Graph-Cut-Based Anaphoricity Determination for Coreference Resolution

Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Noun Phrase Coreference

Identify the noun phrases (NPs) that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify the noun phrases (NPs) that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify the noun phrases (NPs) that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify the noun phrases (NPs) that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into <u>a viable monarch</u>. <u>A renowned speech therapist</u> was summoned to help the King overcome his <u>speech impediment</u>...

Noun Phrase Coreference

Identify the noun phrases (NPs) that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model, $\mathbf{P_C}$, to determine the probability that two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model, $\mathbf{P_C}$, to determine the probability that two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5

- **Step 2: Antecedent selection**
  - find an antecedent for each NP_j

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model, $\mathbf{P_C}$, to determine the probability that two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5
- **Step 2: Antecedent selection**
  - find an antecedent for each NP_j
    - choose the closest preceding noun phrase that is classified as coreferent with NP_j

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model, $\mathbf{P_C}$, to determine the probability that two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5
- **Step 2: Antecedent selection**
  - find an antecedent for each NP_j
    - choose the closest preceding noun phrase that is classified as coreferent with NP_j

coref

[Queen Elizabeth] set about transforming [her] [husband], ...

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model, $\mathbf{P_C}$, to determine the probability that two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5
- **Step 2: Antecedent selection**
  - find an antecedent for each NP_j
    - choose the closest preceding noun phrase that is classified as coreferent with NP_j

coref

[Queen Elizabeth] set about transforming [her] [husband], ...

Standard Machine Learning Approach

- **Step 1: Classification**
  - train a coreference model ... at two NPs are coreferent
  - two NPs are classified as coreferent iff probability ≥ 0.5

find an antecedent for each NPj
that has an antecedent

- **Step 2: Antecedent selection**
  - find an antecedent for each NP_j
    - choose the closest preceding noun phrase that is classified as coreferent with NP_j

coref

[Queen Elizabeth] set about transforming [her] [husband], ...

How to determine whether an NP has an antecedent?

How to determine whether an NP has an antecedent?

Queen Elizabeth set about transforming her husband,
King George VI, into a viable monarch. Logue,
a renowned speech therapist, was summoned to help
the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

~~Queen Elizabeth~~ set about transforming her husband,

King George VI, into a viable monarch.

A renowned speech therapist, was summoned to help

the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

~~Queen Elizabeth~~ set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

Queen Elizabeth set about transforming her husband,
King George VI, into a viable monarch.
A renowned speech therapist was summoned to help
the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

Queen Elizabeth set about transforming her ~~husband~~,
King George VI, into a viable monarch.
A renowned speech therapist was summoned to help
the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

Queen Elizabeth set about transforming her ~~husband~~,

King George VI, into a viable monarch.

A renowned speech therapist was summoned to help

the King overcome his speech impediment...

How to determine whether an NP has an antecedent?

Any NP that is part of a coref chain but is not the head of the chain has an antecedent. It's an anaphoric NP.

Queen Elizabeth set about transforming her ~~husband~~,

King George VI, into a viable monarch.

A renowned speech therapist was summoned to help

the King overcome his speech impediment...

Anaphoricity Determination

- determines whether an NP is anaphoric or not
- helps improve the precision of a coreference system

Goal

Improve learning-based coreference systems using automatically acquired anaphoricity information, by proposing a new approach to anaphoricity determination

Plan for the Talk

- Existing methods for computing and using anaphoricity info
- Our graph-cut-based approach to anaphoricity determination
- Evaluation

Methods for Computing and Using Anaphoricity Information

- Five existing methods
  - Ng & Cardie (2002)
  - Ng (2004)
  - Luo (2007)
  - Denis & Baldridge (2007)
  - Kleener (2007), Finkel & Manning (2008)

What do the methods have in common?

What do the methods have in common?

- Training an anaphoricity model (**P_A**)
  - determines the probability that an NP is anaphoric
  - classifies an NP as anaphoric iff probability ≥ 0.5

What do the methods have in common?

- Training an anaphoricity model (**P_A**)
  - determines the probability that an NP is anaphoric
  - classifies an NP as anaphoric iff probability ≥ 0.5

- Training data creation
  - texts annotated with coreference information
  - one instance for each NP
    - positive if the NP is part of a coref chain but not head of chain
    - negative otherwise

How do the methods differ from each other?

How do the methods differ from each other?

- They differ in terms of
  - whether they improve the output of $\mathbf{P_A}$

How do the methods differ from each other?

- They differ in terms of
  - whether they improve the output of $\mathbf{P_A}$
    - if so, how?

How do the methods differ from each other?

- They differ in terms of
  - whether they improve the output of $\mathbf{P_A}$
    - if so, how?
  - how anaphoricity info is used by the coreference system
    - as hard constraints or as soft constraints?

Ng & Cardie (2002)

Improve P_A's output?	Used as hard constraint?
✗	✓

Ng & Cardie (2002)

Improve $\mathbf{P_A}$'s output?	Used as hard constraint?
✗	✓

NP$_j$ is classified
as anaphoric iff
$\mathbf{P_A}(\underline{NP_j}) \geq 0.5$

Ng & Cardie (2002)

Improve P_A's output?	Used as hard constraint?
✗	✓

NP_j is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq 0.5$

coref system finds an antecedent for NP_j iff it is classified as anaphoric

Ng & Cardie (2002)

Improve P_A's output?	**Used as hard constraint?**
✗	✓
NP$_j$ is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq 0.5$	coref system finds an antecedent for *NP*$_j$ iff it is classified as anaphoric

- Problem
  - many anaphoric NPs are misclassified (as non-anaphoric)

Ng & Cardie (2002)

Improve $\mathbf{P_A}$'s output?	**Used as hard constraint?**
✗	✓
NP_j is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq 0.5$	coref system finds an antecedent for NP_j iff it is classified as anaphoric

- Problem
  - many anaphoric NPs are misclassified (as non-anaphoric)
    - $\mathbf{P_A}$ is overly conservative in classifying an NP as anaphoric

Ng (2004)

Improve P_A's output?	Used as hard constraint?
✓	✓

Ng (2004)

Improve P_A's output?	Used as hard constraint?
✓	✓

coref system finds an antecedent for NP_j iff it is classified as anaphoric

Ng (2004)

Improve P_A's output?	**Used as hard constraint?**
✓	✓

NP$_j$ is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq \boldsymbol{t}$

coref system finds an antecedent for *NP*$_j$ iff it is classified as anaphoric

Ng (2004)

Improve P_A's output?	**Used as hard constraint?**
✓	✓

NP_j is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq \boldsymbol{t}$

coref system finds an antecedent for NP_j iff it is classified as anaphoric

- decreasing ***t*** → more NPs will be classified as anaphoric
- increasing ***t*** → fewer NPs will be classified as anaphoric

Ng (2004)

Improve P_A's output?	**Used as hard constraint?**
✓	✓
NP$_j$ is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq \boldsymbol{t}$	coref system finds an antecedent for *NP*$_j$ iff it is classified as anaphoric

- decreasing ***t*** ⇒ more NPs will be classified as anaphoric
- increasing ***t*** ⇒ fewer NPs will be classified as anaphoric
- ***t*** is the "conservativeness" parameter

Ng (2004)

Improve P_A's output?	**Used as hard constraint?**
✓	✓

NP_j is classified as anaphoric iff $\mathbf{P_A}(\underline{NP_j}) \geq \boldsymbol{t}$

coref system finds an antecedent for *NP_j* iff it is classified as anaphoric

- decreasing ***t*** ⇒ more NPs will be classified as anaphoric
- increasing ***t*** ⇒ fewer NPs will be classified as anaphoric
- select ***t*** to use held-out data to maximize coreference performance (i.e., F-measure)

Luo (2007)

Improve P_A's output?	Used as hard constraint?
✗	✗

Luo (2007)

Improve P_A's output?	Used as hard constraint?
X	X

- Goal
  - score an NP partition
    - by multiplying the probabilities provided by **P_A** and **P_C**

Luo (2007)

Improve P_A's output?	Used as hard constraint?
X	X

- Goal
  - score an NP partition
    - by multiplying the probabilities provided by $\mathbf{P_A}$ and $\mathbf{P_C}$
  - find highest-scored NP partition

Luo (2007)

Improve P_A's output?	Used as hard constraint?
X	X

- Goal
  - score an NP partition
    - by multiplying the probabilities provided by **P_A** and **P_C**
  - find highest-scored NP partition
    - by performing a beam search through the Bell tree

Denis & Baldridge (2007)

Improve P_A's output?	Used as hard constraint?
✓	✓

Denis & Baldridge (2007)

Improve P_A's output?	Used as hard constraint?
✓	✓

coref system finds an antecedent for NP_j iff it is classified as anaphoric

Denis & Baldridge (2007)

Improve P_A's output?	Used as hard constraint?
✓	✓

coref system finds an antecedent for NP_j iff it is classified as anaphoric

- use Integer Linear Programming (ILP) to perform joint inference for anaphoricity determination and coreference

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

  $\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

  $\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

NP 3 is anaphoric!!!

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

$\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

NP 3 is **not** anaphoric!

NP 3 is anaphoric!!!

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

  $\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

NP 3 is **not** anaphoric!

NP 3 is anaphoric!!!

- $\mathbf{P_A}$ and $\mathbf{P_C}$'s outputs don't seem to be consistent. Why???

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

  $\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

NP 3 is **not** anaphoric!

NP 3 is anaphoric!!!

- $\mathbf{P_A}$ and $\mathbf{P_C}$'s outputs don't seem to be consistent. Why???
  - Because they are trained independently of each other

ILP: A Motivating Example

- 3 NPs: 1, 2, 3
- $\mathbf{P_c}(1, 2) = 0.6$, $\mathbf{P_c}(1, 3) = 0.2$, $\mathbf{P_c}(2, 3) = 0.9$

  $\mathbf{P_A}(1) = 0.1$, $\mathbf{P_A}(2) = 0.9$, $\mathbf{P_A}(3) = 0.2$

NP 3 is **not** anaphoric!

NP 3 is anaphoric!!!

- $\mathbf{P_A}$ and $\mathbf{P_C}$'s outputs don't seem to be consistent. Why???
  - Because they are trained independently of each other
- Certain hard constraints need to be enforced
  - If $\mathbf{P_C}$ determines that NP_j is not coreferent with any NP, then $\mathbf{P_A}$ should determine that NP_j is non-anaphoric
  - **…**

ILP for Anaphoricity and Coreference

- Goal
  - jointly determine anaphoricity and coreference decisions such that all the desired constraints are satisfied

- improve anaphoricity decisions with automatically computed coreference information and manually-specified constraints

Kleener (2007), Finkel & Manning (2008)

Improve P_A's output?	Used as hard constraint?
✓	✓

Kleener (2007), Finkel & Manning (2008)

Improve P_A's output?	Used as hard constraint?
✓	✓

- Also employ ILP, but additionally impose the transitivity constraint on the coreference decisions
  - A, B are coref **and** B,C are coref   A,C are coreferent

Summary of the Five Methods

	Improve P_A's output?	Used as hard constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

Summary of the Five Methods

	Improve P_A's output?	Used as hard constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

tune ***t*** on held-out coref-annotated data to optimize F-measure

Summary of the Five Methods

	Improve P_A's output?	Used as hard constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

tune ***t*** on held-out coref-annotated data to optimize F-measure

exploit probabilities provided by $\mathbf{P_C}$

Summary of the Five Methods

	Improve P_A's output?	Used as hard constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

tune ***t*** on held-out coref-annotated data to optimize F-measure

exploit probabilities provided by $\mathbf{P_C}$

do **not** optimize F-measure

Summary of the Five Methods

	Improve P_A's output?	Used as hard constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

tune ***t*** on held-out coref-annotated data to optimize F-measure

does **not** exploit $\mathbf{P_C}$

exploit probabilities provided by $\mathbf{P_C}$

do **not** optimize F-measure

Summary of the Five

Can we have a method that optimizes F-measure and exploits $\mathbf{P_C}$?

	Impr[illegible] $_A$ [illegible] output?	[illegible] constraint?
Ng and Cardie (2002)	✗	✓
Ng (2004)	✓	✓
Luo (2007)	✗	✗
Denis and Baldridge (2007)	✓	✓
Kleener (2007)	✓	✓

tune ***t*** on held-out coref-annotated data to optimize F-measure

exploit probabilities provided by $\mathbf{P_C}$

does **not** exploit $\mathbf{P_C}$

do **not** optimize F-measure

Cut-Based Anaphoricity Determination

- Motivated by our desire to have a method that can
  - optimize the desired coreference evaluation metric
  - exploit probabilities provided by $\mathbf{P_C}$

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T
- Given two types of scores:
  - Membership scores: $mem_S(x_i)$, $mem_T(x_i)$
    - captures the affinity of x_i to S and T, respectively
    - large $mem_S(x_i)$ x_i is likely to be in S
  - Similarity scores: $sim(x_i, x_j)$
    - captures the similarity between x_i and x_j

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T
- Given two types of scores:
  - Membership scores: $mem_S(x_i)$, $mem_T(x_i)$
    - captures the affinity of x_i to S and T, respectively
    - large $mem_S(x_i)$ x_i is likely to be in S
  - Similarity scores: $sim(x_i, x_j)$
    - captures the similarity between x_i and x_j

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T

- Given two types of scores:
  - Membership scores: $mem_S(x_i)$, $mem_T(x_i)$
    - captures the affinity of x_i to S and T, respectively
    - large $mem_S(x_i)$ x_i is likely to be in S
  - Similarity scores: $sim(x_i, x_j)$
    - captures the similarity between x_i and x_j

- Goal:

Maximize $$\sum_{x_i \in S, x_j \in T} sim(x_i, x_j) + \sum_{x \in S} mem_S(x) + \sum_{x \in T} mem_T(x)$$

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T

- Given two types of scores:
  - Membership scores: $mem_S(x_i)$, $mem_T(x_i)$
    - captures the affinity of x_i to S and T, respectively
    - … is likely to be in S
  - S… (x_i, x_j)
    - captures the similarity between x_i and x_j

Put similar objects into the same set

- Goal:

Maximize $$\sum_{x_i \in S, x_j \in T} sim(x_i, x_j) + \sum_{x \in S} mem_S(x) + \sum_{x \in T} mem_T(x)$$

