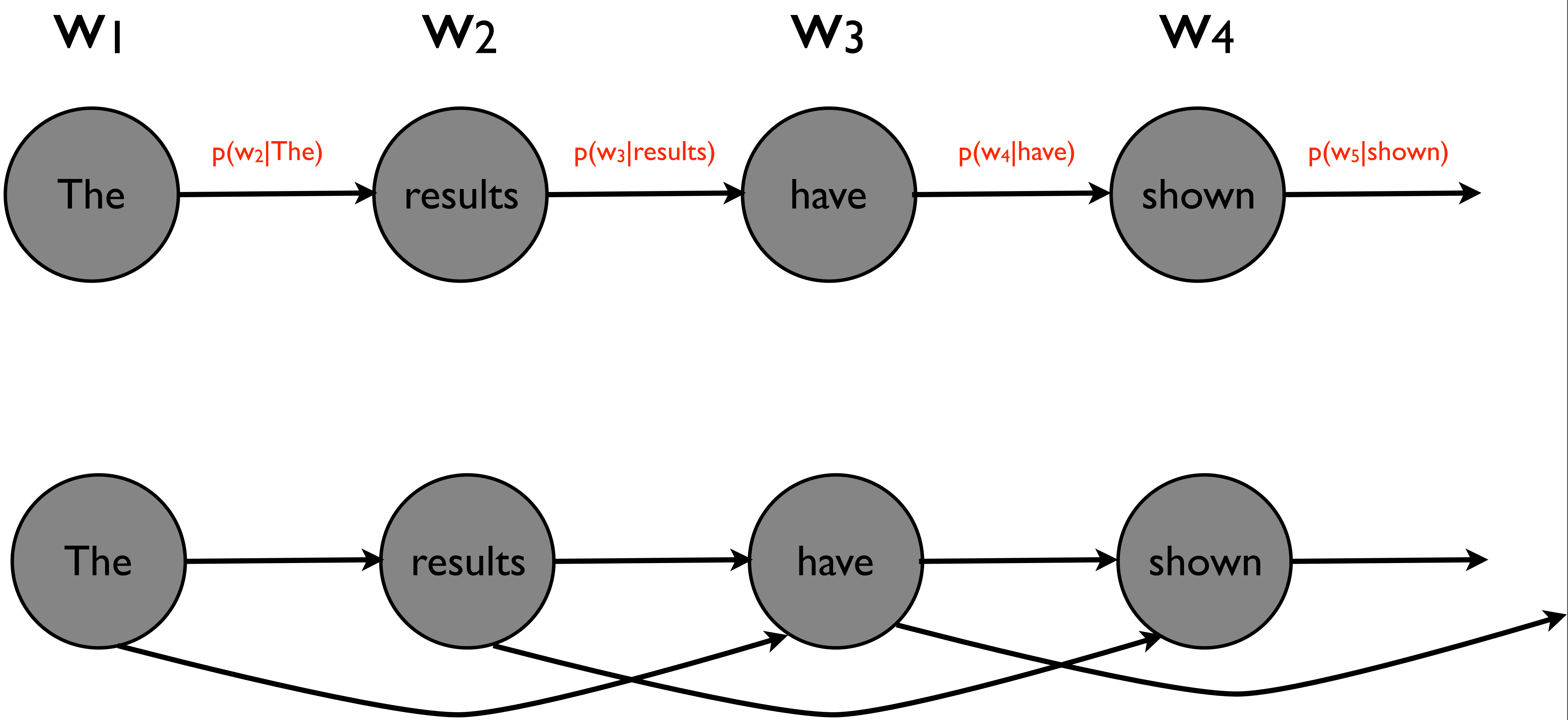
Yet Another View

Directed graphical models: *lack* of edge means conditional independence



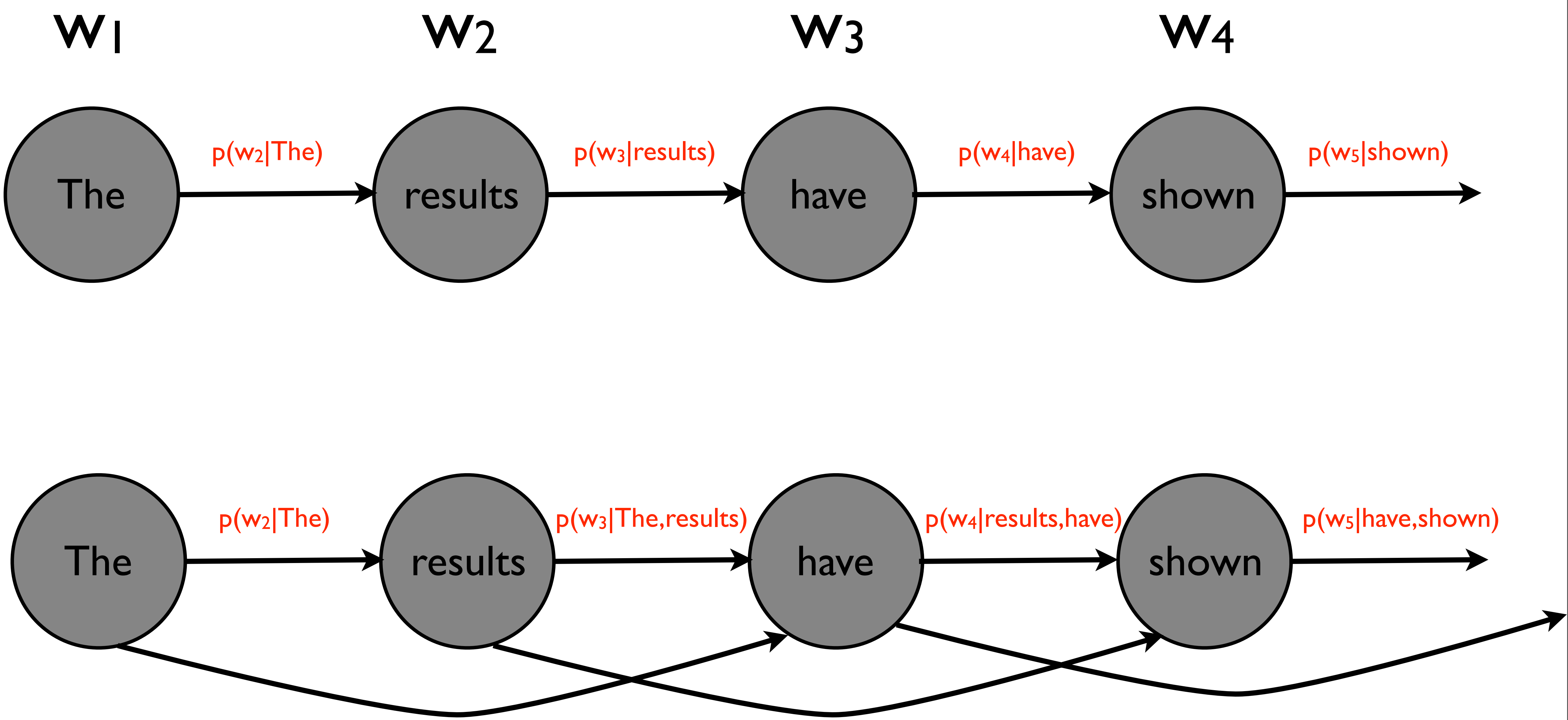
Yet Another View

Directed graphical models: *lack* of edge means conditional independence



Yet Another View

Directed graphical models: *lack* of edge means conditional independence



Classifiers: Language under Different Conditions

Movie Reviews

Movie Reviews

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

Movie Reviews

☺ there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹ seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

the rich legacy of cinema has left us with certain indelible images . the tinkling christmas tree bell in " it ' s a wonderful life . " bogie ' s speech at the airport in " casablanca . " little elliott ' s flying bicycle , silhouetted by the moon in " e . t . " and now , " starship troopers " director paul verhoeven adds one more image that will live in our memories forever : doogie houser doing a vulcan mind meld with a giant slug . " starship troopers , " loosely based on

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

☹

the rich legacy of cinema has left us with certain indelible images . the tinkling christmas tree bell in " it ' s a wonderful life . " bogie ' s speech at the airport in " casablanca . " little elliott ' s flying bicycle , silhouetted by the moon in " e . t . " and now , " starship troopers " director paul verhoeven adds one more image that will live in our memories forever : doogie houser doing a vulcan mind meld with a giant slug . " starship troopers , " loosely based on

Setting up a Classifier

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

- What we know how to build:

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

- What we know how to build:
  - A language model for each class

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

- What we know how to build:
  - A language model for each class
    - $p(w_1, w_2, ..., w_n \mid ☺)$

Setting up a Classifier

- What we want:

  p(☺ | $w_1, w_2, \ldots, w_n$) > p(☹ | $w_1, w_2, \ldots, w_n$) ?

- What we know how to build:
  - A language model for each class
    - p($w_1, w_2, \ldots, w_n$ | ☺)
    - p($w_1, w_2, \ldots, w_n$ | ☹)

Bayes' Theorem

By the definition of conditional probability:

$$P(A, B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

we can show:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Seemingly trivial result from 1763; interesting consequences...

