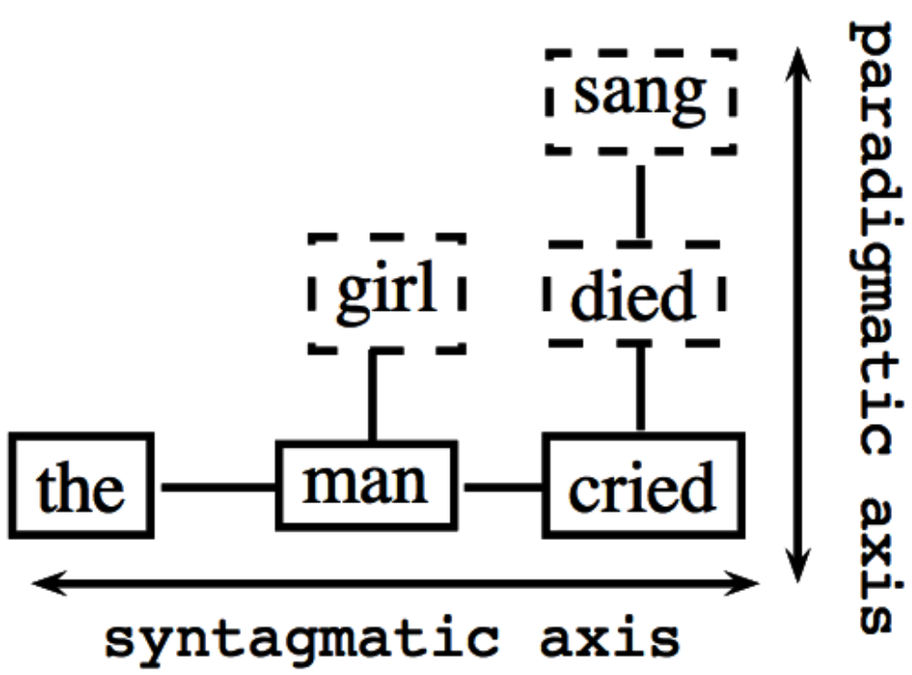
Linguistic Category Induction and Tagging Using the Paradigmatic Context Representations with Substitute Words

Mehmet Ali Yatbaz



Natural Language Processing

Natural Language Processing

- Natural Languages has ambiguities

Natural Language Processing

- Natural Languages has ambiguities
  - **I love spicy dishes .**

Natural Language Processing

- Natural Languages has ambiguities
  - **I love spicy dishes .**

- S: (n) a piece of dishware normally used as a container for holding or serving food
- S: (n) a particular item of prepared food
- S: (n) the quantity that a dish will hold
- S: (n) a very attractive or seductive looking woman
- S: (n) directional antenna consisting of a parabolic reflector for microwave or radio frequency radiation
- S: (n) an activity that you like or at which you are superior

Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways

Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways
- it is important to:

Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways
- it is important to:
  - disambiguate

Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways
- it is important to:
  - disambiguate
  - find words that are similar to each other

Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways
- it is important to:
  - disambiguate
  - find words that are similar to each other
- to do that one can use words features

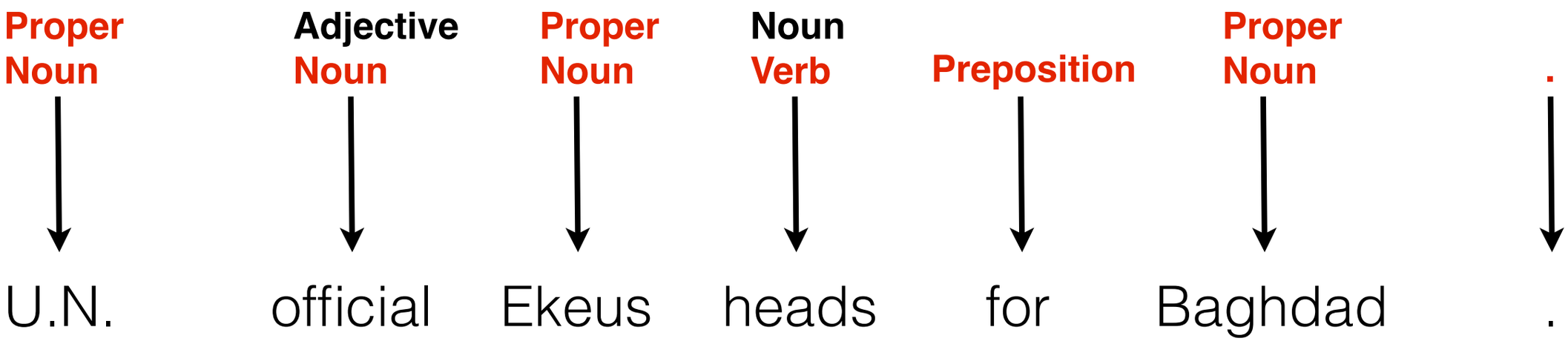
Natural Language Processing

- Natural Languages has ambiguities
  - a word can be used in different ways
- it is important to:
  - disambiguate
  - find words that are similar to each other
- to do that one can use words features
  - Word context is one of the word features

Outline

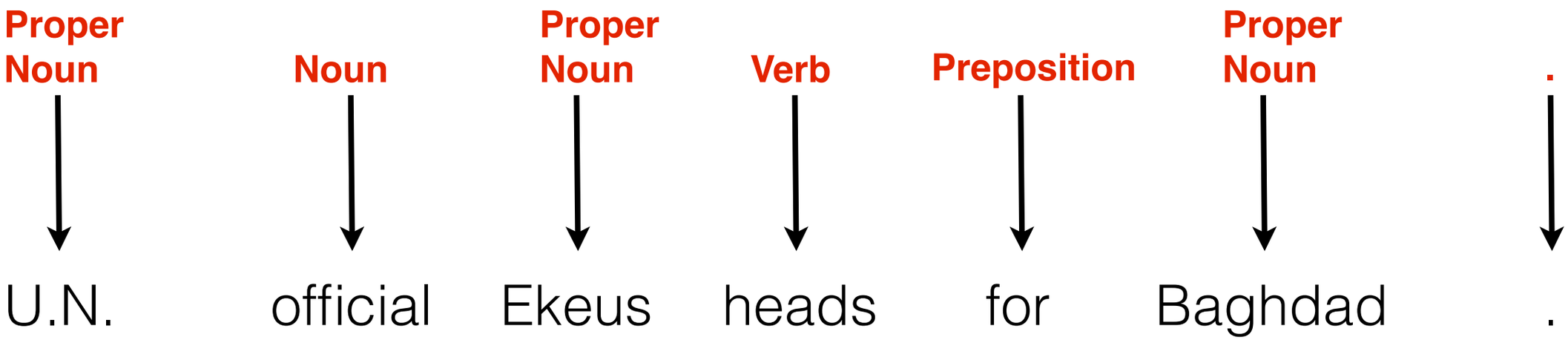
- Tagging
- Paradigmatic Context representation
- Clustering Model
- Co-occurrence Modeling
- Probabilistic Voting
- HMM based Model
- Noisy Channel Model

Part of speech disambiguation



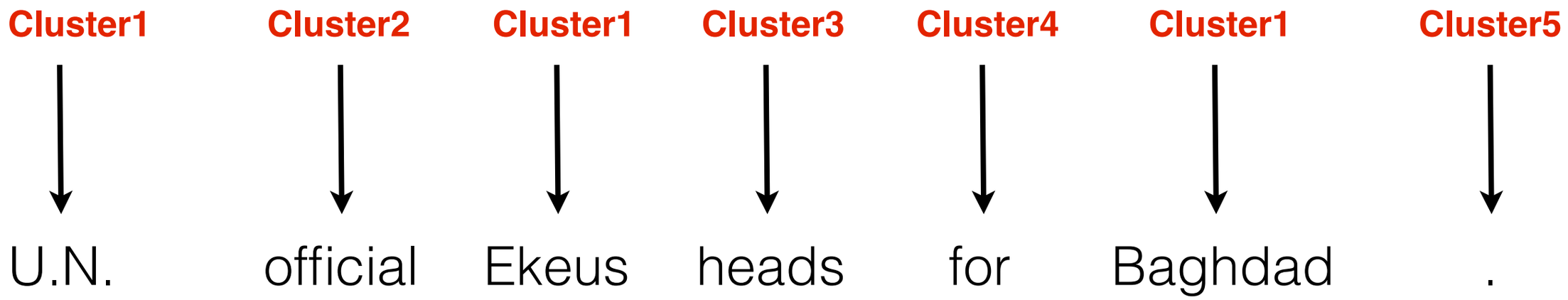
- Part-of-speech represents groups of words that are substitutable
  - without altering the grammaticality of a sentence

Part of speech disambiguation



- Part-of-speech represents groups of words that are substitutable
  - without altering the grammaticality of a sentence

Part of speech induction



- Clusters represents the groups of words that are substitutable
- No tag information is available

Word-sense disambiguation

U.N. official Ekeus heads for Baghdad .

- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Word-sense disambiguation

noun(1): United Nations (an organization of independent states formed in 1945 to promote international peace and security)

U.N. official Ekeus heads for Baghdad .

- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Word-sense disambiguation

noun(1): United Nations (an organization of independent states formed in 1945 to promote international peace and security)

noun(1): a worker who holds or is invested with and office.
noun(2): someone who administers the rules of an organization.

X

U.N. official Ekeus heads for Baghdad .

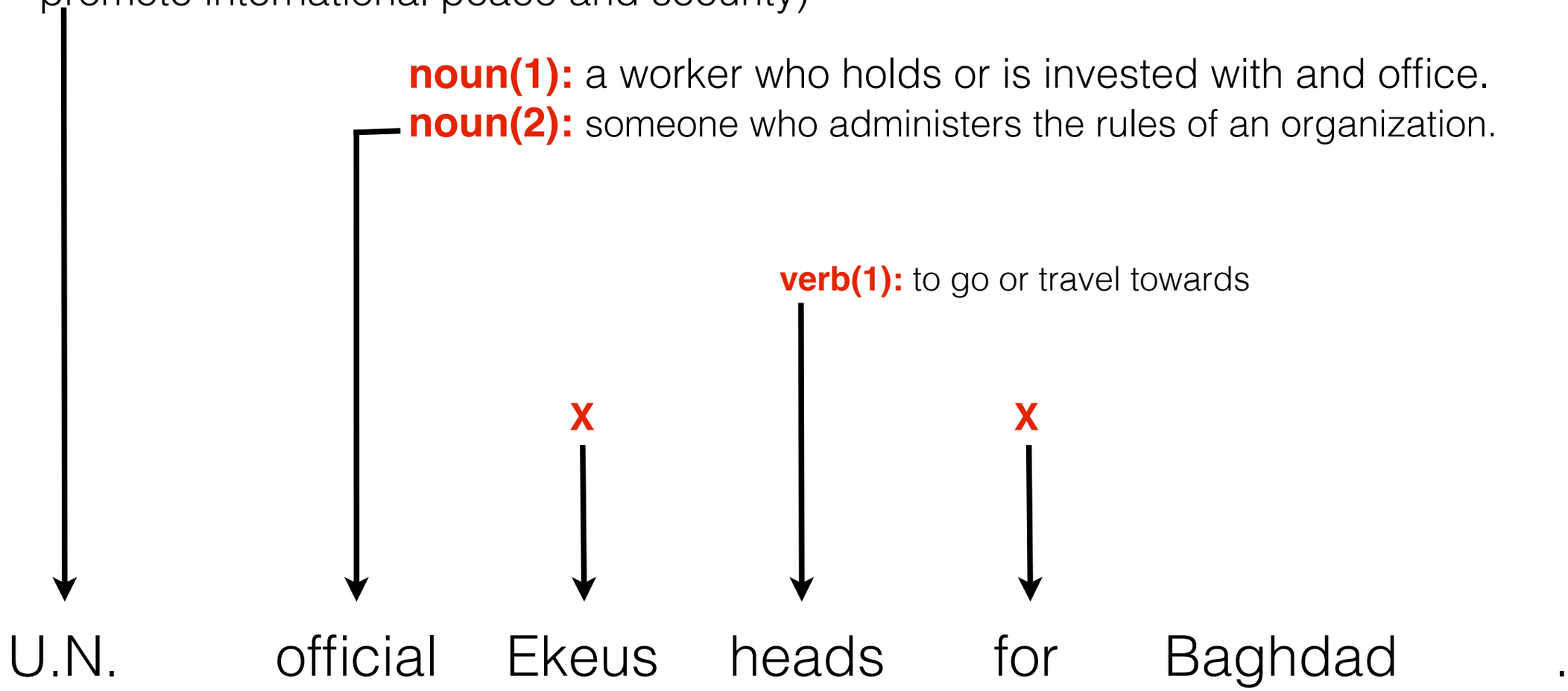
- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Word-sense disambiguation

noun(1): United Nations (an organization of independent states formed in 1945 to promote international peace and security)

noun(1): a worker who holds or is invested with and office.
noun(2): someone who administers the rules of an organization.

verb(1): to go or travel towards



- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Word-sense disambiguation

noun(1): United Nations (an organization of independent states formed in 1945 to promote international peace and security)

noun(1): a worker who holds or is invested with and office.
noun(2): someone who administers the rules of an organization.

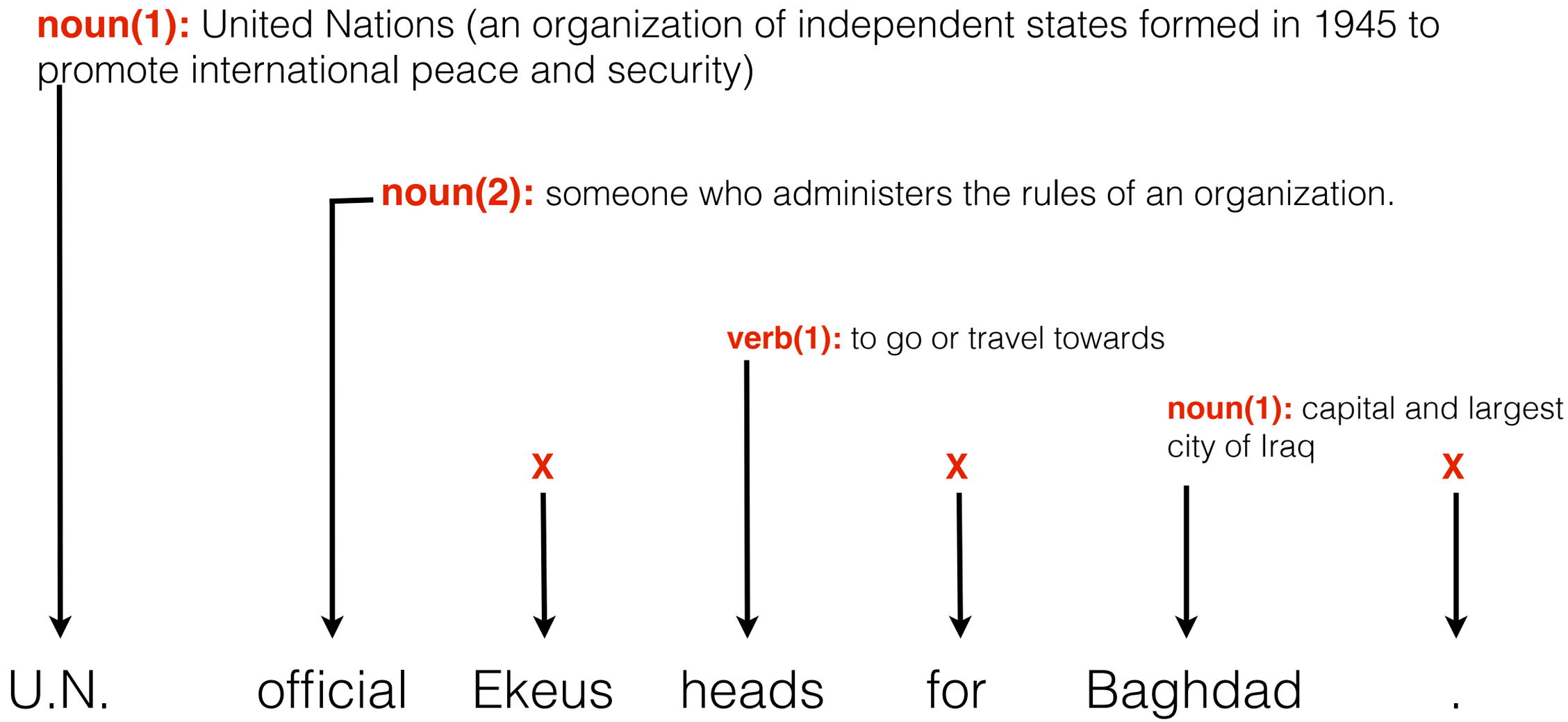
verb(1): to go or travel towards

noun(1): capital and largest city of Iraq

U.N.	official	X	heads	X	for	Baghdad	.

- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Word-sense disambiguation



- identifying which sense of a word (i.e. meaning) is used
- **X** : has no entry in **WordNet**

Tagging in General

Things that are common?

Tagging in General

Things that are common?

- **input:** Sequence of words

w_1 w_2 w_3 w_4 w_5 w_6 w_7

Tagging in General

Things that are common?

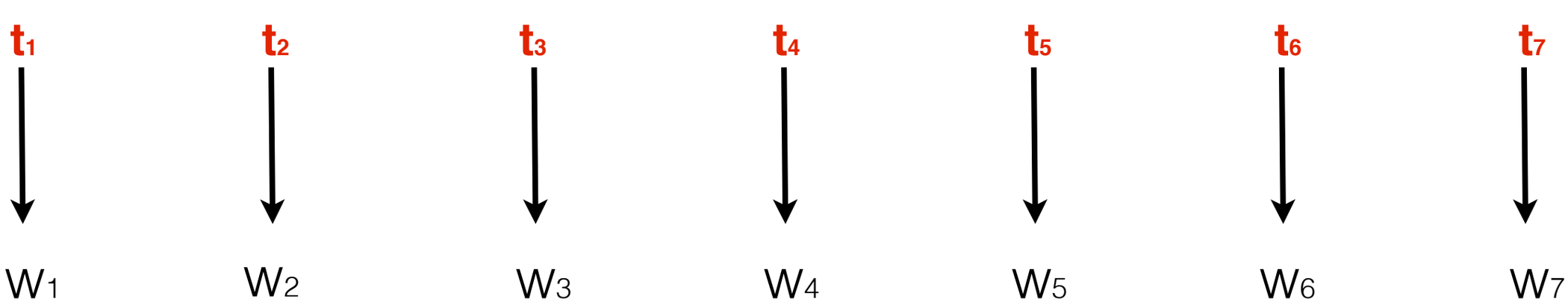
- **input:** Sequence of words
- **output:** Sequence of tags

t_1	t_2	t_3	t_4	t_5	t_6	t_7
w_1	w_2	w_3	w_4	w_5	w_6	w_7

Tagging in General

Things that are common?

- **input:** Sequence of words
- **output:** Sequence of tags
- **objective:** best tag sequence (depends on the task)



Why tagging is not trivial?

Disambiguation

Some words have multiple tags **(ambiguous words)**
Ex: in POS tagging **offer** can be **verb** or **noun**

VERB: ... it will also **offer** buyers the option ...

NOUN: The **offer** is begin launched ....

Correct tag depends on the context

Why tagging is not trivial?

Induction

put similar words into same clusters

Ex: instances of **verb offer** and **noun offer** should be in different clusters.

cluster i: ... it will also **offer** buyers the option ...

cluster j: The **offer** is begin launched ....

Cluster id depends on the context

Disambiguation **vs** Induction

Disambiguation vs Induction

- Requires some level of **tag information**
  - Tagging is expensive

- no annotation **(no tag information)**
  - good for resource poor languages

Disambiguation vs Induction

- Requires some level of **tag information**
  - Tagging is expensive

- disambiguates the **correct tag** of an **ambiguous word**

- no annotation **(no tag information)**
  - good for resource poor languages

- puts **similar words** into same **cluster**

Disambiguation vs Induction

- Requires some level of **tag information**
  - Tagging is expensive

- disambiguates the **correct tag** of an **ambiguous word**

- Used by higher level **NLP tools** (ex: parsing, translation)

- no annotation **(no tag information)**
  - good for resource poor languages

- puts **similar words** into same **cluster**

- More relevant to **Child Language Acquisition**

Tagging Models:

Unsupervised Models (induction)

Supervised Models

- word sequence

- word sequence
- tag sequence

Tagging Models:

Unsupervised Models (induction)

Supervised Models

Level of annotation during training

- word sequence

- word sequence
- word-tag dictionary

word	tags
of	Preposition
	Adverb
	Particle
....	....

- word sequence
- tag sequence

Tagging Models:

Unsupervised Models (induction)

Supervised Models

Level of annotation during training

- word sequence

- word sequence
- word-tag dictionary

word	tags
of	Preposition
	Adverb
	Particle
....	....

- word sequence
- word-tag distribution

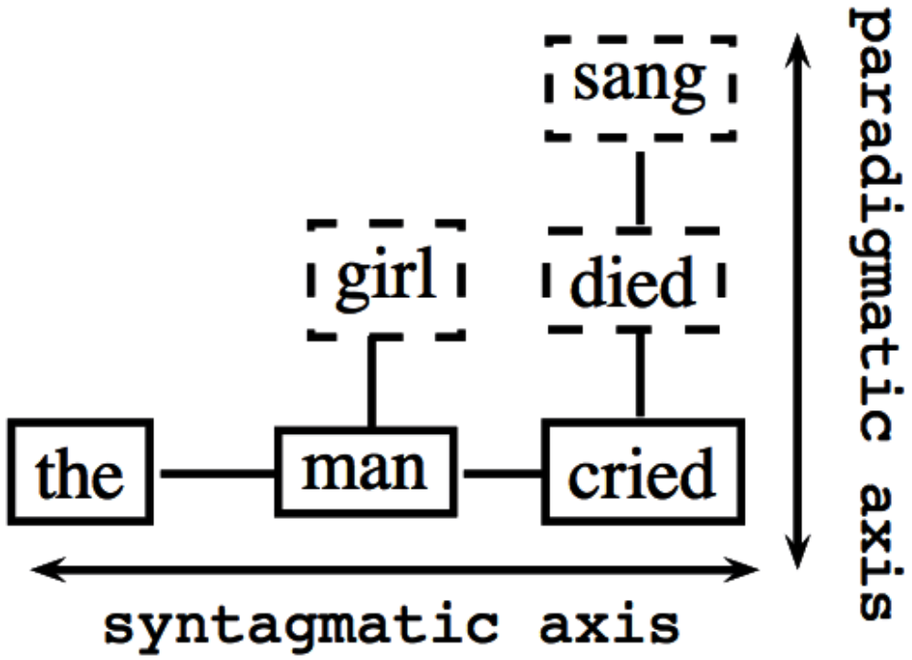
word	tags
of	Preposition .99
	Adverb .005
	Particle .005
....	**....**

- word sequence
- tag sequence

Outline

- **Paradigmatic Context representation**
- Clustering Model
- Co-occurrence Modeling
- Probabilistic Voting
- HMM based Model
- Noisy Channel Model
- Conclusion

Representations of Word Context



Syntagmatic Representation

- Similar words share similar neighbors.
- Context is represented by the **neighboring words** of the target word

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

2-gram context **the** **as**

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

2-gram context	**the**	**as**
3-gram context	**join the**	**as a**

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

2-gram context	**the**	**as**
3-gram context	**join the**	**as a**
4-gram context	**will join the**	**as a nonexecutive**

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

2-gram context	**the**	**as**
3-gram context	**join the**	**as a**
4-gram context	**will join the**	**as a nonexecutive**
5-gram context	**, will join the**	**as a nonexecutive director**

Syntagmatic Context Representation

Pierre Vinken, 61 years old , will join the **board** as a nonexecutive director Nov. 29 .

2-gram context	**the**	**as**
3-gram context	**join the**	**as a**
4-gram context	**will join the**	**as a nonexecutive**
5-gram context	**, will join the**	**as a nonexecutive director**
⋮	⋮	⋮
Pierre Vinken, 61 years old , will join the		**as a nonexecutive director Nov. 29 .**

Syntagmatic Context Representation

let’s relate two similar words

Pierre Vinken, 61 years old , will join the board as a nonexecutive **director** Nov. 29 .

... Joseph Corr was succeeded by Frank Lorenzo , **chief** of parent Texas Air .

Syntagmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the board as a nonexecutive **director** Nov. 29 .

5-gram context board as a nonexecutive Nov. 29 .

... Joseph Corr was succeeded by Frank Lorenzo , **chief** of parent Texas Air .

5-gram context by Frank Lorenzo , of parent Texas Air

Syntagmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the board as a nonexecutive **director** Nov. 29 .

5-gram context board as a nonexecutive Nov. 29 .

... Joseph Corr was succeeded by Frank Lorenzo , **chief** of parent Texas Air .

5-gram context by Frank Lorenzo , of parent Texas Air

- suffers from **sparsity** as the context becomes larger

Syntagmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the board as a nonexecutive **director** Nov. 29 .

5-gram context board as a nonexecutive Nov. 29 .

... Joseph Corr was succeeded by Frank Lorenzo , **chief** of parent Texas Air .

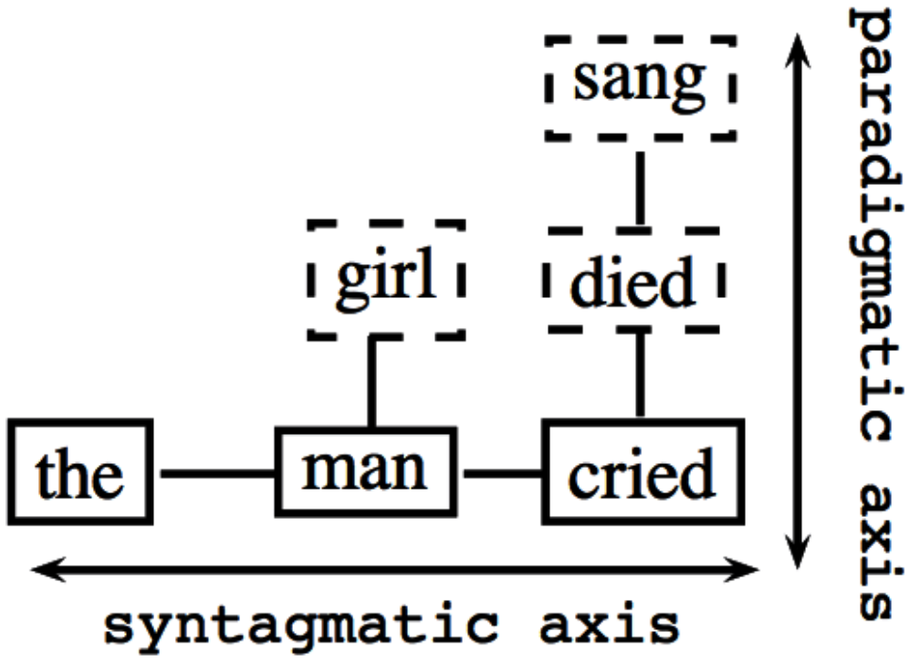
5-gram context by Frank Lorenzo , of parent Texas Air

- suffers from **sparsity** as the context becomes larger

↓

Chance of having same context becomes smaller

Representations of Word Context



Paradigmatic Representation

- Similar words have similar substitute distributions
- Context is represented by the **distribution of substitutes**.

Paradigmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the **board as a nonexecutive director Nov. 29 .**

... Joseph Corr was succeeded **by Frank Lorenzo , chief of parent Texas Air** .

Paradigmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the **board as a nonexecutive** _____ **Nov. 29 .**

... Joseph Corr was succeeded **by Frank Lorenzo ,** _____ **of parent Texas Air** .

Paradigmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the **board as a nonexecutive** _____ **Nov. 29 .**

chairman	**0.8242**
director	**0.0127**
directors	**0.0127**
....	**....**

... Joseph Corr was succeeded **by Frank Lorenzo ,** _____ **of parent Texas Air** .

chairman	**0.9945**
president	**0.0031**
directors	**0.0012**
....	**....**

Paradigmatic Context Representation

let's relate two similar words

Pierre Vinken, 61 years old , will join the **board as a nonexecutive** _____ **Nov. 29 .**

chairman	**0.8242**
director	**0.0127**
directors	**0.0127**
....	**....**

Do not suffer from sparsity

... Joseph Corr was succeeded **by Frank Lorenzo ,** _____ **of parent Texas Air** .

chairman	**0.9945**
president	**0.0031**
directors	**0.0012**
....	**....**

✓Given the context, substitute distribution is independent of the **target word!**

Paradigmatic Representations of Word Context

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

  ✓ **Morphological disambiguation (NIPS, 2009.)**

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

  - ✓ **Morphological disambiguation (NIPS, 2009.)**
  - ✓ **POS disambiguation(COLING, 2010.)**

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

  - ✓ **Morphological disambiguation (NIPS, 2009.)**
  - ✓ **POS disambiguation(COLING, 2010.)**
  - ✓ **Word sense disambiguation (Computational Linguistics, 2010.)**

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

  ✓ **Morphological disambiguation (NIPS, 2009.)**

  ✓ **POS disambiguation(COLING, 2010.)**

  ✓ **Word sense disambiguation (Computational Linguistics, 2010.)**

  ✓ **Learning Syntactic Categories Using Paradigmatic Representations of Word Context (EMNLP, 2012)**

Paradigmatic Representations of Word Context

- Substitute distributions are successfully applied to

  - ✓ **Morphological disambiguation (NIPS, 2009.)**
  - ✓ **POS disambiguation(COLING, 2010.)**
  - ✓ **Word sense disambiguation (Computational Linguistics, 2010.)**
  - ✓ **Learning Syntactic Categories Using Paradigmatic Representations of Word Context (EMNLP, 2012)**
  - ✓ **Unsupervised Instance-Based Part of Speech Induction Using Probable Substitutes (submitted to ACL2014)**

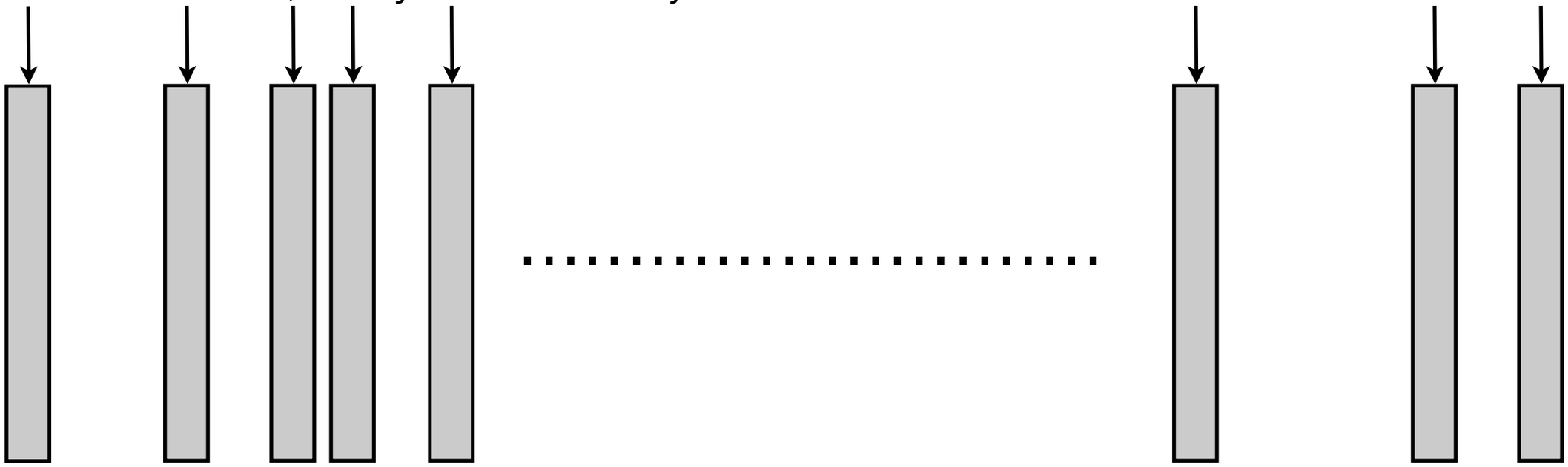
Outline

- Paradigmatic Context representation
- **Clustering Model**
- Co-occurrence Modeling

- Probabilistic Voting
- HMM based Model
- Noisy Channel Model
- Conclusion

Clustering of the substitute distributions:

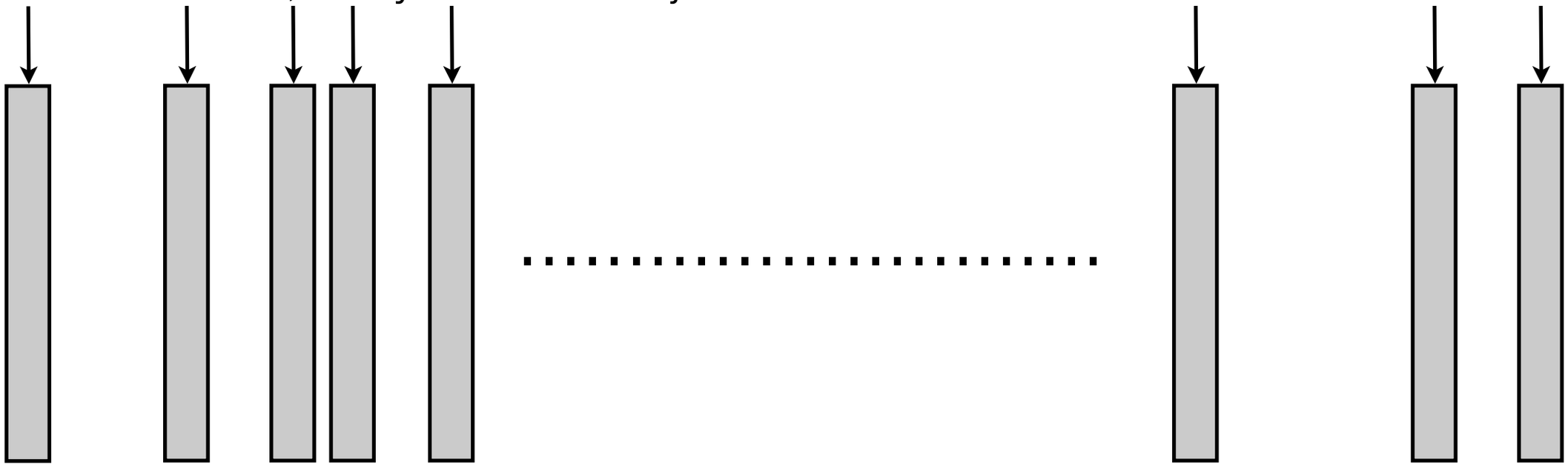
- Pierre Vinken , 61 years old will join the board as a nonexecutive director .



- Domain of the substitute distributions is the vocabulary
- Entries of the substitute distributions are probabilities

Clustering of the substitute distributions:

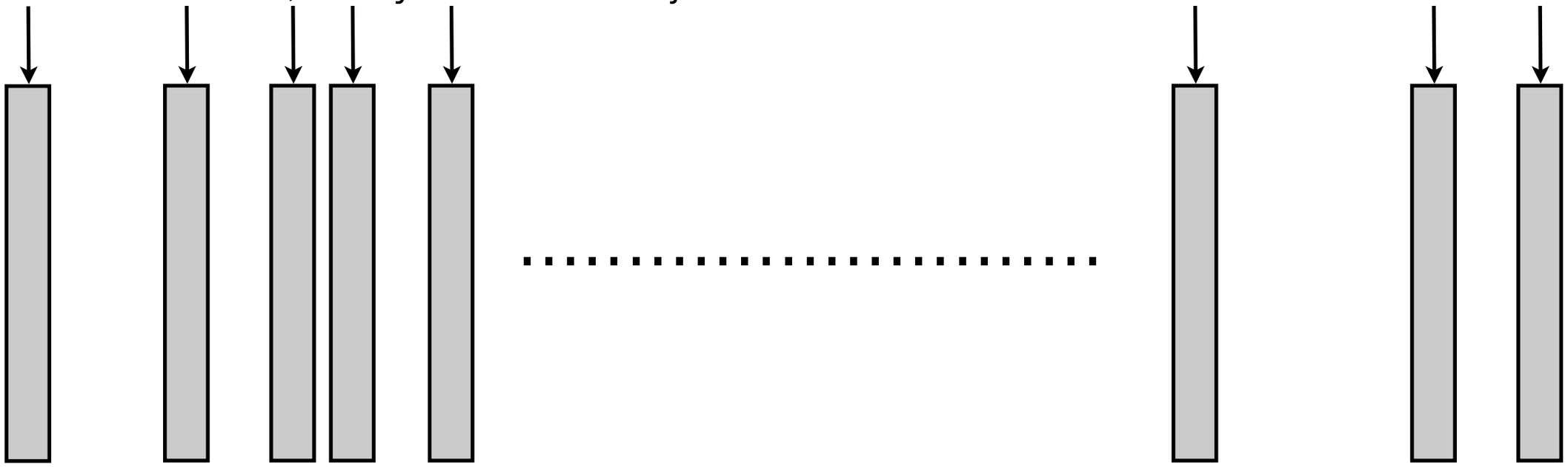
- Pierre Vinken , 61 years old will join the board as a nonexecutive director .



- Apply clustering model on **substitute distributions**.

Clustering of the substitute distributions:

- Pierre Vinken , 61 years old will join the board as a nonexecutive director .



- Apply clustering model on **substitute distributions**.

achieves **~60 %** accuracy on POS induction

Clustering of the substitute distributions:

Clustering of the substitute distributions:

- Assumes a **w**ord is independent of the **t**ag given the **c**ontext.

Clustering of the substitute distributions:

- Assumes a **w**ord is independent of the **t**ag given the **c**ontext.
- Ignores **word features**
  - Different instances of the same word can not share information

Clustering of the substitute distributions:

- Assumes a **w**ord is independent of the **t**ag given the **c**ontext.
- Ignores **word features**
  - Different instances of the same word can not share information
- Tags of consecutive words are independent of each other given the contexts:
  - Ex: determiner “a” usually followed by a singular noun (ex: cat).

Outline

- Paradigmatic Context representation
- Clustering Model (POS induction)
- **Co-occurrence Modeling (POS induction)**
- Probabilistic Voting (POS disambiguation)
- HMM based Model (POS disambiguation)
- Noisy Channel Model (WSD disambiguation)
- Conclusion

Co-occurrence Modeling

- Incorporates **word-features** by modeling co-occurrence of words and their substitutes
- Maps co-occurrence data to embeddings on n-dimensional sphere
  - Transforms co-occurrence probabilities to distances on sphere

Modeling Co-occurrence

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	**....**

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	**....**

- **sample k** substitutes from **substitute distribution** .

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	**....**

- **sample k** substitutes from **substitute distribution** .
  - "will join the _____ as a nonexecutive"

when k = 1

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	**....**

- **sample k** substitutes from **substitute distribution** .
  - "will join the _____ as a nonexecutive"

board

when k = 1

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	....

- **sample k** substitutes from **substitute distribution** .
  - "will join the _____ as a nonexecutive"
    board
  - "will join the _____ as a nonexecutive"

when k = 1

when k = 4

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...

board	**0.4288**
company	**0.2584**
firm	**0.2024**
bank	**0.0731**
....	....