The Problem Setting

- Want to partition a set of objects, $\{x_1, x_2, \ldots, x_n\}$, into two sets, S and T

- Given two types of scores:
  - Membership scores: $mem_S(x_i)$, $mem_T(x_i)$
    - captures the affinity of x_i to S and T, respectively
    - is likely to be
  - S (x_i, x_j)
    - captures the similarity between x_i and x_j

Put similar objects into the same set

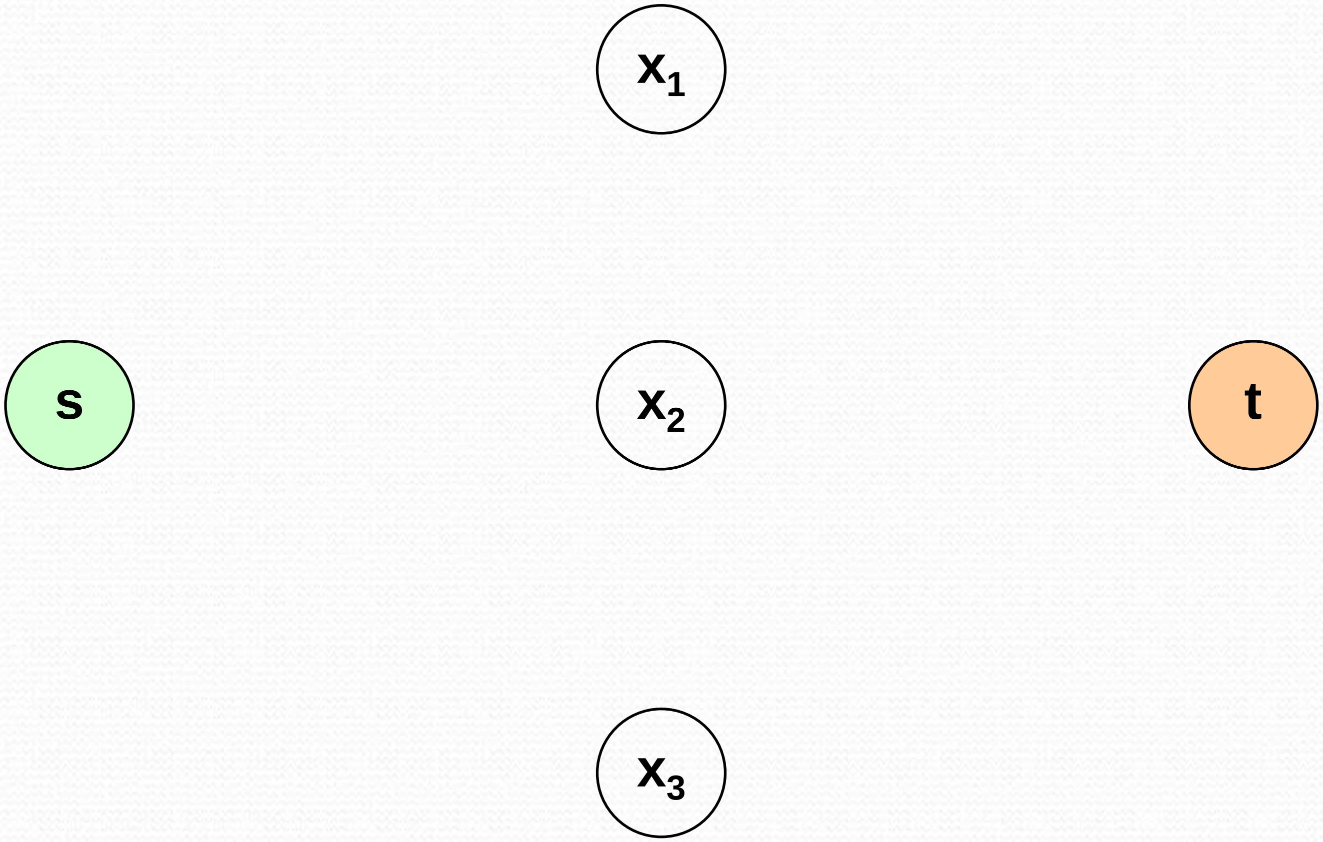
Put an object to the set where its membership score is high

- Goal:

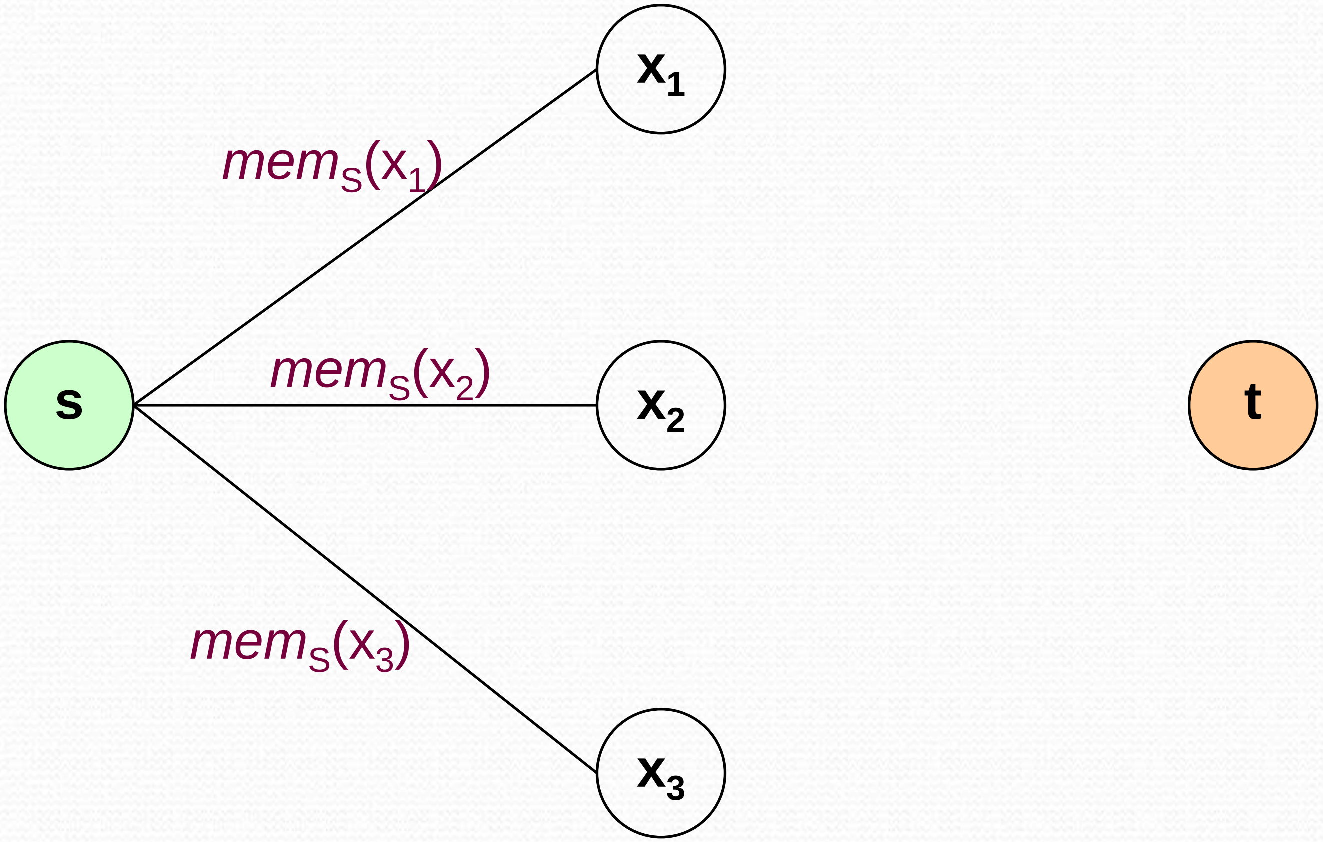
Maximize $$\sum_{x_i \in S, x_j \in T} sim(x_i, x_j) + \sum_{x \in S} mem_S(x) + \sum_{x \in T} mem_T(x)$$