REV. T. BAYES

A "Bayesian" Classifier

$$p(R \mid w_1, w_2, \ldots, w_n) = \frac{p(R)p(w_1, w_2, \ldots, w_n \mid R)}{p(w_1, w_2, \ldots, w_n)}$$

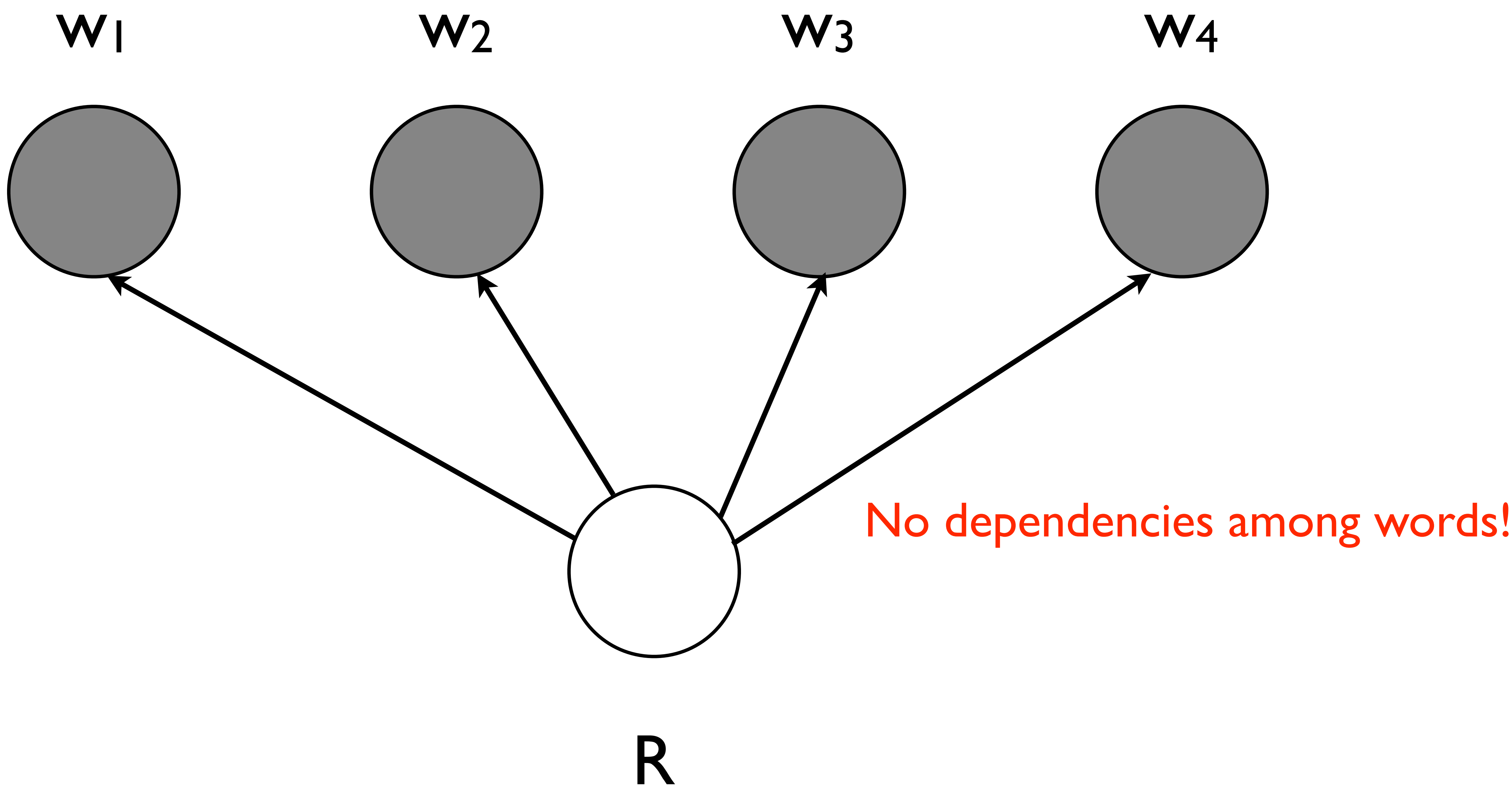
$$\max_{R \in \{\ddot\smile, \ddot\frown\}} p(R \mid w_1, w_2, \ldots, w_n) = \max_{R \in \{\ddot\smile, \ddot\frown\}} p(R)p(w_1, w_2, \ldots, w_n \mid R)$$

Posterior

Prior

Likelihood

Naive Bayes Classifier



NB on Movie Reviews

- Train models for positive, negative
- For each review, find higher posterior
- Which word probability ratios are highest?

```
>>> classifier.show_most_informative_features(5)

classifier.show_most_informative_features(5)
Most Informative Features
   contains(outstanding) = True              pos : neg    =     14.1 : 1.0
         contains(mulan) = True              pos : neg    =      8.3 : 1.0
        contains(seagal) = True              neg : pos    =      7.8 : 1.0
   contains(wonderfully) = True              pos : neg    =      6.6 : 1.0
         contains(damon) = True              pos : neg    =      6.1 : 1.0
```

What's Wrong With NB?

- What happens for word dependencies are strong?
- What happens when some words occur only once?
- What happens when the classifier sees a new word?

Summing Up

- Exploit rules of probability to condition events
- Exploit Bayes rule for classification
- Smooth to avoid zeroes
- Read Manning & Schütze 2.1 and chap. 6