- **sample k** substitutes from **substitute distribution** .
  - "will join the _____ as a nonexecutive"

    board

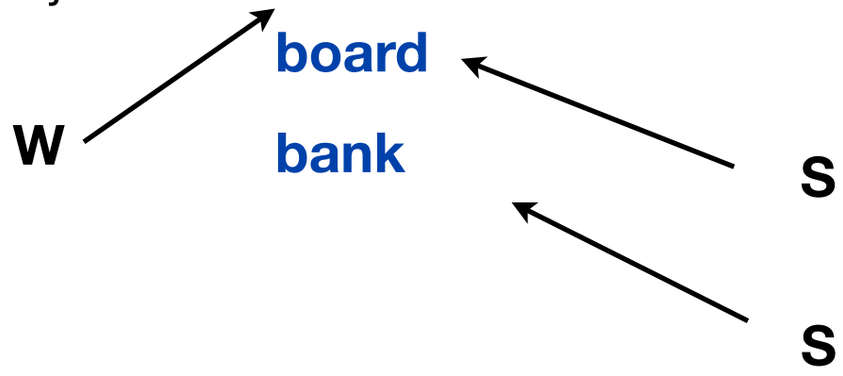
    when k = 1
  - "will join the _____ as a nonexecutive"

    firm
    board
    company
    board

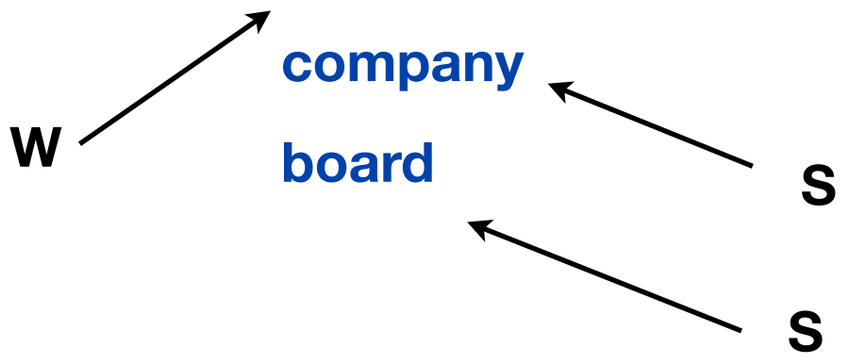
    when k = 4

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...



- ... 25 % of the seats on the **council** . </s>

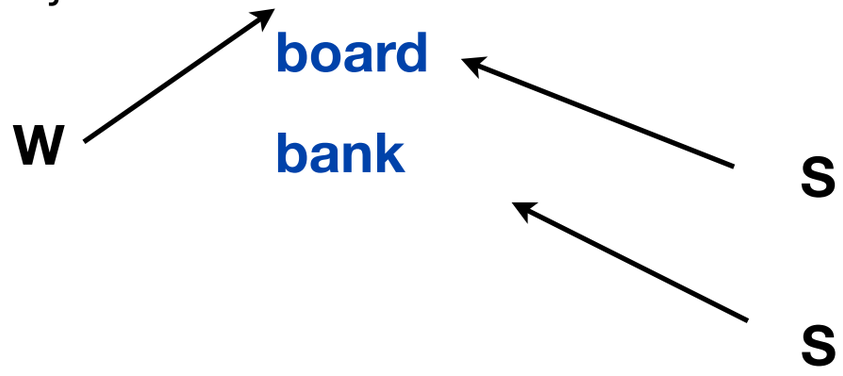


Example Co-occurrences	
Words	**Substitutes**
board	board
board	bank
council	company
council	board

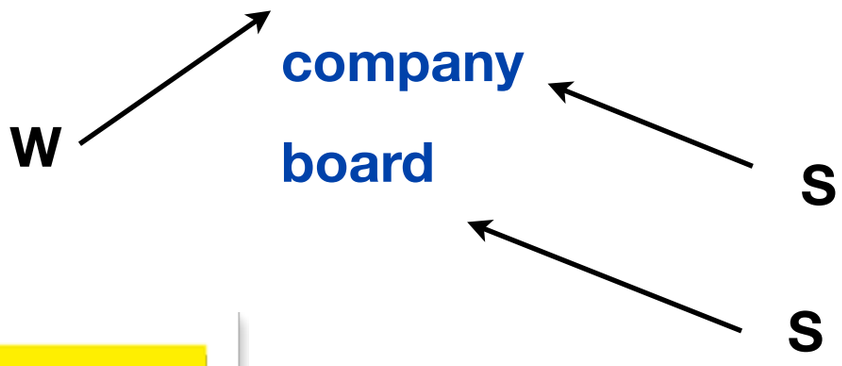
when k = 2

Modeling Co-occurrence

- ... will join the **board** as a nonexecutive ...



- ... 25 % of the seats on the **council** . </s>



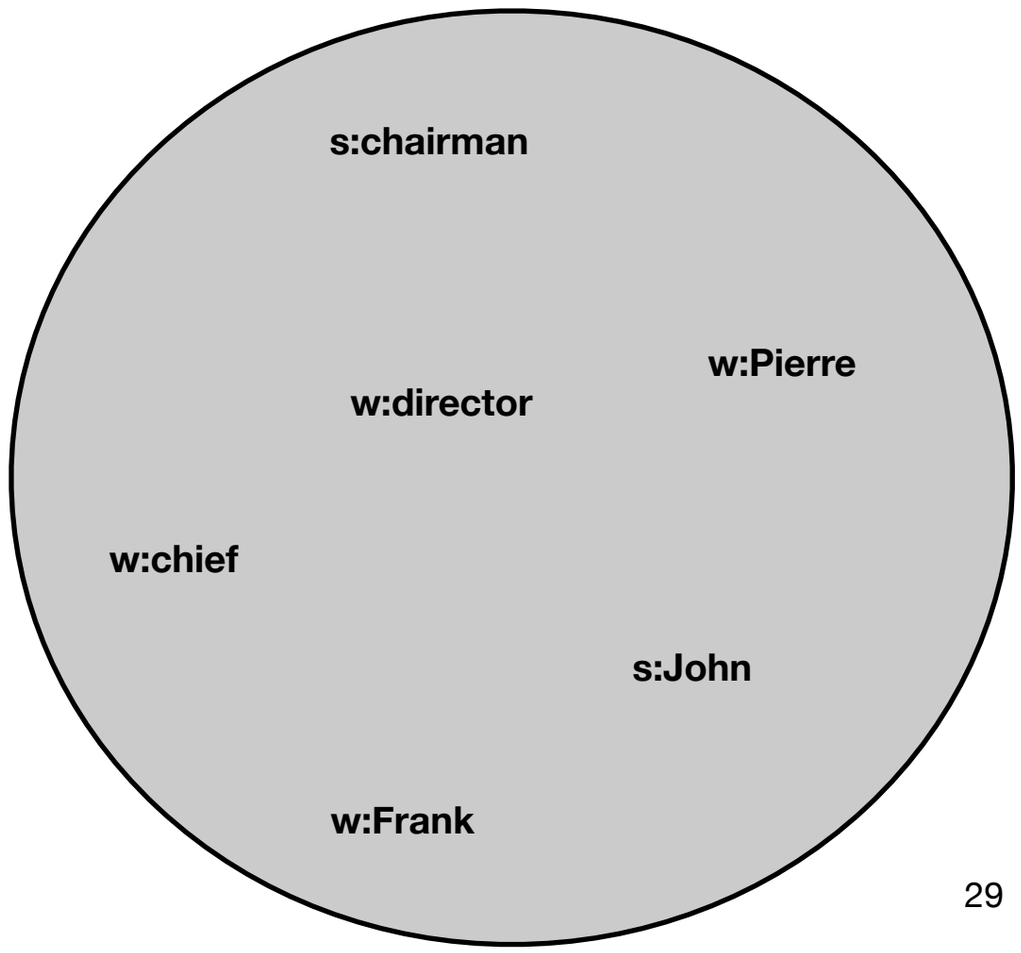
Different **W** values are pulled together by shared **S** values.

Example Co-occurrences	
Words	**Substitutes**
board	board
board	bank
council	company
council	board

when k = 2

Modeling Co-occurrence

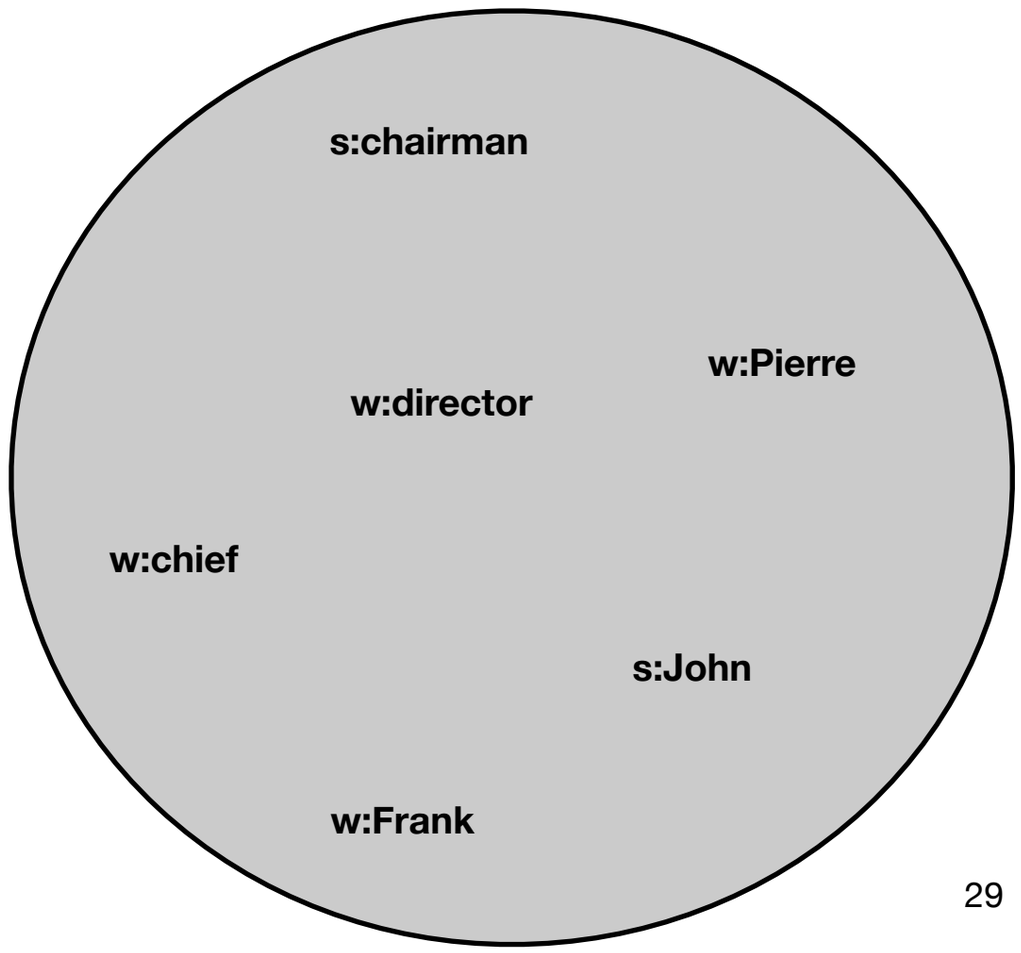
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)

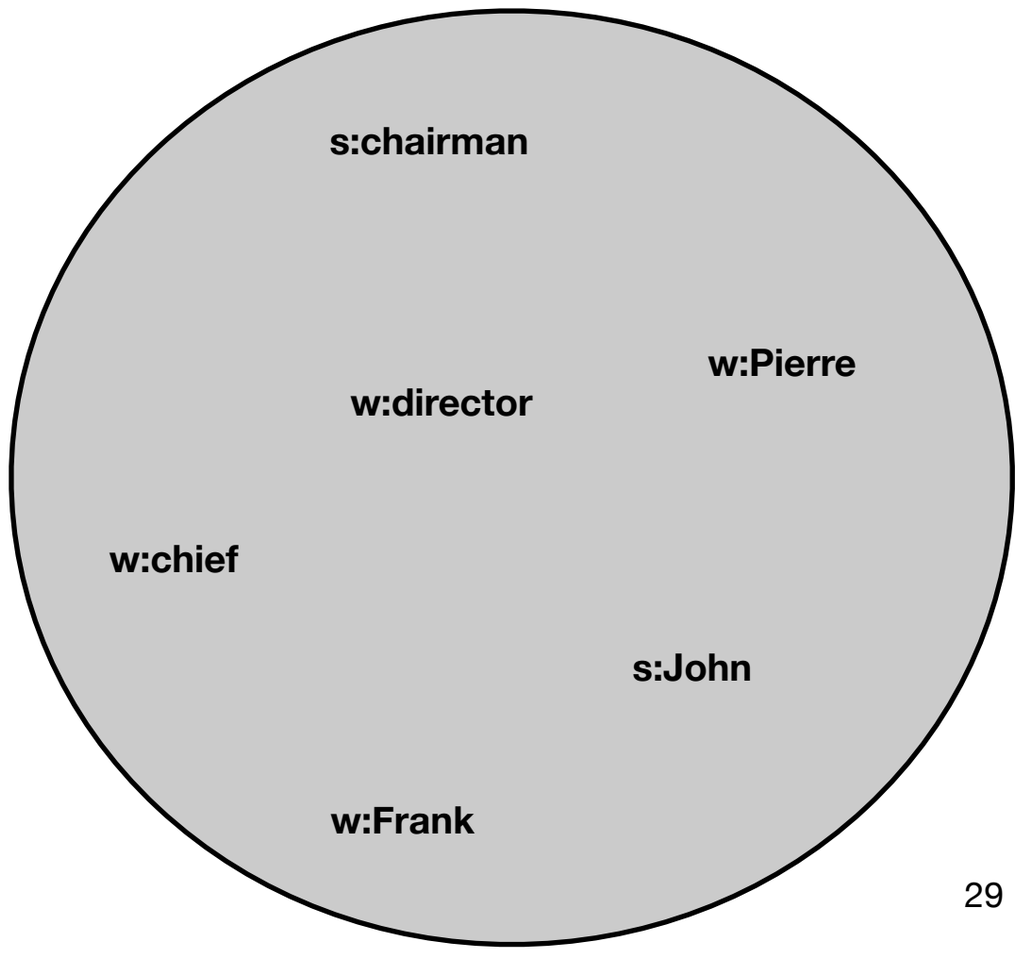
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.

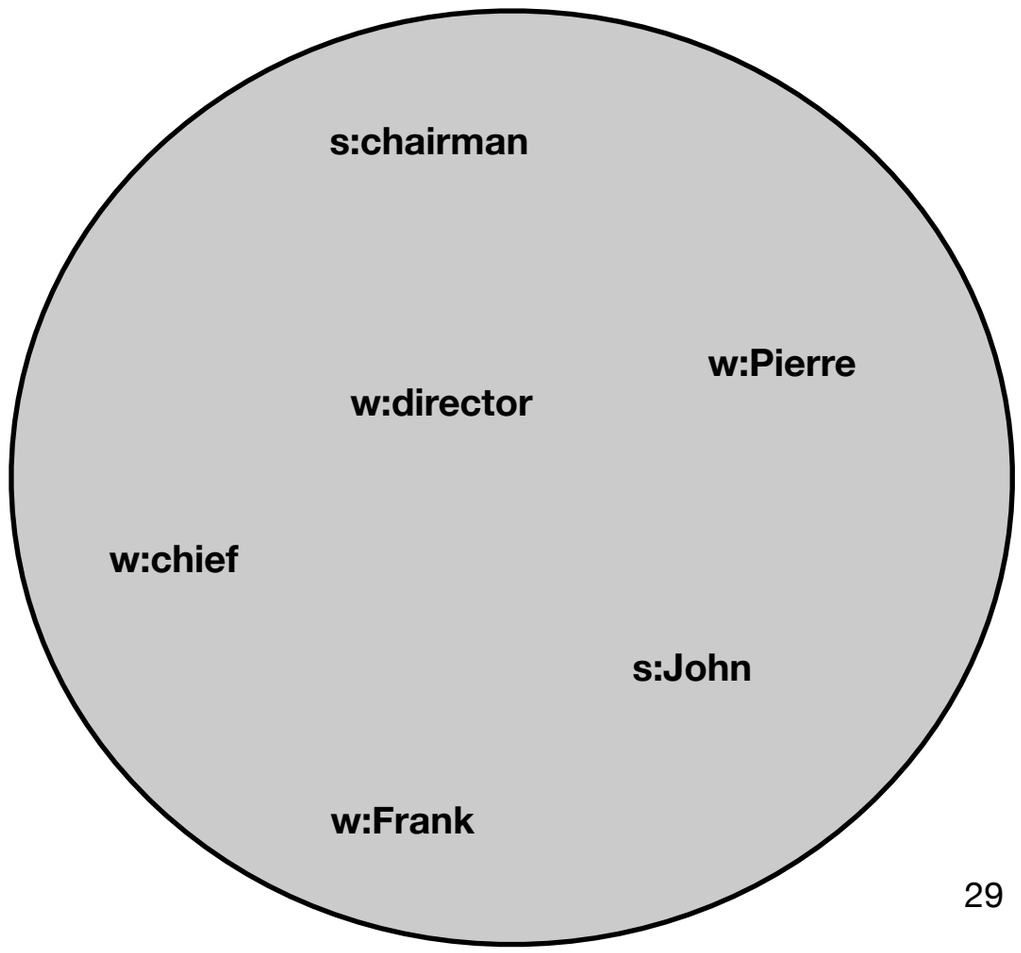
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)

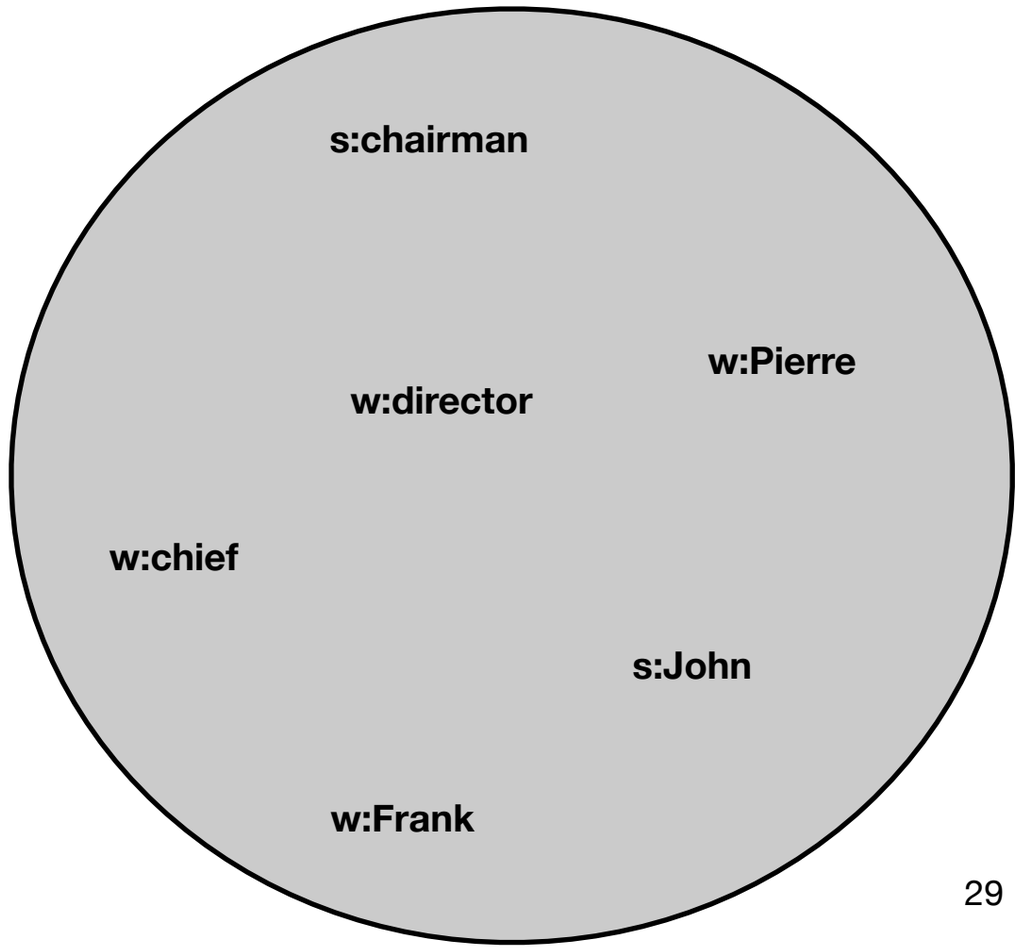
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)

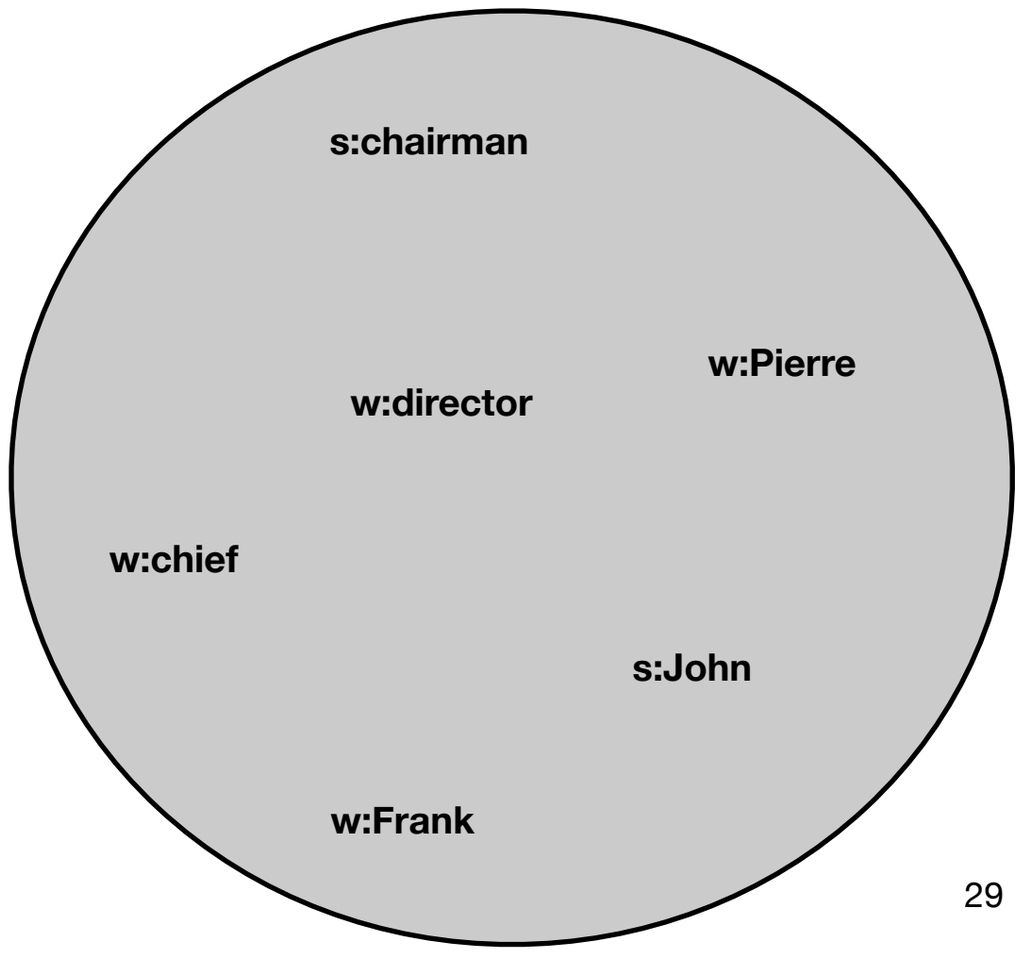
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere

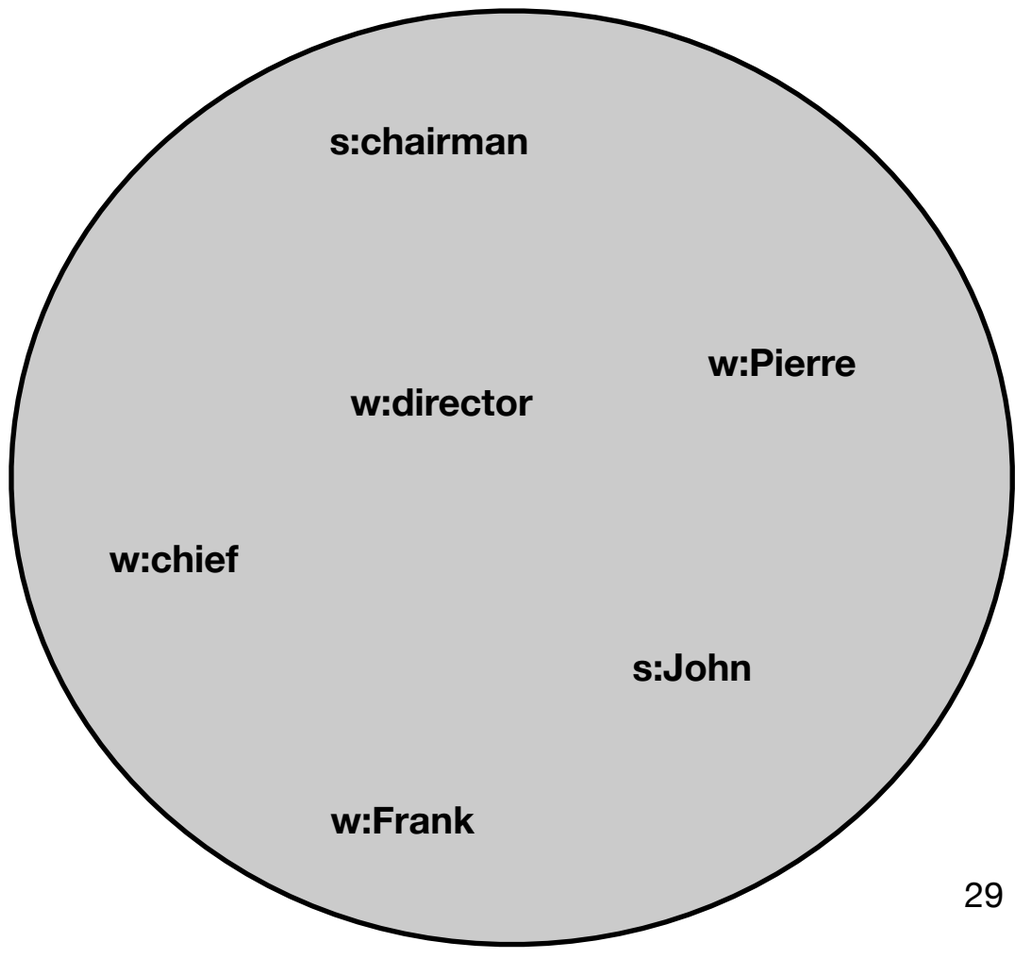
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere

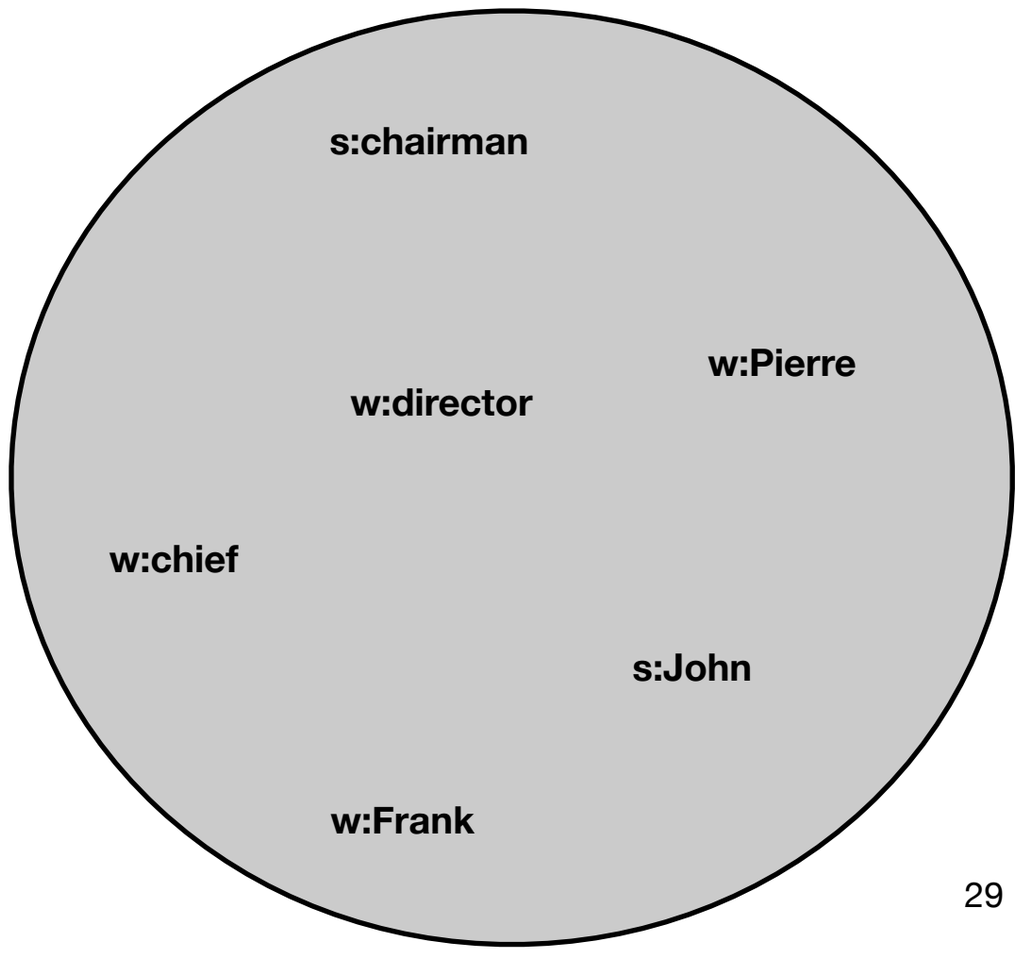
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

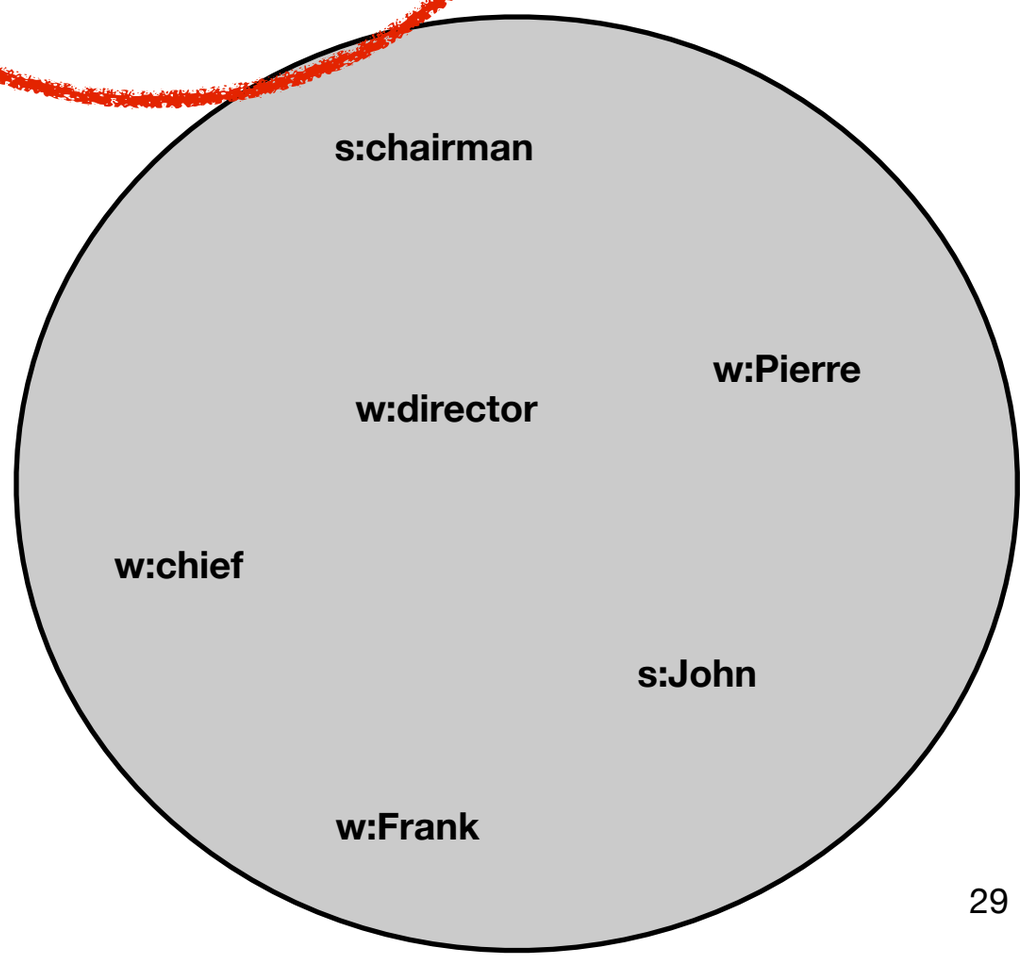
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

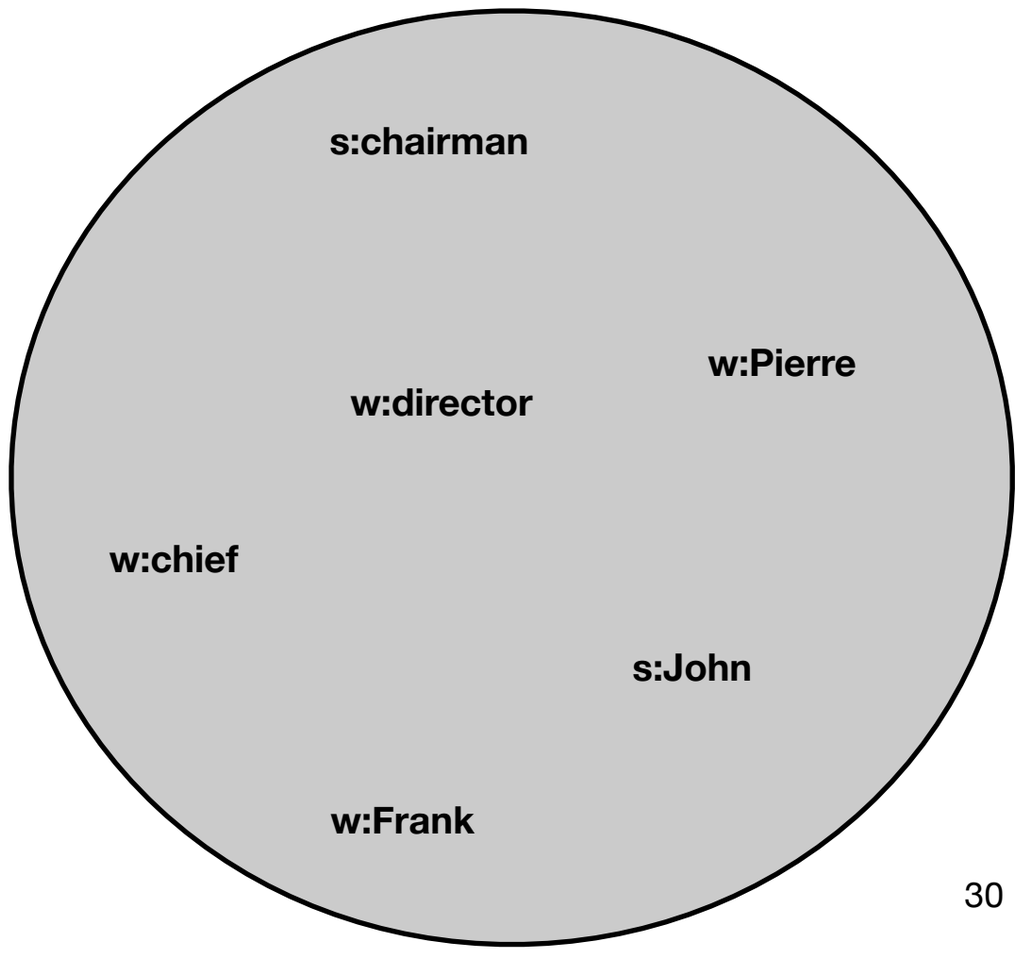
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

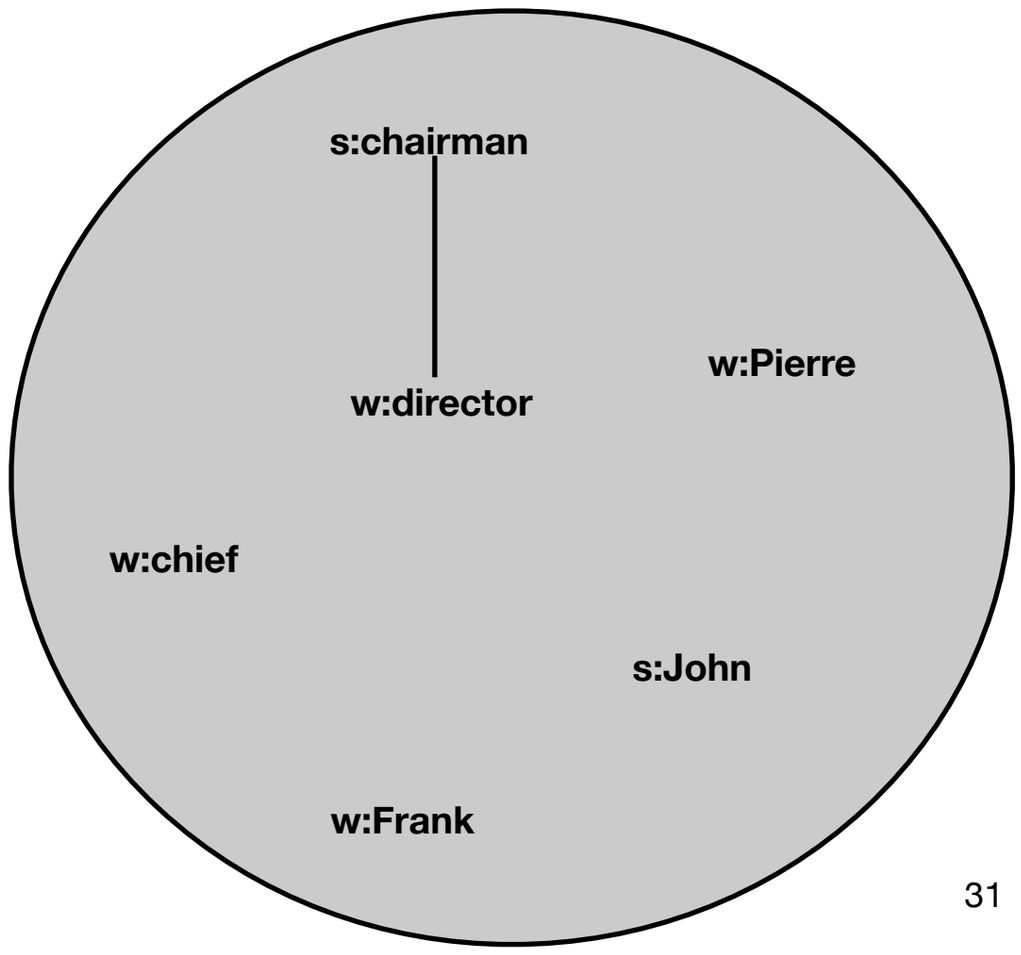
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