Solving this Problem Using MinCut

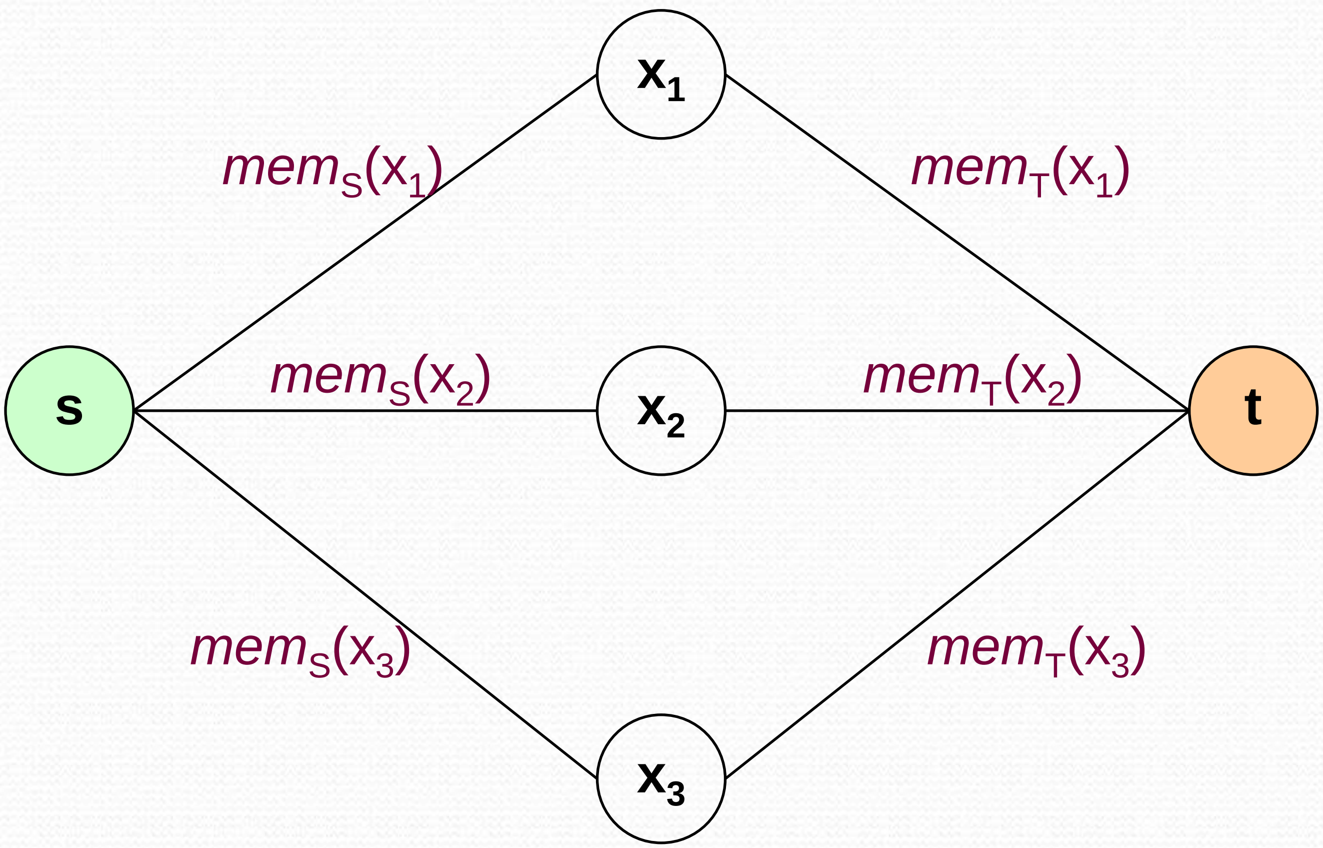
Solving this Problem Using MinCut



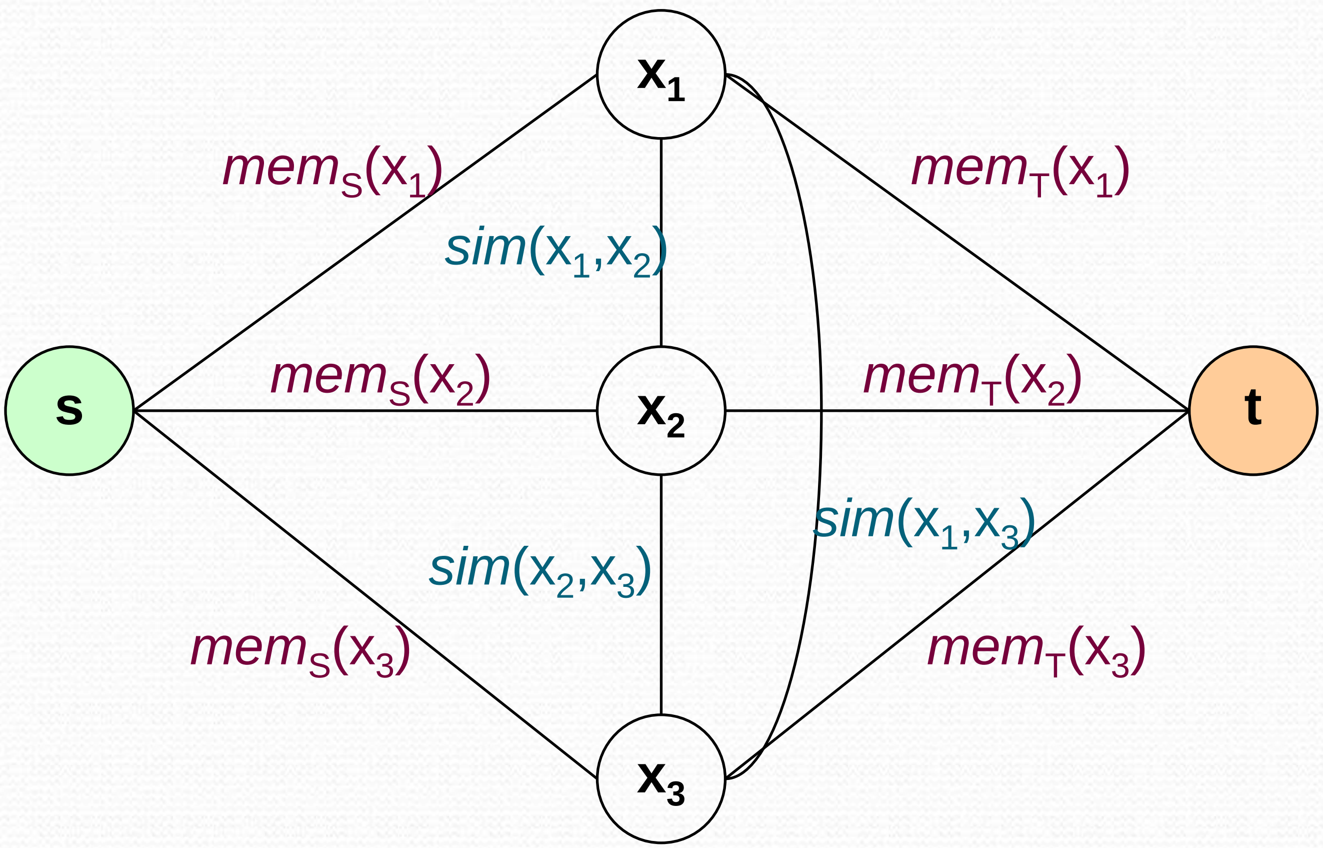
Solving this Problem Using MinCut



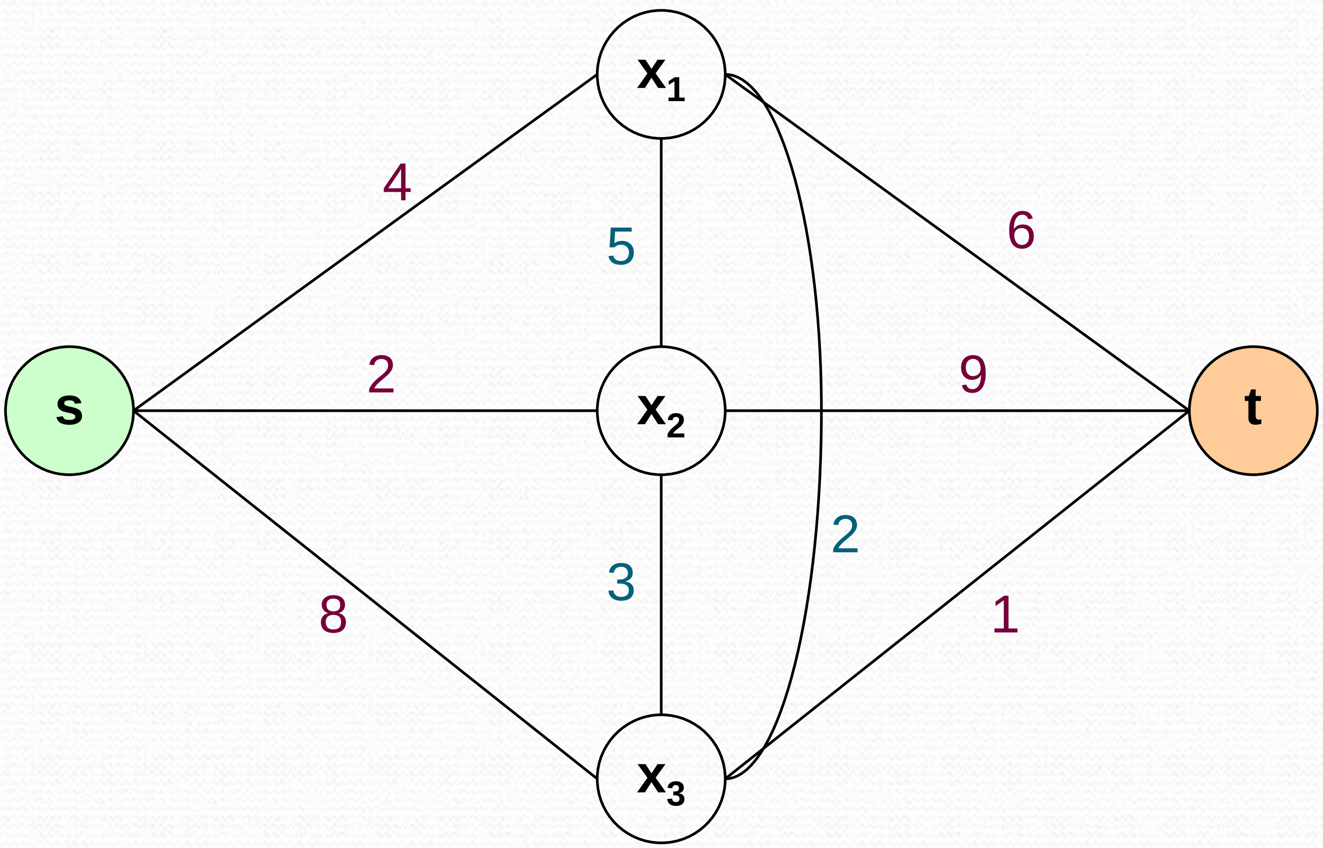
Solving this Problem Using MinCut



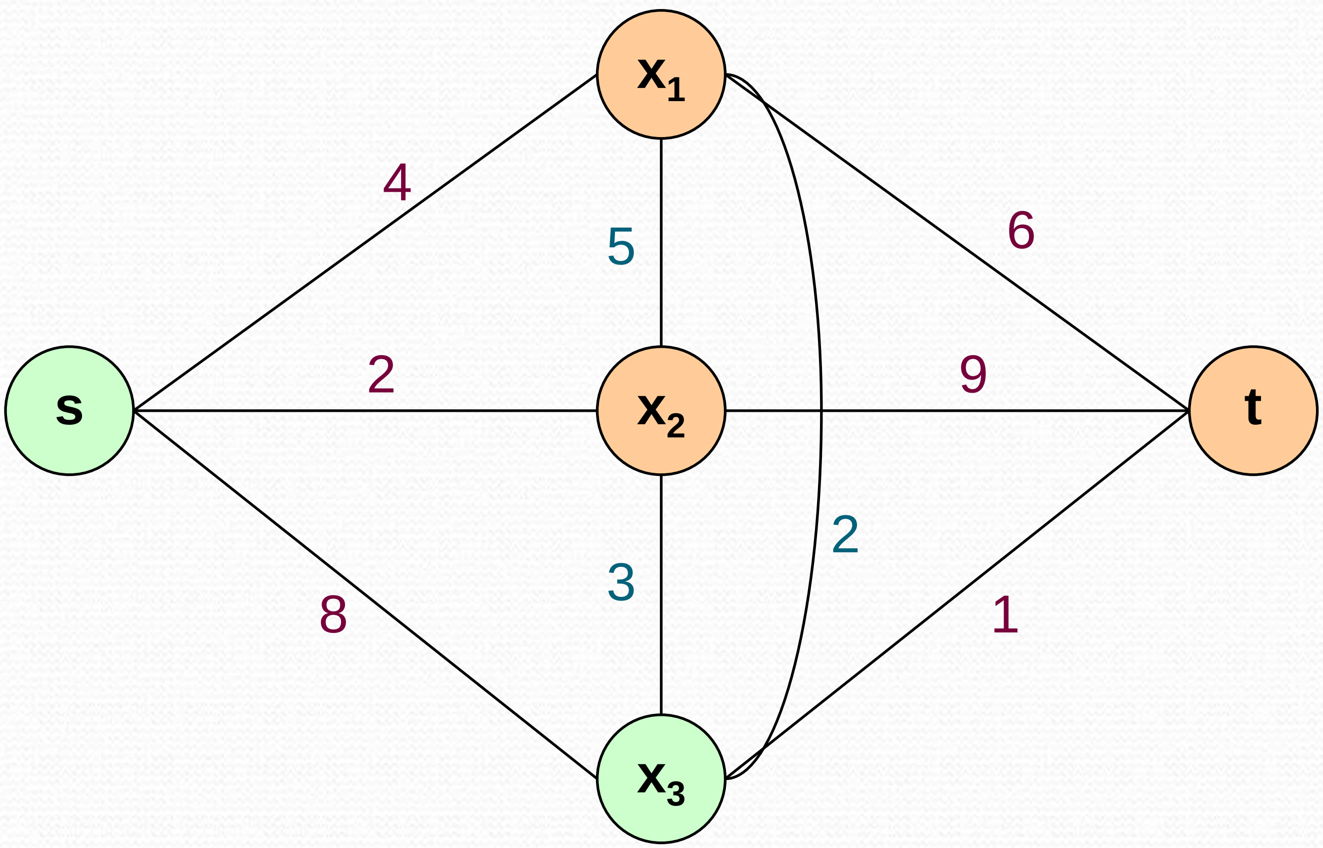
Solving this Problem Using MinCut



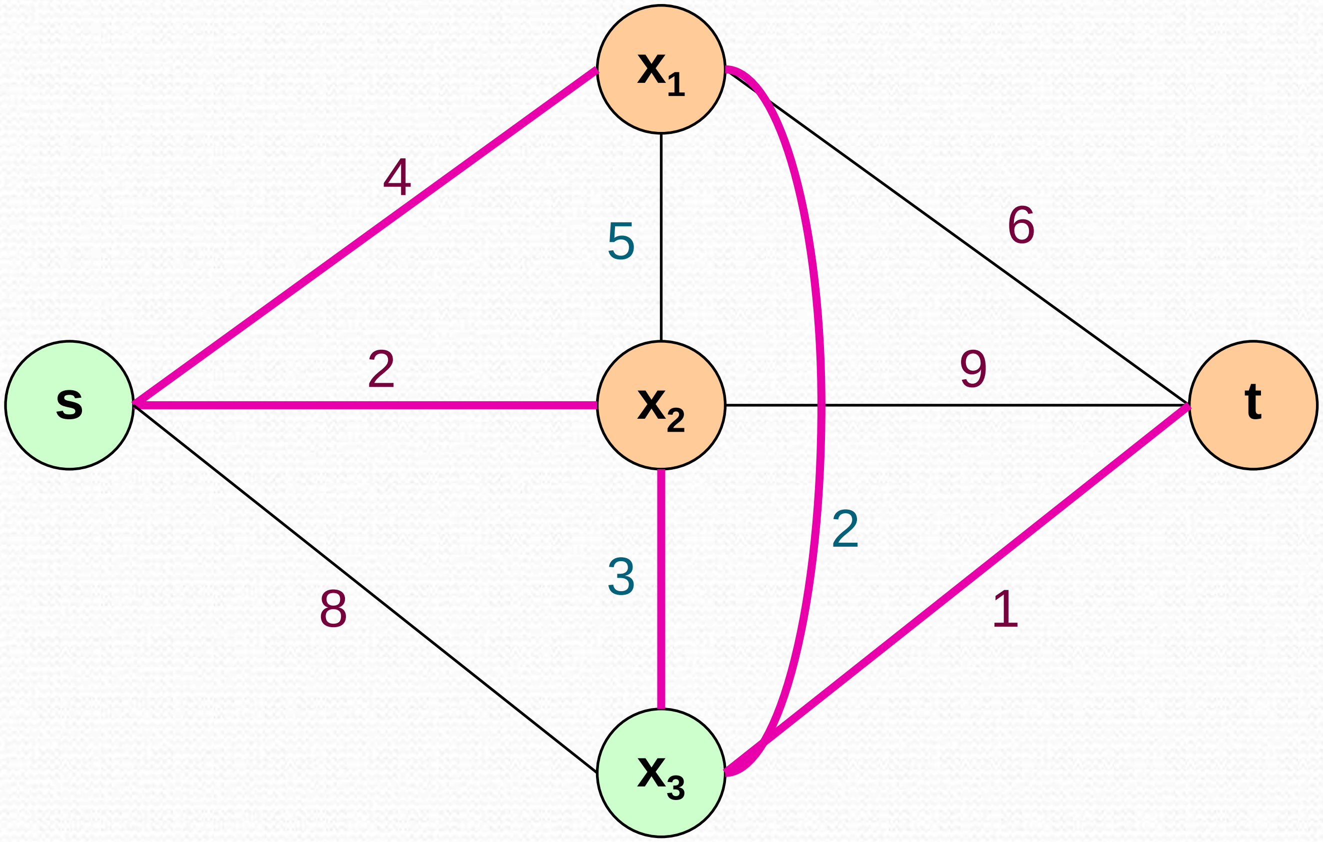
Solving this Problem Using MinCut



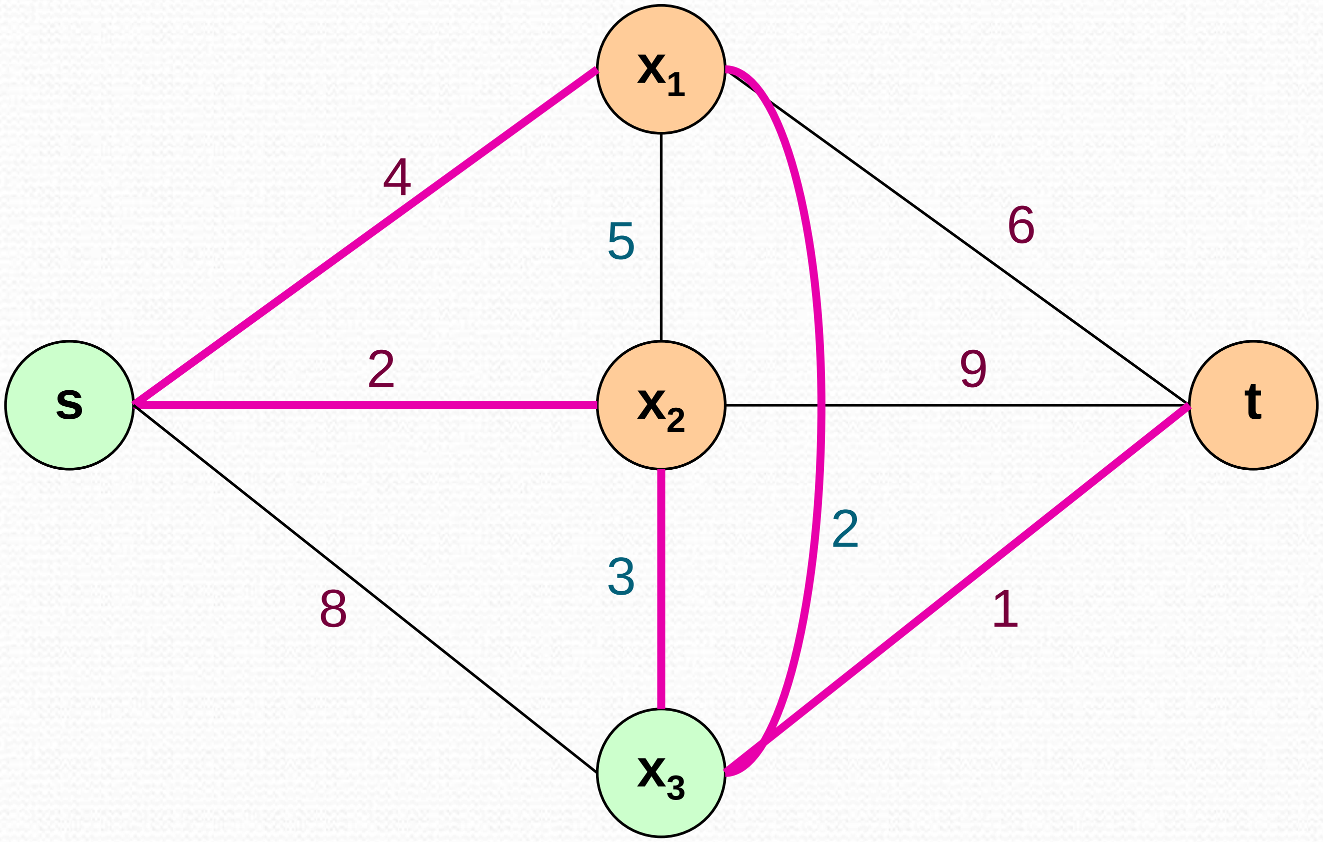
Solving this Problem Using MinCut



Solving this Problem Using MinCut



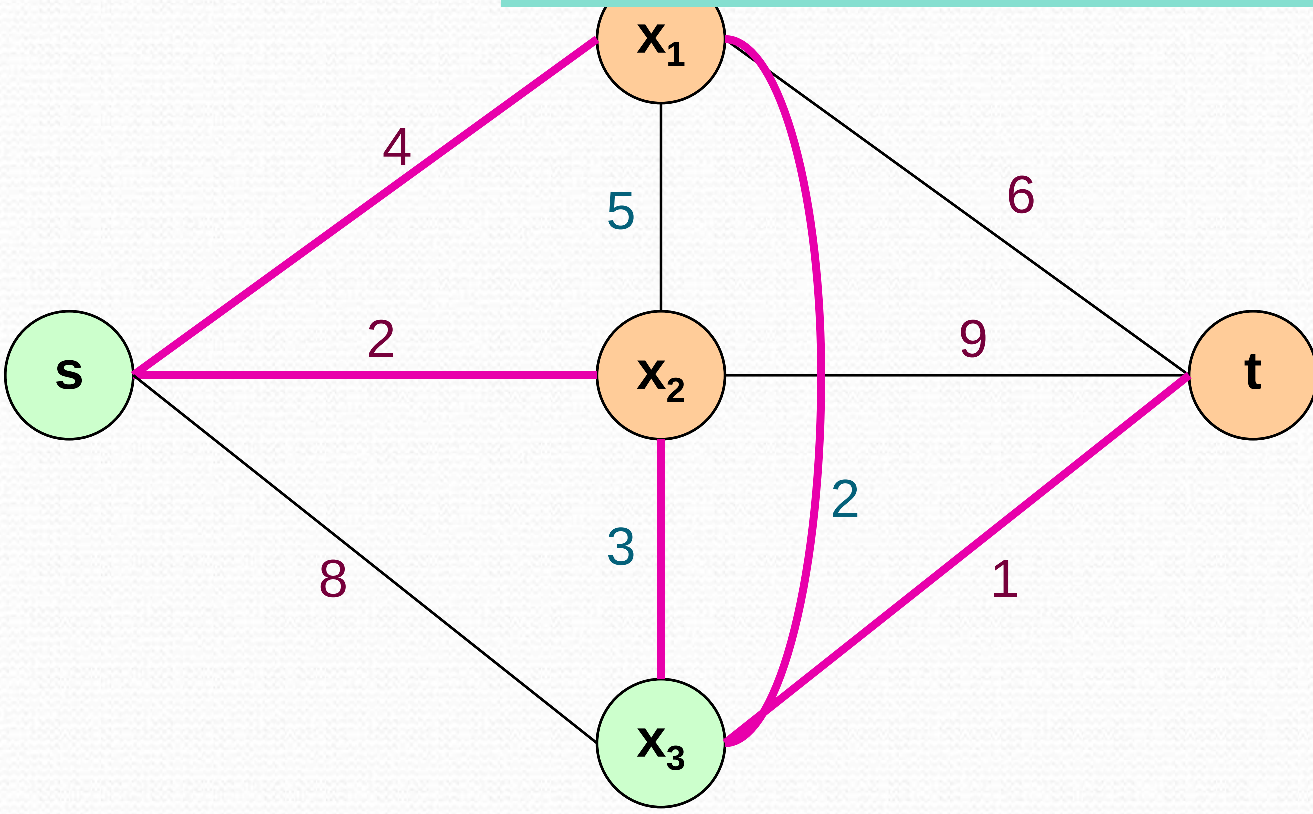
Solving this Problem Using MinCut



- Efficient algorithms for finding the mincut exist

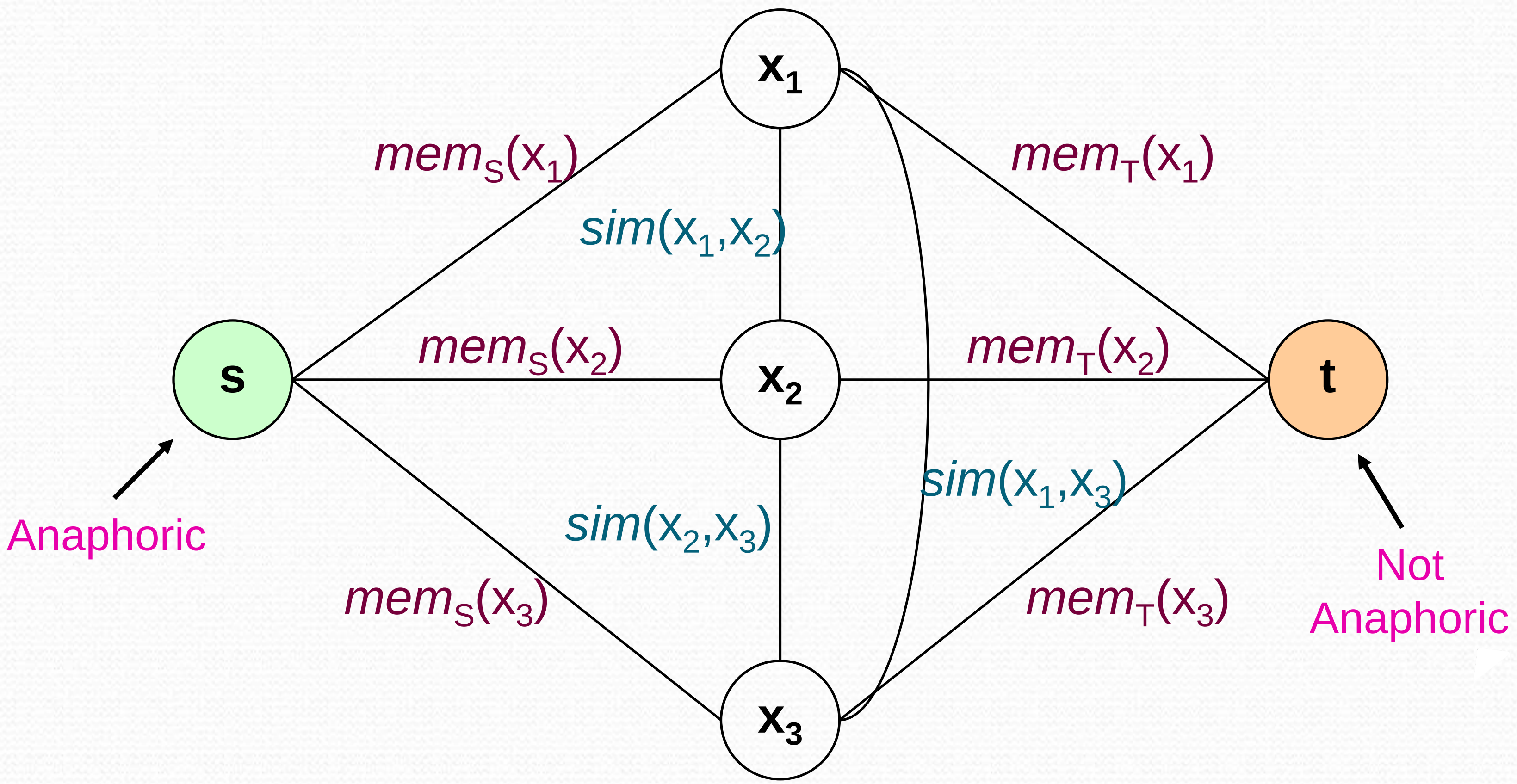
Solving this Problem Using MinCut

How to recast anaphoricity determination as a graph mincut problem?