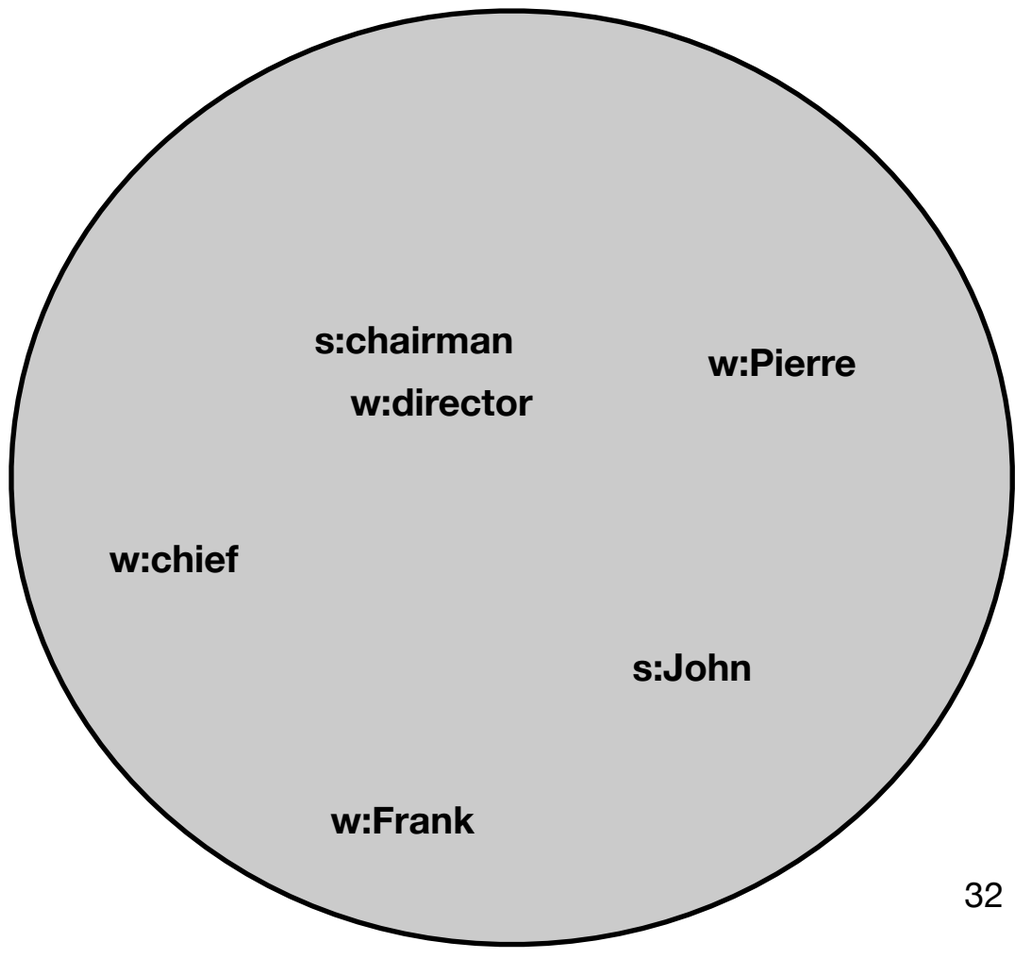
W	**S**
w:director	**s:chairman**
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from Pr(**W**,**S**)
  - Want to model Pr(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

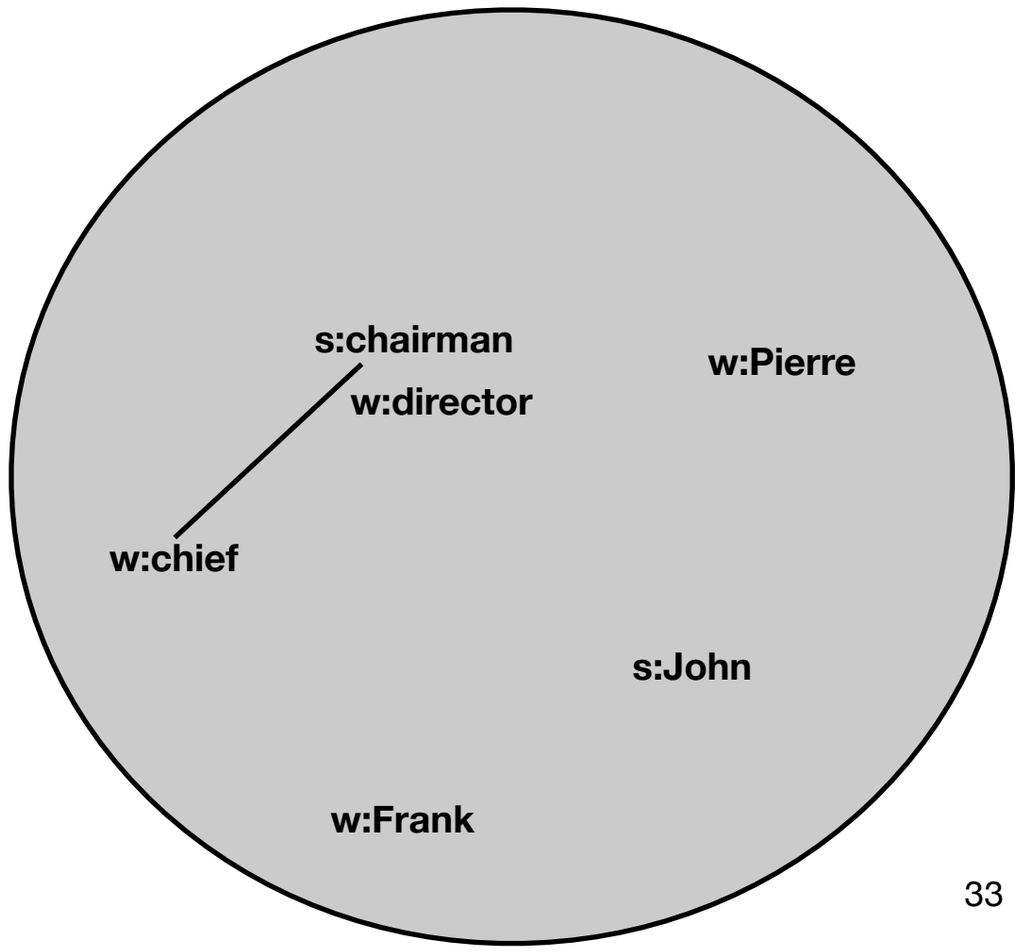
W	**S**
w:director	**s:chairman**
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

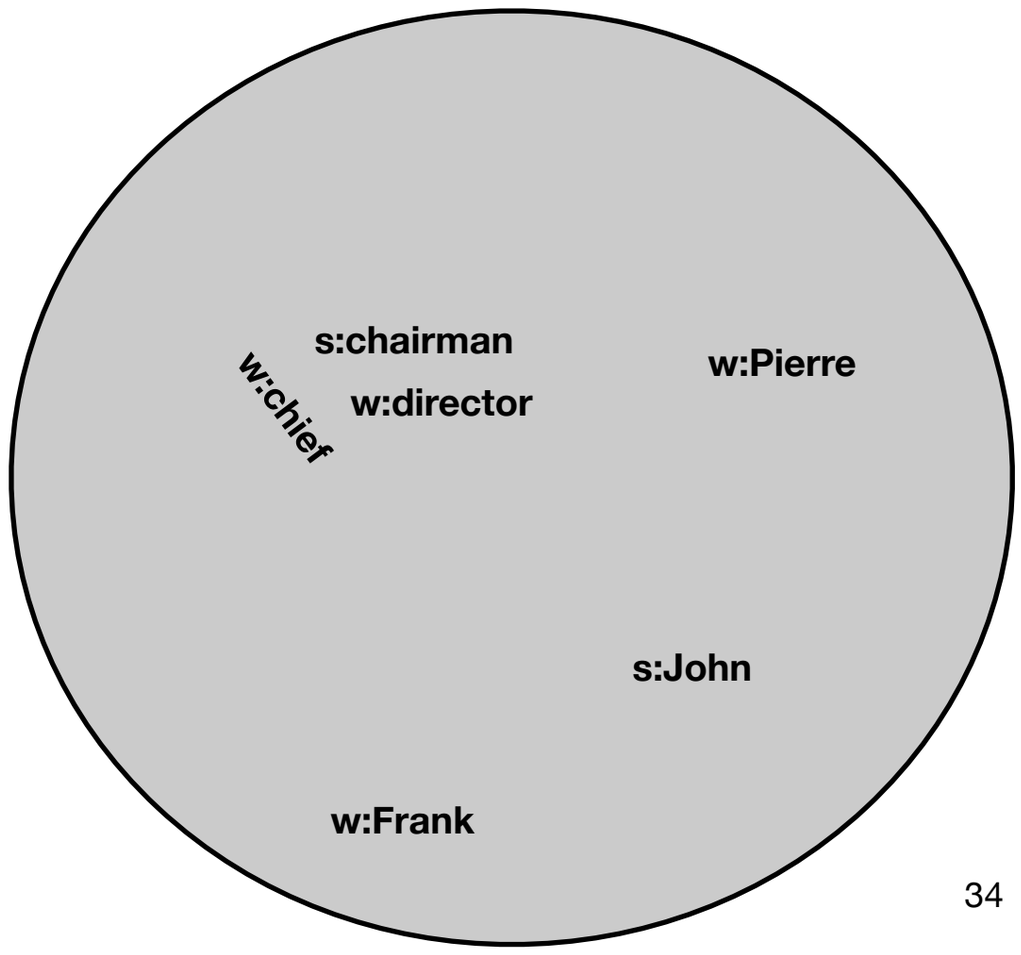
W	S
w:director	s:chairman
w:chief	**s:chairman**
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

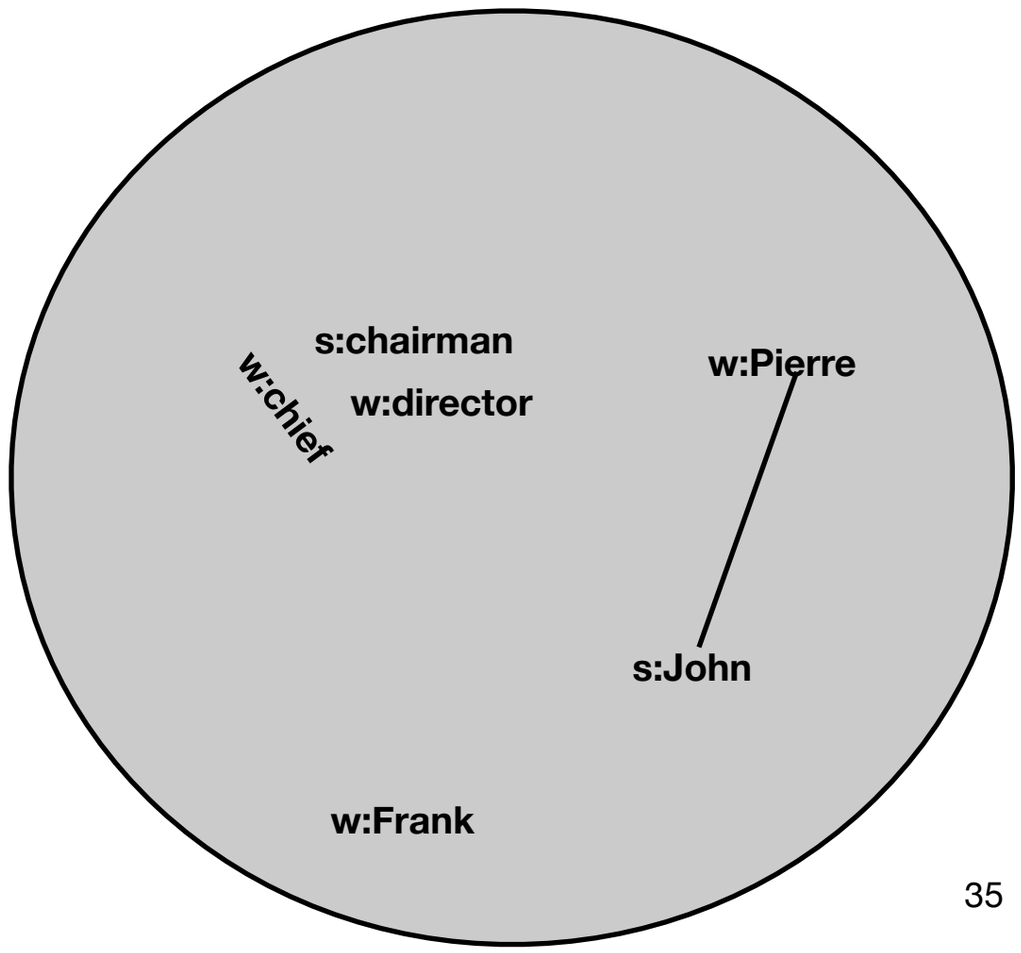
W	**S**
w:director	s:chairman
w:chief	**s:chairman**
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- S-CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

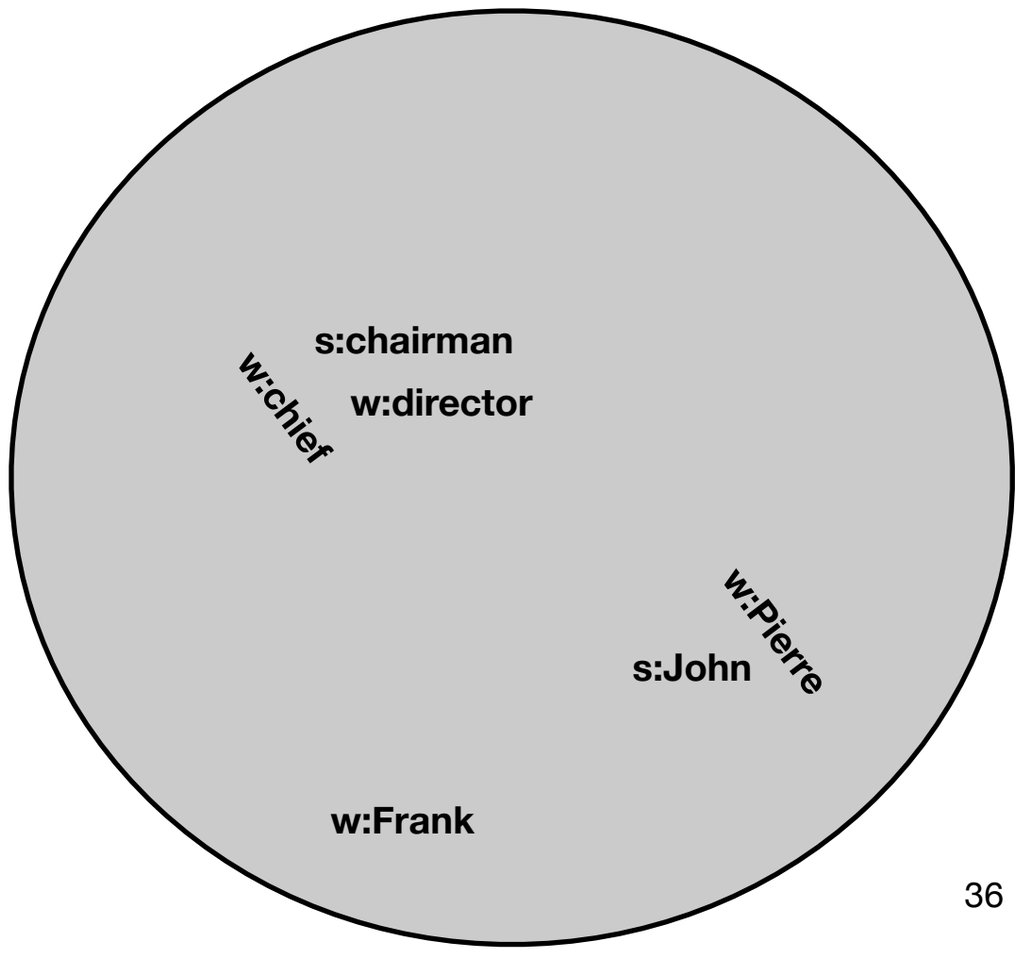
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	**s:John**
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

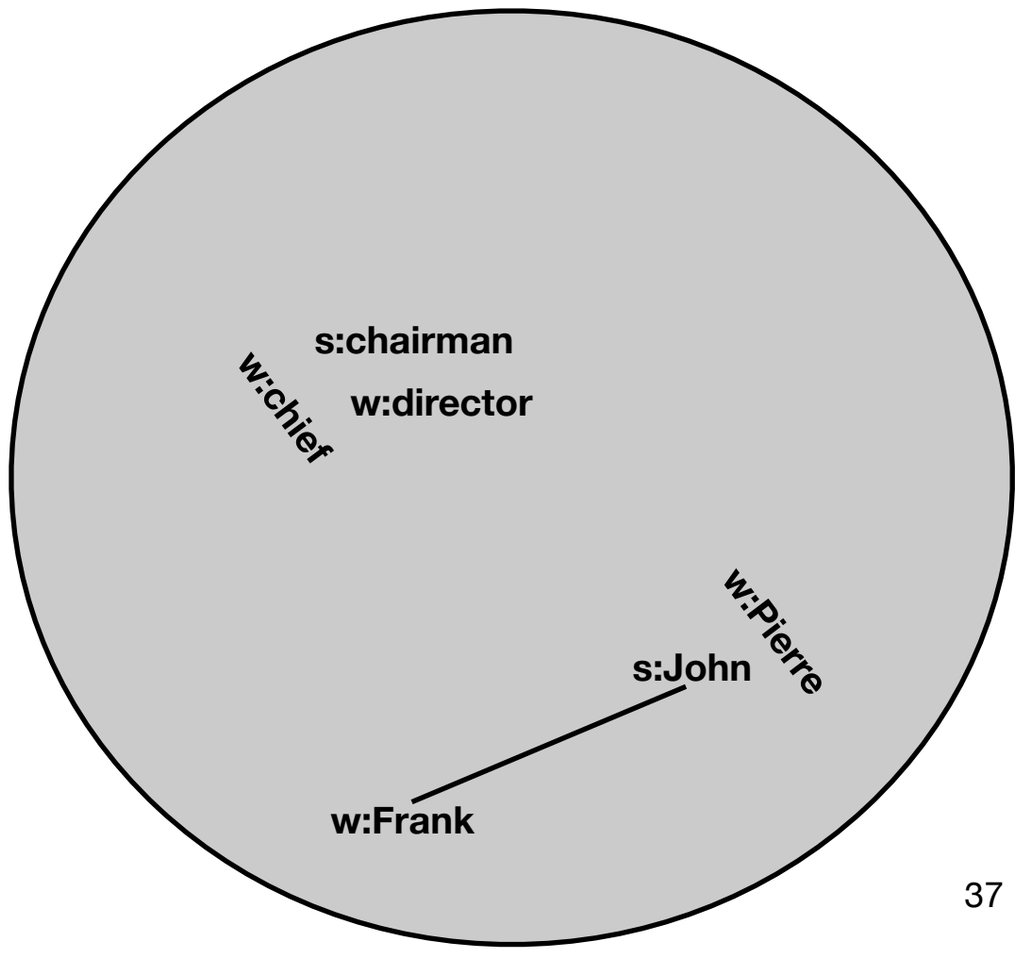
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	**s:John**
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

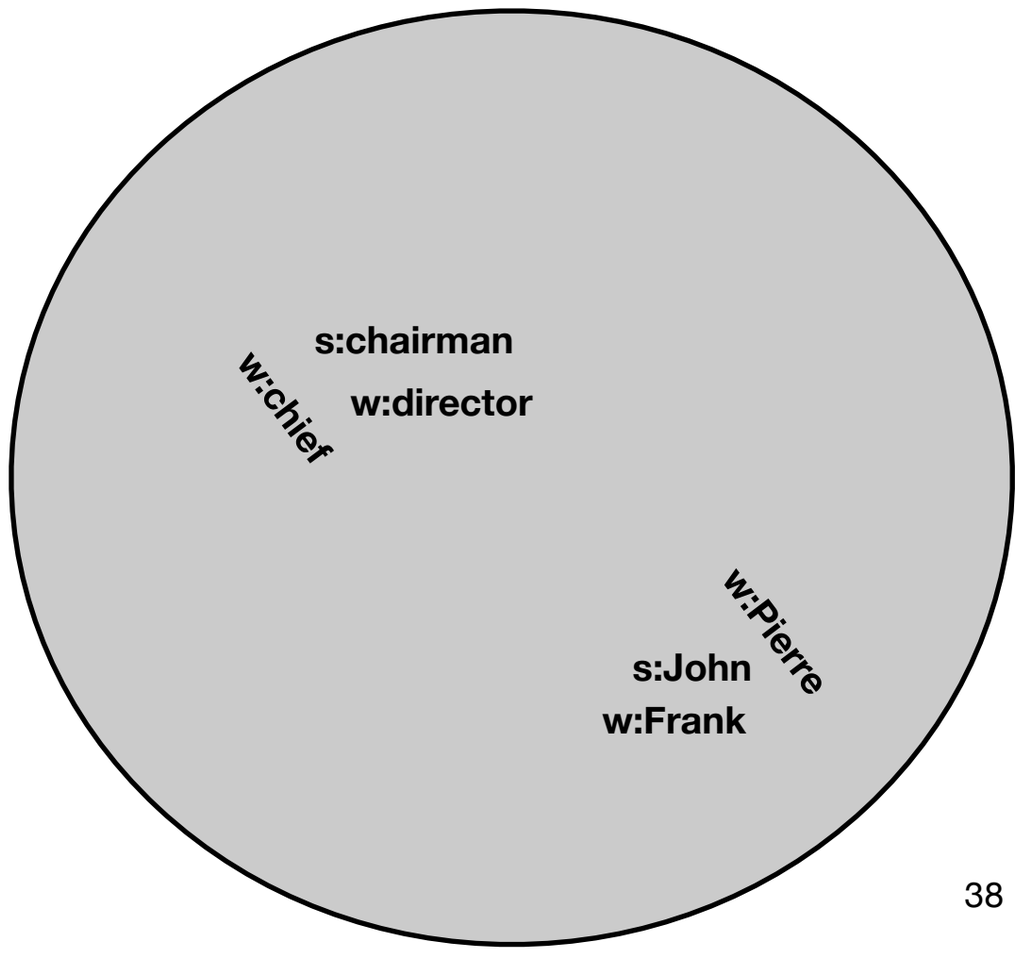
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	**s:John**



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
  - Want to model p(**W**,**S**)
  - Map W and S values to points on N-Sphere
  - Transform probabilities to distances on n-sphere
  - Frequently co-occurring values should map to nearby points.

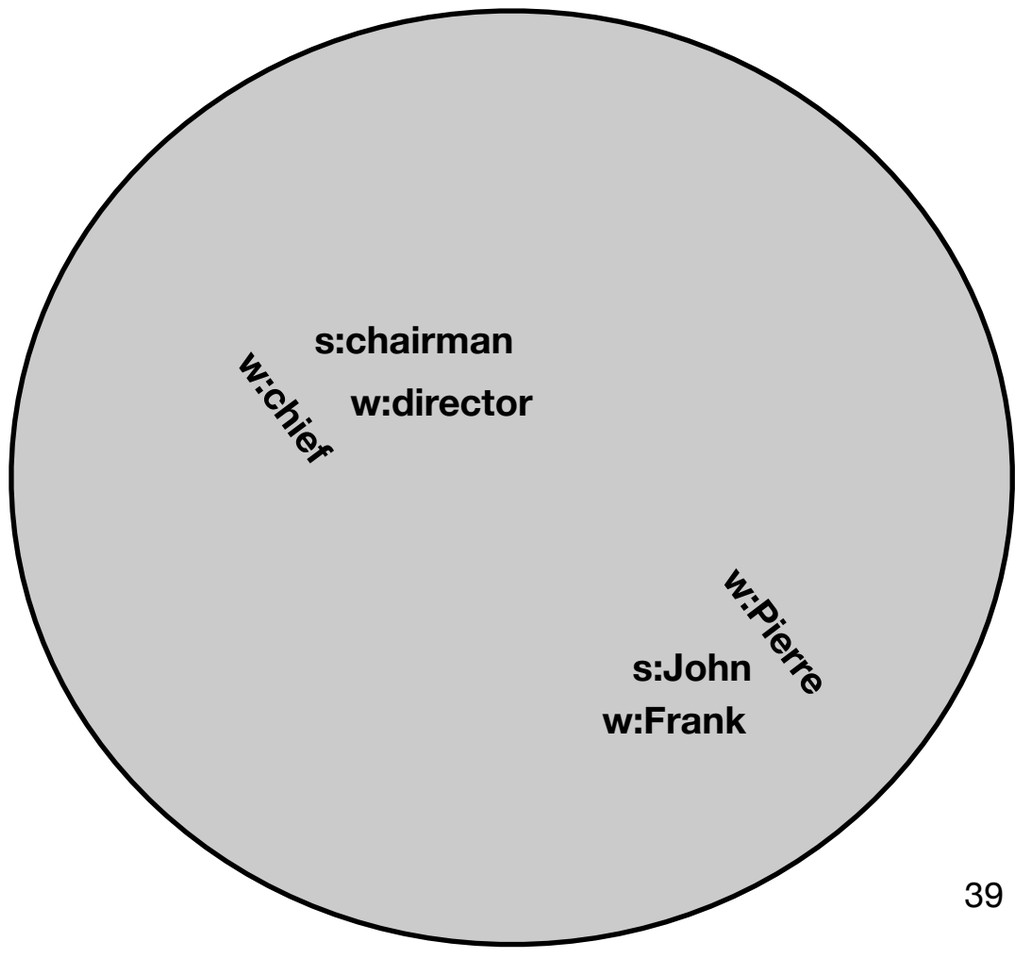
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
    - Want to model p(**W**,**S**)
    - Map W and S values to points on N-Sphere
    - Frequently co-occurring values should map to nearby points.
- Cluster the **W** points with k-means

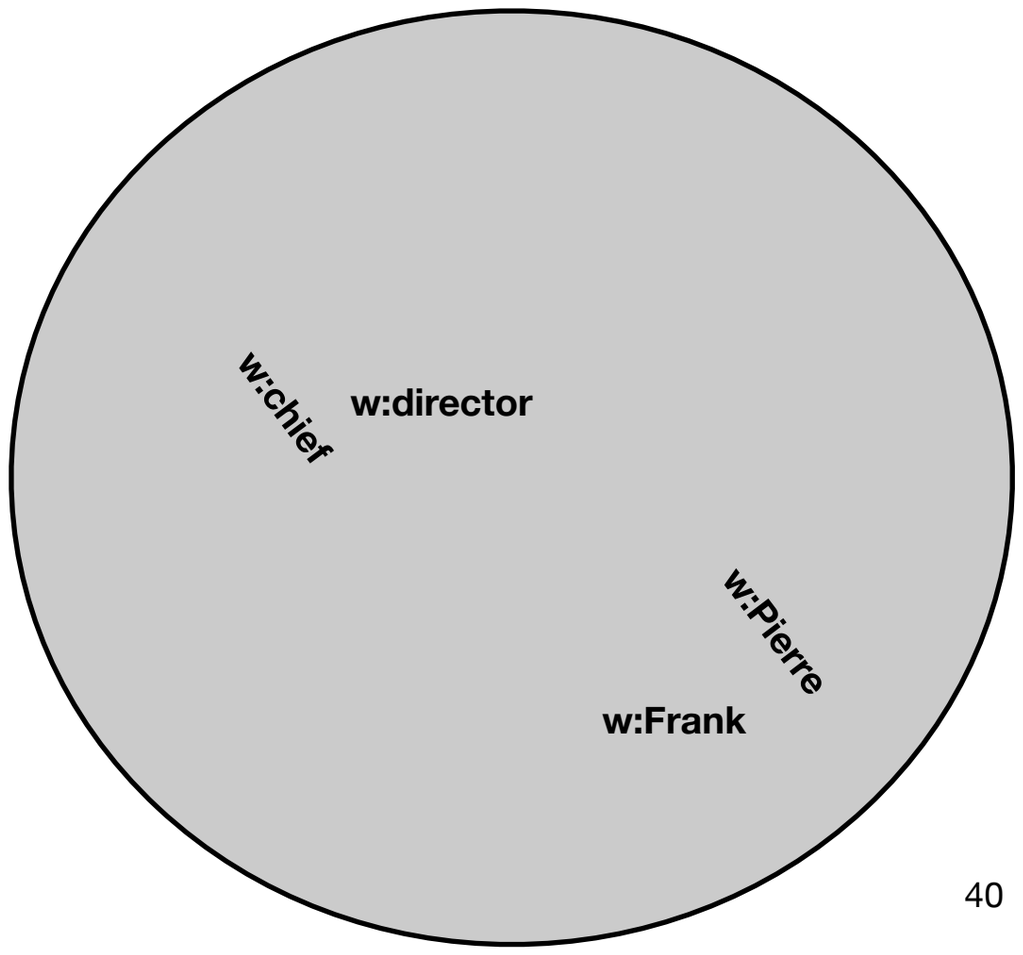
W	**S**
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
    - Want to model p(**W**,**S**)
    - Map W and S values to points on N-Sphere
    - Frequently co-occurring values should map to nearby points.
- Cluster the **W** points with k-means

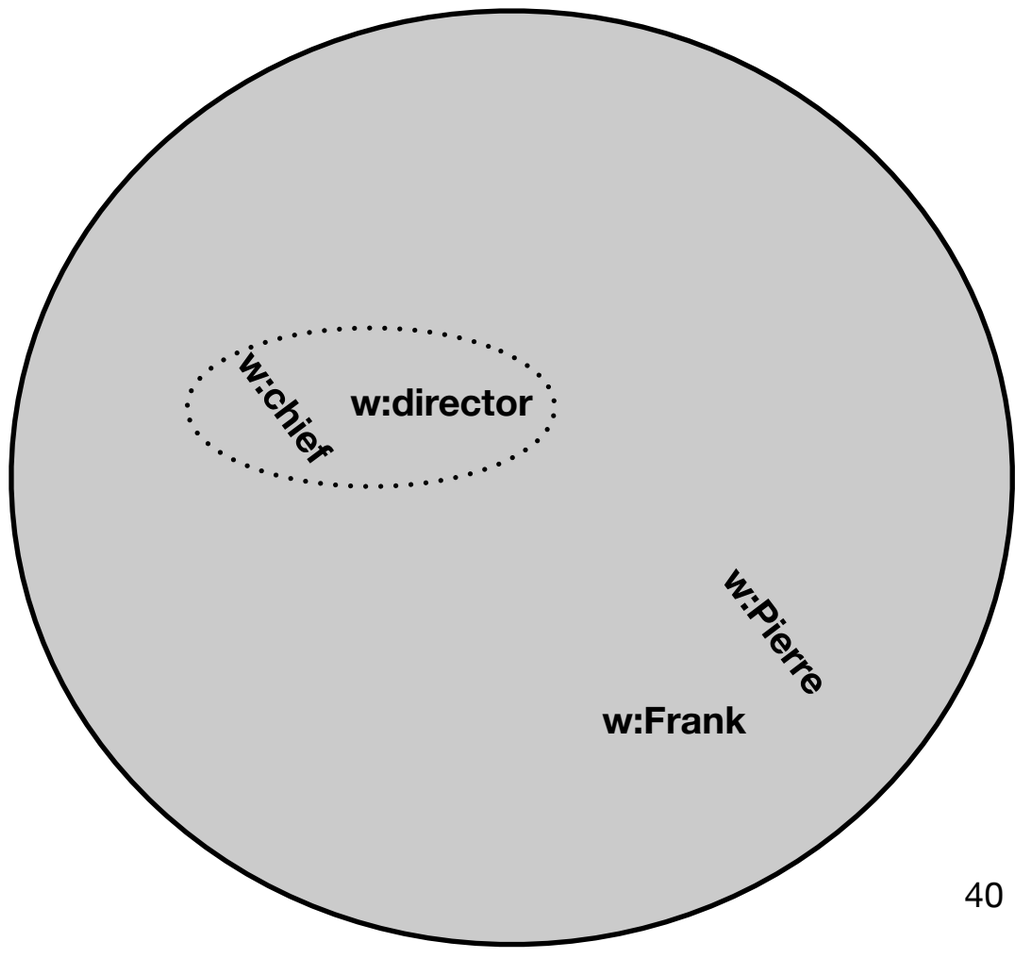
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
    - Want to model p(**W**,**S**)
    - Map W and S values to points on N-Sphere
    - Frequently co-occurring values should map to nearby points.
- Cluster the **W** points with k-means

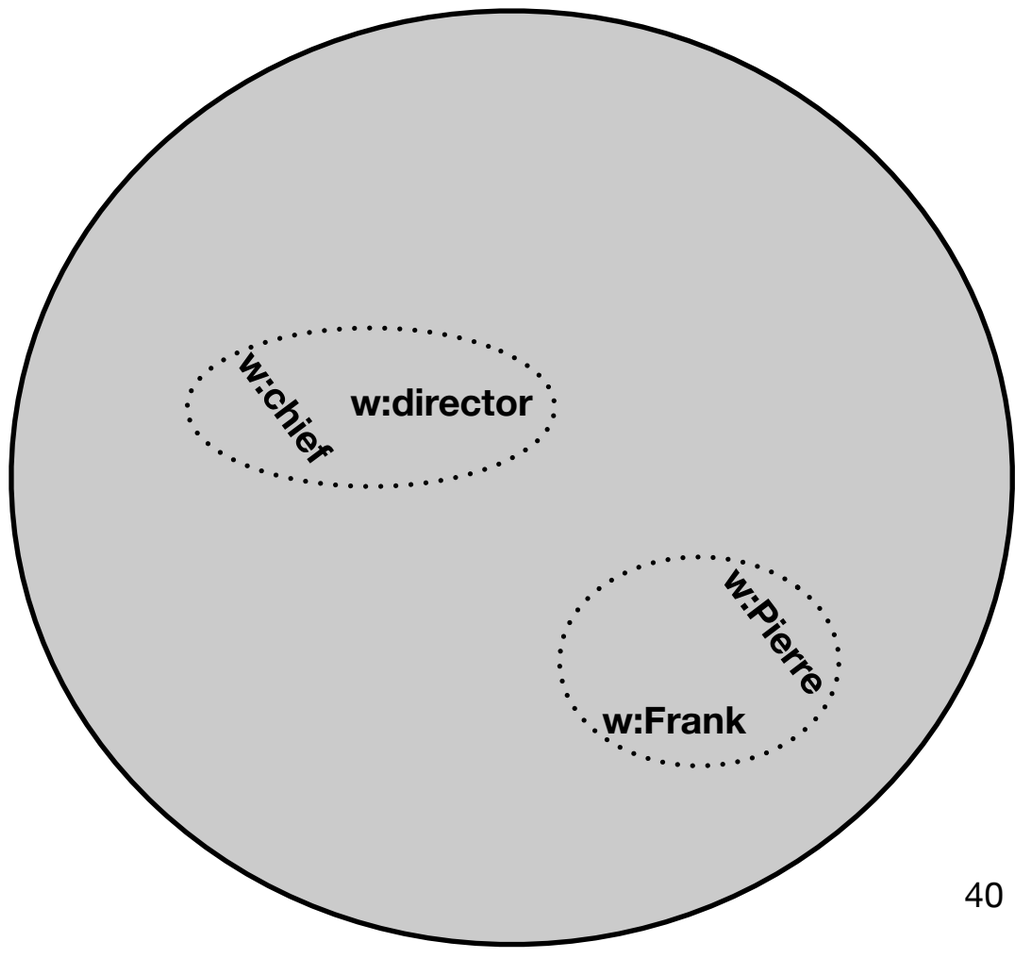
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- CODE (Maron et al. 2010)
  - **W**,**S** two categorical random variables.
  - Observe **W**,**S** pairs drawn i.i.d. from p(**W**,**S**)
    - Want to model p(**W**,**S**)
    - Map W and S values to points on N-Sphere
    - Frequently co-occurring values should map to nearby points.
- Cluster the **W** points with k-means

W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



CODE

- CODE defines the model **joint probability** of **W** and **S** as

$$p(w,s) = \frac{1}{Z}\overline{p}(w)\overline{p}(s)e^{-d^2(w,s)}$$

CODE

- CODE defines the model **joint probability** of **W** and **S** as

$$p(w,s) = \frac{1}{Z}\overline{p}(w)\overline{p}(s)e^{-d^2(w,s)}$$

Empirical marginals of W and S

CODE

- CODE defines the model **joint probability** of **W** and **S** as

Euclidean distance between embeddings of w and s

$$d^2_{w,s} = \| \Phi_w - \Phi_s \|^2$$

$$p(w,s) = \frac{1}{Z} \bar{p}(w) \bar{p}(s) e^{-d^2(w,s)}$$

Empirical marginals of W and S

CODE

- CODE defines the model **joint probability** of **W** and **S** as

Euclidean distance between embeddings of w and s

$$p(w,s) = \frac{1}{Z}\overline{p}(w)\overline{p}(s)e^{-d^2(w,s)}$$

$$d^2_{w,s} = \| \Phi_w - \Phi_s \|^2$$

Empirical marginals of W and S

Embeddings of w and s

CODE

- CODE defines the model **joint probability** of **W** and **S** as

Euclidean distance between embeddings of w and s

$$d^2_{w,s} = \| \Phi_w - \Phi_s \|^2$$

$$p(w,s) = \frac{1}{Z} \bar{p}(w) \bar{p}(s) e^{-d^2(w,s)}$$

Normalization Constant

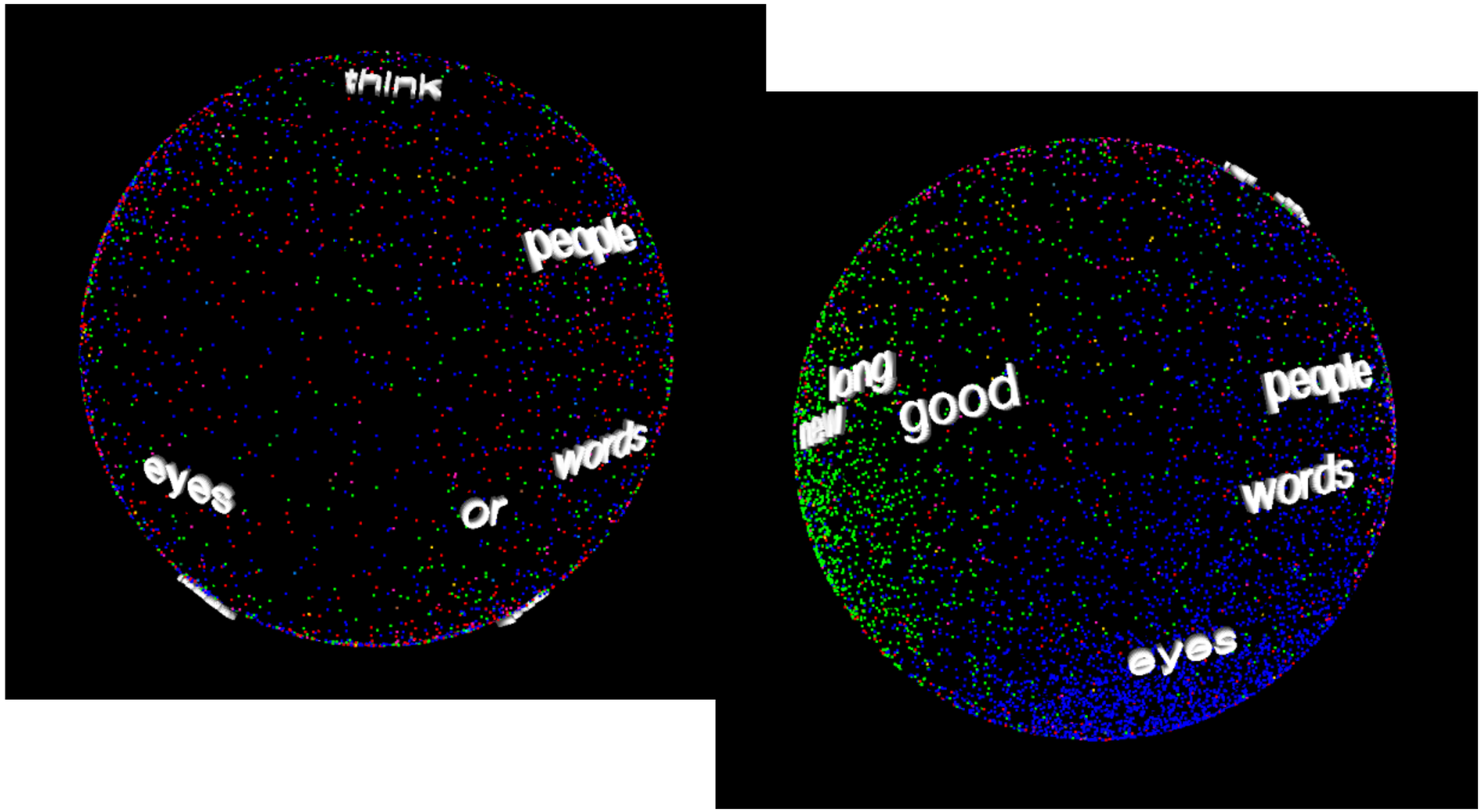
$$Z = \sum_{(w,s)} \bar{p}(w) \bar{p}(s) e^{-d^2_{w,s}}$$