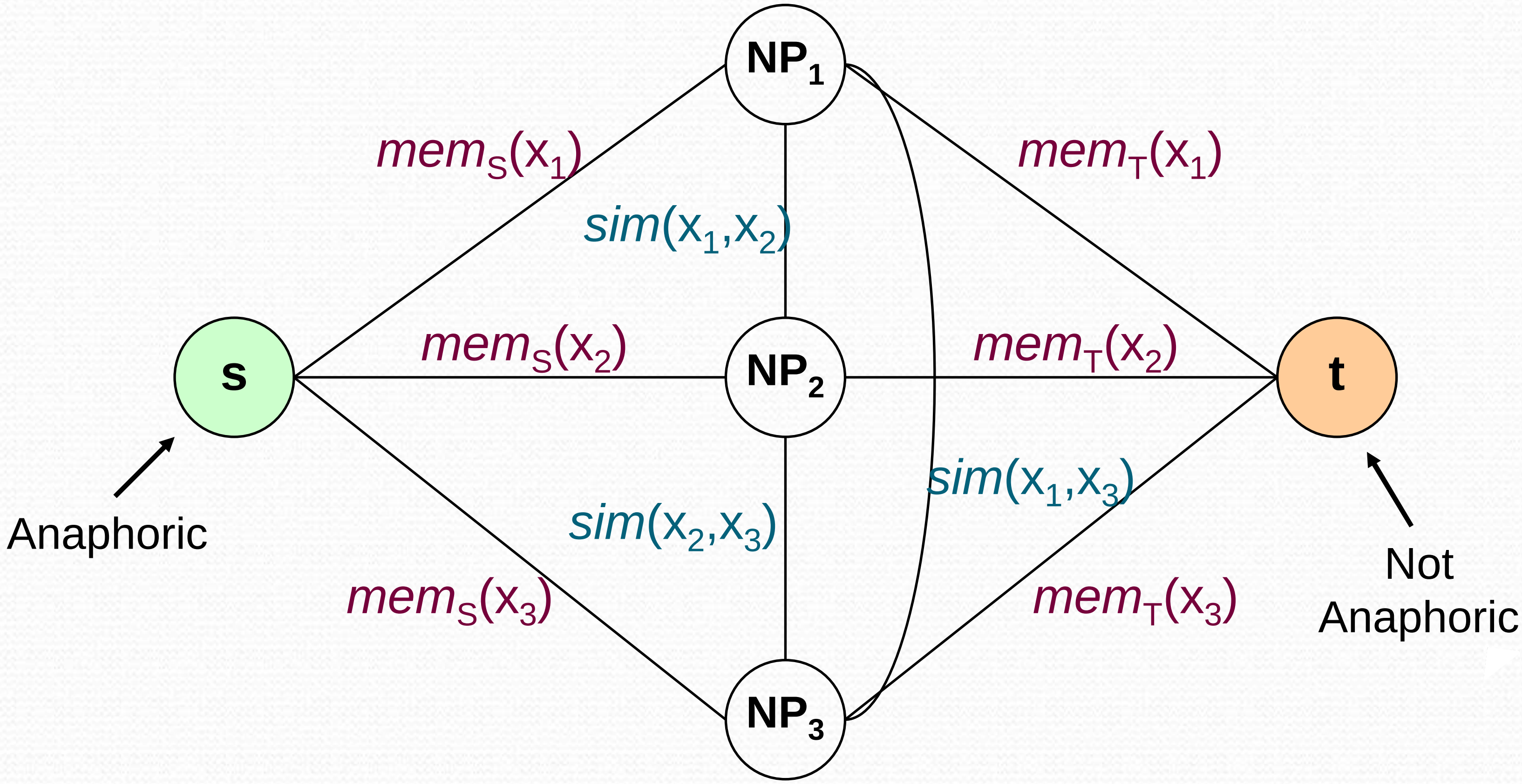


- Efficient algorithms for finding the mincut exist

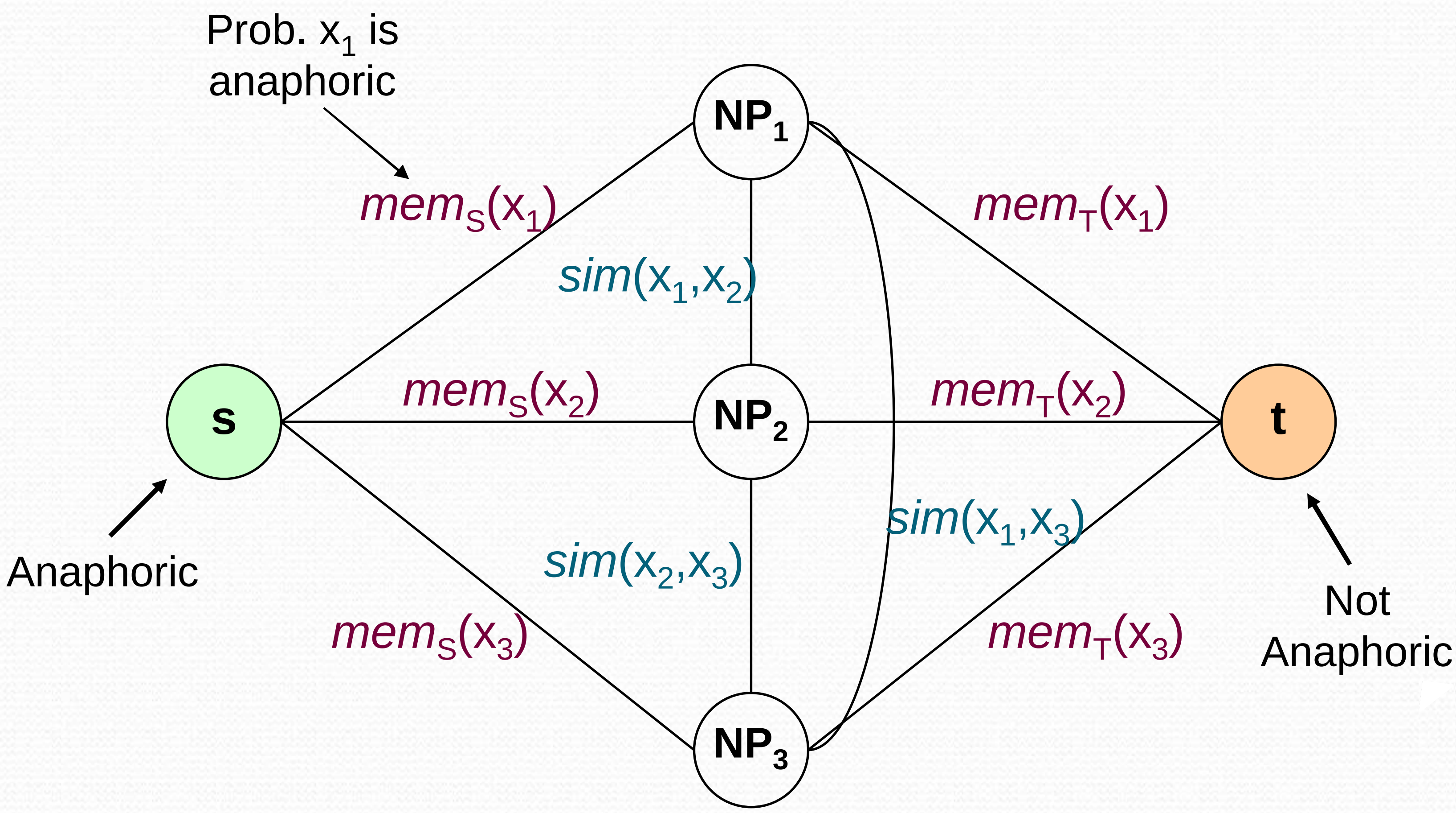
Cut-Based Anaphoricity Determination



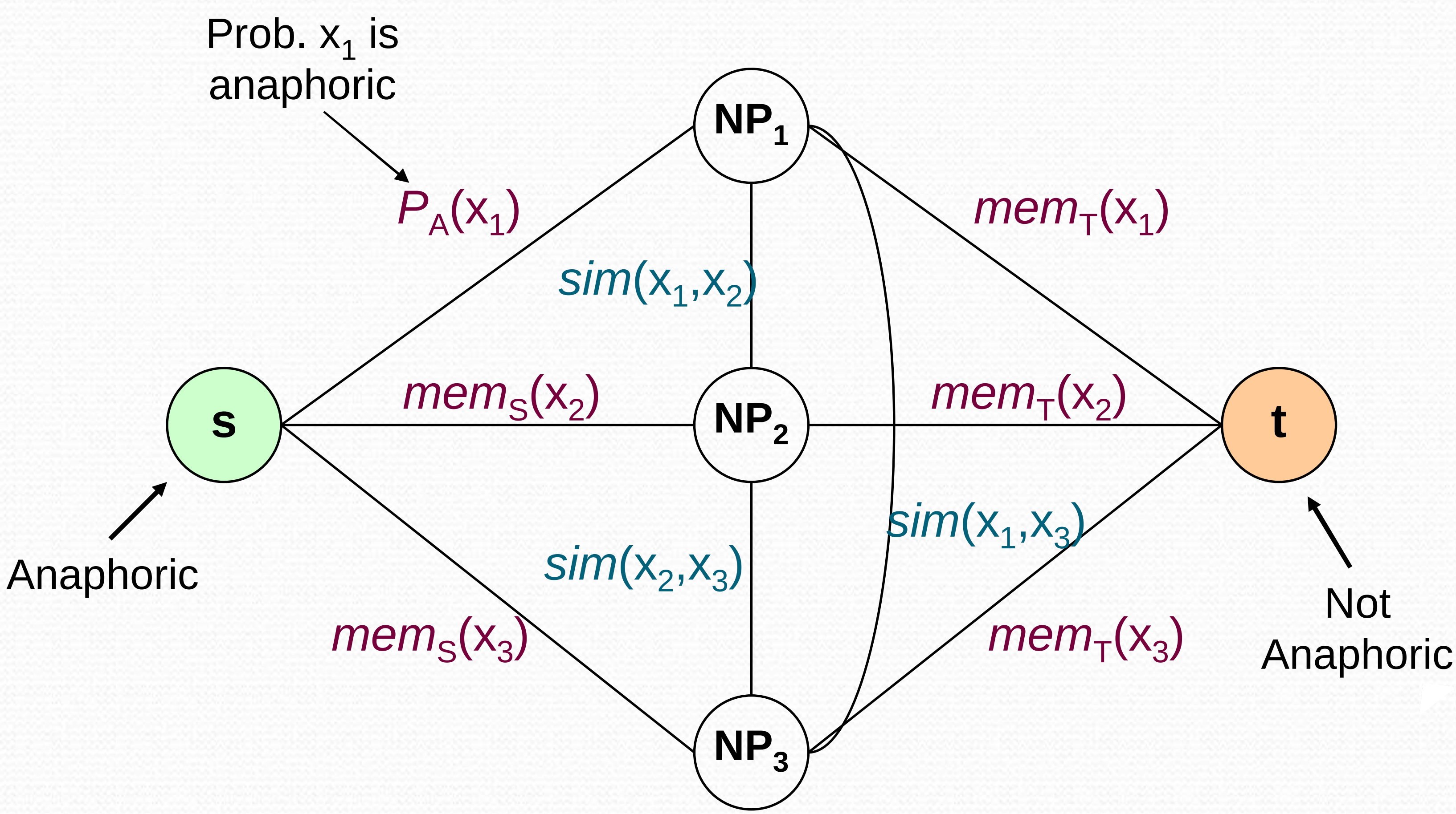
Cut-Based Anaphoricity Determination



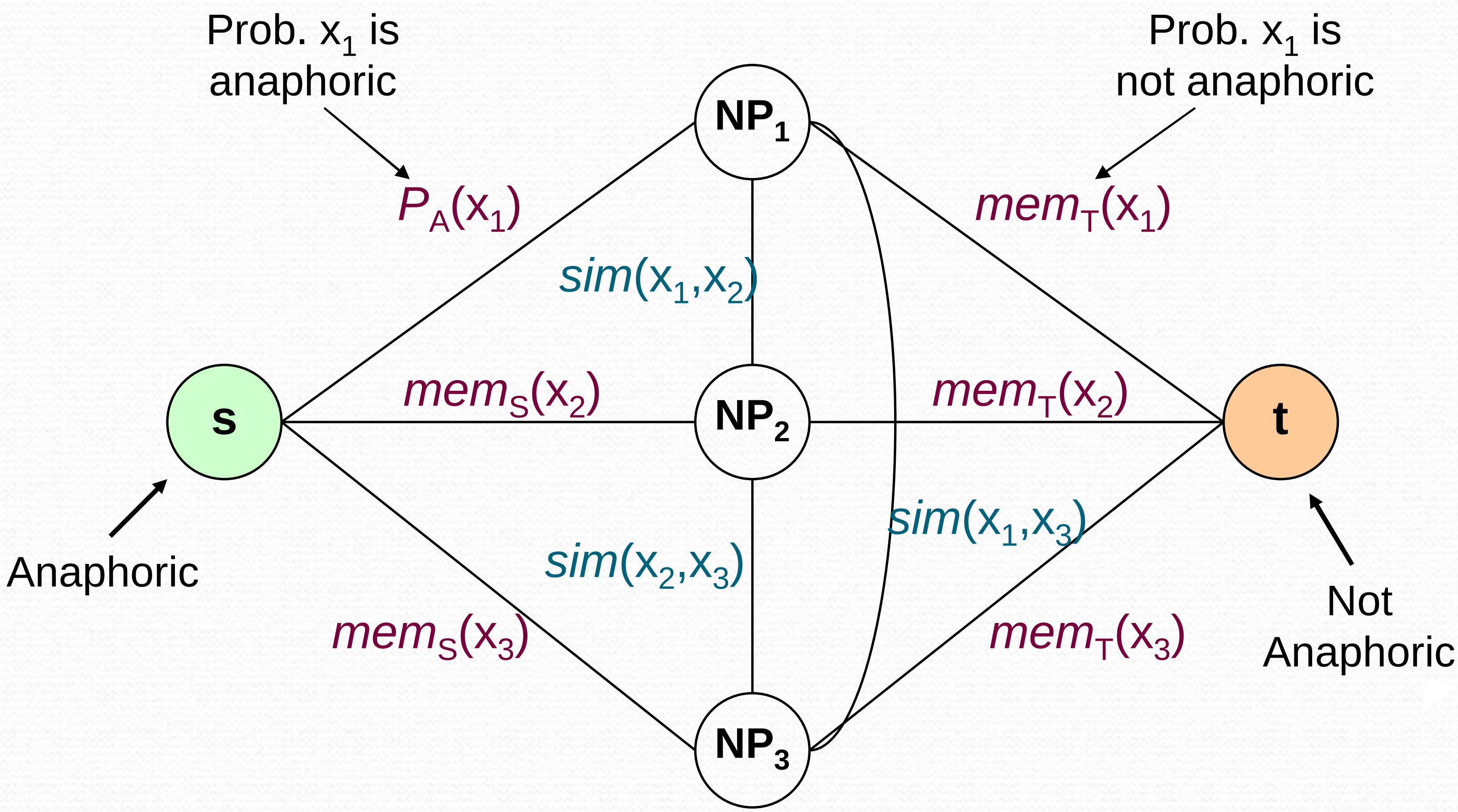
Cut-Based Anaphoricity Determination



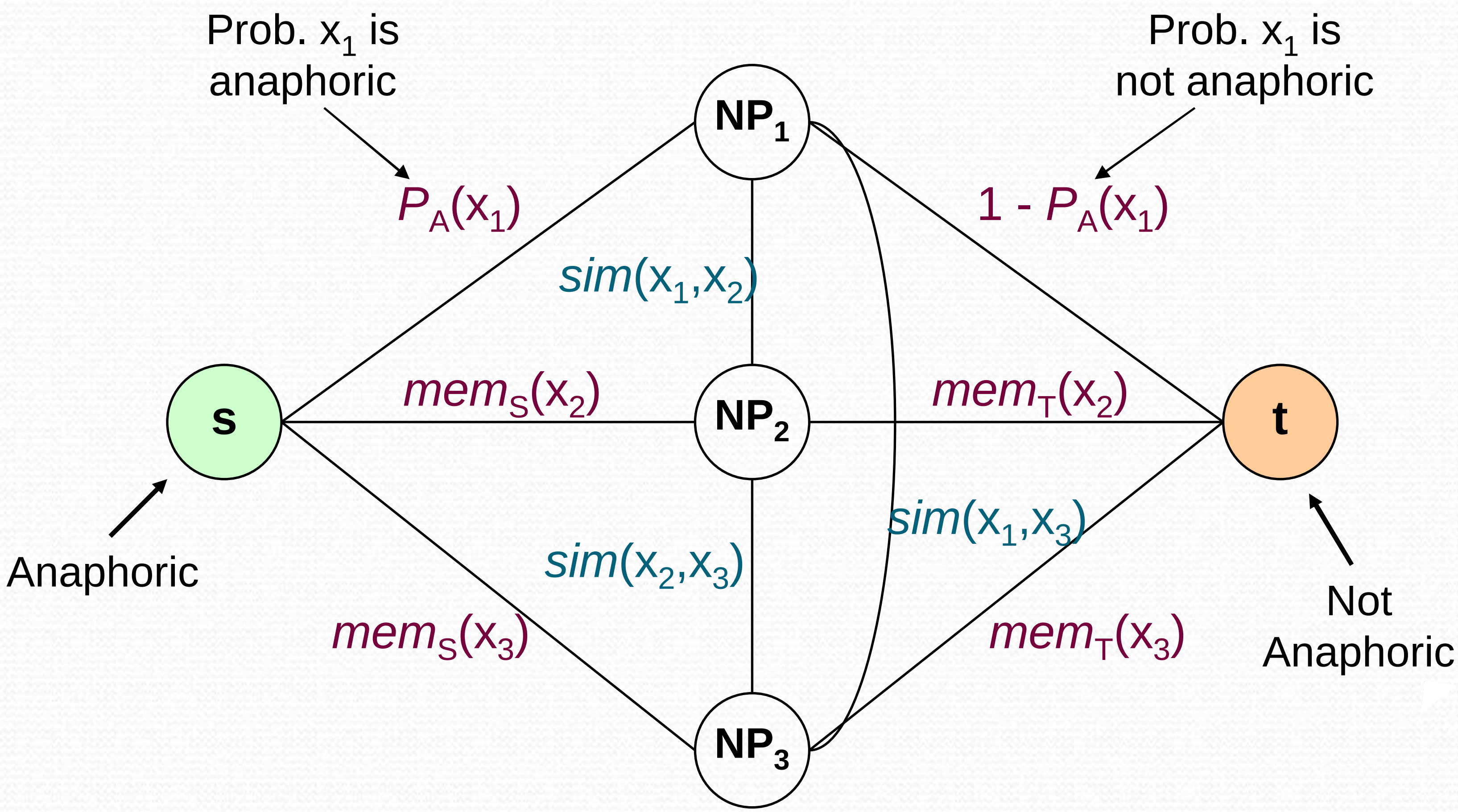
Cut-Based Anaphoricity Determination