Empirical marginals of W and S

Embeddings of w and s

Modeling Co-occurrence

- Click here for a demo (may take a few minutes to load)

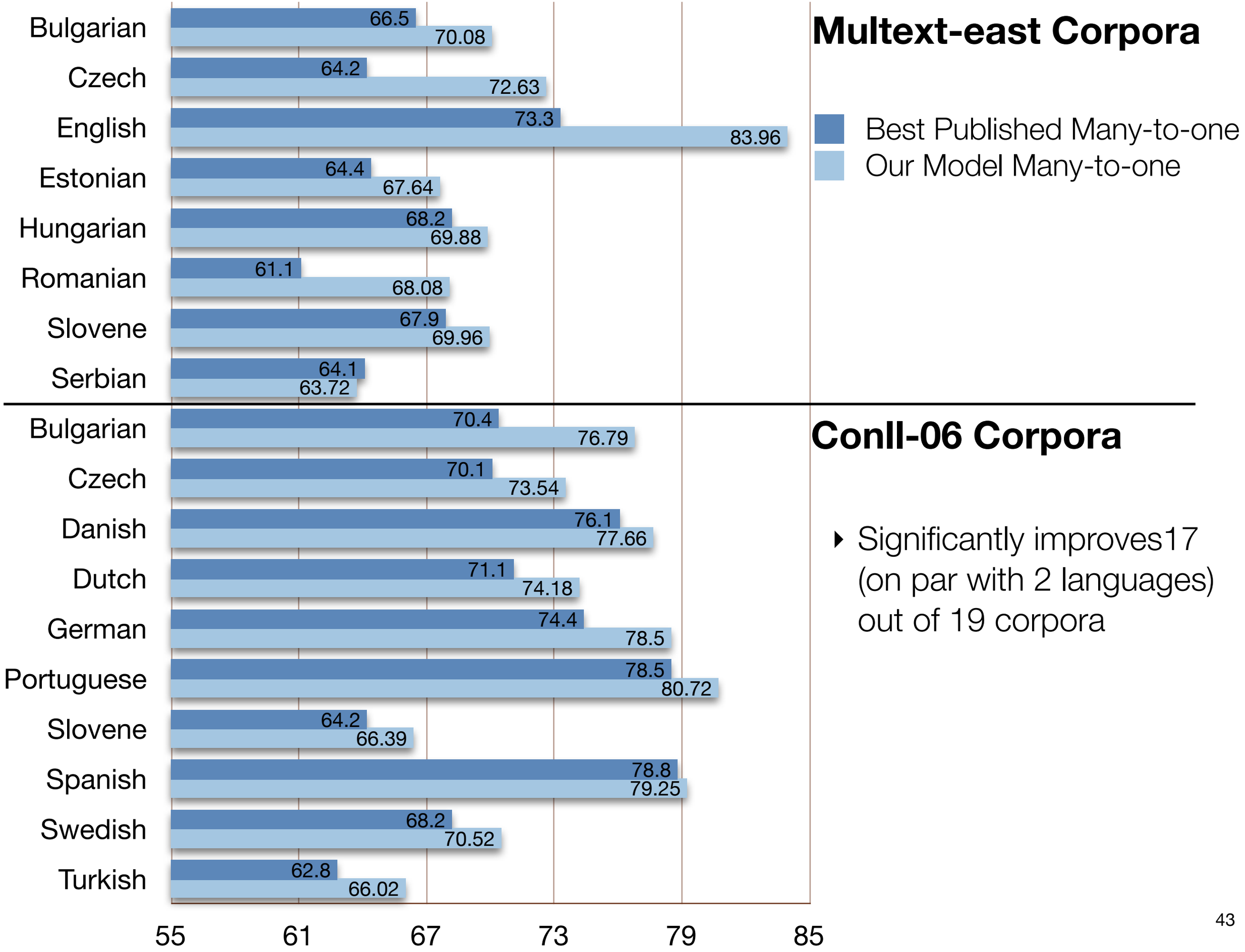


Multext-east Corpora

Best Published Many-to-one
Our Model Many-to-one

Conll-06 Corpora

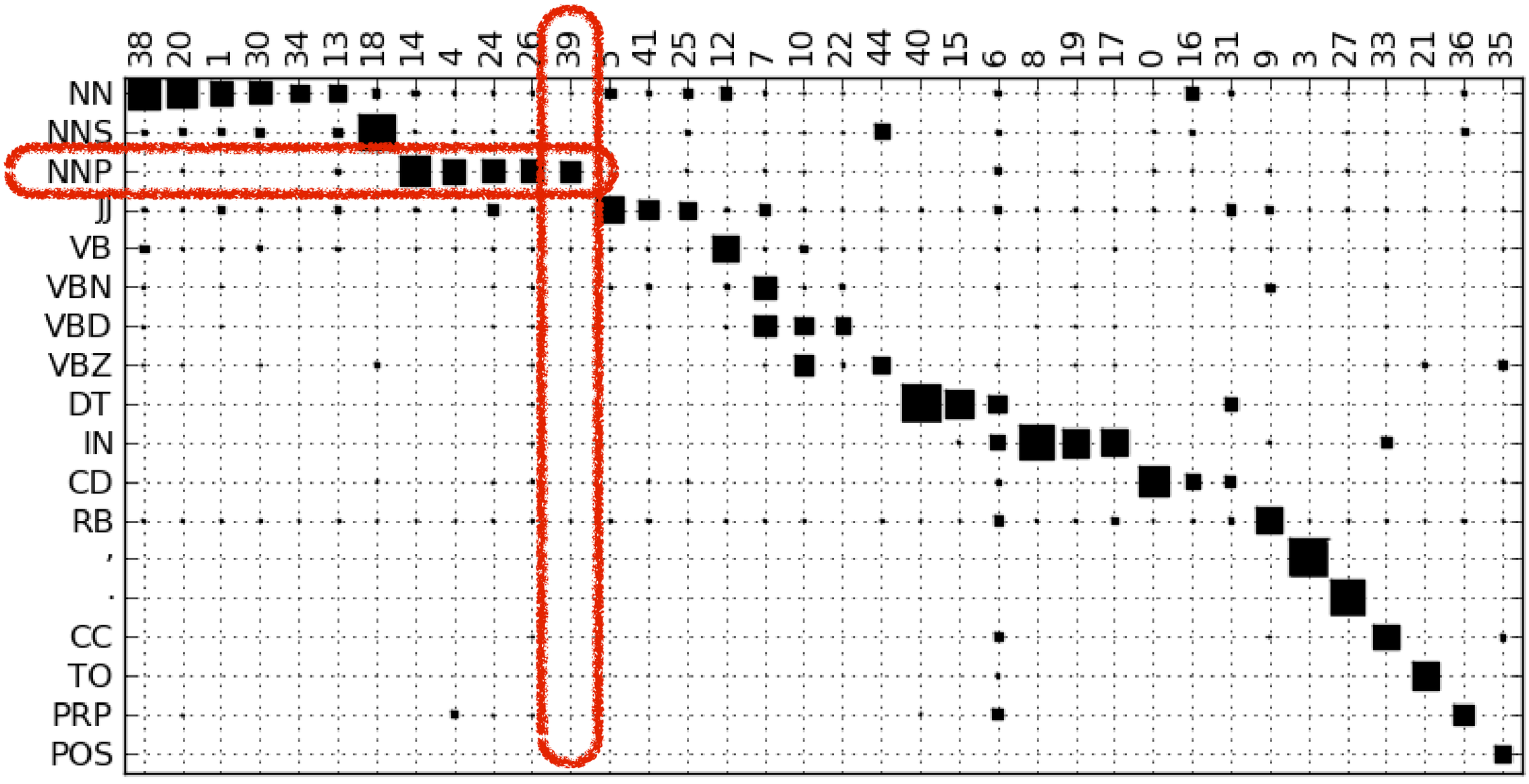
- Significantly improves17 (on par with 2 languages) out of 19 corpora



- Significantly improves17 (on par with 2 languages) out of 19 corpora

Experiments and Results

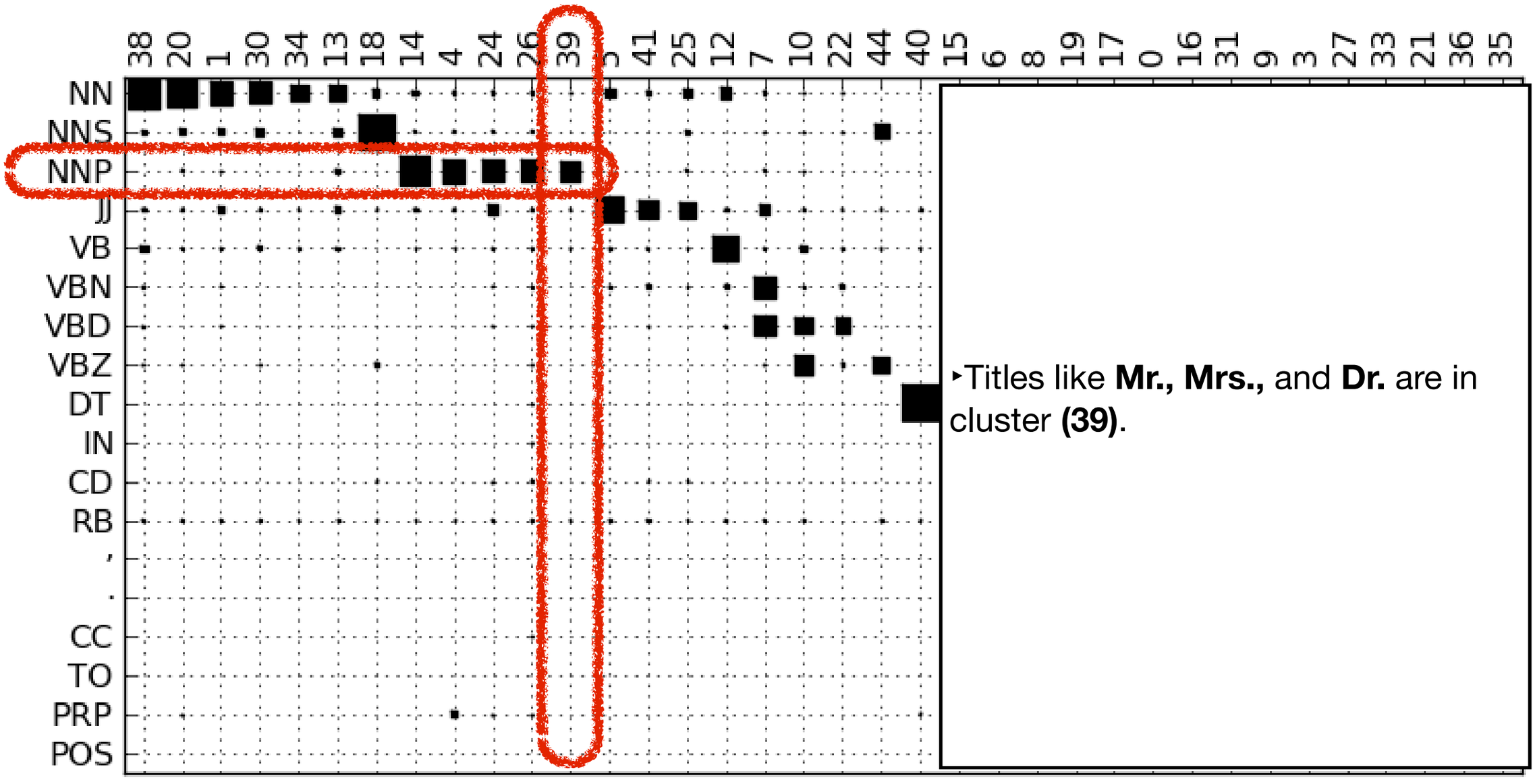
- Hinton Graph



number of clusters is set to number of gold-tags

Experiments and Results

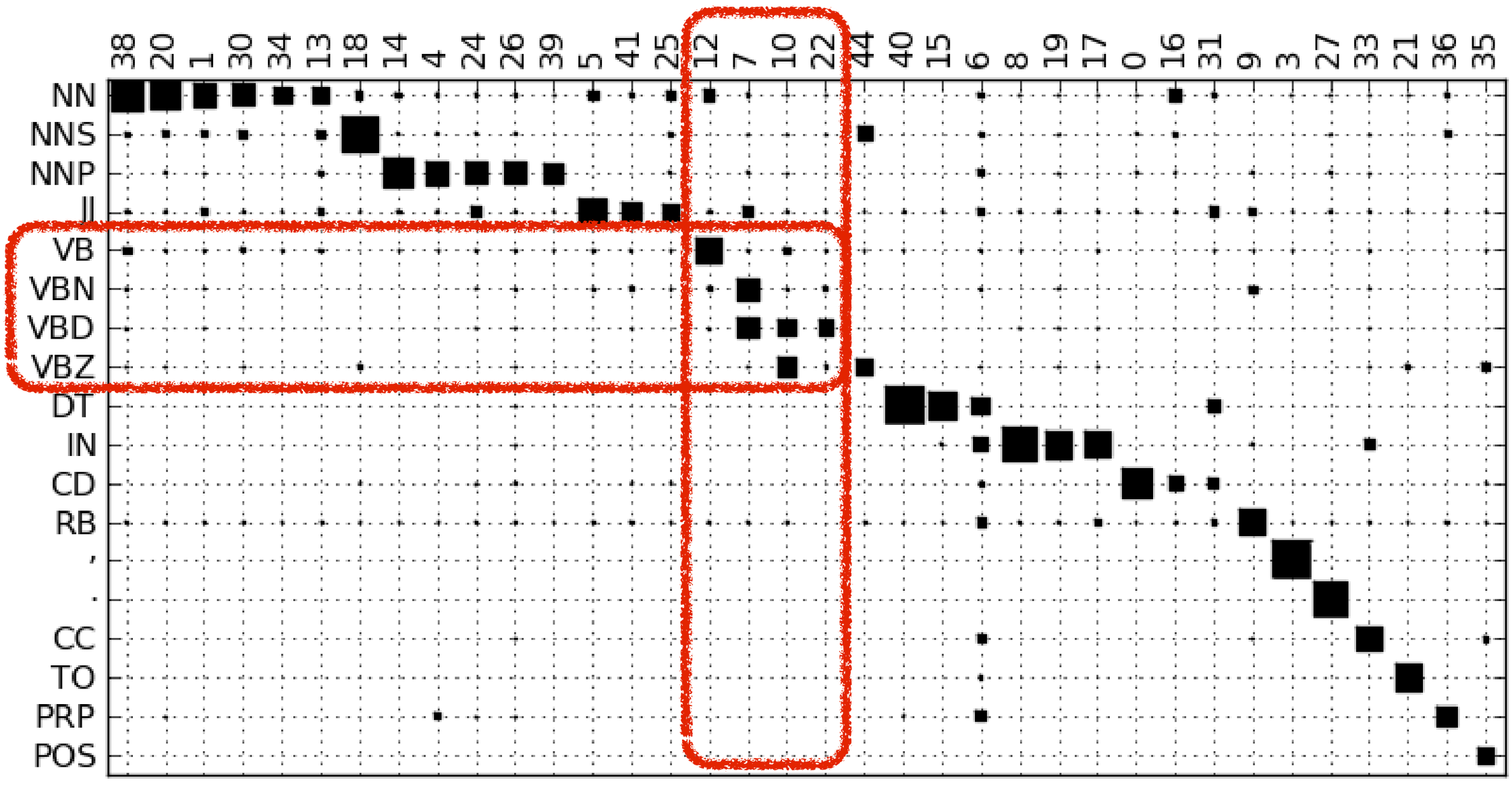
- Hinton Graph



number of clusters is set to number of gold-tags

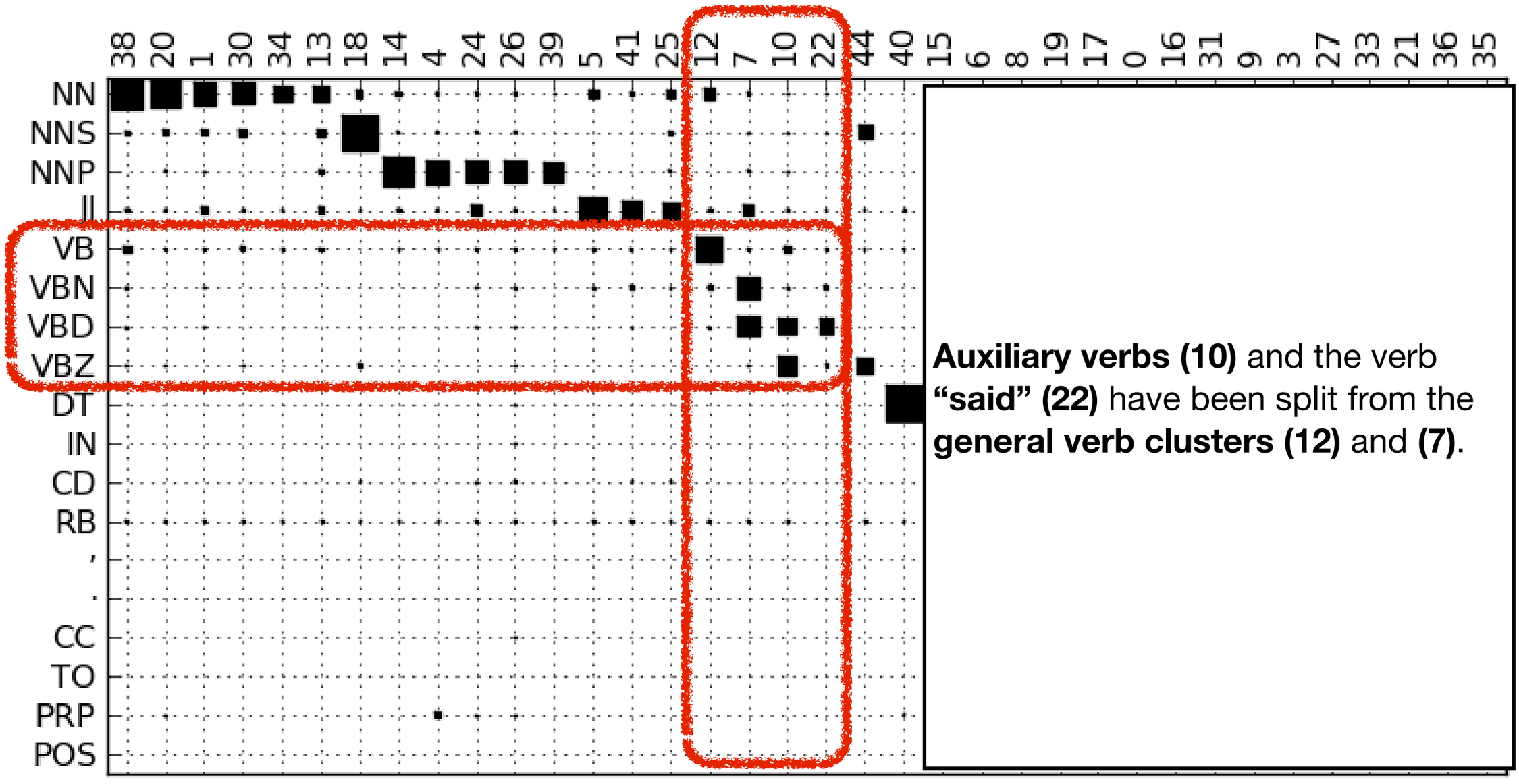
Experiments and Results

- Hinton Graph



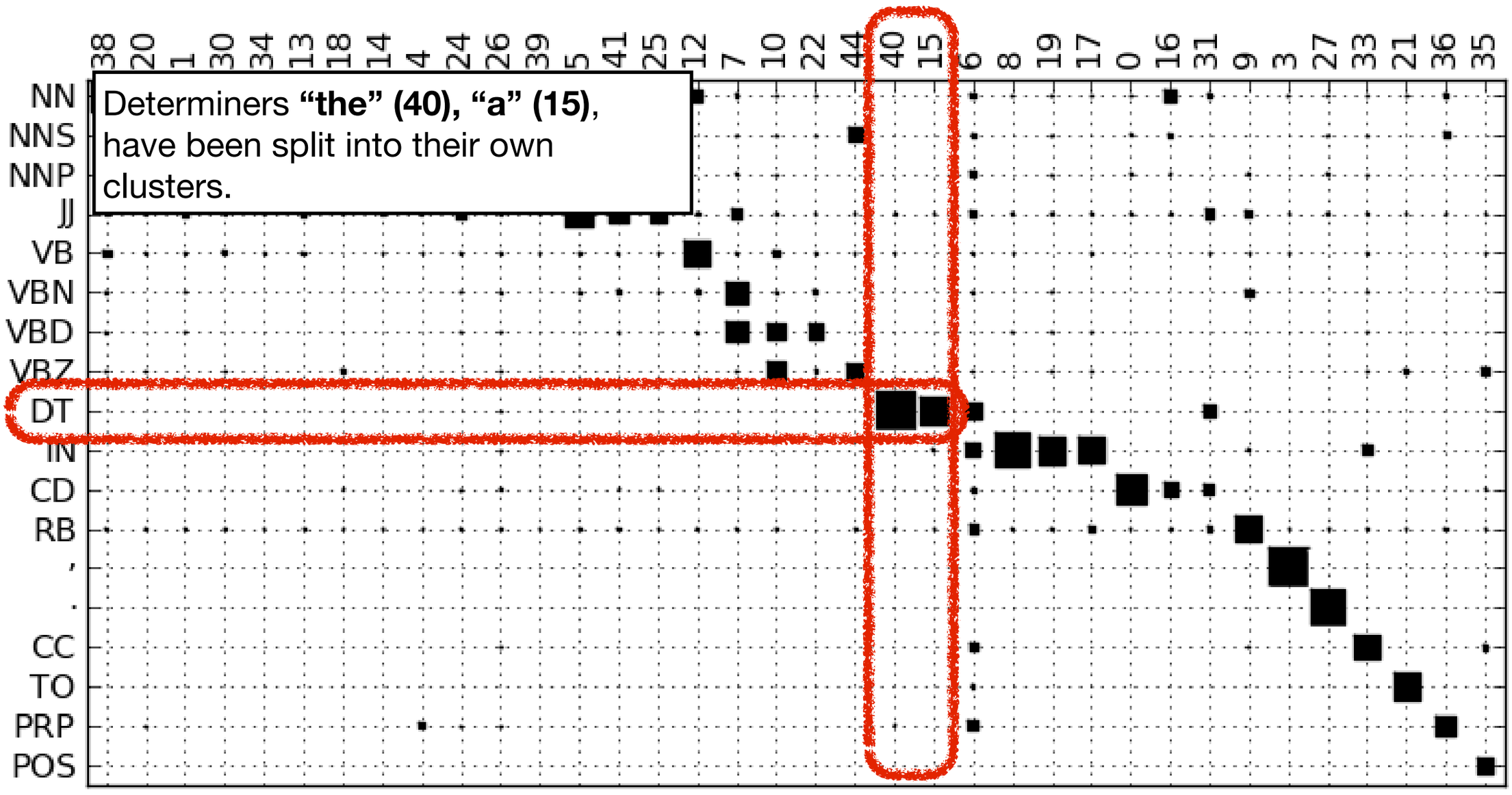
Experiments and Results

- Hinton Graph



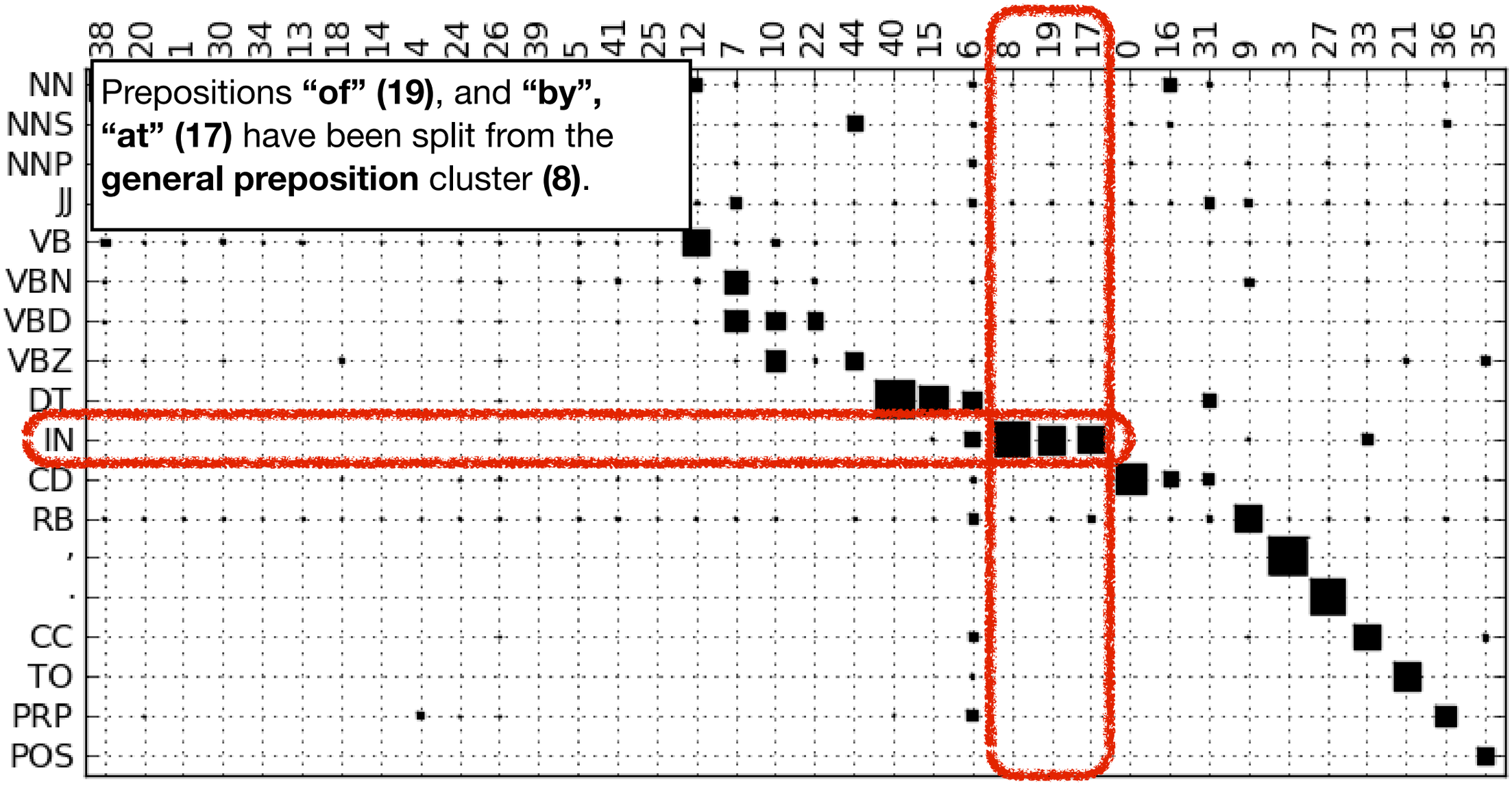
Experiments and Results

- Hinton Graph



Experiments and Results

- Hinton Graph



Induction **vs** Disambiguation

Induction **vs** Disambiguation

do not have tag info

Put similar words into

same clusters

given the **contexts.**

Induction **vs** Disambiguation

do not have tag info

Put similar words into

same clusters

given the **contexts.**

Have possible tags

Disambiguate the

correct tag

of an ambiguous word

given the **context.**

Outline

- Paradigmatic Context representation
- Clustering Model (POS induction)
- Co-occurrence Modeling (POS induction)
- **Probabilistic Voting (POS disambiguation)**
- HMM based Model (POS disambiguation)
- Noisy Channel Model (WSD disambiguation)
- Conclusion

Probabilistic Voting Model

- Word sequence and a word-tag dictionary is available
  - ... it will also **offer** buyers the option ...

Verb

- The **offer** is begin launched ....

Noun

Probabilistic Voting Model

- Word sequence and a word-tag dictionary is available

  - ... it will also **offer** buyers the option ...
    give
    help
    attract
    ...

    Verb

- The **offer** is begin launched ....
  campaign
  project
  scheme
  ...

  Noun

Probabilistic Voting Model

- Word sequence and a word-tag dictionary is available

  - ... it will also **offer** buyers the option ...
    give
    help
    attract
    ...

    Verb

  - The **offer** is begin launched ....
    campaign
    project
    scheme
    ...

    Noun

- Substitutes can disambiguate the correct tag.

Probabilistic Voting Model

- Estimates the tag distribution in a given word context

$$\begin{aligned}\Pr(t|c) &= \sum_{s \in S} \Pr(t|s,c) \Pr(s|c) \\ &= \sum_{s \in S} \Pr(t|s) \Pr(s|c)\end{aligned}$$

- where t is the tag, S is the set of substitutes and c is the context

Probabilistic Voting Model on the Morphological Disambiguation of Turkish

- Turkish word has multiple morphological parses (tags)
  - Example word: masalı

masal + Noun+A3sg+Pnon+acc	the story
masal + Noun+A3sg+P3sg+Nom	his story
masa + Noun+A3sg+Pnon+Nom+^DG+Adj+With	with tables

Accuracy on ambiguous words (45 % of our test corpus)

- Random unsupervised baseline is 39.4%
- Most frequent tag with word-tag distribution baseline is 71.0%
- I achieved 64.5%.

Probabilistic Voting Model

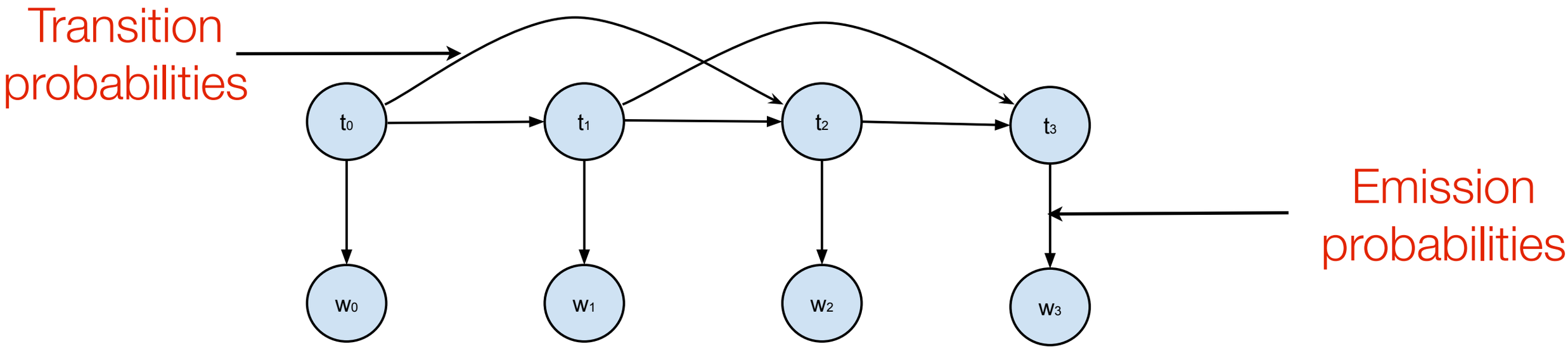
- This model ignores the **target word features** and only uses substitute words and their tags

- One limitation is it ignores the tags of the consecutive words.

Outline

- Paradigmatic Context representation
- Clustering Model
- Co-occurrence Modeling
- Probabilistic Voting
- **HMM based Model**
- Noisy Channel Model
- Conclusion

Constraining HMM-Based Models



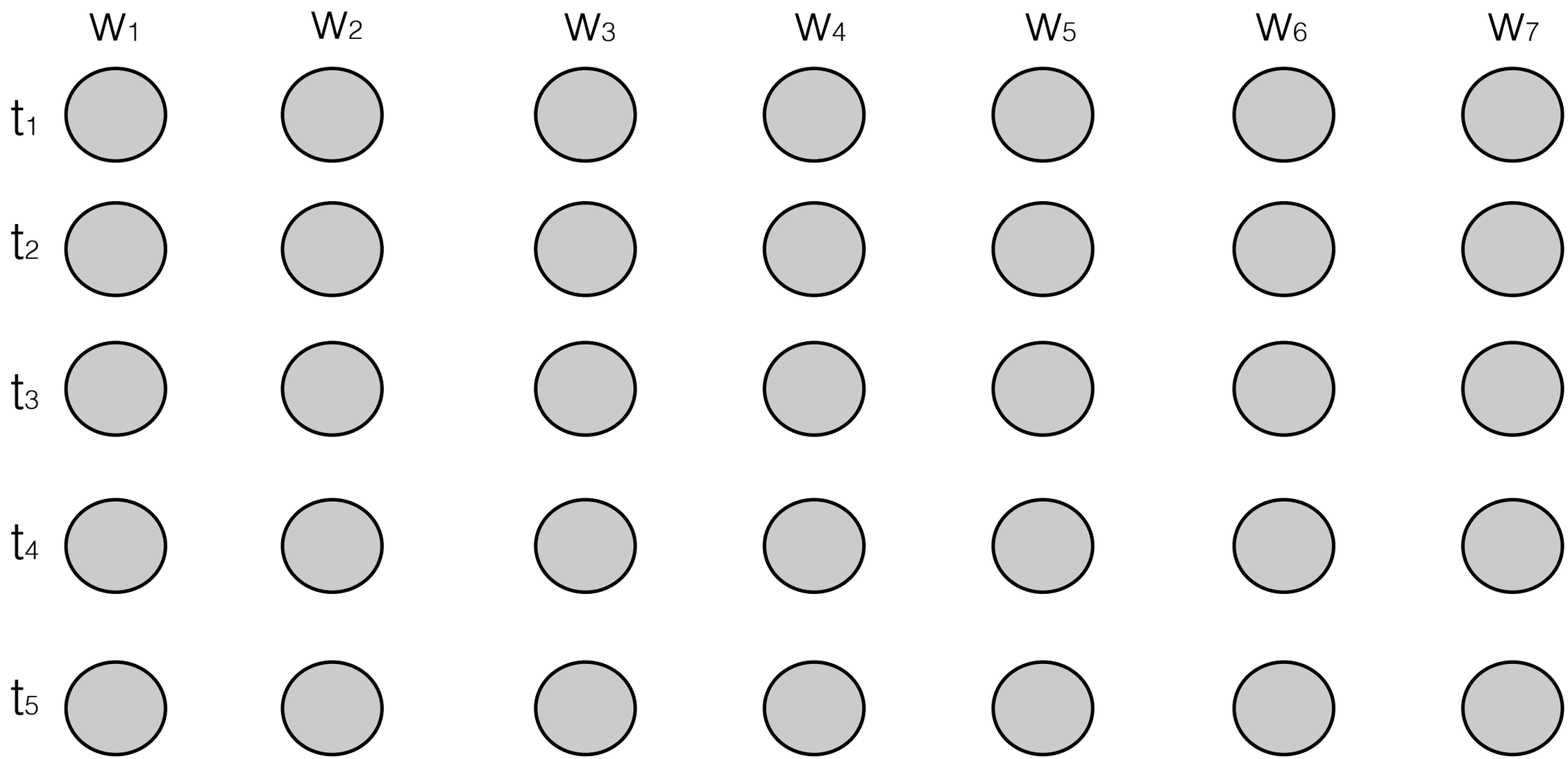
- The word sequence is generated by the hidden tag sequence.
- Each tag depends on the previous n-1 tags
- Each word is independent of each other given the tag
- Likelihood of an n-th order HMM model:

$$P(w_1|t_1)\prod_{i=2}^{N} P(w_i|t_i)P(t_i|t_{i-1},\ldots,t_{i-n+1})$$

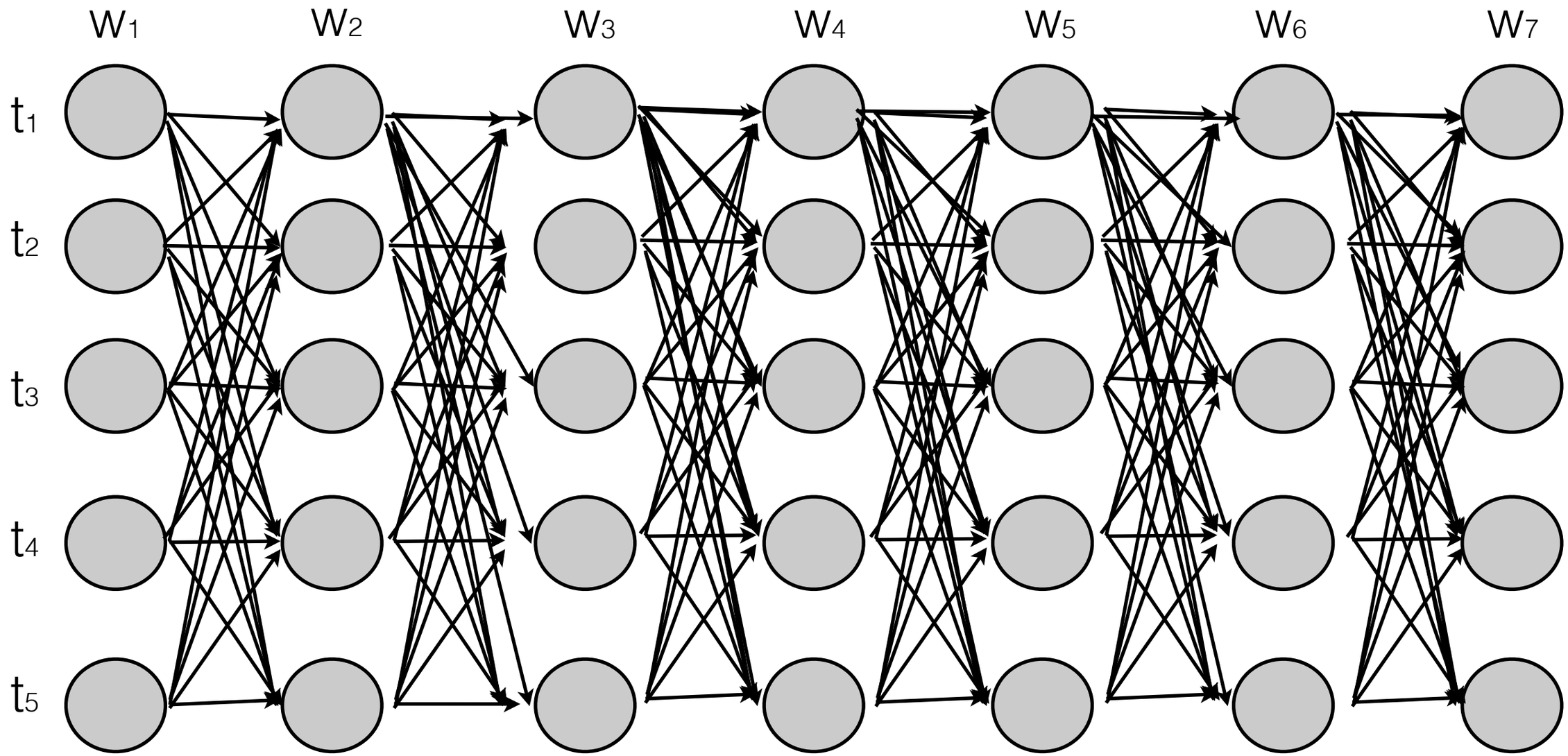
Constraining HMM-Based Models

- The model parameters are estimated using the Expectation Maximization (EM) algorithm.
  - Transition Probabilities $\Pr(t_i \mid t_{i-1})$
  - Emission Probabilities $\Pr(w_i \mid t_i)$
- Viterbi search algorithm finds the best tag sequence