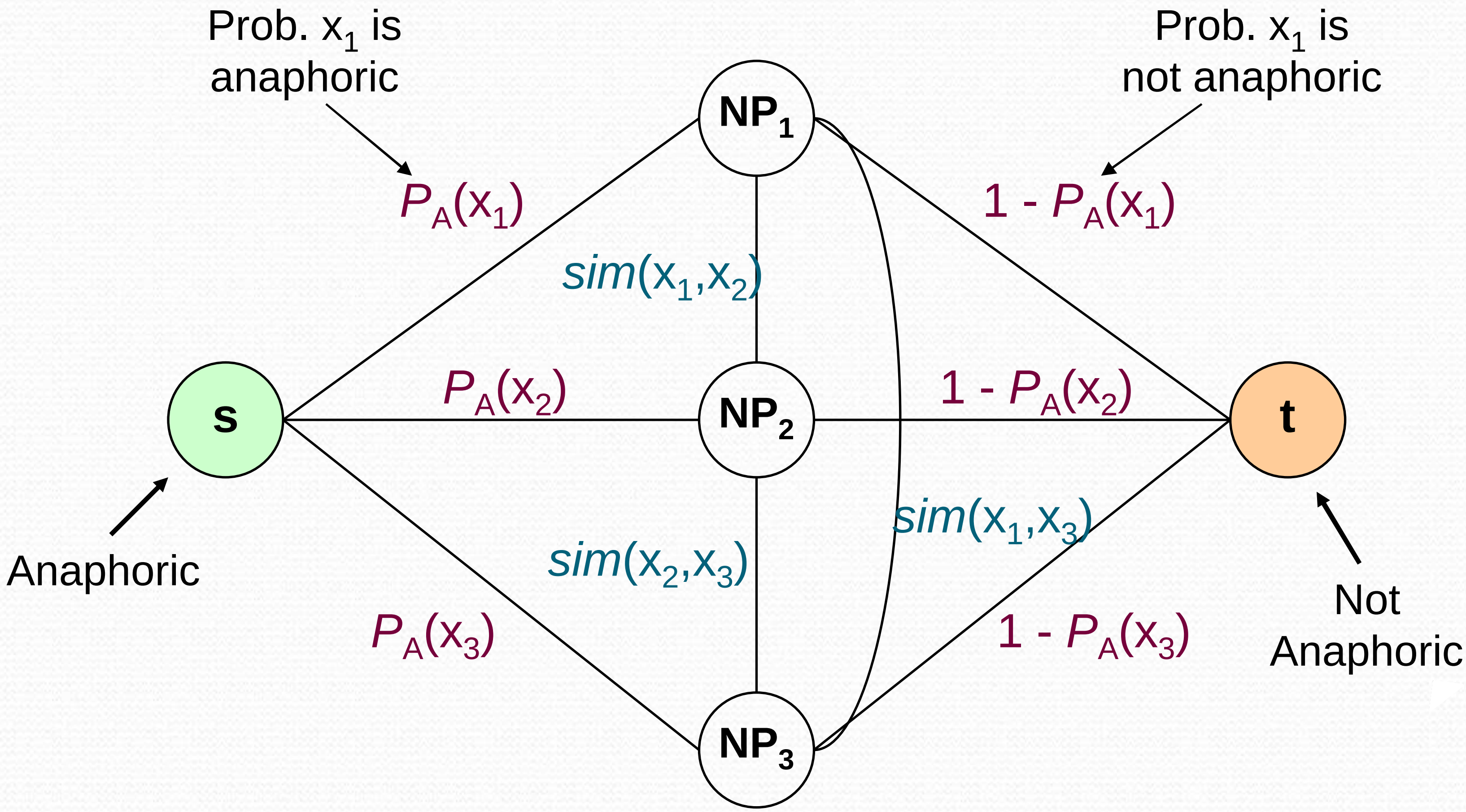
Cut-Based Anaphoricity Determination



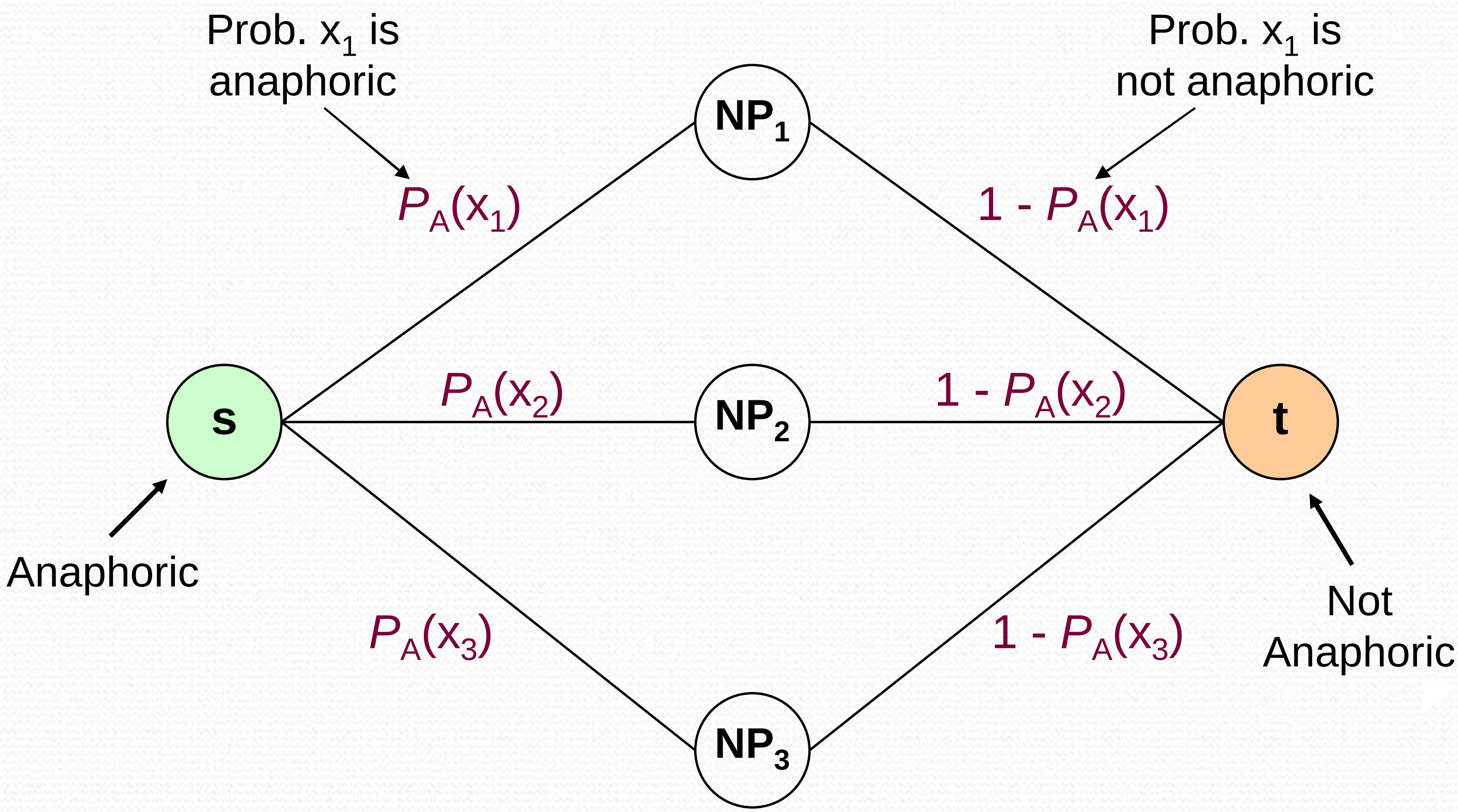
Cut-Based Anaphoricity Determination



Cut-Based Anaphoricity Determination

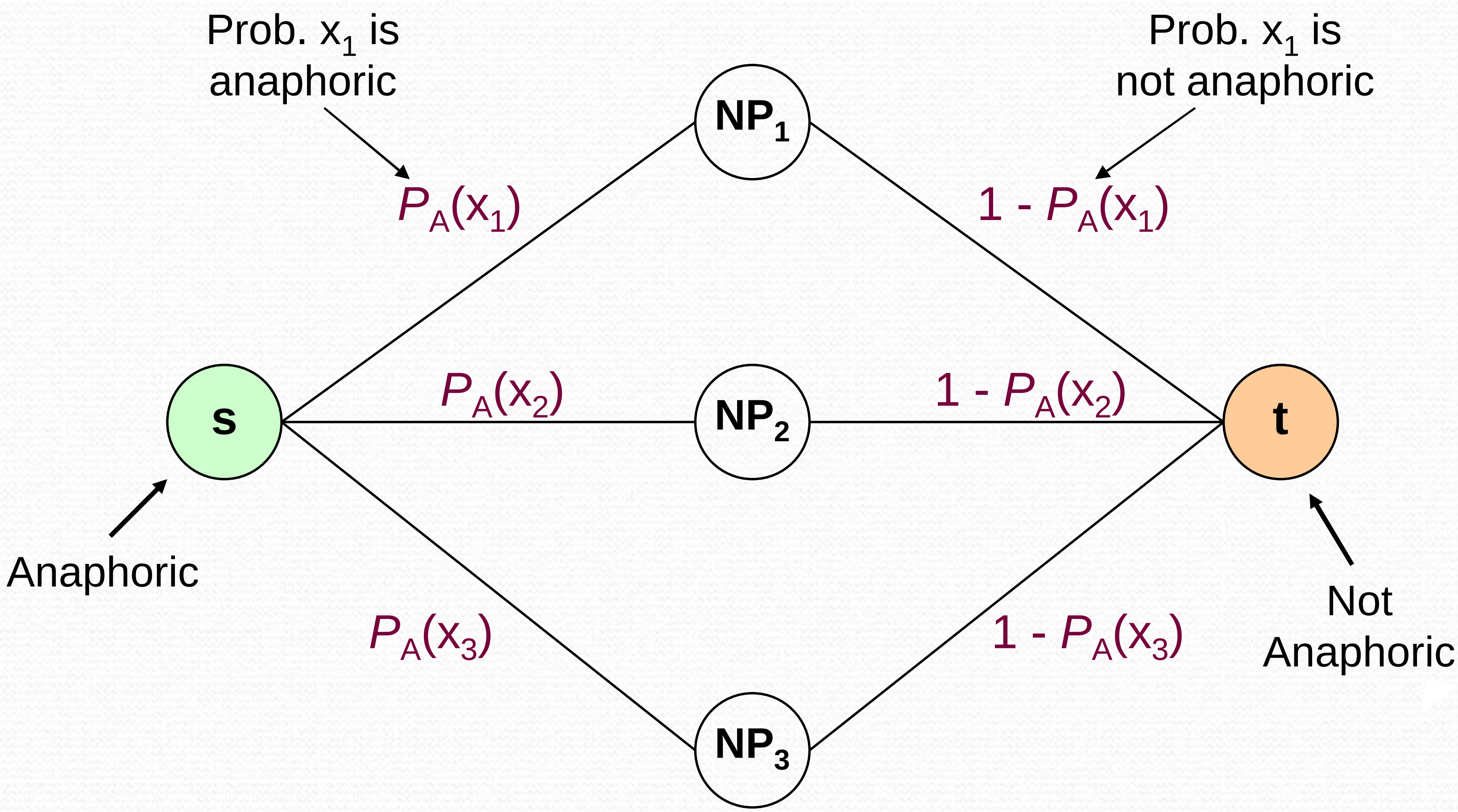


Cut-Based Anaphoricity Determination



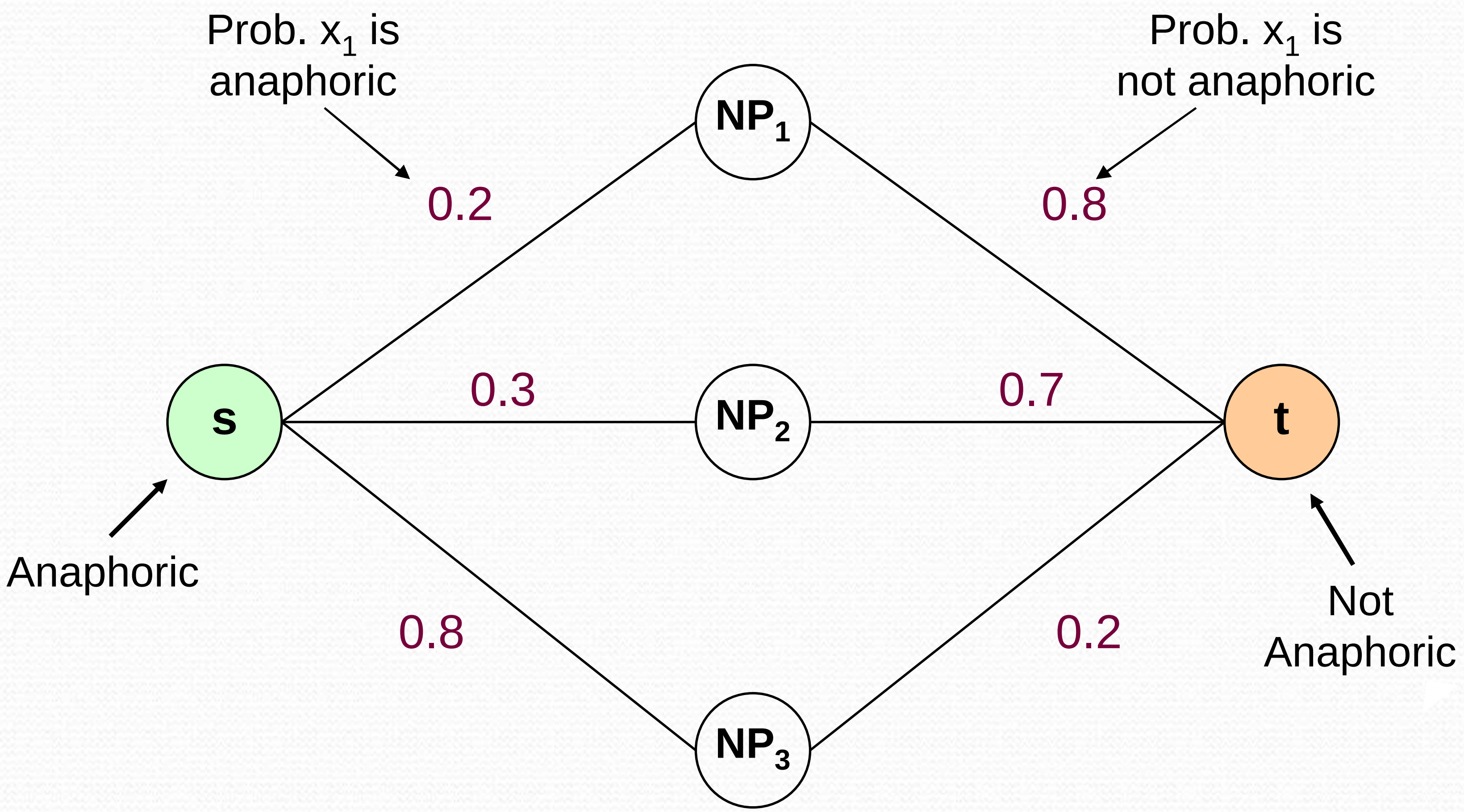
Cut-Based Anaphori

What if we run the mincut finding algorithm on this graph?