Constraining HMM-Based Models: POS tagging

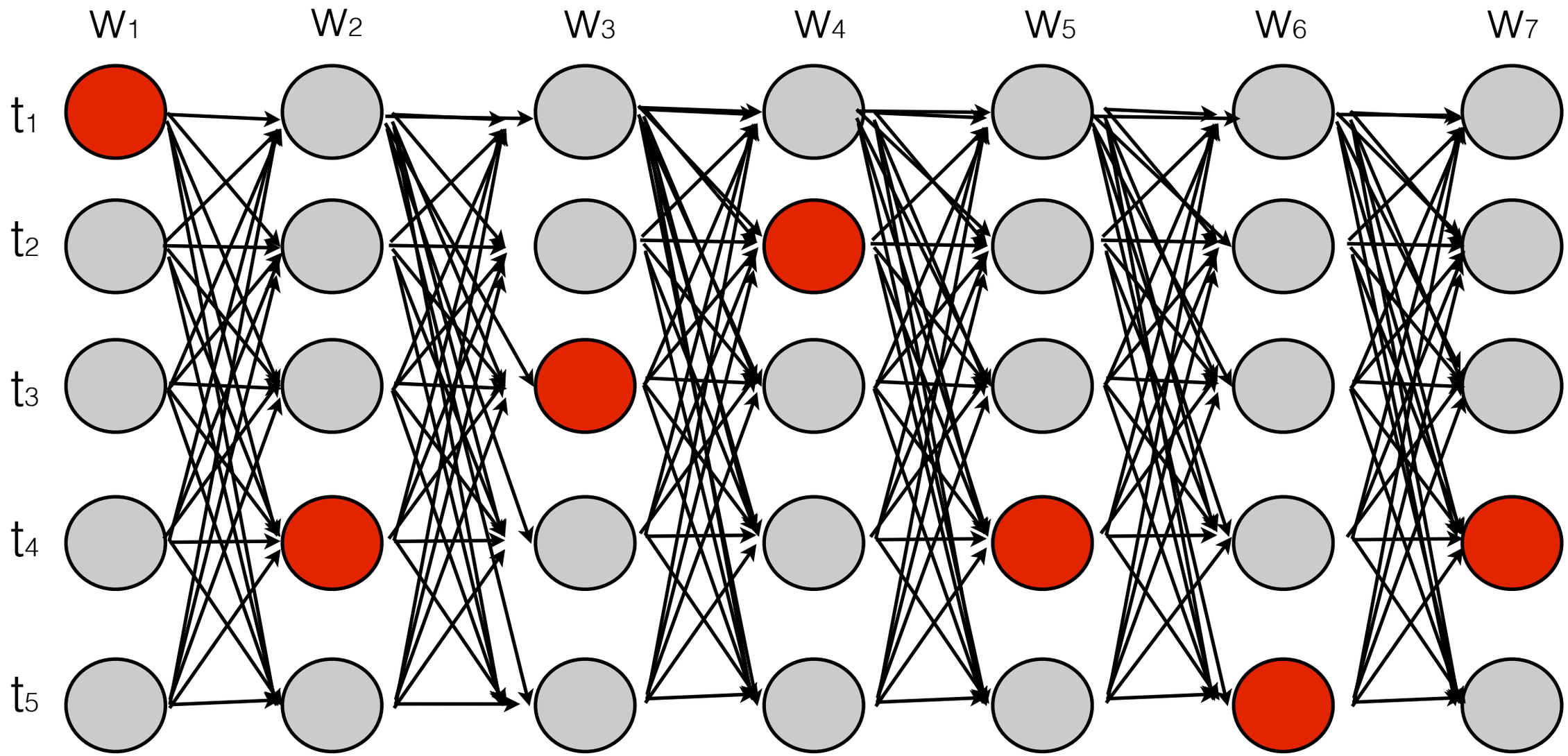


Constraining HMM-Based Models: POS tagging

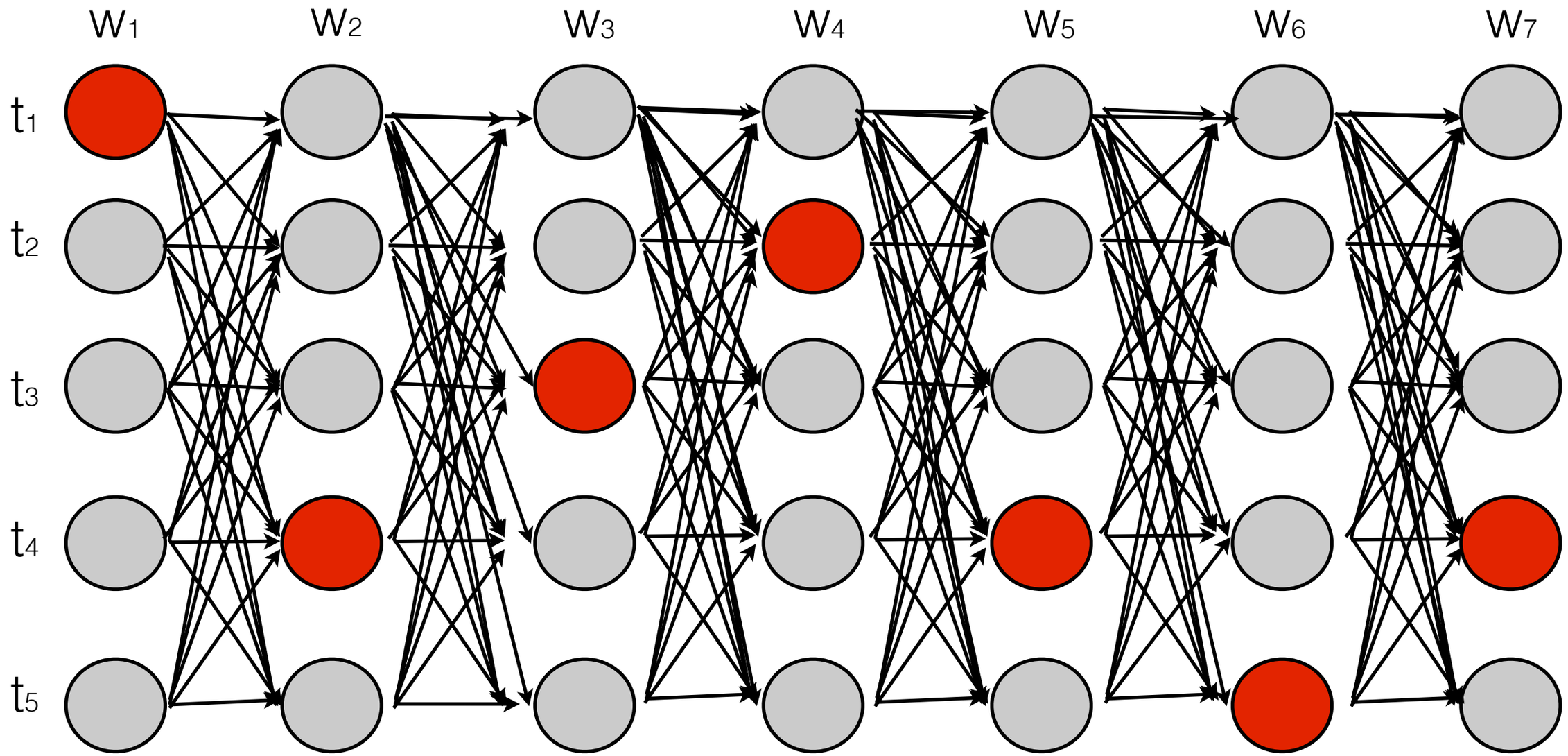


- Nodes are emission probabilities
- Arrows are transition probabilities

Constraining HMM-Based Models: POS tagging



Constraining HMM-Based Models: POS tagging



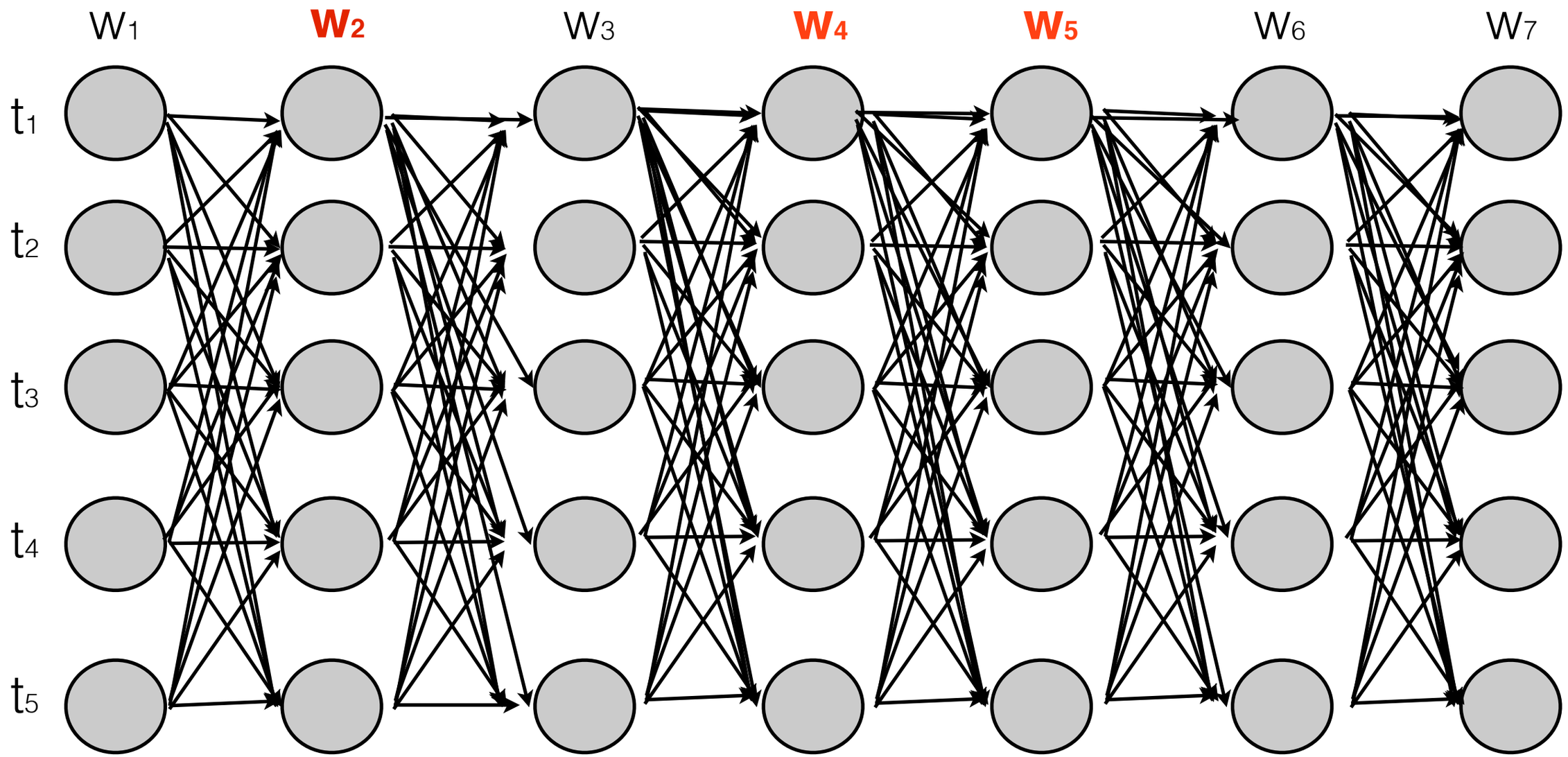
- After EM, the Viterbi algorithm finds the best tag sequence
- t1 t4 t3 t2 t4 t6 t7

Constraining HMM-Based Models: POS tagging

- In POS tagging we have a word-tag dictionary
  - Example dictionary

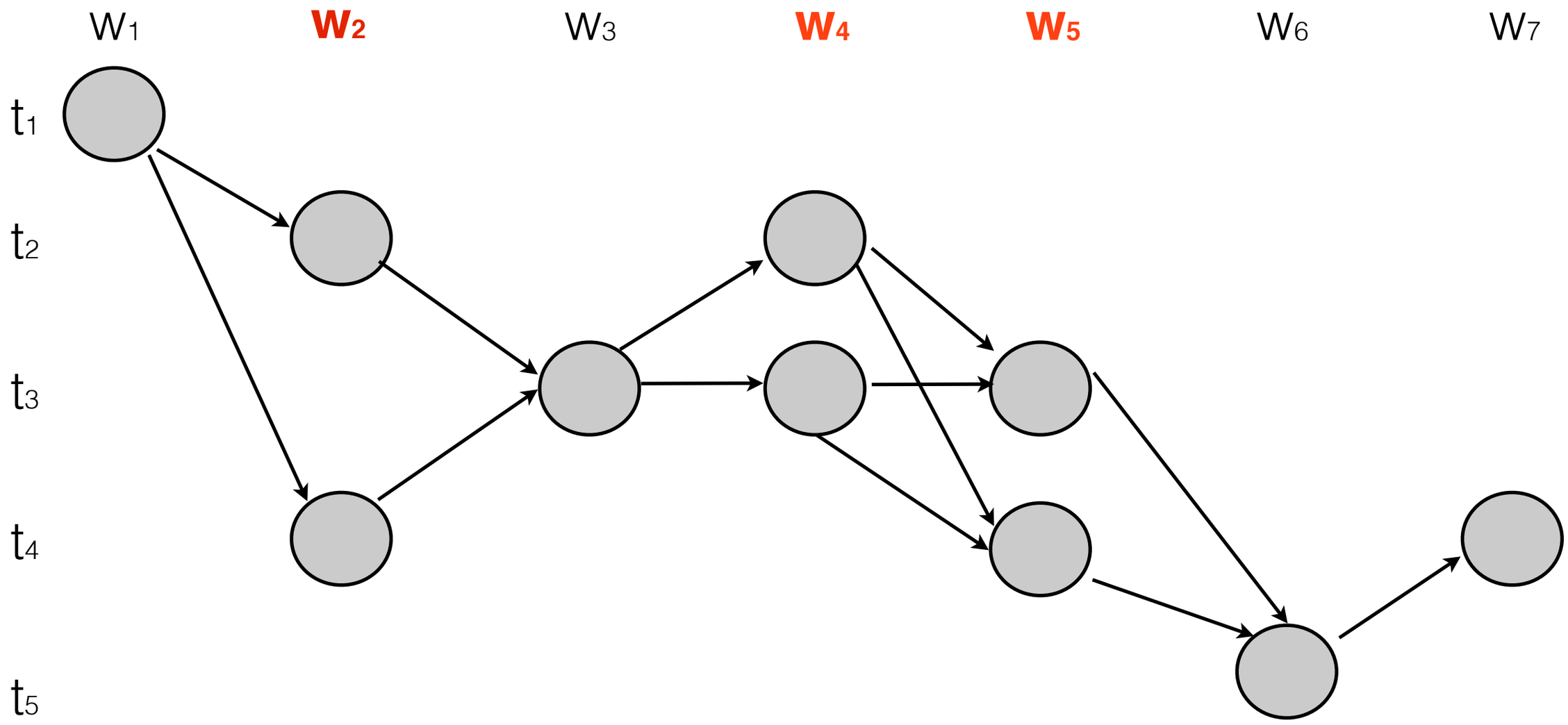
Word	Tags
w1	t1
w2	t2, t4
w3	t3
w4	t2, t3
w5	t3, t4
w6	t5
w7	t4

Constraining HMM-Based Models: POS tagging

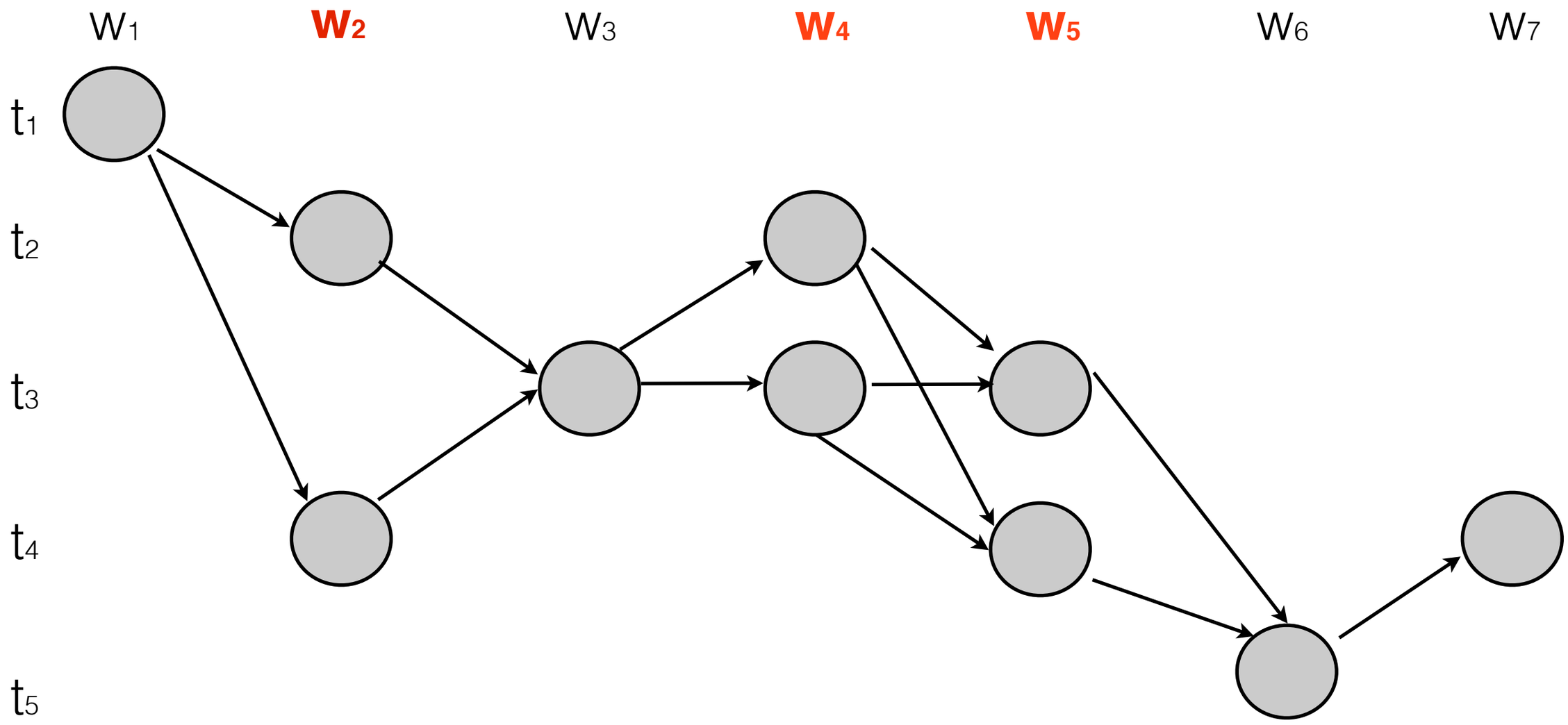


- In POS tagging we have a word-tag dictionary

Constraining HMM-Based Models: POS tagging



Constraining HMM-Based Models: POS tagging



- Constrain using the word-tag dictionary

Constraining HMM-Based Models: POS tagging

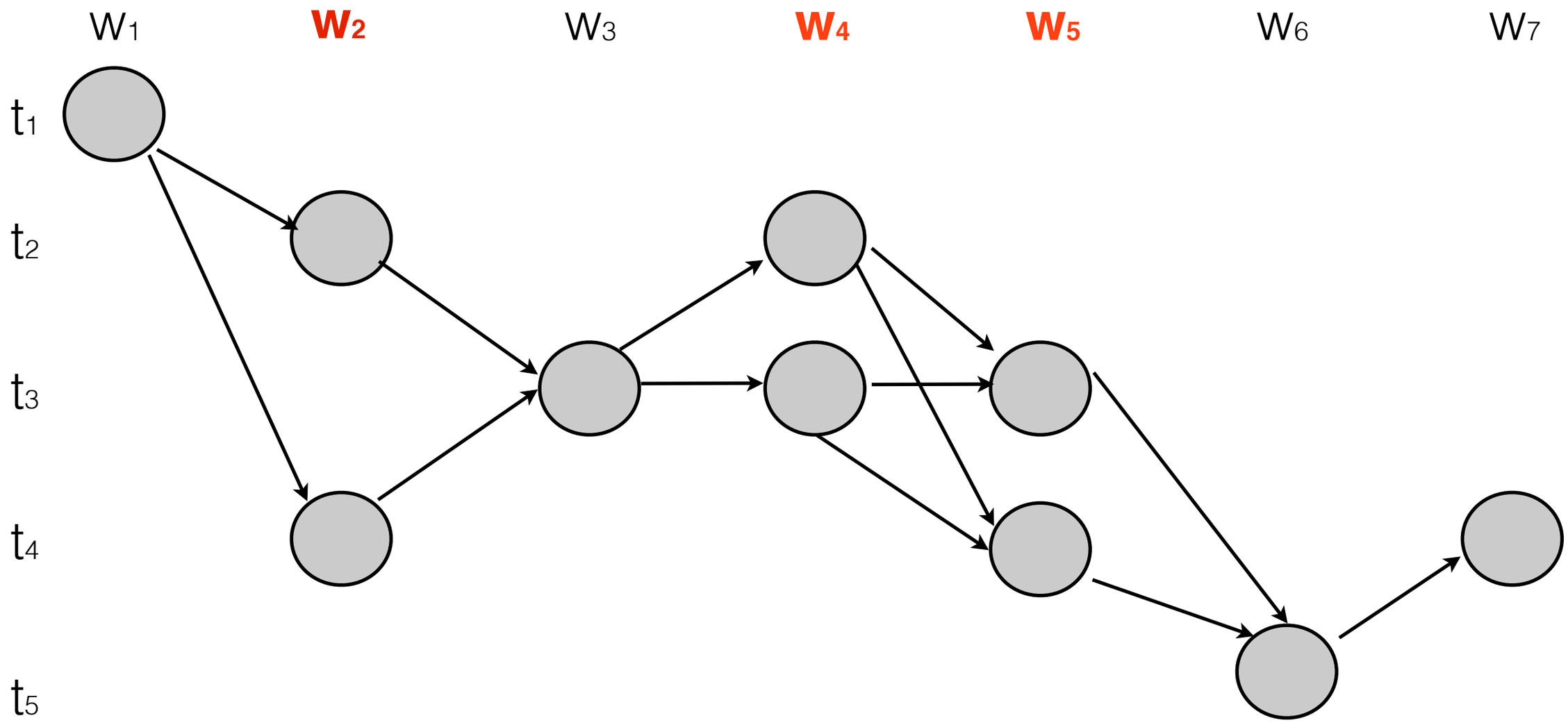
- Johnson (2007b) showed that HMM-EM has tendency of assigning similar number of words to each tag.
  - However POS tags have skewed distributions

Constraining HMM-Based Models: POS tagging

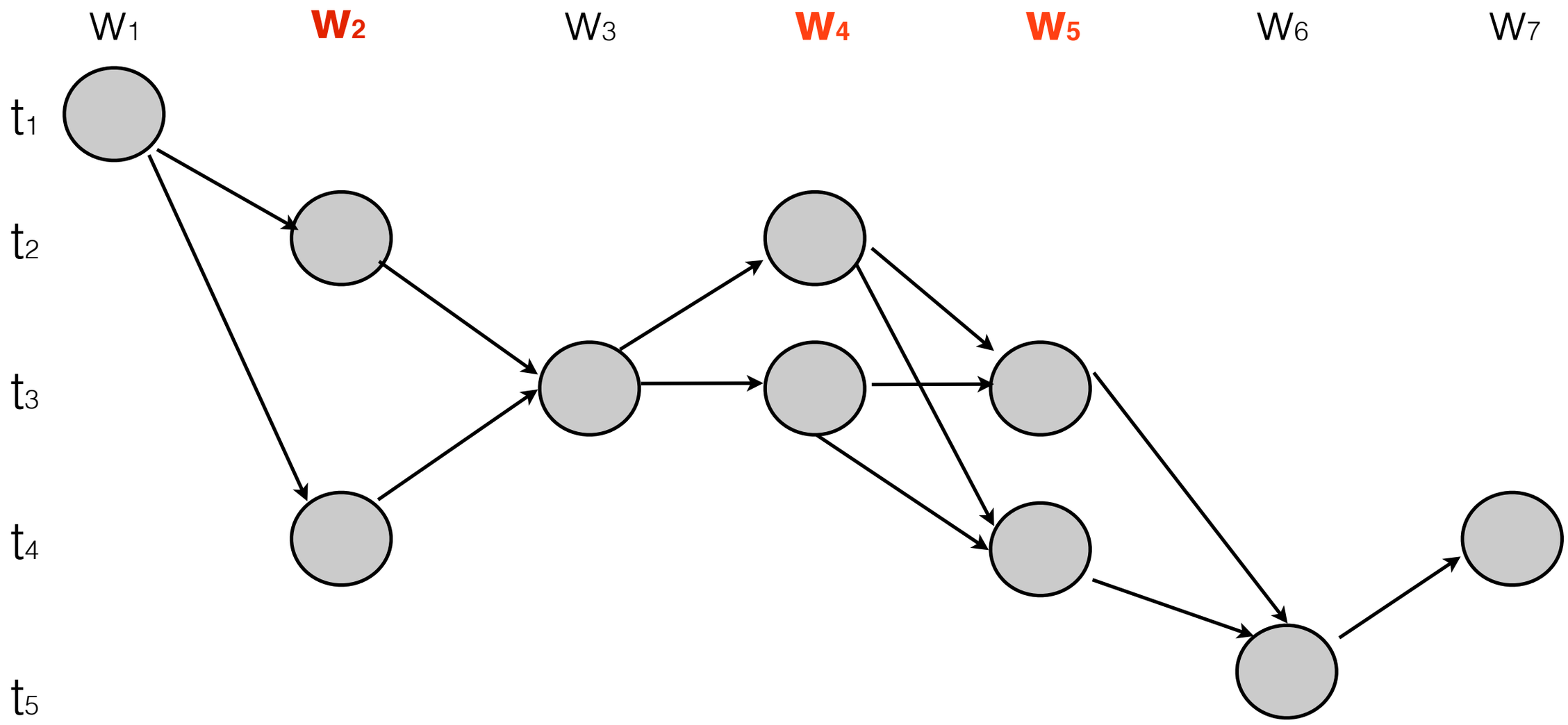
- Johnson (2007b) showed that HMM-EM has tendency of assigning similar number of words to each tag.
  - However POS tags have skewed distributions

How to constrain more?

Constraining HMM-Based Models: POS tagging

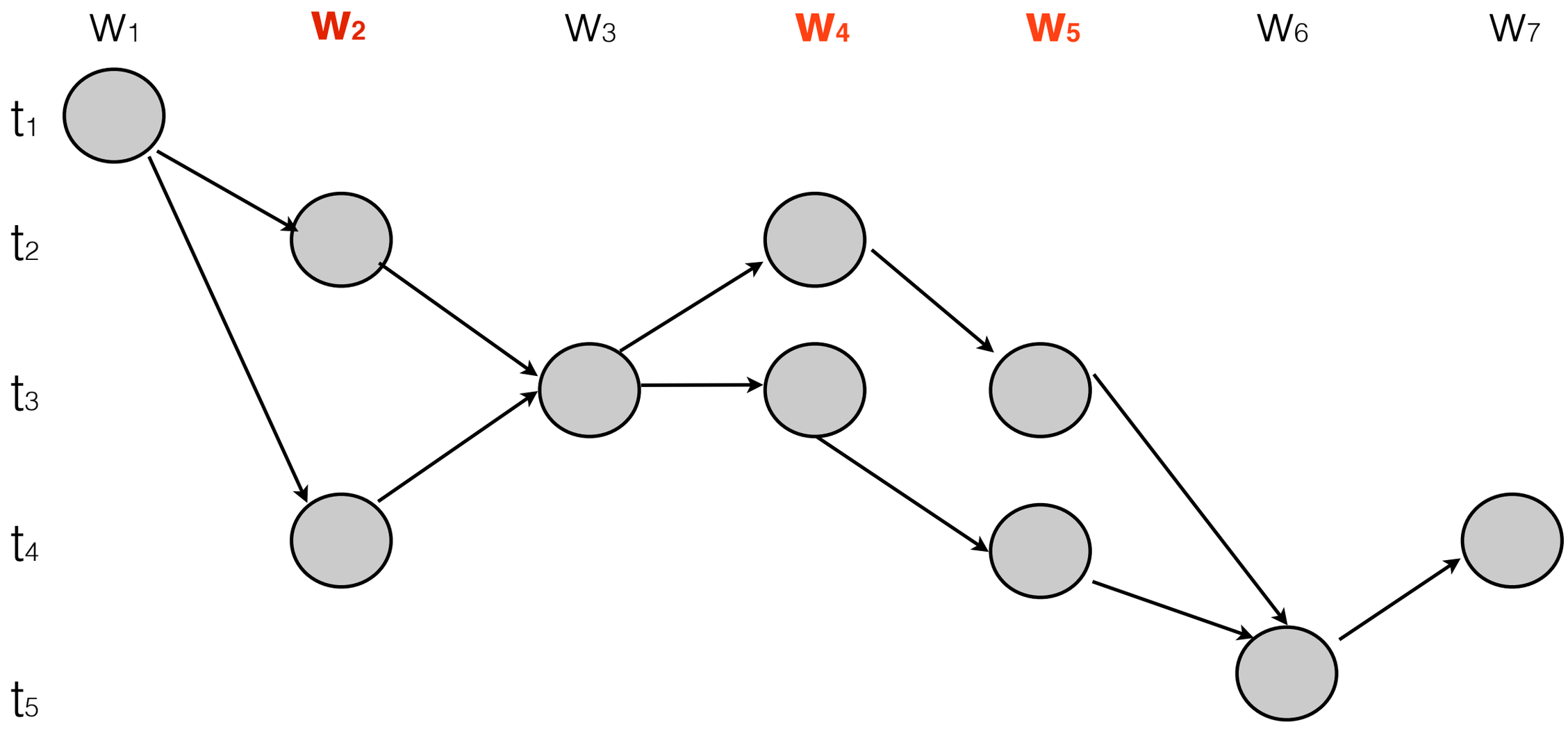


Constraining HMM-Based Models: POS tagging



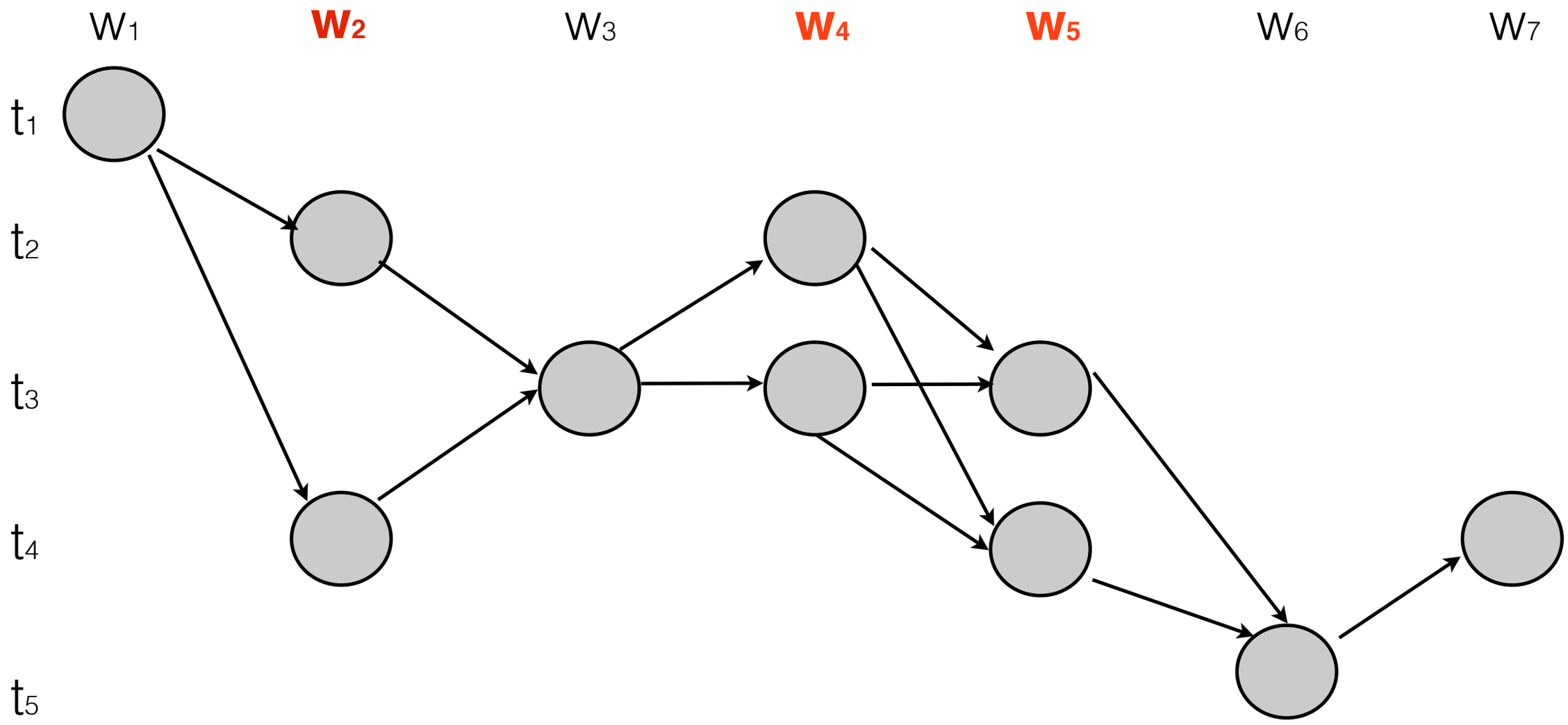
- Ravi and Knight (2009a) improve the accuracy by constraining the number of **non-zero transition probabilities**

Constraining HMM-Based Models: POS tagging

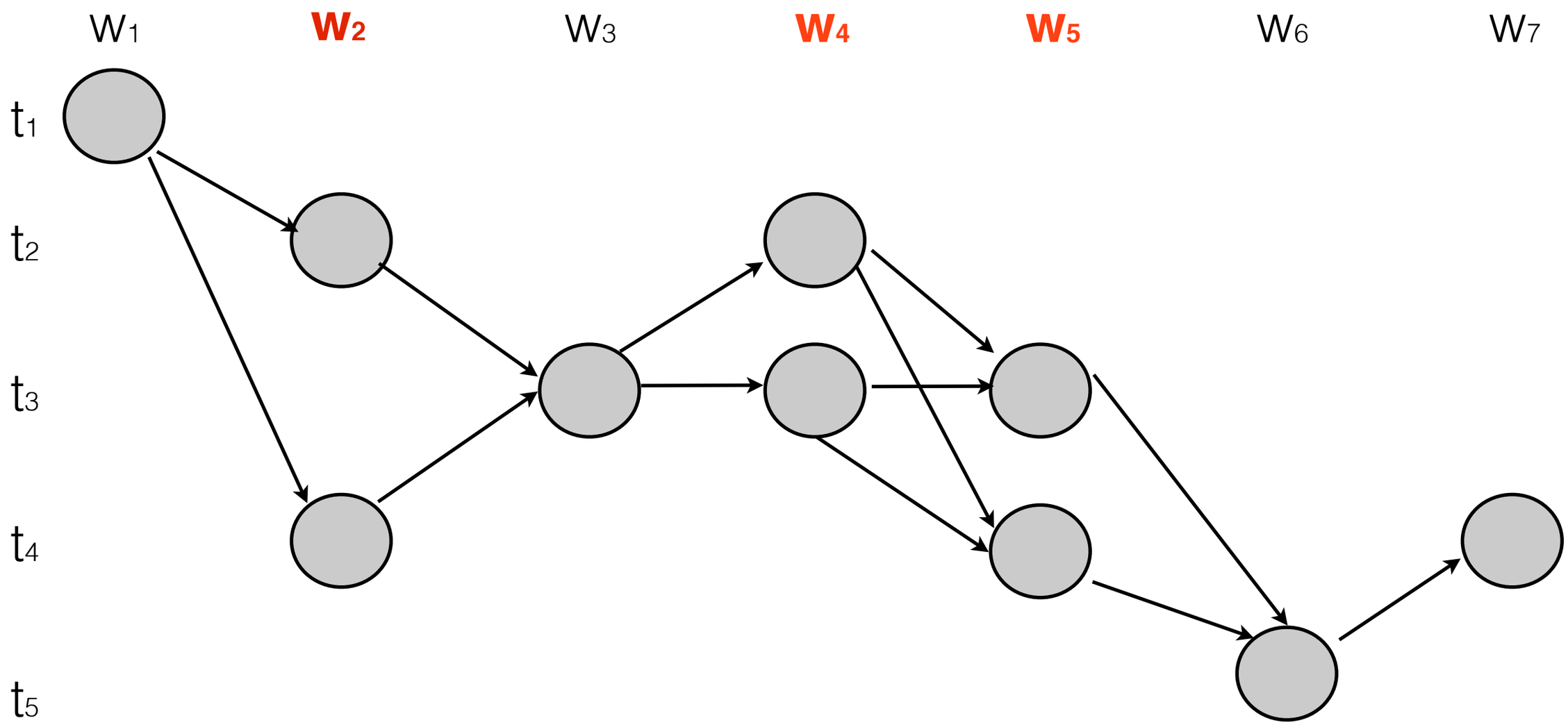


- Ravi and Knight (2009a) improve the accuracy by constraining the number of **non-zero transition probabilities**

Constraining HMM-Based Models: POS tagging

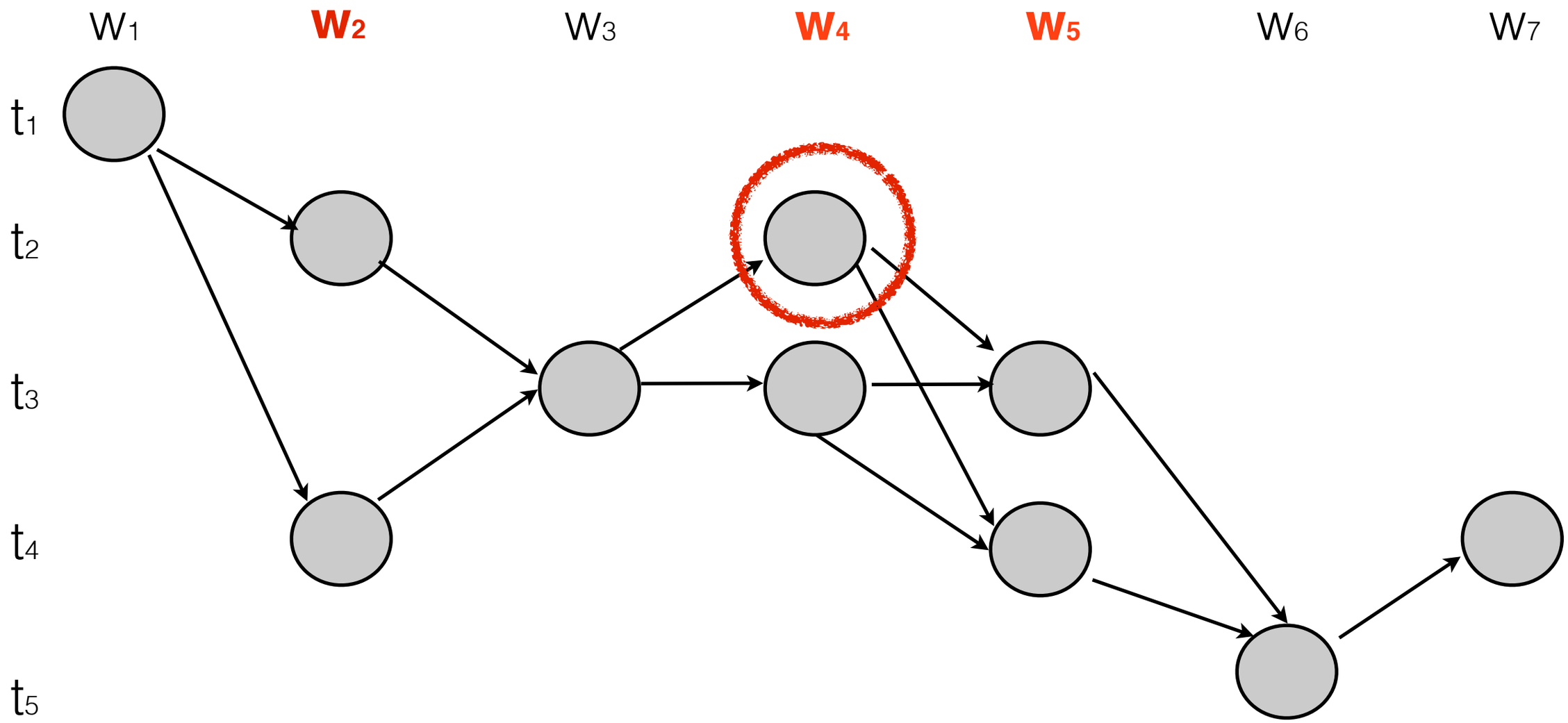


Constraining HMM-Based Models: POS tagging

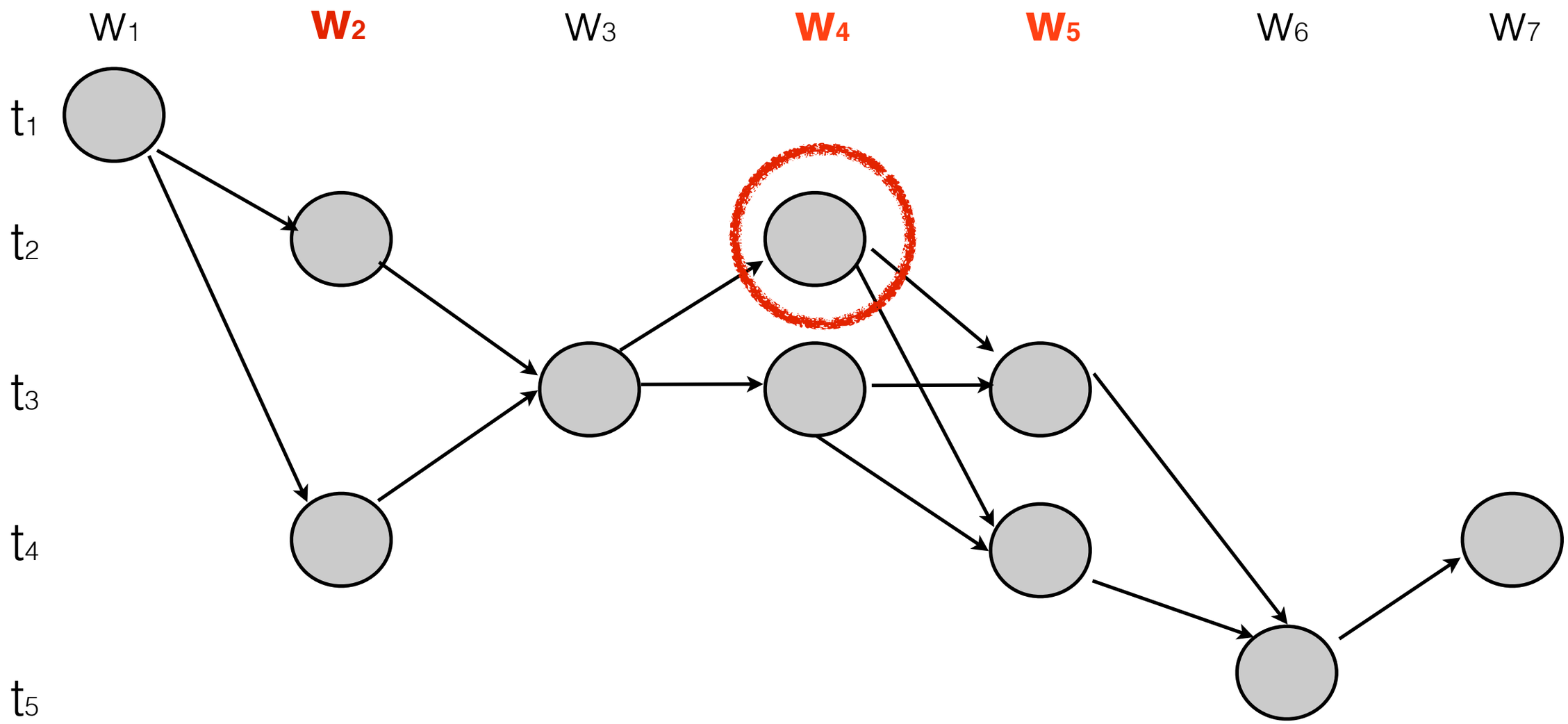


- Constraint the emission probabilities by reducing dictionary size

Constraining HMM-Based Models: POS tagging

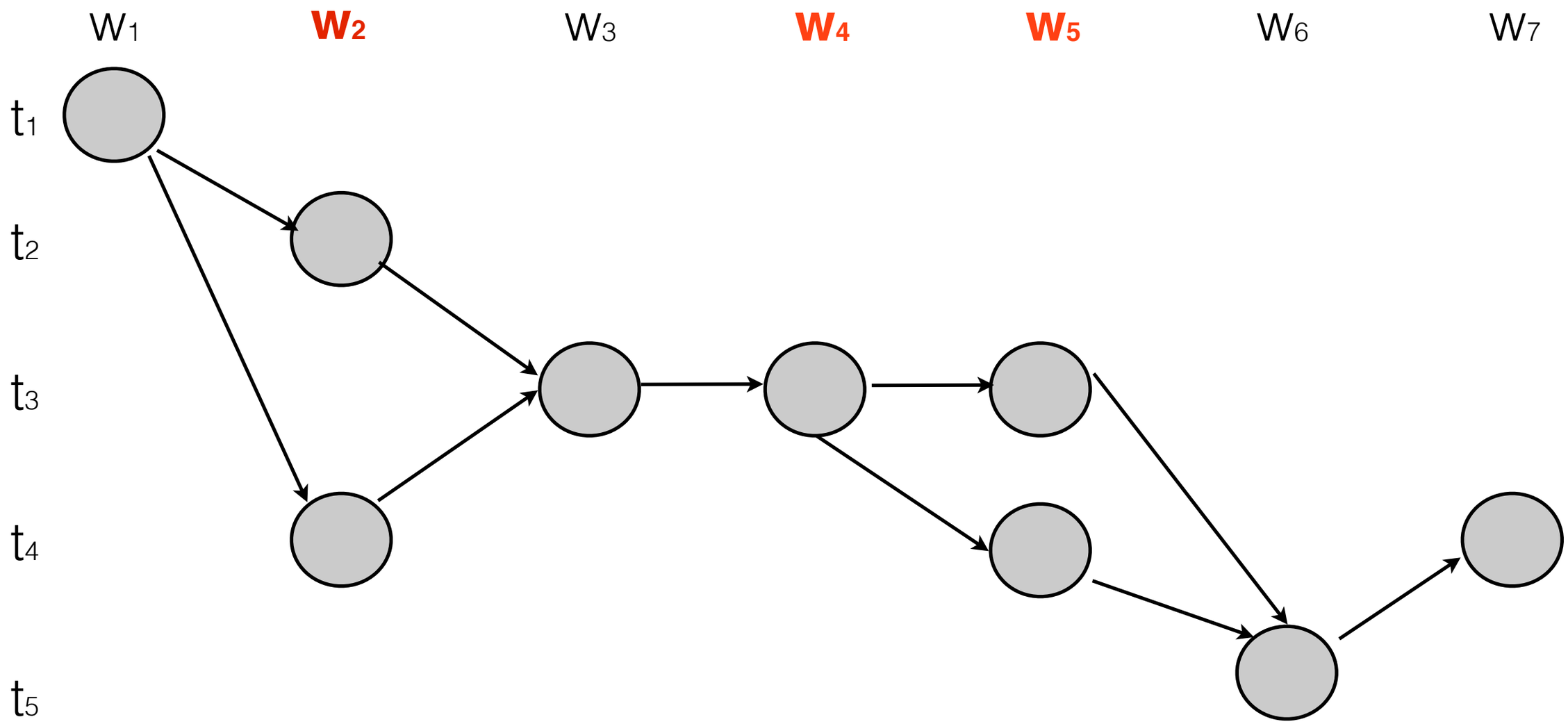


Constraining HMM-Based Models: POS tagging



- Estimate P(T|C) of all instances of w4 with probabilistic voting
  - Average them and remove the unlikely tags

Constraining HMM-Based Models: POS tagging



- Constraint the emission probabilities by reducing dictionary size

Dictionary reduction on POS tagging

POS Groups	2-gram HMM Accuracy (%)	2-gram HMM RD Accuracy (%)
Noun	92.22	94.01
Verb	83.84	84.90
Adj	85.22	89.52
Adv	83.96	85.18
Pronoun	95.56	95.92
Content	89.42	91.18
Function	70.49	92.92
All	82.05	91.85

State-of-the-art tagging results when only word-tag dictionary is available.

Outline

- Paradigmatic Context representation
- Clustering Model
- Co-occurrence Modeling
- Probabilistic Voting
- HMM based Model
- **Noisy Channel Model**
- Conclusion

Noisy Channel Model

- **Input:**
  - word sequence
  - word-tag distribution

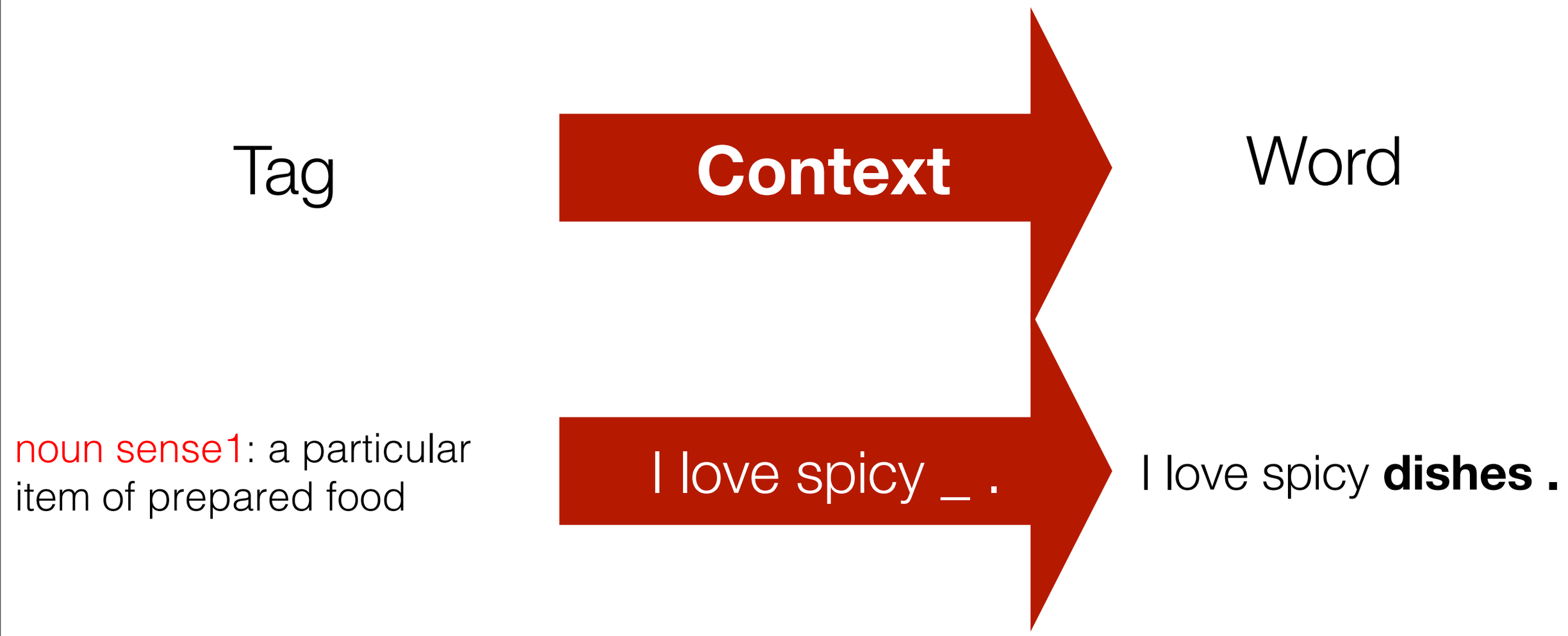
Noisy Channel Model

Noisy Channel Model

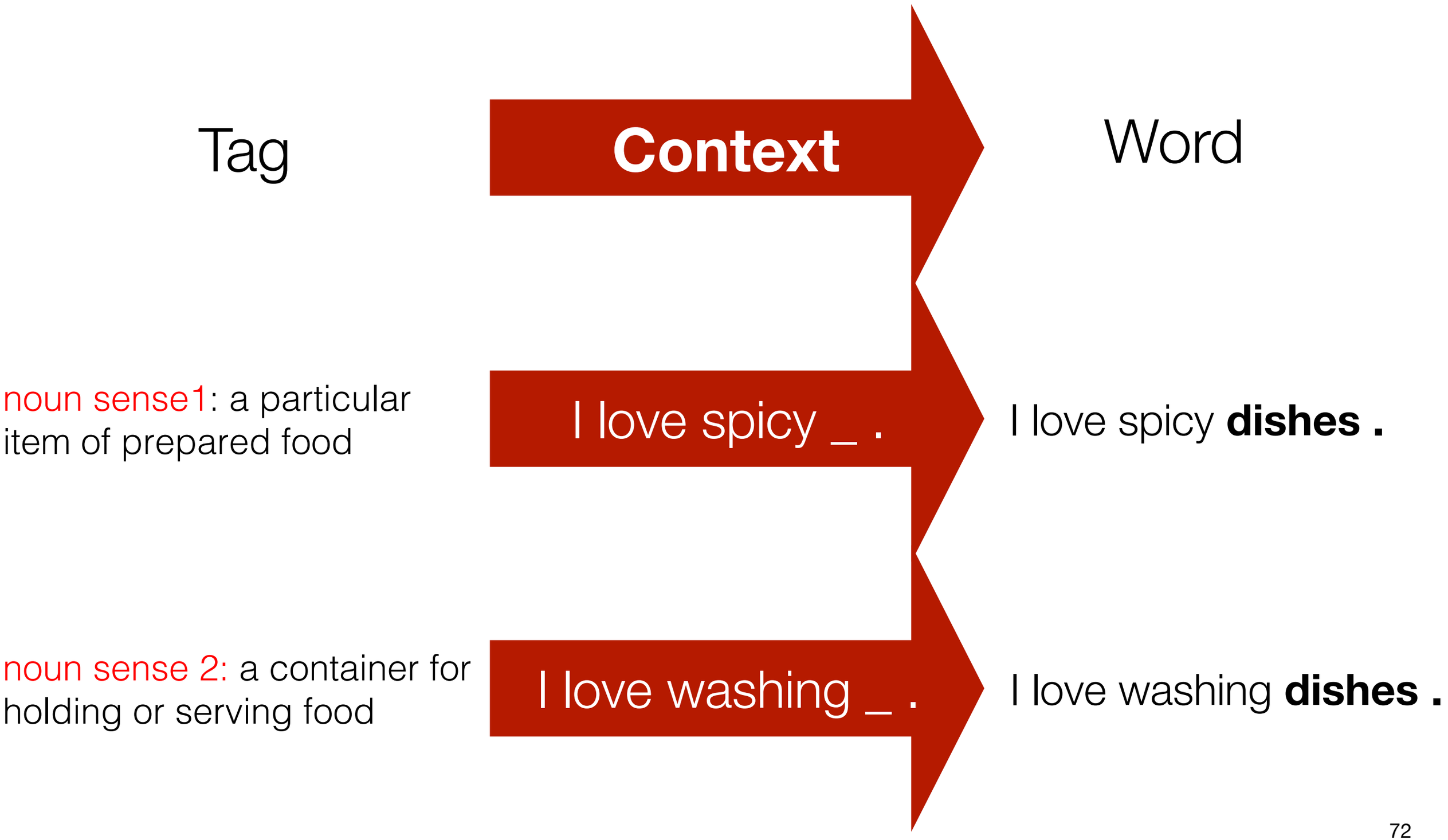
In tagging problems

Noisy Channel Model in Tagging Problems

Noisy Channel Model in Tagging Problems



Noisy Channel Model in Tagging Problems



Noisy Channel Model

$$\Pr(T|W,C) = \frac{\Pr(W|T,C)\Pr(T|C)}{\Pr(W|C)}$$

Noisy Channel Model

$$\Pr(T|W,C) = \frac{\Pr(W|T,C)\Pr(T|C)}{\Pr(W|C)}$$

- find the tag, **T** that maximizes the **Pr(T I W, C)**

Noisy Channel Model

$$\Pr(T|W,C) = \frac{\Pr(W|T,C)\Pr(T|C)}{\Pr(W|C)}$$

- find the tag, **T** that maximizes the **Pr(T I W, C)**

- **Pr(WIC)** does not depend on **T**

Noisy Channel Model

$$\Pr(T|W,C) = \frac{\Pr(W|T,C)\Pr(T|C)}{\Pr(W|C)}$$

- find the tag, **T** that maximizes the **Pr(T I W, C)**

- **Pr(WIC)** does not depend on **T**

- need to estimate **Pr(W I T, C)** and **Pr(TIC)**

  - Assume **W** is independent of **C** given **T** so **Pr(W I T,C) = Pr(W I T)**

Noisy Channel Model

$$\Pr(T \mid W, C) \propto \Pr(W \mid T)\ \Pr(T \mid C)$$

Noisy Channel Model

$$\mathbf{Pr(\ T \mid W,\ C) \propto Pr(W|T)\ Pr(T|C)}$$

- need to estimate **Pr(W I T)** and **Pr(TIC)**

Noisy Channel Model

$$\mathbf{Pr(\ T \mid W,\ C) \propto Pr(W|T)\ Pr(T|C)}$$

- need to estimate **Pr(W I T)** and **Pr(TIC)**
- Estimate **Pr(W I T)** from word tag distribution

Noisy Channel Model

$$\mathbf{Pr(\ T \mid W,\ C) \propto Pr(W|T)\ Pr(T|C)}$$

- need to estimate **Pr(W I T)** and **Pr(T|C)**
- Estimate **Pr(W I T)** from word tag distribution
- We have **Pr(W I C)** and **Pr(W I T)**, how to estimate **Pr(T I C)**

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

- need to estimate **Pr(W I T)** and **Pr(TIC)**
- Estimate **Pr(W I T)** from word tag distribution
- We have **Pr(W I C)** and **Pr(W I T)**, how to estimate **Pr(T I C)**

$$\Pr(W|C) = \sum_T \Pr(T|C) \Pr(W|T, C)$$

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

For every $w_i \in$ **W** in a fixed context (channel) **C**

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

For every $w_i \in$ **W** in a fixed context (channel) **C**

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T)\Pr(T \mid C)$$

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

For every $w_i \in$ **W** in a fixed context (channel) **C**

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T)\Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T)\Pr(T \mid C)$$

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

For every $w_i \in$ **W** in a fixed context (channel) **C**

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T)\Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T)\Pr(T \mid C)$$

Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\Pr(W|C) = \sum_T \Pr(W|T,C)\Pr(T|C)$$

$$\Pr(W|C) = \sum_T \Pr(W|T)\Pr(T|C)$$

For every $w_i \in \mathbf{W}$ in a fixed context (channel) **C**

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T)\Pr(T \mid C)$$

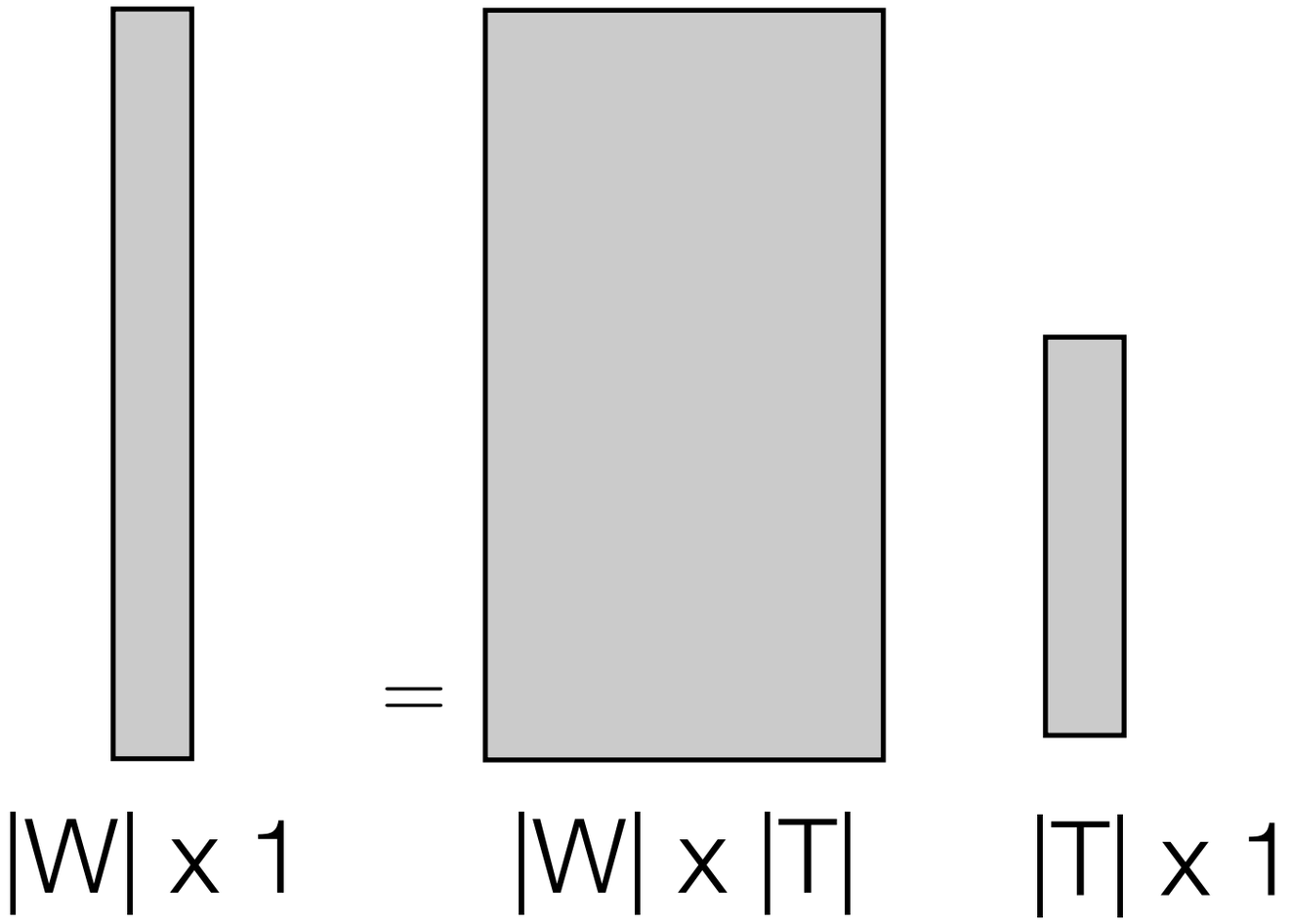
$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T)\Pr(T \mid C)$$

$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T)\Pr(T \mid C)$$

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T) \Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T) \Pr(T \mid C)$$

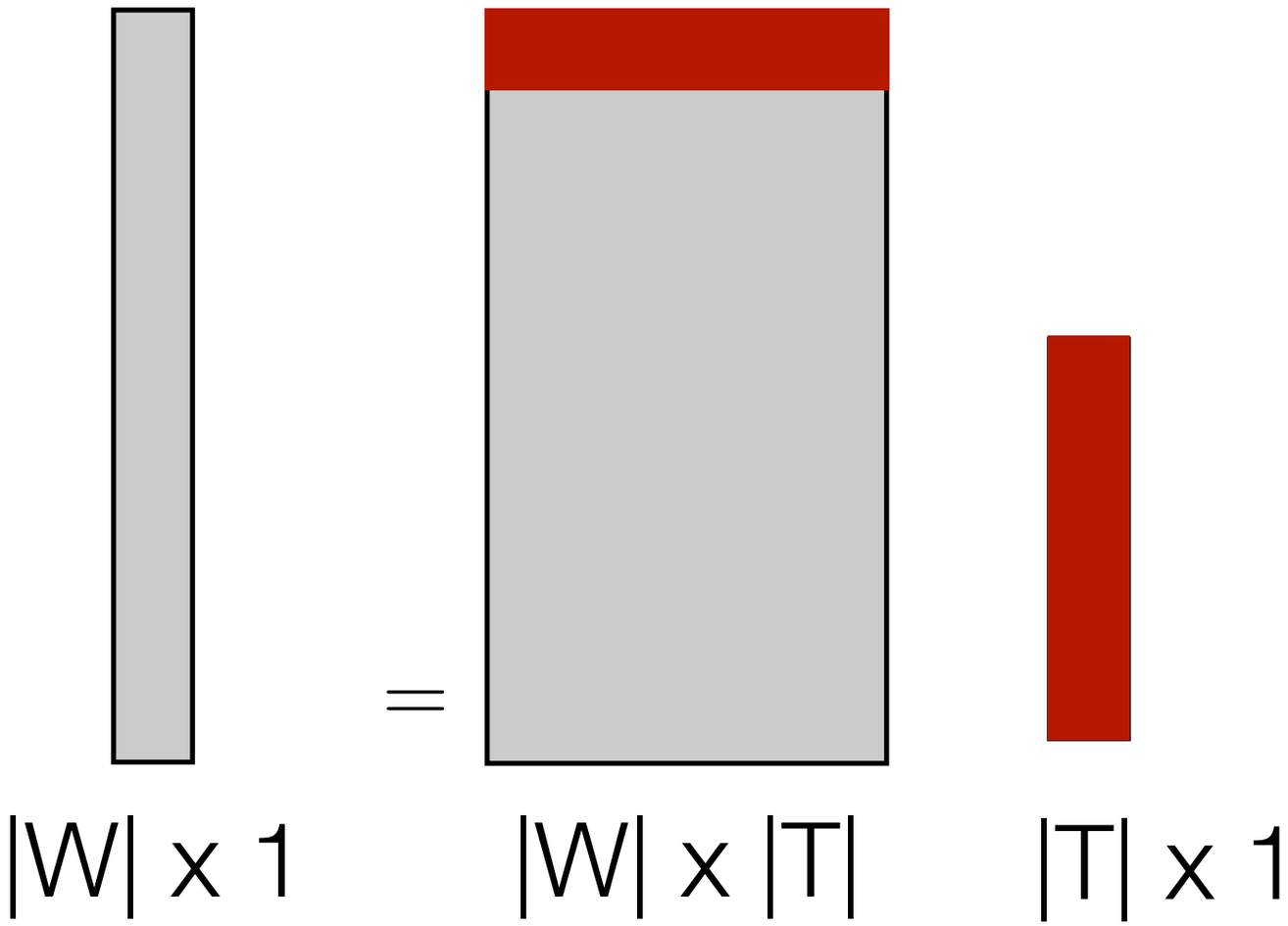
$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T) \Pr(T \mid C)$$



$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T)\Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T)\Pr(T \mid C)$$

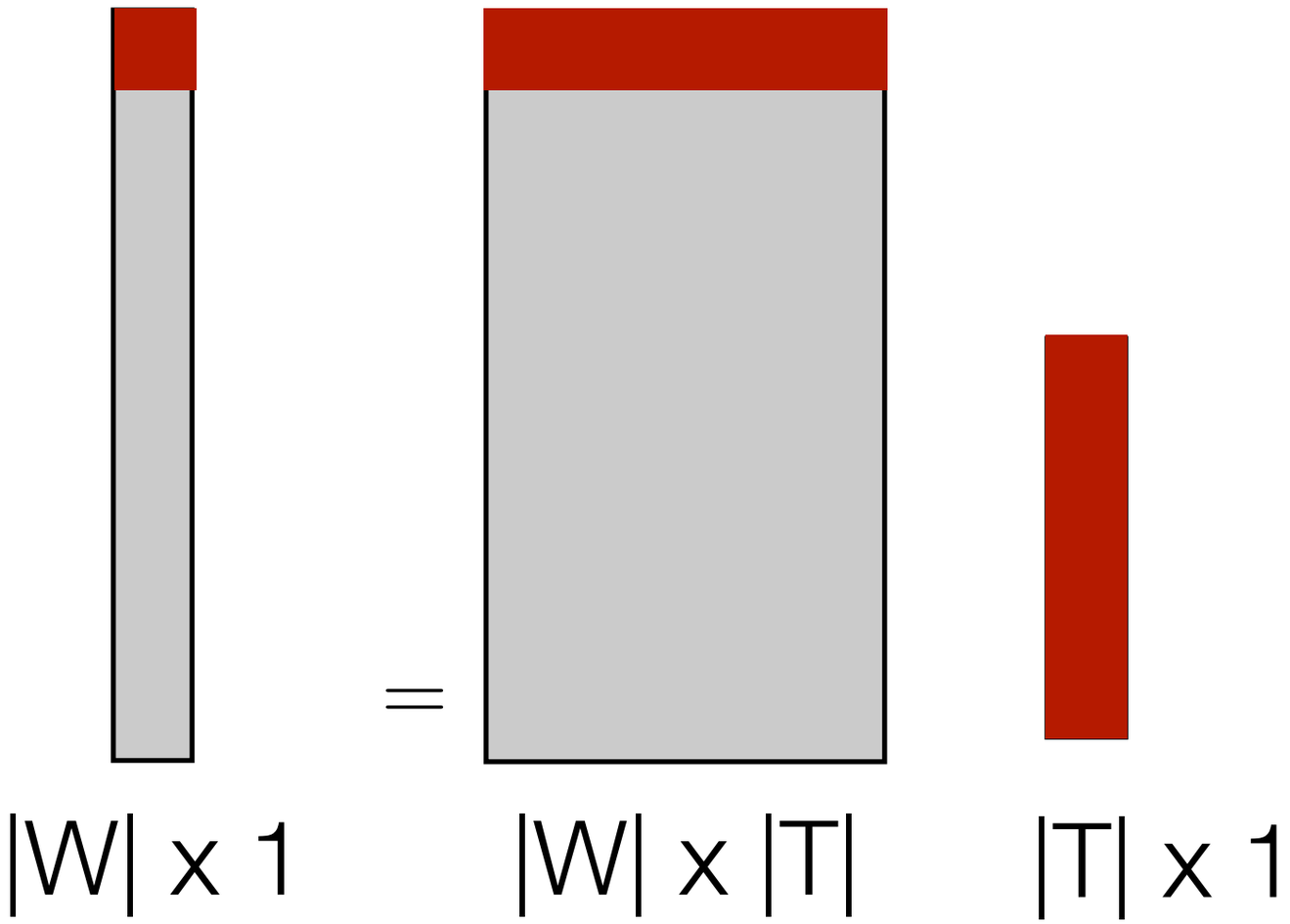
$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T)\Pr(T \mid C)$$



$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T) \Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T) \Pr(T \mid C)$$

$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T) \Pr(T \mid C)$$

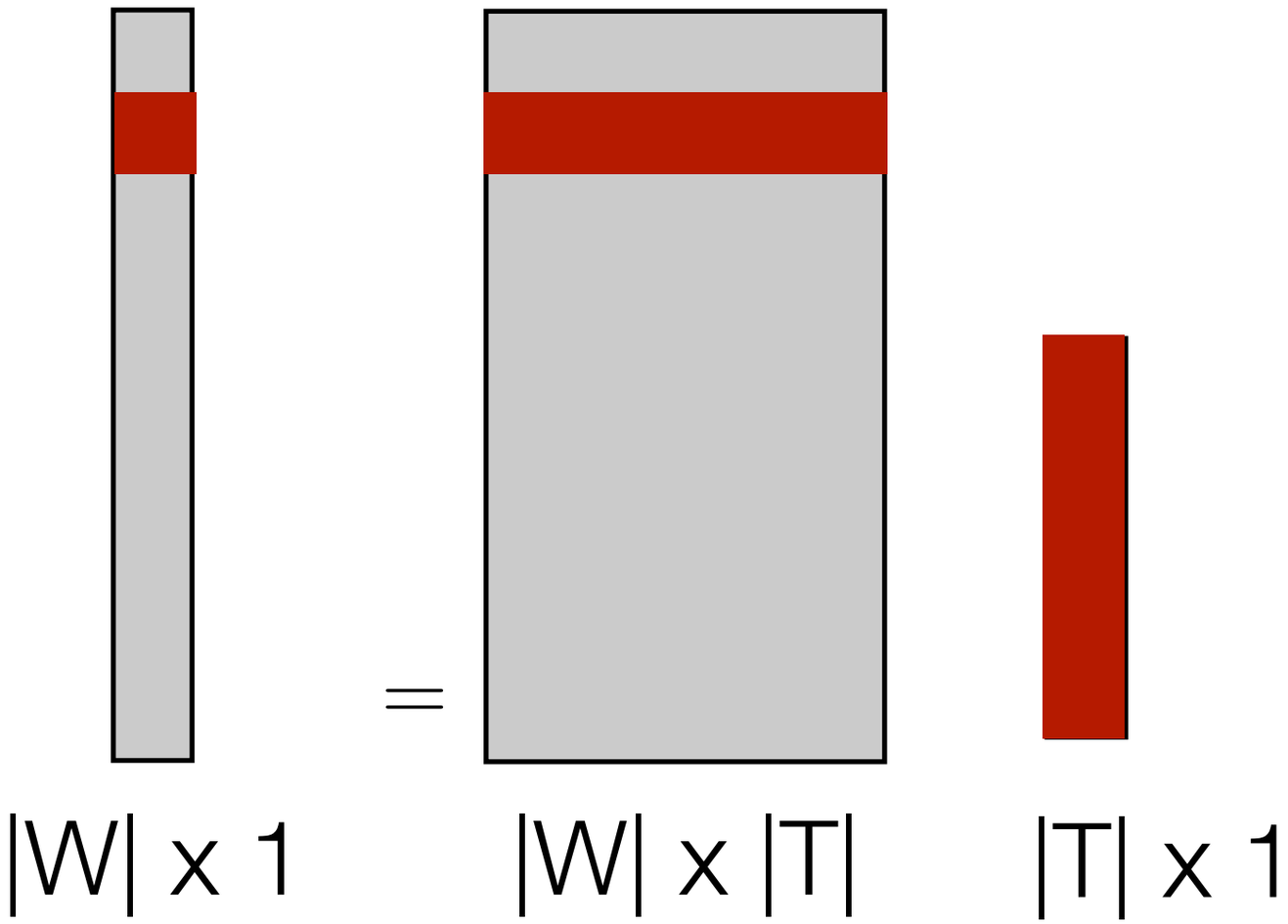


|W| x 1 |W| x |T| |T| x 1

$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T) \Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T) \Pr(T \mid C)$$

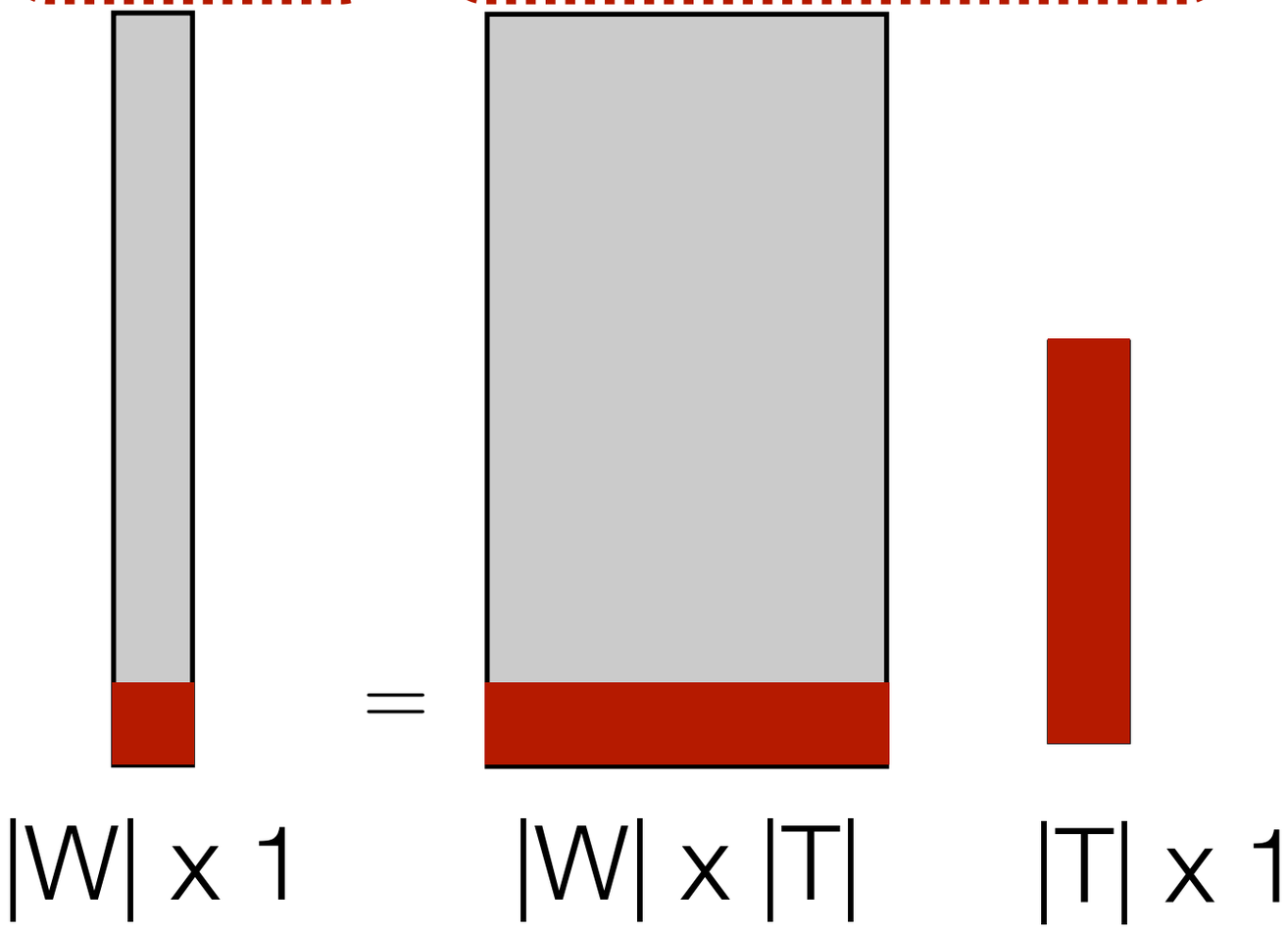
$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T) \Pr(T \mid C)$$



$$\Pr(w_1 \mid C) = \sum_T \Pr(w_1 \mid T) \Pr(T \mid C)$$

$$\Pr(w_2 \mid C) = \sum_T \Pr(w_2 \mid T) \Pr(T \mid C)$$

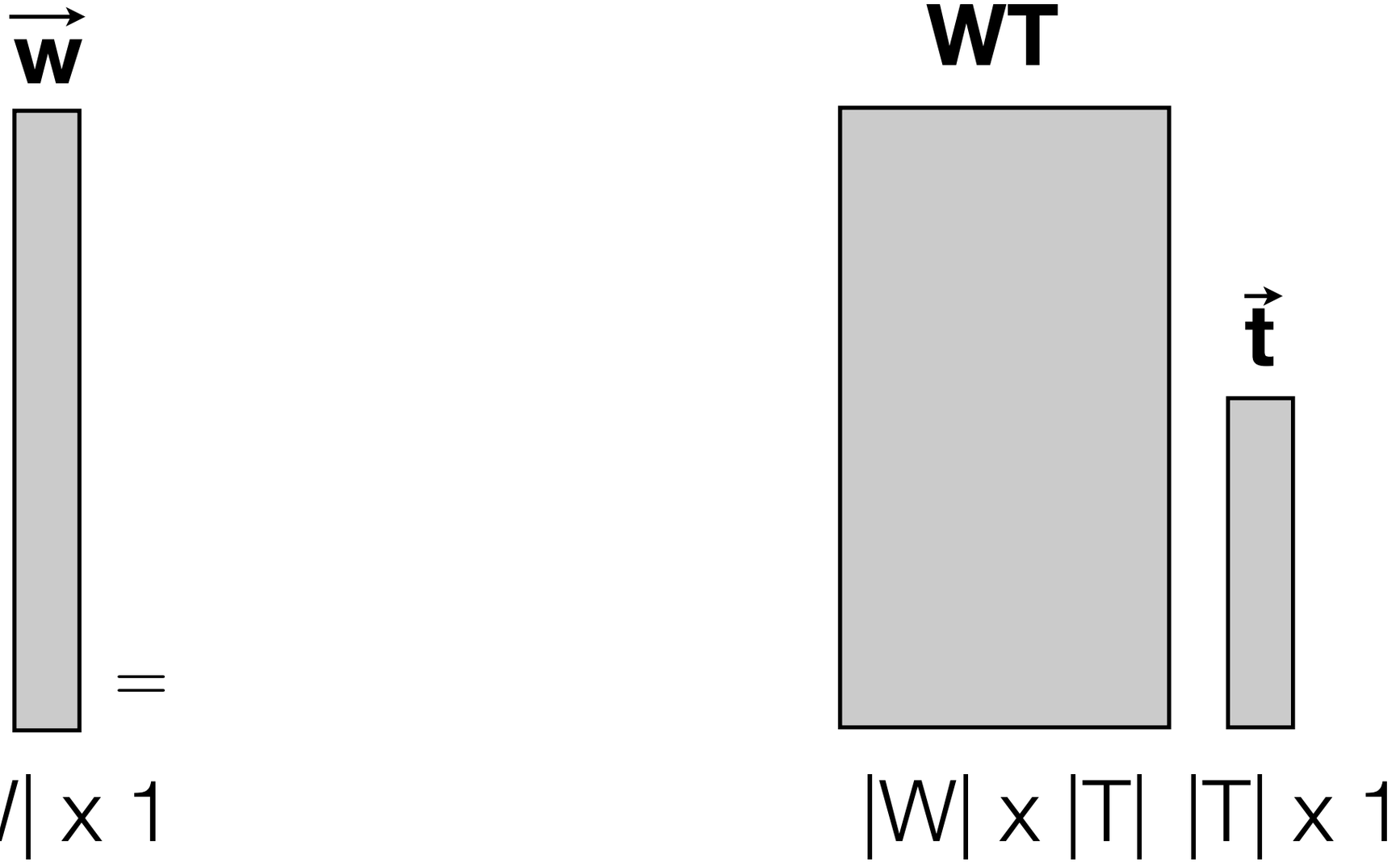
$$\Pr(w_{|W|} \mid C) = \sum_T \Pr(w_{|W|} \mid T) \Pr(T \mid C)$$