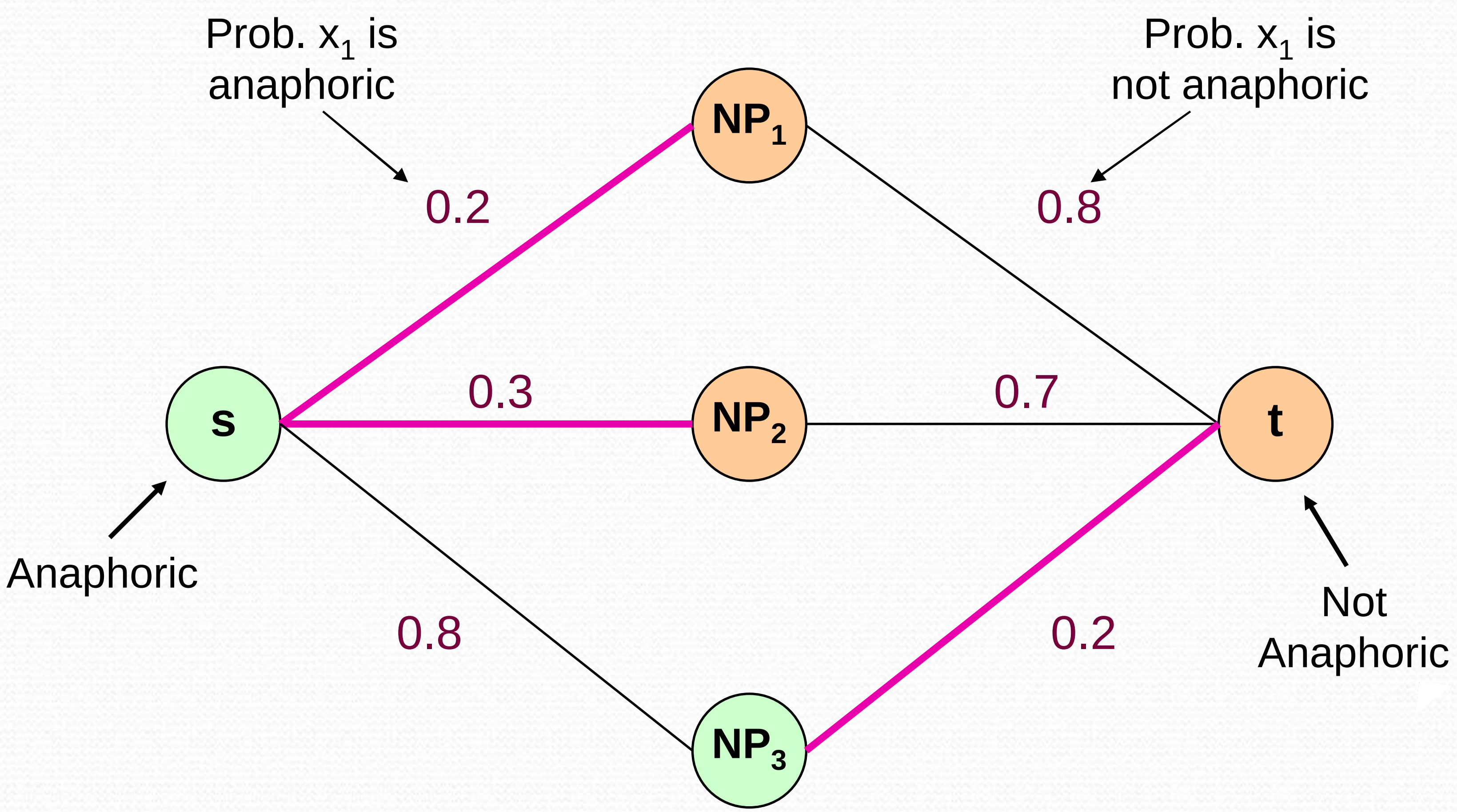
Cut-Based Anaphori

What if we run the mincut finding algorithm on this graph?



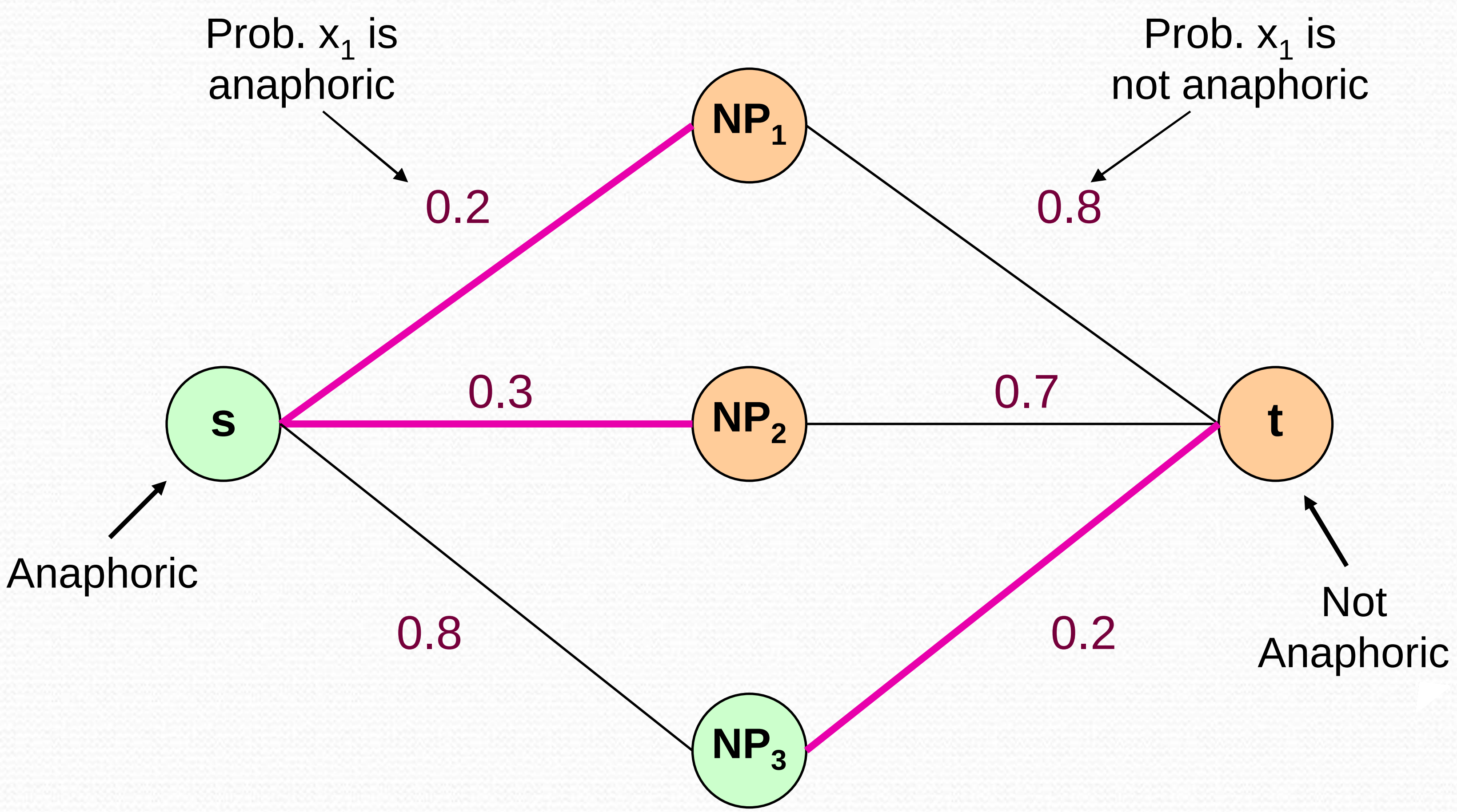
Cut-Based Anaphori

What if we run the mincut finding algorithm on this graph?



Cut-Based Anaphori

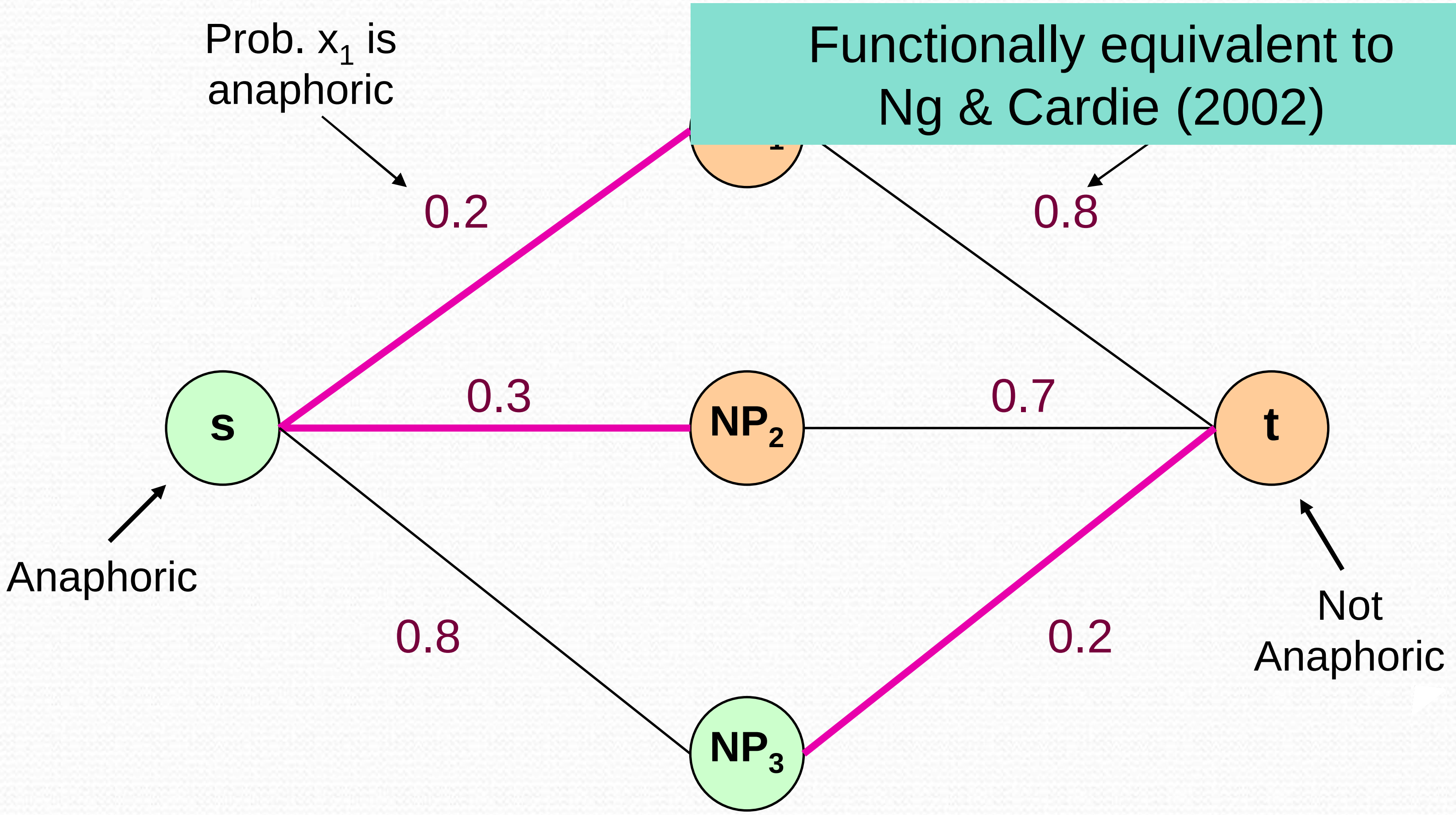
NP$_i$ will be assigned to the Anaphoric class iff $P_A(NP_i) \geq 0.5$



Cut-Based Anaphori

NP$_i$ will be assigned to the Anaphoric class iff $P_A(NP_i) \geq 0.5$

Functionally equivalent to Ng & Cardie (2002)

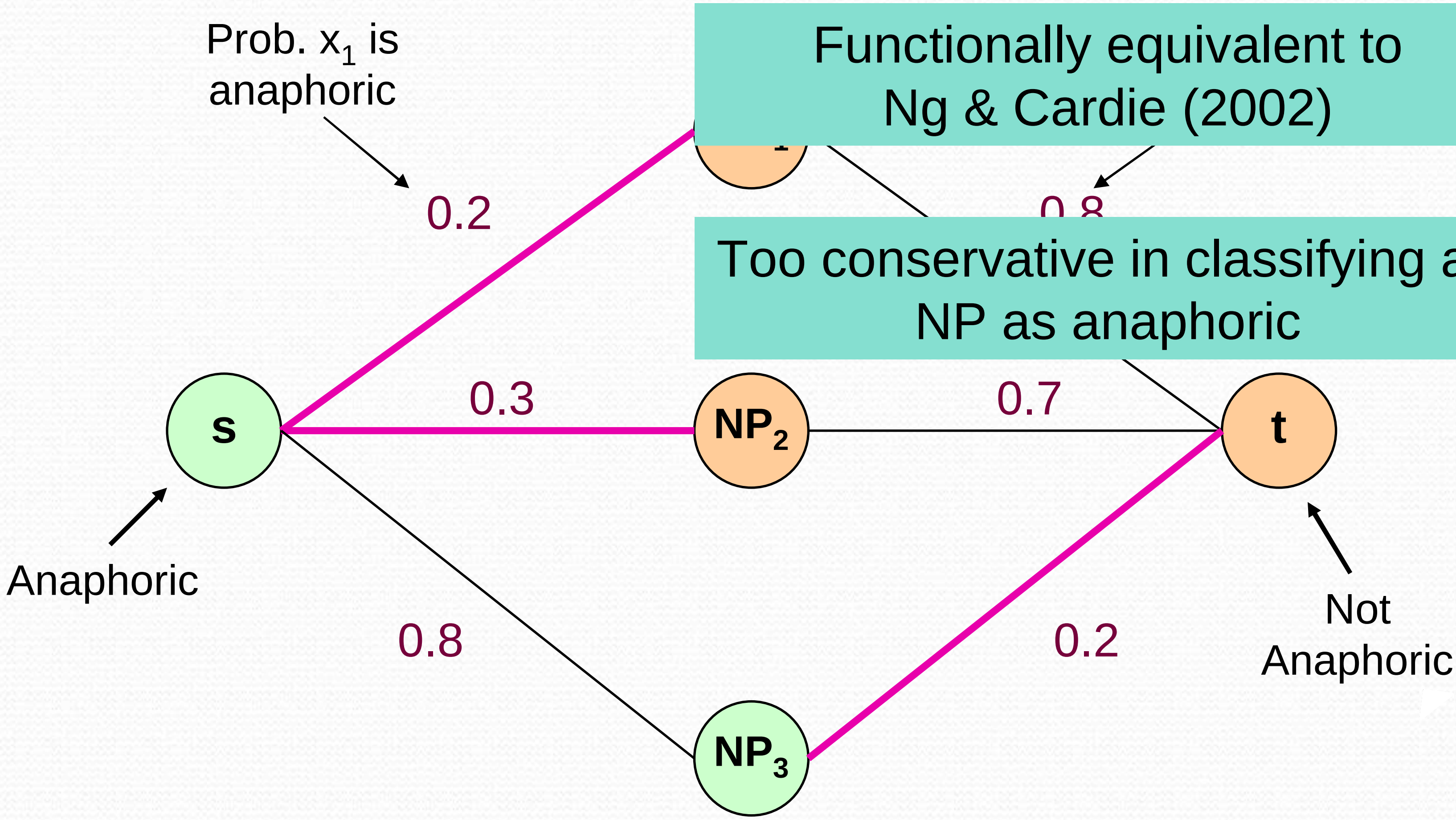


Cut-Based Anaphori

NP$_i$ will be assigned to the Anaphoric class iff $P_A(NP_i) \geq 0.5$

Functionally equivalent to Ng & Cardie (2002)