$|W| \times 1$ $\quad$ $|W| \times |T|$ $\quad$ $|T| \times 1$

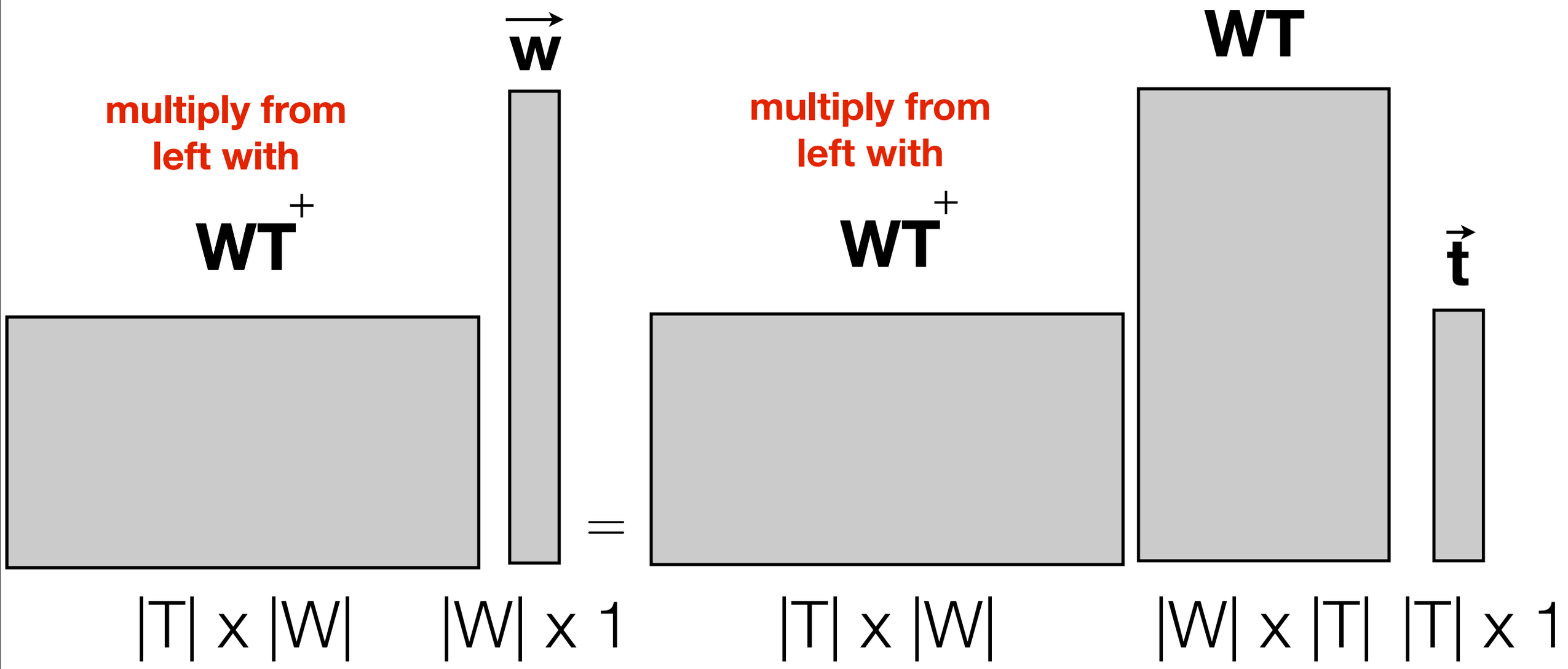
Noisy Channel Model

$$
\begin{aligned}
\mathbf{WT}_{ij} &= \Pr(W = i | T = j) \\
\vec{t}_j &= \Pr(T = j | C = k) \\
\vec{w}_i &= \Pr(W = i | C = k)
\end{aligned}
$$



Noisy Channel Model

$$\begin{aligned}
\mathbf{WT}_{ij} &= \Pr(W = i | T = j) \\
\vec{t}_j &= \Pr(T = j | C = k) \\
\vec{w}_i &= \Pr(W = i | C = k)
\end{aligned}$$



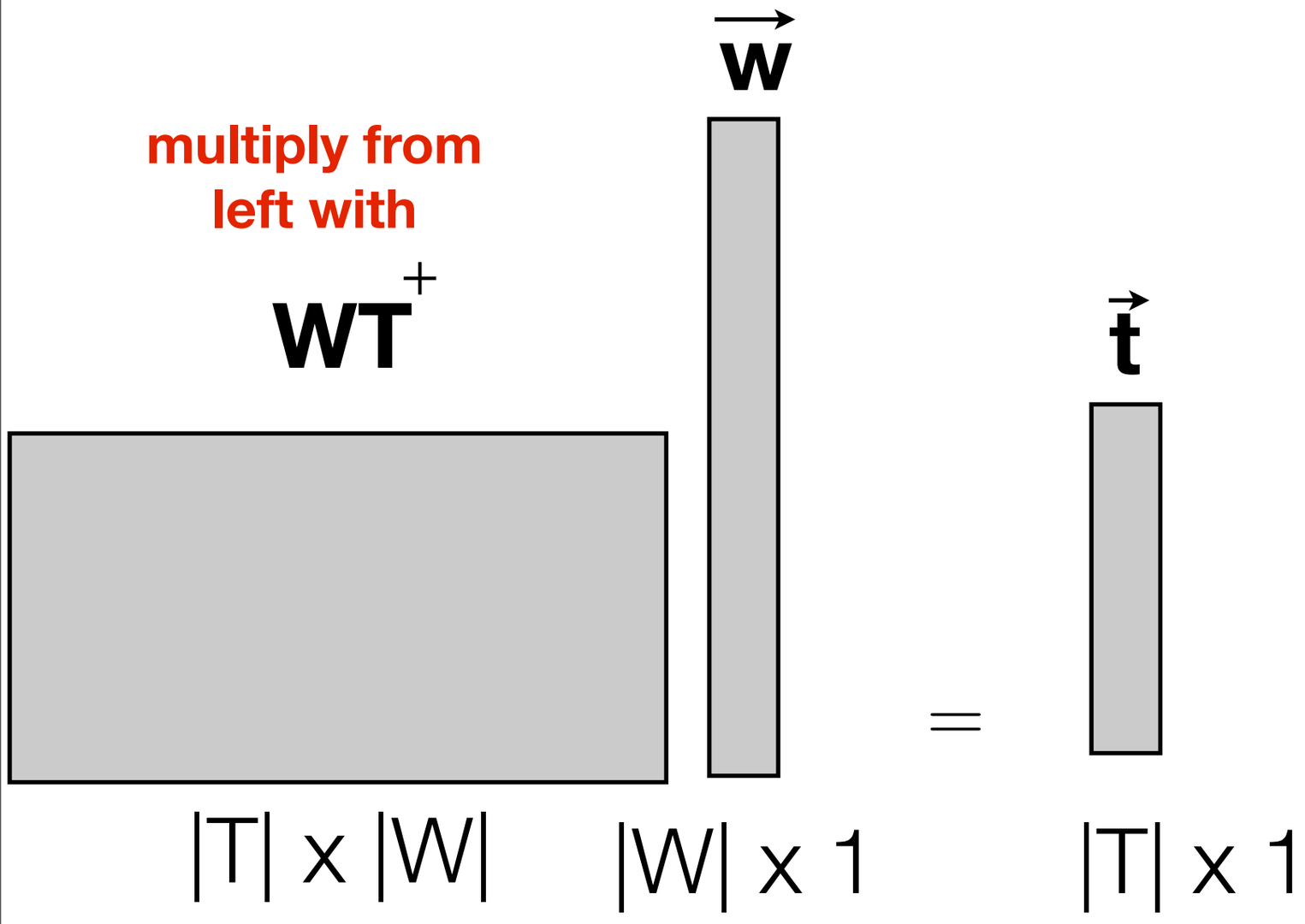
Noisy Channel Model

Pr(T | W, C) ∝ Pr(W|T) Pr(T|C)

$$\begin{aligned}
\mathbf{WT}_{ij} &= \Pr(W = i | T = j) \\
\vec{t}_j &= \Pr(T = j | C = k) \\
\vec{w}_i &= \Pr(W = i | C = k)
\end{aligned}$$

multiply from left with

WT^{+}



|T| x |W| |W| x 1 |T| x 1

Pr(T|C) is estimated

Noisy channel model on WSD

Applied on WSD of English nouns

- State of the art results among the models with word-tag distribution or dictionary available

- Comparable results with supervised systems (tag sequence is available)

Conclusion

- I propose a new context representation
- I propose 5 models that can use this representations
- Achieve the state-of-the-art results on 19 corpus in 15 languages languages in POS induction problem
- Applied to the probabilistic voting to morphological disambiguation of Turkish and achieve promising results
- Achieve the state-of-the art results on POS disambiguation of English when a word-tag dictionary is available
- Achieve the state-of-the-art results on WSD disambiguation of English nouns when a word-tag distribution is available

Any Questions?

Thanks to _____ .

Dad

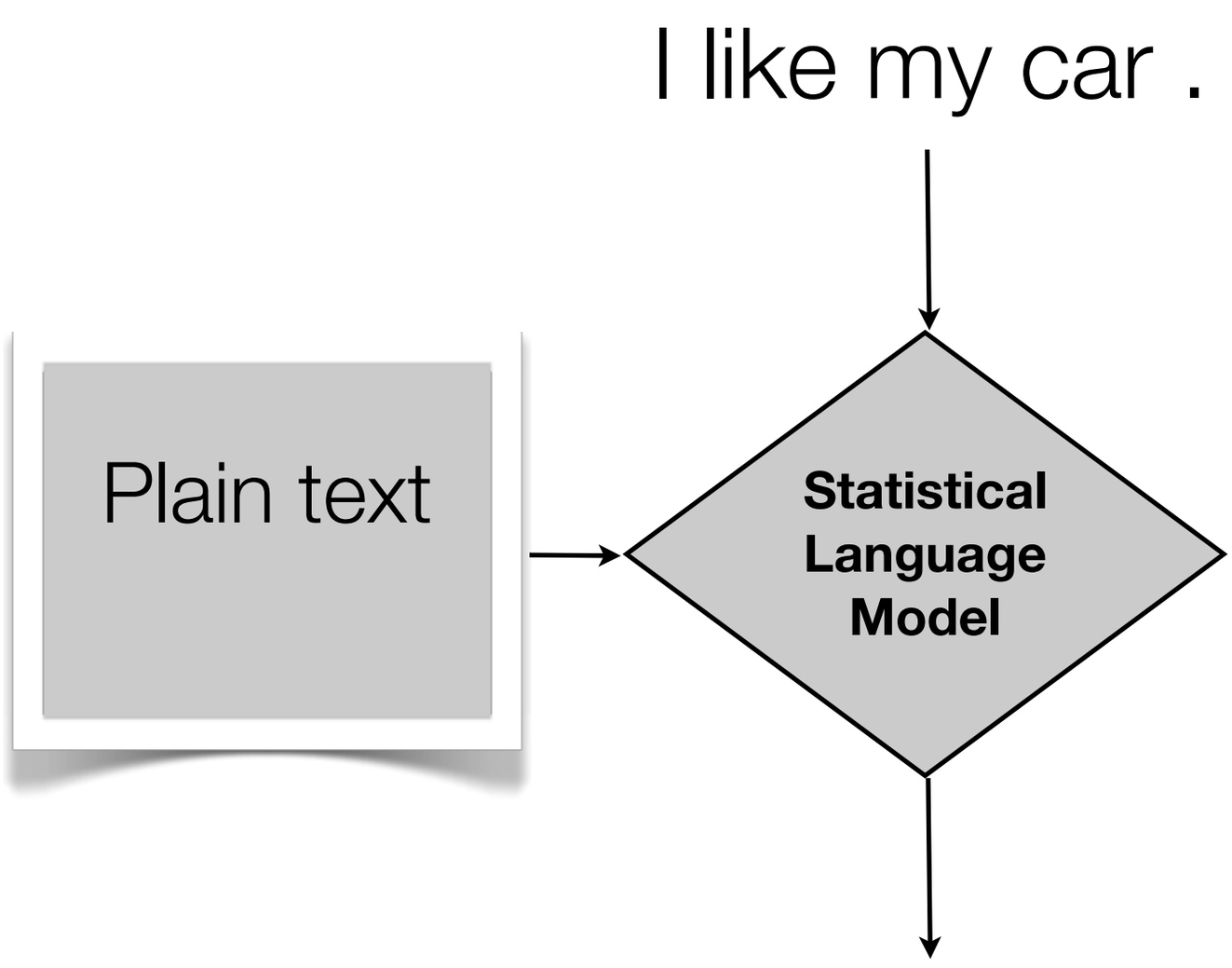
Mom

Deniz Yuret

Professors

Friends

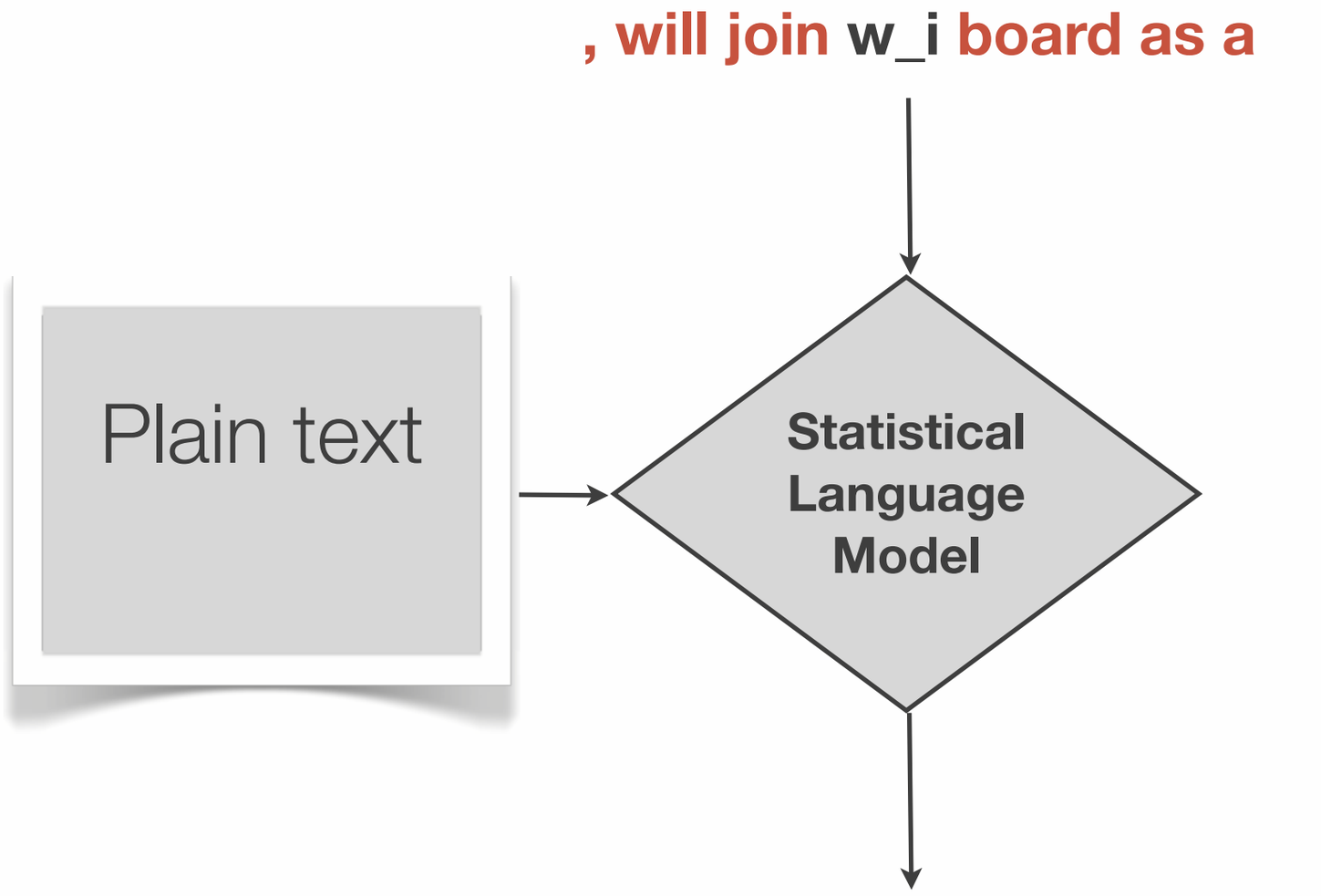
How to calculate substitute distrbitions?



- Statistical Language Models
  - read large amount of plain text
  - assign probabilities to a given word sequence

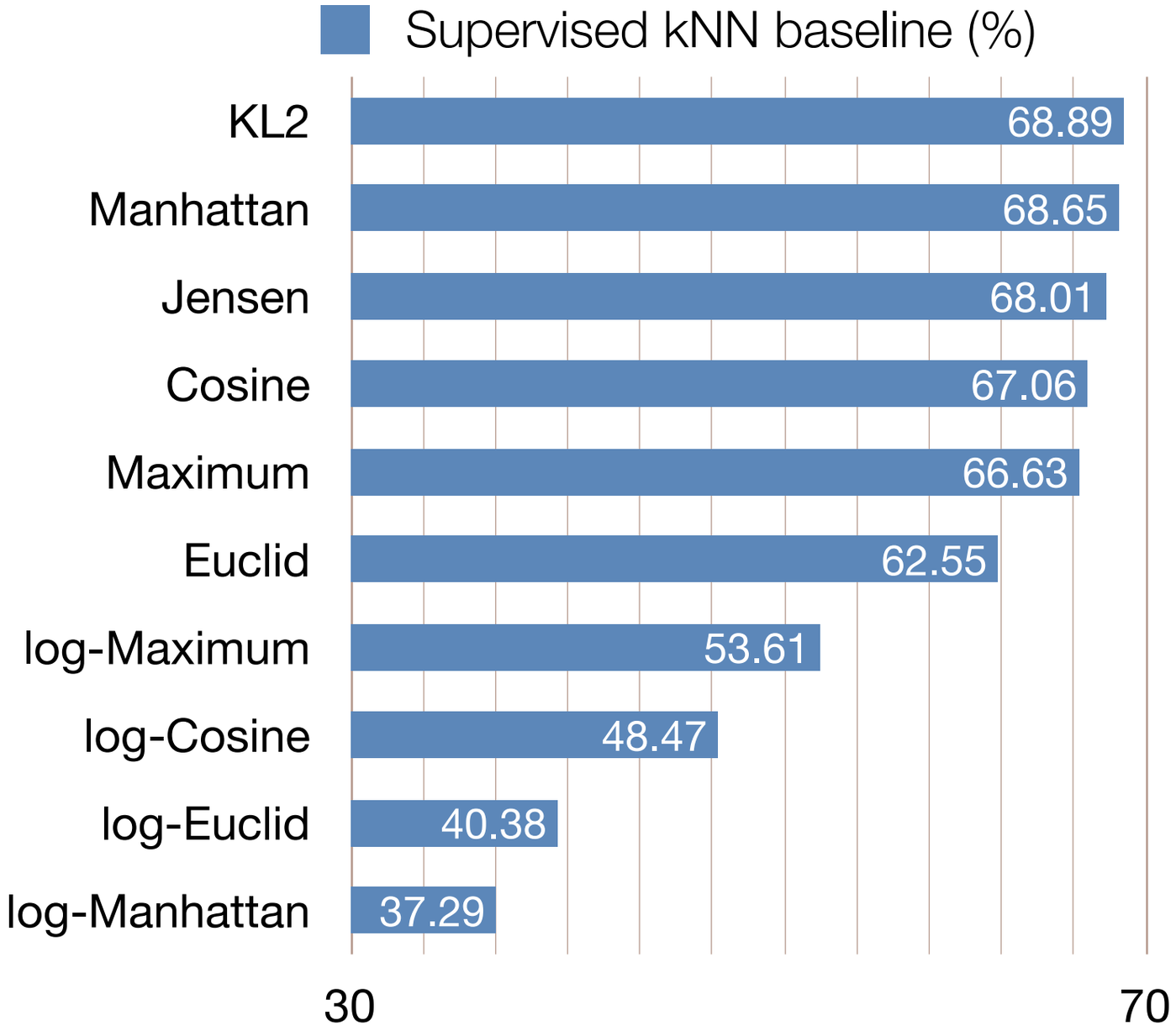
How to calculate substitute distributions?

for **w_i** in LM **vocabulary** perform:



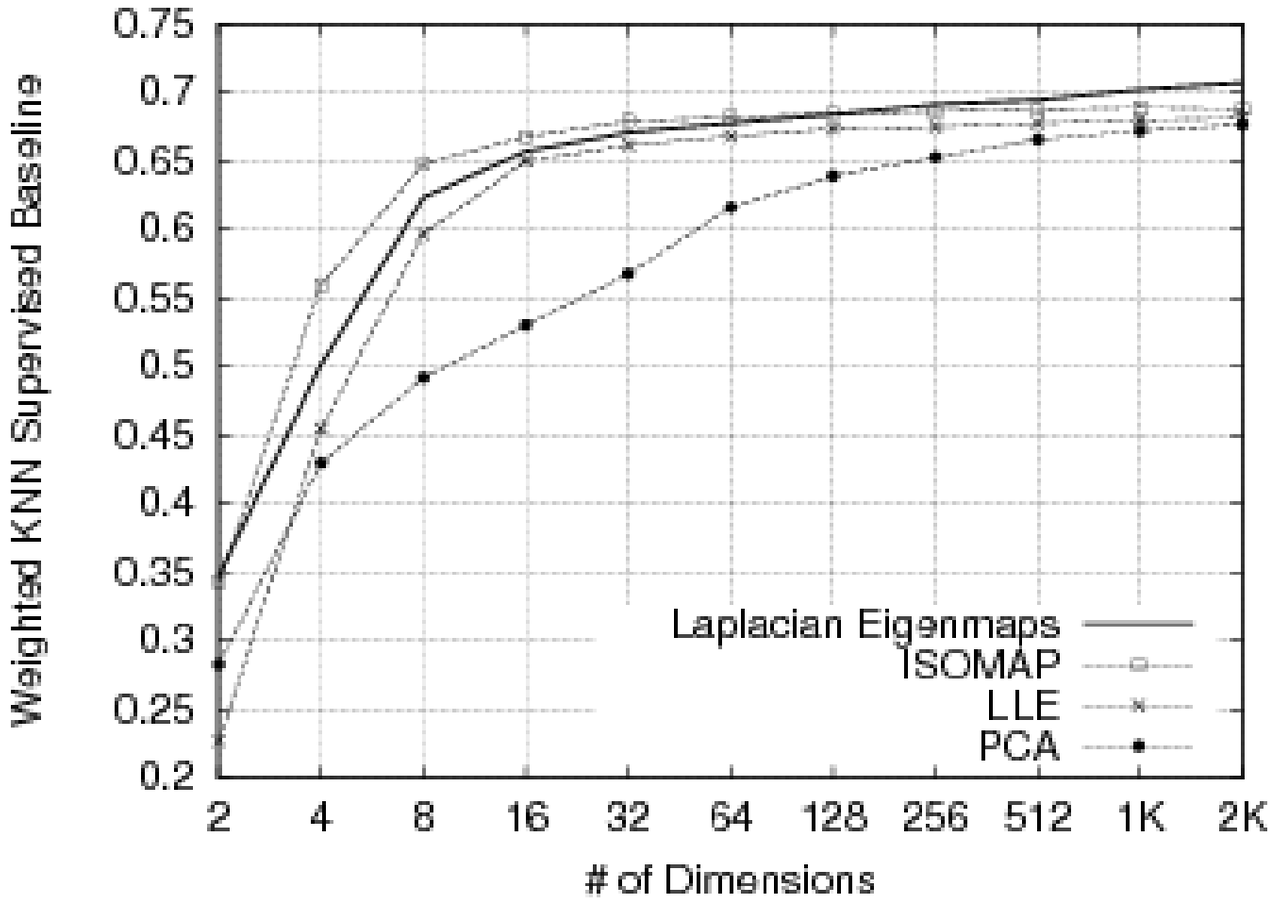
- For all w_i in vocabulary
  - Replace with target word
  - use LM to get probability of the new sequence
- will have a |vocabulary| dimensional vector, **s**
- normalize **s** to make it probability distribution

Paradigmatic Representations of Word Context



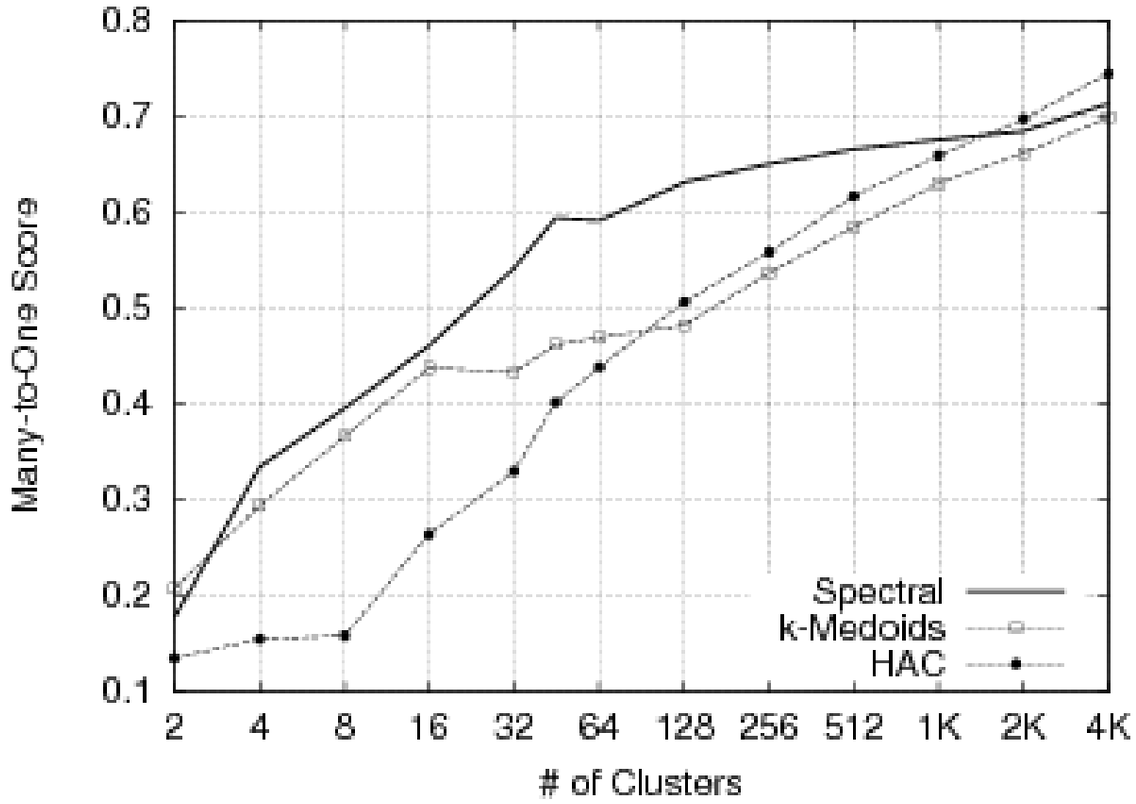
- Distance metrics on log probability vectors performed poorly compared to their regular counterparts
- The differences in low probability words are relatively unimportant
- High probability substitutes determine syntactic category.

Paradigmatic Representations of Word Context



- Supervised KNN baselines for POS accuracy using various dimensionality reduction algorithms on substitute vectors.

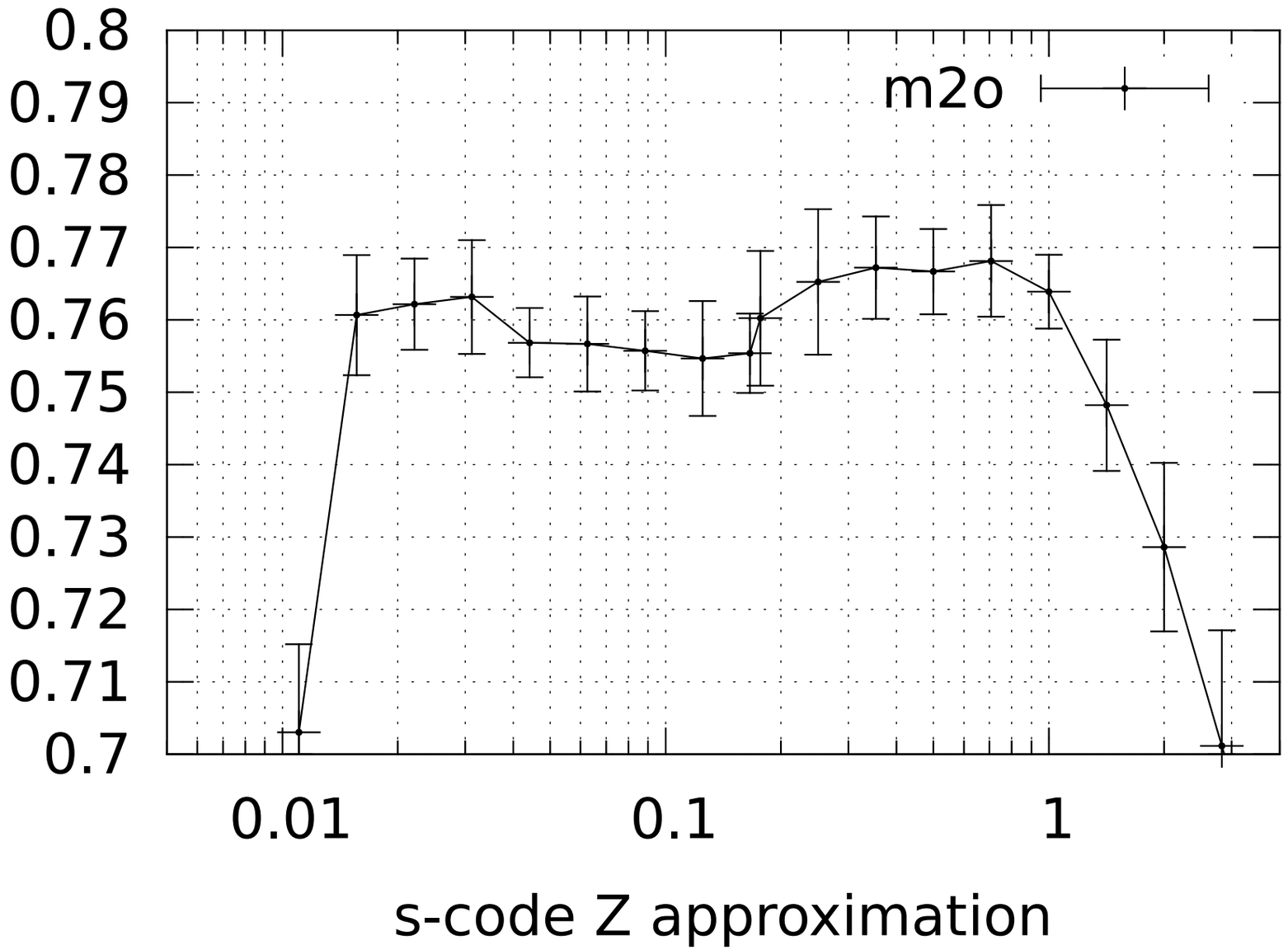
Paradigmatic Representations of Word Context



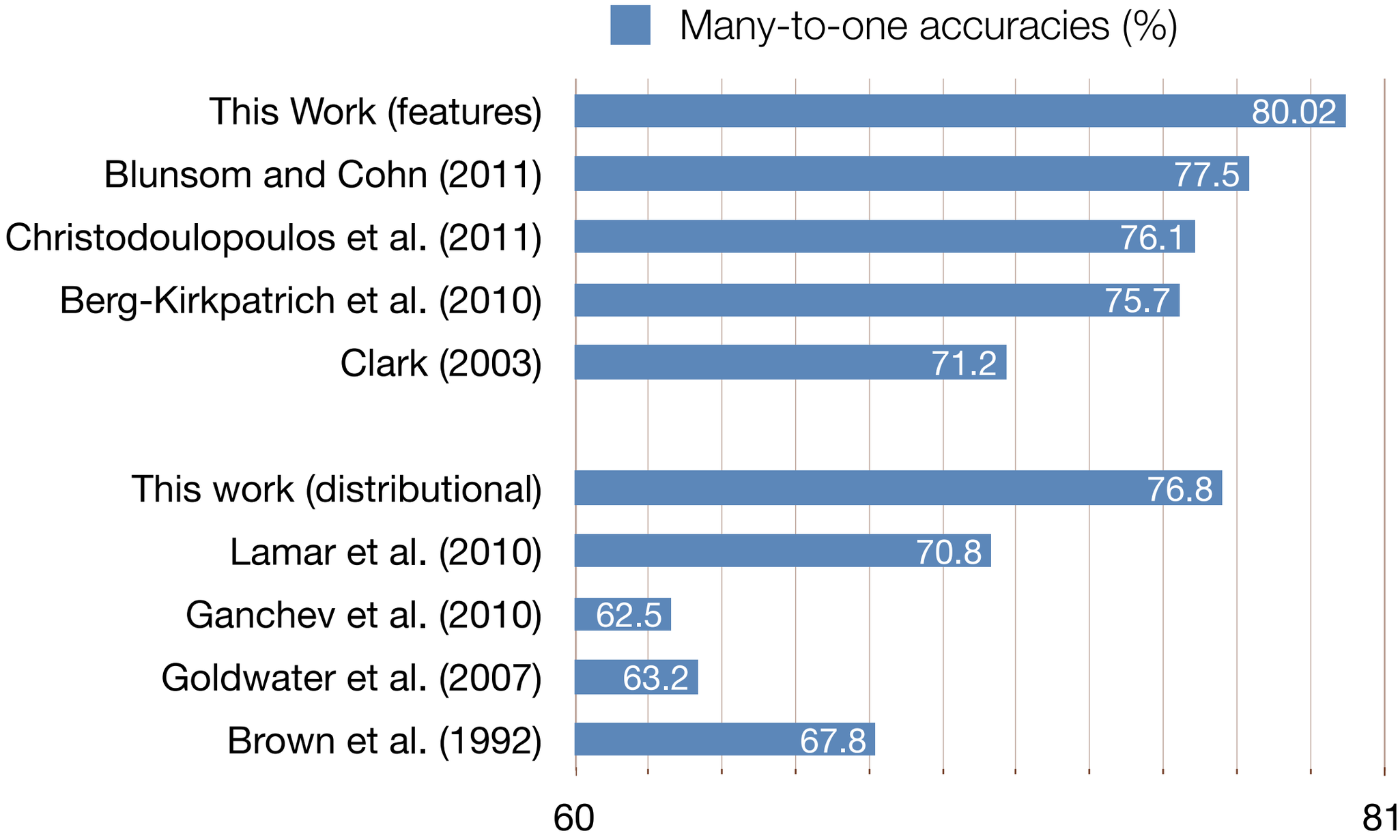
- Supervised KNN baselines for POS accuracy using various dimensionality reduction algorithms on substitute vectors.

Experiments and Results

- Sensitivity Analysis

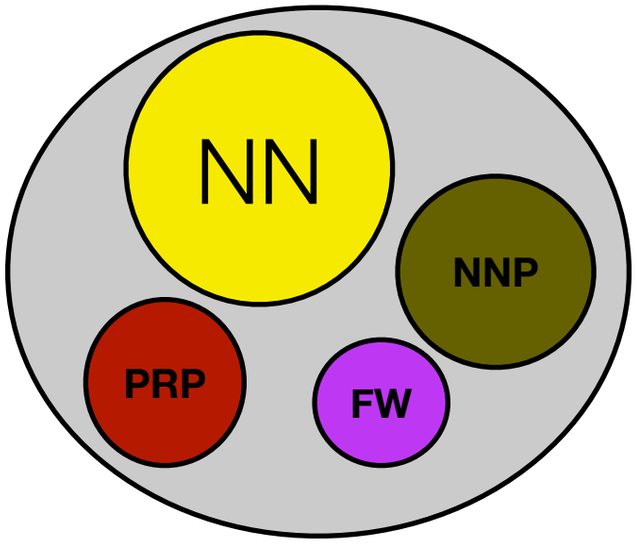


POS induction related work



How to **evaluate** Unsupervised Results

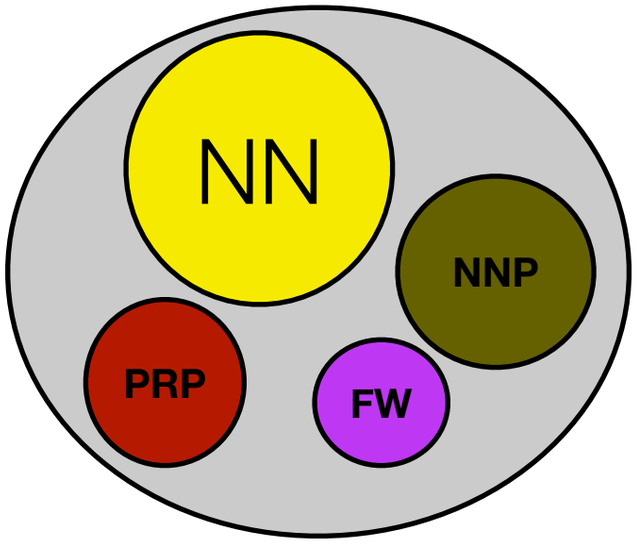
The gold tag distribution of Cluster C



- compare their consistency with **answer (gold)** tags:
  - Many-to-one Score:
    - Label each word in a cluster with the **most observed** gold tag in that cluster.

How to **evaluate** Unsupervised Results

The gold tag distribution of Cluster C

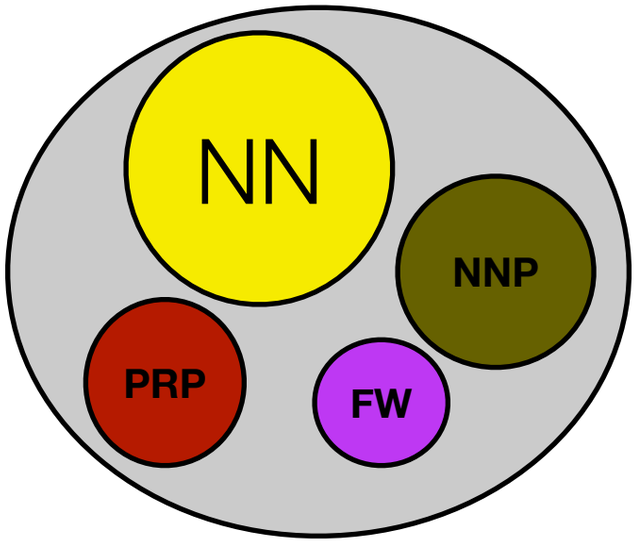


Cluster C is mapped to **NN**

- compare their consistency with **answer (gold)** tags:
  - Many-to-one Score:
    - Label each word in a cluster with the **most observed** gold tag in that cluster.

How to **evaluate** Unsupervised Results

The gold tag distribution of Cluster C



Cluster C is mapped to **NN**

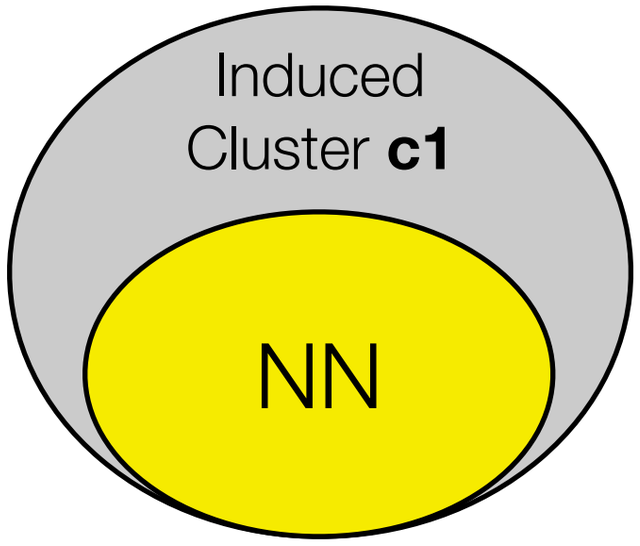
- compare their consistency with **answer (gold)** tags:
  - Many-to-one Score:
    - Label each word in a cluster with the **most observed** gold tag in that cluster.

$$\text{m2o Accuracy} = \frac{\text{NN}}{\text{FW} + \text{PRP} + \text{NNP} + \text{NN}}$$

How to **evaluate** Unsupervised Results

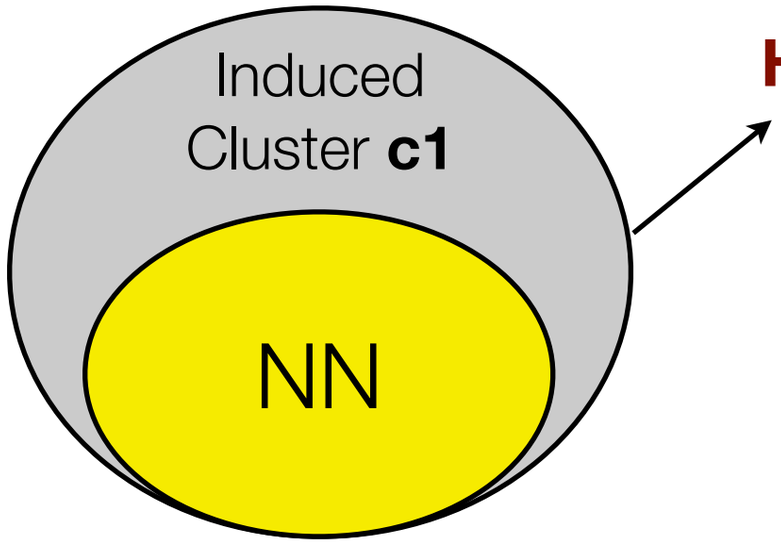
- VM Score
  - Entropy based measure
  - Analogous to F measure
    - harmonic mean of **homogeneity** and **completeness**.

How to **evaluate** Unsupervised Results



- VM Score
  - Entropy based measure
  - Analogous to F measure
    - harmonic mean of **homogeneity** and **completeness**.

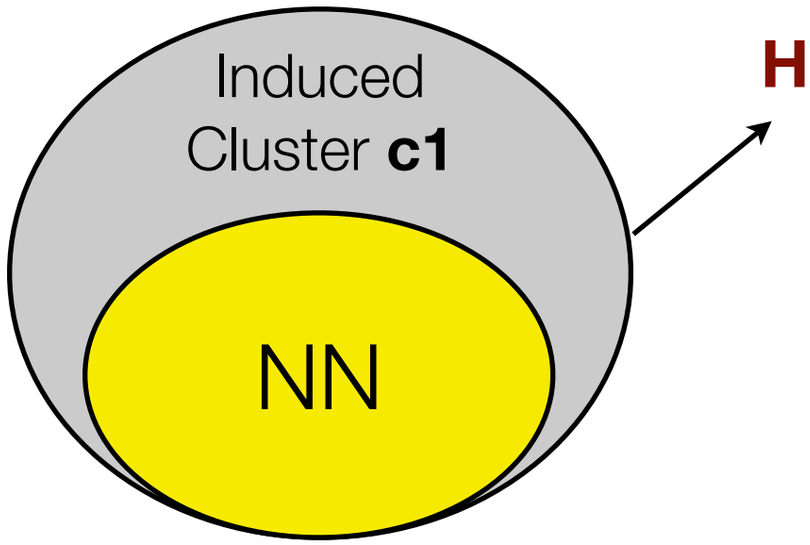
How to **evaluate** Unsupervised Results



High homogeneity

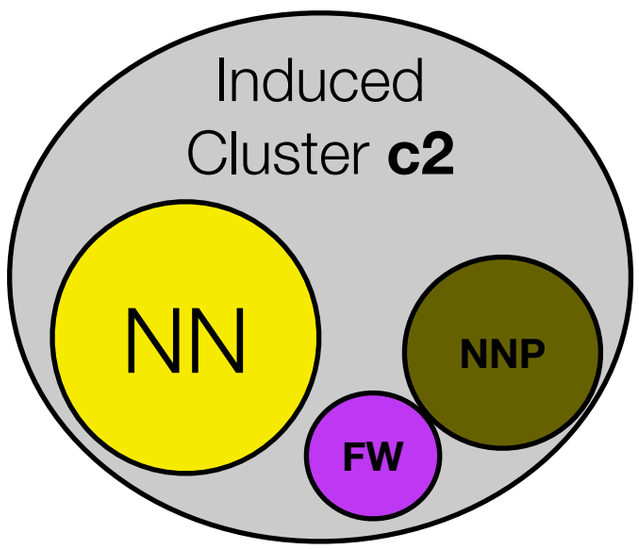
- VM Score
  - Entropy based measure
  - Analogous to F measure
    - harmonic mean of **homogeneity** and **completeness**.

How to **evaluate** Unsupervised Results

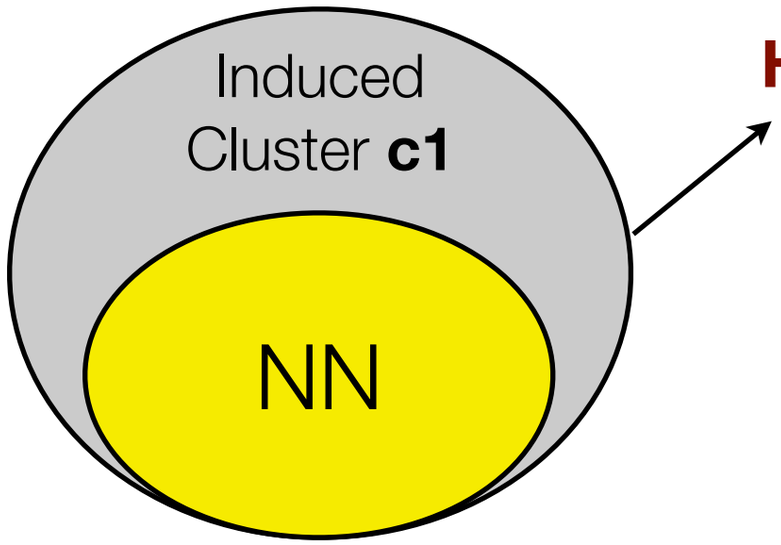


High homogeneity

- VM Score
  - Entropy based measure
  - Analogous to F measure
  - harmonic mean of **homogeneity** and **completeness**.

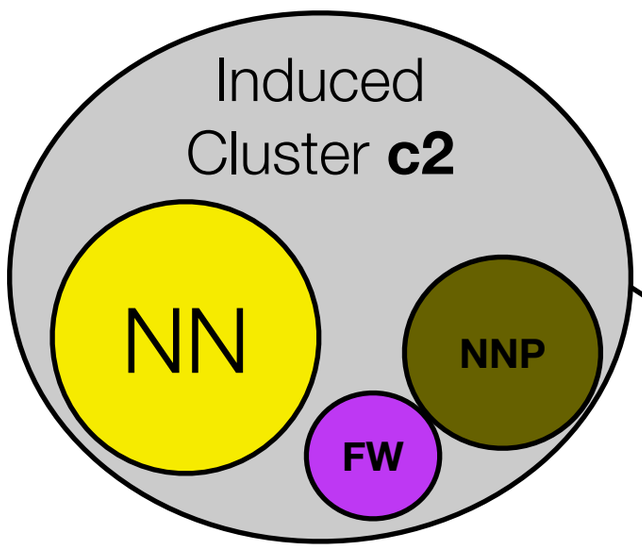


How to **evaluate** Unsupervised Results



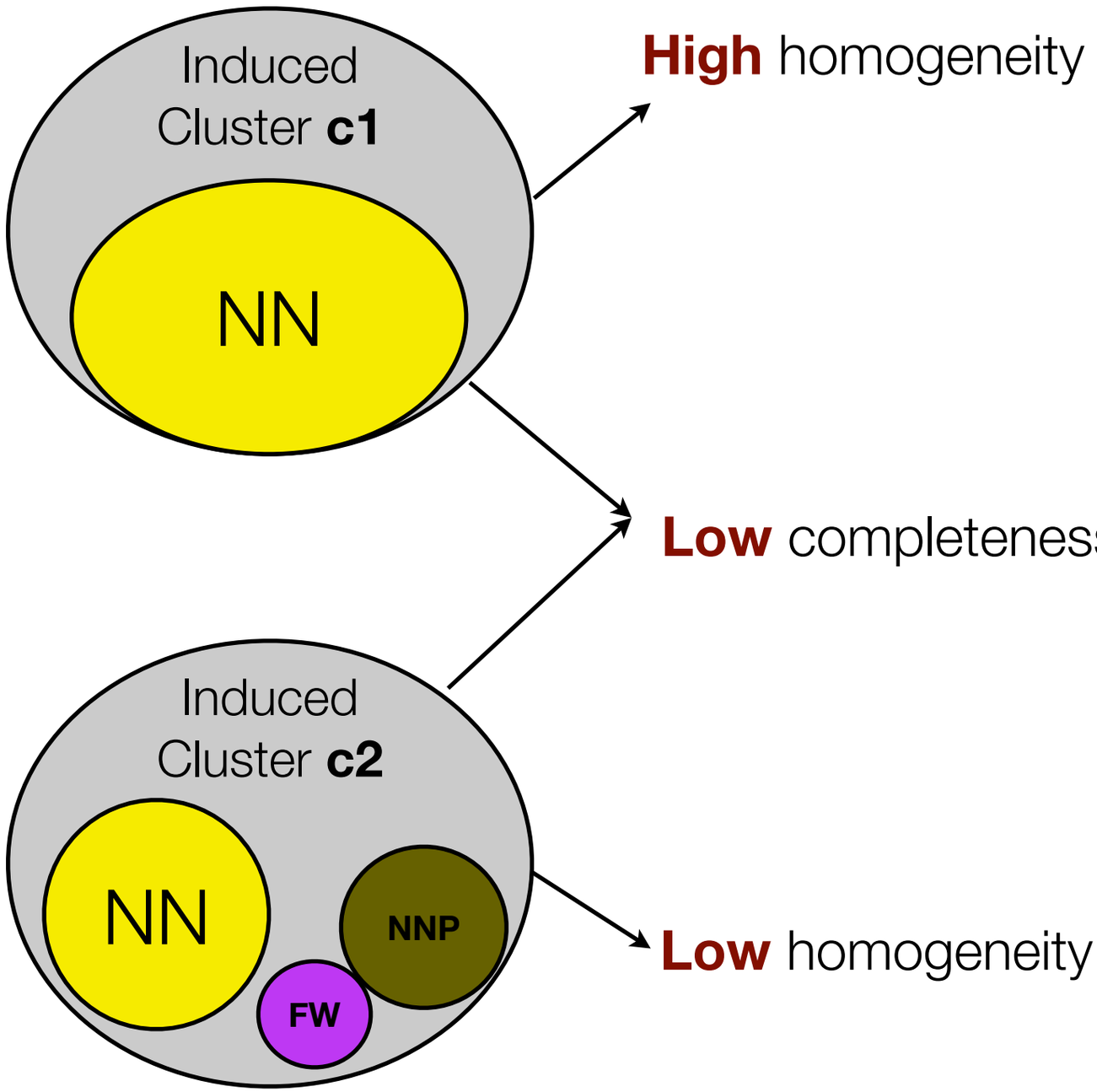
High homogeneity

- VM Score
  - Entropy based measure
  - Analogous to F measure
  - harmonic mean of **homogeneity** and **completeness**.



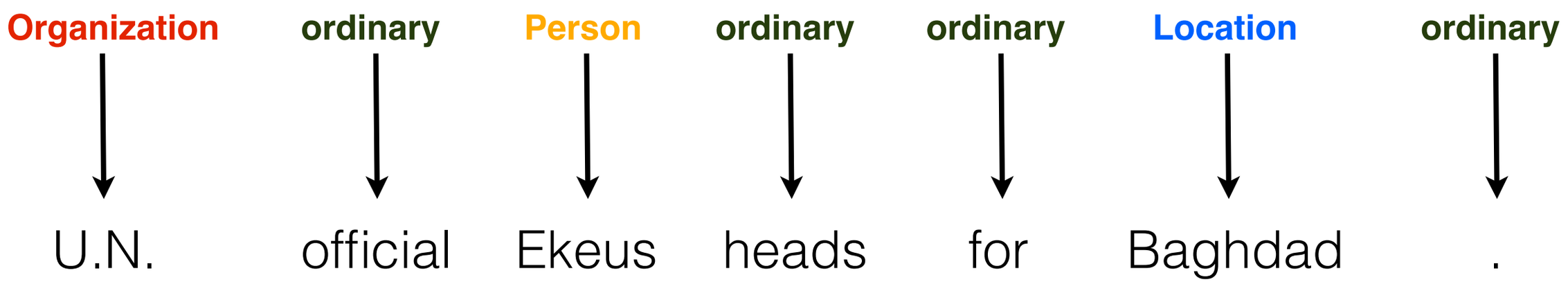
Low homogeneity

How to **evaluate** Unsupervised Results



- VM Score
  - Entropy based measure
  - Analogous to F measure
  - harmonic mean of **homogeneity** and **completeness**.

Named entity tagging



- tag words with predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages,

Clustering Substitute Distributions

- Ignores target-word identities or features
  - 93% of the tags in words in English is assigned to most frequent POS tag (one-tag-per-word assumption)
  - After clustering the substitute distributions, we assigned each word to the majority cluster of its instances
  - Ex: If instances of the word **W** distributed as c1(10), c2(20), c3(60) and c4(10) then all instances will be moved to majority cluster **c3**.

achieves **~71 % m2o** accuracy

S-CODE (Maron et al. 2010)

- Distance is proportional to mutual information

$$\frac{\Pr(w,s)}{\Pr(w)\Pr(s)} \propto e^{-d_{w,s}^2} \qquad d_{w,s}^2 = (w-s)^2$$

$$\Pr(w,s) \propto \Pr(s)\Pr(s)e^{-d_{w,s}^2} \qquad Z = \sum_{(w,s)} \Pr(w)\Pr(s)e^{-d_{w,s}^2}$$

$$\Pr(w,s) = \frac{1}{Z}\Pr(x)\Pr(y)e^{-d_{w,s}^2}$$

$$\log likelihood = \sum_{(w,s)} \overline{p}(w,s)\log(p(w,s))$$

S-CODE (Maron et al. 2010)

- S-CODE defines the log likelihood of co-occurrence data

$$l = \sum_{(w,s)} \overline{p}(w,s) \log(p(w,s))$$

S-CODE (Maron et al. 2010)

- S-CODE defines the log likelihood of co-occurrence data

$$l = \sum_{(w,s)} \overline{p}(w,s) \log(p(w,s))$$

$$l = \sum_{(w,s)} \overline{p}(w,s) \left(-\log Z + \log \overline{p}(w) \overline{p}(s) - d_{w,s}^2 \right)$$

S-CODE (Maron et al. 2010)

- S-CODE defines the log likelihood of co-occurrence data

$$l = \sum_{(w,s)} \overline{p}(w,s) \log(p(w,s))$$

$$l = \sum_{(w,s)} \overline{p}(w,s) \left(-\log Z + \log \overline{p}(w) \overline{p}(s) - d_{w,s}^2 \right)$$

$$l = -\log Z + \sum_{w,s} \overline{p}(w,s) \log \overline{p}(w) \overline{p}(s) - \overline{p}(w,s) \| \Phi_w - \Phi_s \|^2$$

S-CODE (Maron et al. 2010)

- S-CODE defines the log likelihood of co-occurrence data

$$l = \sum_{(w,s)} \overline{p}(w,s) \log(p(w,s))$$

$$l = \sum_{(w,s)} \overline{p}(w,s) \left(-\log Z + \log \overline{p}(w) \overline{p}(s) - d_{w,s}^2 \right)$$

$$l = -\log Z + \sum_{w,s} \quad \textbf{constant} \quad - \overline{p}(w,s) \| \Phi_w - \Phi_s \|^2$$

S-CODE (Maron et al. 2010)

- Take derivative according to embeddings of x and y

$$l = -\log Z + const - \sum_{w,s} \bar{p}(w,s) \| \Phi_w - \Phi_s \|^2 \qquad \text{Eq1}$$

97

derivative of y's embedding is similar

S-CODE (Maron et al. 2010)

- Take derivative according to embeddings of x and y

$$l = -\log Z + const - \sum_{w,s} \bar{p}(w,s) \| \Phi_w - \Phi_s \|^2 \qquad \text{Eq1}$$

$$\frac{\delta l}{\delta \Phi_w} = \sum_s \frac{1}{Z} \bar{p}_w \bar{p}_s e^{-d^2_{w,s}} (\Phi_w - \Phi_s) + \sum_s \bar{p}_{w,s} (\Phi_s - \Phi_w) \qquad \text{Eq2}$$

derivative of y's embedding is similar

S-CODE (Maron et al. 2010)

- Take derivative according to embeddings of x and y

$$l = -\log Z + const - \sum_{w,s} \bar{p}(w,s) \| \Phi_w - \Phi_s \|^2 \qquad \text{Eq1}$$

$$\frac{\delta l}{\delta \Phi_w} = \sum_s \frac{1}{Z} \bar{p}_w \bar{p}_s e^{-d^2_{w,s}} (\Phi_w - \Phi_s) + \sum_s \bar{p}_{w,s} (\Phi_s - \Phi_w) \qquad \text{Eq2}$$

$$\frac{\delta l}{\delta \Phi_w} = \sum_s p_{w,s} (\Phi_w - \Phi_s) + \sum_s \bar{p}_{w,s} (\Phi_s - \Phi_w) \qquad \text{Eq3}$$

97

derivative of y's embedding is similar

S-CODE (Maron et al. 2010)

- Take derivative according to embeddings of x and y

$$l = -\log Z + const - \sum_{w,s} \bar{p}(w,s) \| \Phi_w - \Phi_s \|^2 \qquad \text{Eq1}$$

$$\frac{\delta l}{\delta \Phi_w} = \sum_s \frac{1}{Z} \bar{p}_w \bar{p}_s e^{-d^2_{w,s}} (\Phi_w - \Phi_s) + \sum_s \bar{p}_{w,s} (\Phi_s - \Phi_w) \qquad \text{Eq2}$$

by definition

$$\frac{\delta l}{\delta \Phi_w} = \sum_s p_{w,s} (\Phi_w - \Phi_s) + \sum_s \bar{p}_{w,s} (\Phi_s - \Phi_w) \qquad \text{Eq3}$$

97

derivative of y's embedding is similar

Noisy channel model on WSD

- 25 WordNet Semantic Categories for nouns are used
- For a given word **w** the model
  1. find the correct sense class
  2. selects the most frequent sense of **w** in that class

Task	WN	Nouns	FSB	1st	2nd	3rd	Unsup	Score
senseval2	1.7	1067	71.9	78.0	74.5	70.0	61.8	77.7
senseval3	1.7.1	892	71.0	72.0	71.2	71.0	62.6	70.1
semeval07	2.1	159	64.2	68.6	66.7	66.7	63.5	64.8
total		2118	70.9	74.4	72.5	70.2	62.2	73.5

Noisy Channel Model

$$\begin{aligned}\Pr(W|C) &= \sum_T \Pr(W|T,C)\Pr(T|C) \\ \Pr(W|C) &= \sum_T \Pr(W|T)\Pr(T|C)\end{aligned}$$

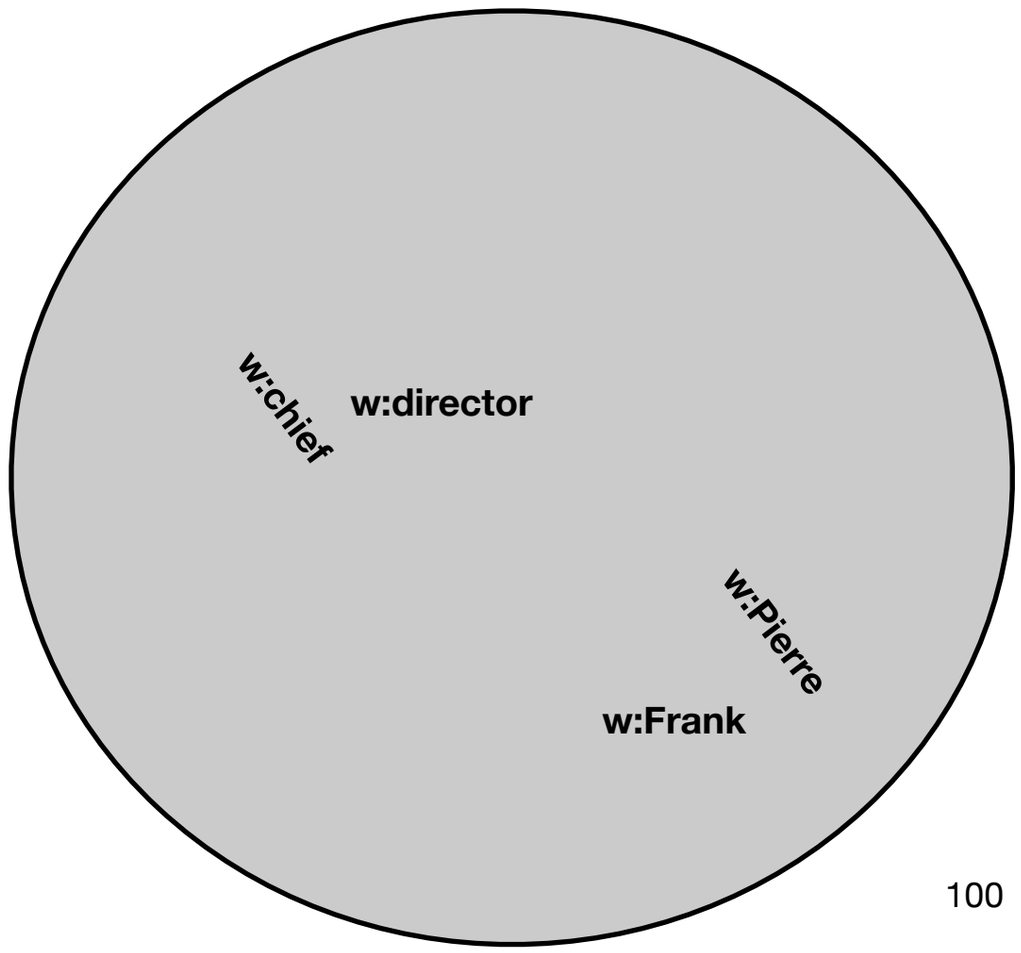
For a fixed context (channel) **k**

$$\begin{aligned}\mathbf{WT}_{ij} &= \Pr(W=i|T=j) \\ \vec{t}_j &= \Pr(T=j|C=k) \\ \vec{w}_i &= \Pr(W=i|C=k)\end{aligned}$$

Modeling Co-occurrence

- Handle ambiguity by clustering the **S** or concatenation of **W** and **S.**
- We represent each instance with the concatenation of correspond W and the average of S embeddings
  - con(w:director s:chairman)
- We achieve comparable results with the best published systems on 15 out of 19 corpora

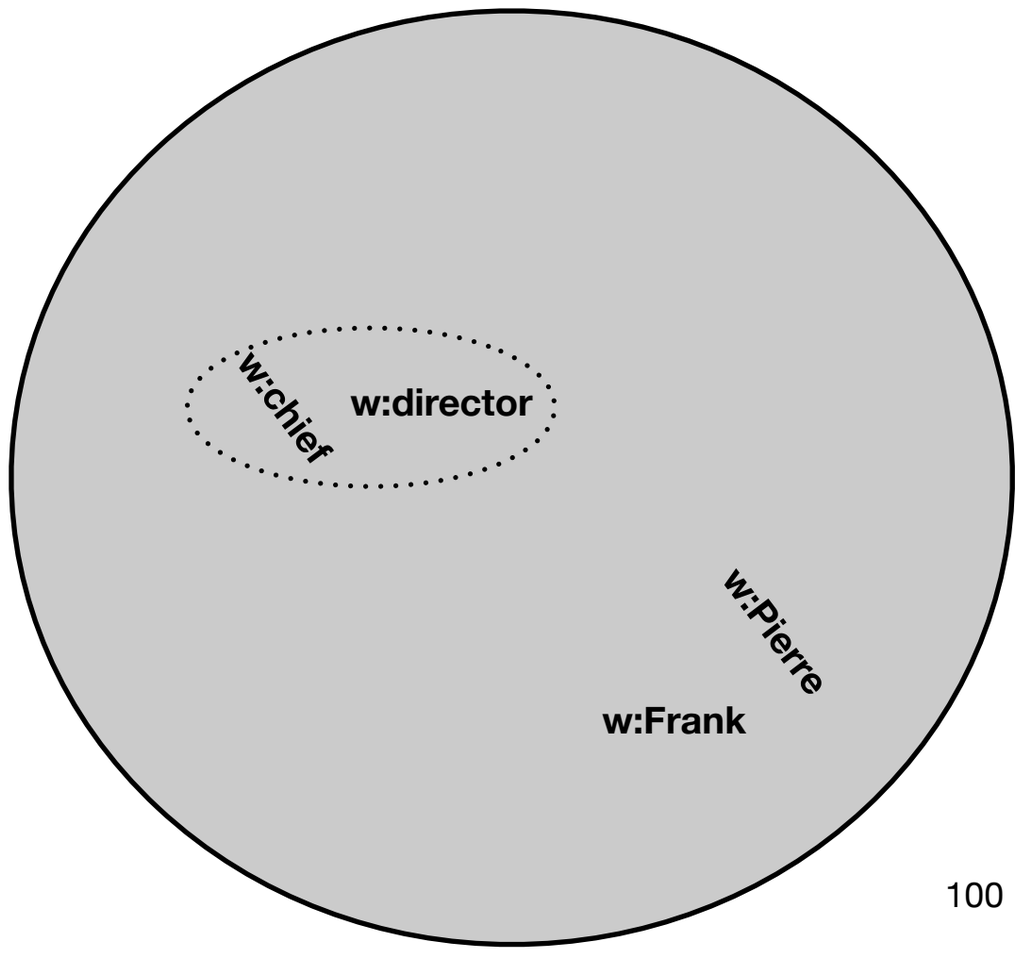
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence

- Handle ambiguity by clustering the **S** or concatenation of **W** and **S.**
- We represent each instance with the concatenation of correspond W and the average of S embeddings
  - con(w:director s:chairman)
- We achieve comparable results with the best published systems on 15 out of 19 corpora

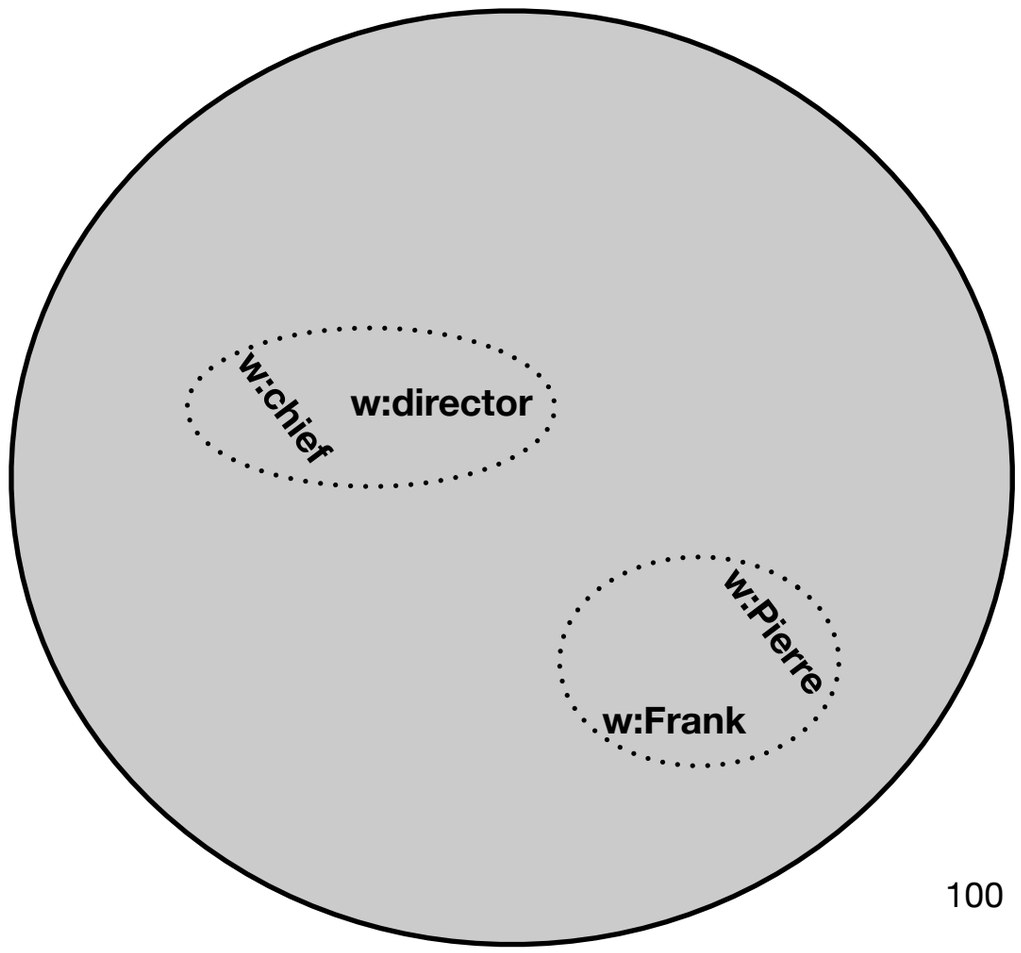
W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



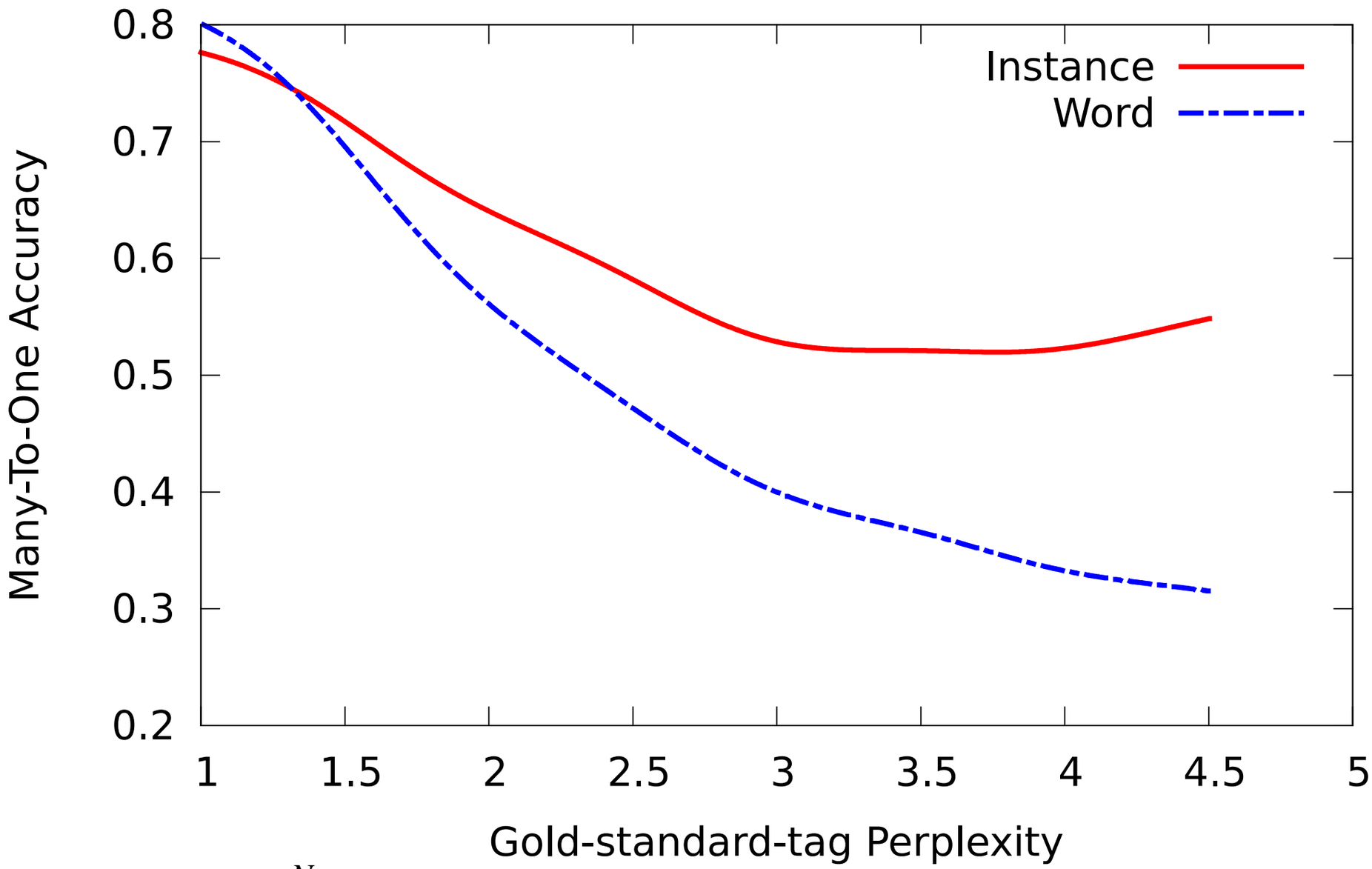
Modeling Co-occurrence

- Handle ambiguity by clustering the **S** or concatenation of **W** and **S.**
- We represent each instance with the concatenation of correspond W and the average of S embeddings
  - con(w:director s:chairman)
- We achieve comparable results with the best published systems on 15 out of 19 corpora

W	S
w:director	s:chairman
w:chief	s:chairman
w:Pierre	s:John
w:Frank	s:John



Modeling Co-occurrence



$$GP(w) = 2^{-\sum_{i=1}^{N} \Pr(t|w) \log_2 \Pr(t|w)}$$

GP=1 when word is unambiguous

Dictionary reduction on POS tagging

Word	Tag dictionary	Gold tagging	EM tagging	Substitutes POS counts
of	{RB, RP, IN}	IN(632) RP(0) RB(0)	IN(0) RP(632) RB(0)	IN(2377) **RP(0)** RB(850)
a	{LS, SYM, NNP, FW, JJ, IN, DT}	DT(458) IN(1) JJ(2) SYM(1) LS(0)	DT(0) IN(0) JJ(0) SYM(258) LS(230)	DT(513) IN(317) JJ(1329) **SYM(0)** **LS(0)**

Dictionary reduction on POS tagging

Word	Tag dictionary	Gold tagging	EM tagging	Substitutes POS counts
of	{RB, RP, IN}	IN(632) RP(0) RB(0)	IN(0) RP(632) RB(0)	IN(2377) **RP(0)** RB(850)
a	{LS, SYM, NNP, FW, JJ, IN, DT}	DT(458) IN(1) JJ(2) SYM(1) LS(0)	DT(0) IN(0) JJ(0) SYM(258) LS(230)	DT(513) IN(317) JJ(1329) **SYM(0)** **LS(0)**

Inverse

- w = WS x t
- t = inv(WS) x w
  - This solution minimizes the distance |WS x t - w|
    - might violate non-negativity
    - add up to 1
- D(P||Q) = sum_i ln(P(i)/Q(i)) P(i) where P = WS x t , P = w

		Language Model			Test Corpus			
	Language	Source	Instance Count	Word Count	Instance Count	Word Count	Unknown Word	Perplexity (ppl)
WSJ	English	ukWaC	2,303,225,131	4,254,946	1,173,766	49,206	0.0081	303.477
MULTEXT-East	Bulgarian	TenTen	849,023,297	1,965,178	101,173	16,353	.0151	295.704
	Czech	TenTen	1,791,613,805	4,758,807	100,368	19,121	.0038	294.022
	English	ukWaC	2,303,225,131	4,254,946	118,424	9,774	.0046	143.451
	Estonian	TenTen	330,671,558	2,526,585	94,898	17,847	.0166	477.805
	Hungarian	Wikipedia	66,069,788	1,065,897	98,426	20,323	.0449	654.086
	Romanian	TenTen	53,456,650	310,366	118,328	15,192	.0070	126.596
	Slovene	Wikipedia	18,969,864	363,251	112,278	17,873	.0389	648.347
	Serbian	Wikipedia	17,129,679	368,778	108,809	18,113	.0580	804.962
CoNLL-X Shared Task	Bulgarian	TenTen	849,023,297	1,965,178	190,217	32,439	.0196	168.592
	Czech	TenTen	1,791,613,805	4,758,807	1,249,408	130,208	.0050	476.434
	Danish	TenTen	1,857,746,600	5,304,957	94,386	18356	.0218	185.325
	Dutch	WaC	127,580,512	774,965	195,069	28,493	.0465	261.709
	German	TenTen	1,810,802,875	6,513,804	699,610	72,326	.0227	417.676
	Portuguese	TenTen	3,267,166,367	3,434,834	206,678	28,932	.0493	364.92
	Slovene	Wikipedia	18,969,864	363,251	28,750	7,128	.0414	596.678
	Spanish	TenTen	2,445,878,830	3,067,682	89,334	16,458	.0343	193.94
	Swedish	TenTen	113,975,094	926,875	191,467	20,057	.0179	288.16
	Turkish	TenTen	1,804,606,896	5,308,241	47,605	17,563	.0550	600.632