Too conservative in classifying an NP as anaphoric



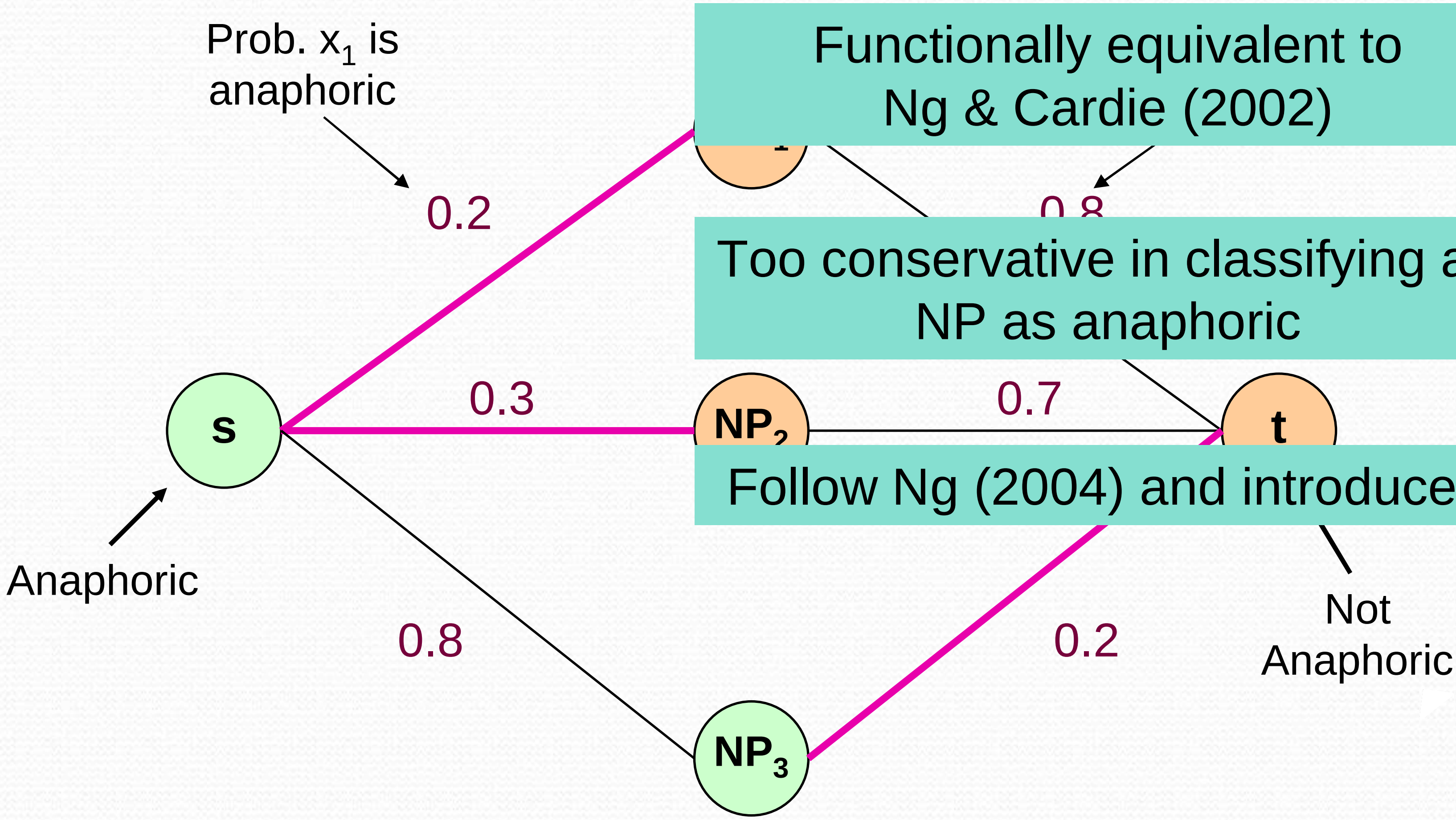
Cut-Based Anaphori

NP$_i$ will be assigned to the Anaphoric class iff $P_A(NP_i) \geq 0.5$

Functionally equivalent to Ng & Cardie (2002)

Too conservative in classifying an NP as anaphoric

Follow Ng (2004) and introduce ***t***



Mimicking Ng (2004)

- NP_i is classified as anaphoric iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$

Mimicking Ng (2004)

- NP_i is classified as anaphoric iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$
  - $\boldsymbol{t}$ is the "conservativeness" parameter
    - tuned on held-out data to maximize coreference F-measure

Mimicking Ng (2004)

- NP_i is classified as anaphoric iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$
  - $\boldsymbol{t}$ is the "conservativeness" parameter
    - tuned on held-out data to maximize coreference F-measure

- Goal: modify the edge weights s.t. NP_i will be assigned to the Anaphoric class iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$

Mimicking Ng (2004)

- NP_i is classified as anaphoric iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$
  - $\boldsymbol{t}$ is the "conservativeness" parameter
    - tuned on held-out data to maximize coreference F-measure

- Goal: modify the edge weights s.t. NP_i will be assigned to the Anaphoric class iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$

- How? Do a linear transformation

Mimicking Ng (2004)

- NP_i is classified as anaphoric iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$
  - $\boldsymbol{t}$ is the “conservativeness” parameter
    - tuned on held-out data to maximize coreference F-measure
- Goal: modify the edge weights s.t. NP_i will be assigned to the Anaphoric class iff $\mathbf{P_A}(NP_i) \geq \boldsymbol{t}$
- How? Do a linear transformation
- But … we are not happy with just mimicking Ng (2004)

Incorporating Similarity Scores

- What is a good candidate similarity function?

Incorporating Similarity Scores

- What is a good candidate similarity function?

- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)

Incorporating Similarity Scores

- What is a good candidate similarity function?
- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)
    - want mincut finder to assign the NPs to the same set

Incorporating Similarity Scores

- What is a good candidate similarity function?
- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)
    - want mincut finder to assign the NPs to the same set
  - If NP_i and NP_j are unlikely to be coreferent according to $\mathbf{P_C}$, it's hard to claim anything regarding their anaphoricity

Incorporating Similarity Scores

- What is a good candidate similarity function?
- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)
    - want mincut finder to assign the NPs to the same set
  - If NP_i and NP_j are unlikely to be coreferent according to $\mathbf{P_C}$, it's hard to claim anything regarding their anaphoricity
- Use $\mathbf{P_C}(NP_i,NP_j)$ as the $sim(NP_i,NP_j)$, but only if $\mathbf{P_C}(NP_i,NP_j) > \boldsymbol{t_2}$

Incorporating Similarity Scores

- What is a good candidate similarity function?
- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)
    - want mincut finder to assign the NPs to the same set
  - If NP_i and NP_j are unlikely to be coreferent according to $\mathbf{P_C}$, it's hard to claim anything regarding their anaphoricity
- Use $\mathbf{P_C}(NP_i,NP_j)$ as the $sim(NP_i,NP_j)$, but only if $\mathbf{P_C}(NP_i,NP_j) > \boldsymbol{t_2}$ [tuned jointly with $\boldsymbol{t}$ on held-out data]

Incorporating Similarity Scores

- What is a good candidate similarity function?
- Two observations:
  - If NP_i and NP_j are likely to be coreferent according to $\mathbf{P_C}$, then NP_i and NP_j are likely to be both anaphoric (except if NP_i is the head of a coreference chain)
    - want mincut finder to assign the NPs to the same set
  - If NP_i and NP_j are unlikely to be coreferent according to $\mathbf{P_C}$, it's hard to claim anything regarding their anaphoricity
- Use $\mathbf{P_C}(NP_i,NP_j)$ as the $sim(NP_i,NP_j)$, but only if $\mathbf{P_C}(NP_i,NP_j) > \boldsymbol{t_2}$ [tuned jointly with ***t*** on held-out data]
- Otherwise, set $sim(NP_i, NP_j)$ to 0

Cut-Based Anaphoricity Determination

- An anaphoricity determination method that can
  - maximize the desired coreference metric (by tuning $\boldsymbol{t}$ and $\boldsymbol{t}_2$)
  - exploit probabilities provided by $\mathbf{P_C}$ (via the *sim* function)

Plan for the Talk

- Existing methods for computing and using anaphoricity info
- Our graph-cut-based approach to anaphoricity determination
- Evaluation

Evaluation: Goal

- compare our cut-based method for anaphoricity determination with existing methods w.r.t. their effectiveness in improving a learning-based coreference system

Experimental Setup

- Coreference system [Ng, 2007]
  - implements the standard machine learning framework
  - 34 features per instance
- Features for anaphoricity determination [Ng & Cardie, 2002]
  - 37 features per instance
- Learning algorithm
  - Maximum entropy for training $\mathbf{P_A}$ and $\mathbf{P_C}$

Experimental Setup (Cont')

- The ACE coreference corpus
  - 3 data sets (Broadcast News, Newspaper, Newswire)
  - each data set comprises a training set and a test set
- NPs extracted automatically
- Scoring programs
  - MUC (Vilain et al., 1995) and CEAF (Luo, 2005)
    - recall, precision, F-measure

CEAF Results: “No Anaphoricity” Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**

CEAF Results: “No Anaphoricity” Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**

CEAF Results: “No Anaphoricity” Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**

CEAF Results: “No Anaphoricity” Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**

CEAF Results: Ng & Cardie (2002) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**

CEAF Results: Ng & Cardie (2002) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**

- F-measure drops slightly in all cases
  - large drops in recall accompanied by smaller gains in precision
  - many anaphoric NPs were misclassified

CEAF Results: Ng (2004) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**

CEAF Results: Ng (2004) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**

- requires tuning of ***t***
  - reserve 1/3 of the training data for parameter tuning
  - train $\mathbf{P}_A$ and $\mathbf{P}_C$ on remaining 2/3 of the training data

CEAF Results: Ng (2004) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**

- requires tuning of ***t***
  - reserve 1/3 of the training data for parameter tuning
  - train $\mathbf{P_A}$ and $\mathbf{P_C}$ on remaining 2/3 of the training data
- mixed results in comparison to “No Anaphoricity” baseline
  - F-measure gains by 0.2% for BNEWS, 1.1% for NPAPER, but drops by 0.6% for NWIRE

CEAF Results: Luo (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**

CEAF Results: Luo (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**

- in comparison to "No Anaphoricity" baseline
  - F-measure improves insignificantly (by 0.3-1.0%)

Results: Denis & Baldridge (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**

Results: Denis & Baldridge (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**

- mixed results in comparison to “No Anaphoricity” baseline
  - F-measure rises significantly for BNEWS, drop insignificantly for NPAPER, and rises insignificantly for NWIRE

CEAF Results: Kleener (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**

CEAF Results: Kleener (2007) Baseline

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**

- in comparison to Denis & Baldridge baseline,
  - F-measure never improves, recall slightly deteriorates
  - transitivity constraints tend to produce smaller clusters
  - enforcing transitivity does not improve coreference performance

CEAF Results: Graph Minimum Cut

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**
Graph Minimum Cut	61.4	57.6	**59.4**	64.1	59.4	**61.7**	65.7	61.9	**63.8**

- 1/3 of training data for joint tuning of t and t_2; 2/3 for training $\mathbf{P_A}$ and $\mathbf{P_C}$

CEAF Results: Graph Minimum Cut

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**
Graph Minimum Cut	61.4	57.6	**59.4**	64.1	59.4	**61.7**	65.7	61.9	**63.8**

- 1/3 of training data for joint tuning of t and t_2; 2/3 for training $\mathbf{P_A}$ and $\mathbf{P_C}$
- significant improvement over "No Anaphoricity" baseline
  - large gains in precision and smaller drops in recall

CEAF Results: Graph Minimum Cut

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**
Graph Minimum Cut	61.4	57.6	**59.4**	64.1	59.4	**61.7**	65.7	61.9	**63.8**

- 1/3 of training data for joint tuning of t and t_2; 2/3 for training $\mathbf{P_A}$ and $\mathbf{P_C}$
- significant improvement over "No Anaphoricity" baseline
  - large gains in precision and smaller drops in recall
- significant improvement over best baseline (D&B)

CEAF Results: Graph Minimum Cut

	Broadcast News			Newspaper			Newswire		
	R	P	F	R	P	F	R	P	F
No anaphoricity	63.2	49.2	**55.3**	64.5	54.3	**59.0**	67.3	56.1	**61.2**
Ng & Cardie (2002)	55.9	53.3	**54.5**	60.7	56.3	**58.3**	60.6	58.2	**59.4**
Ng (2004)	62.5	49.9	**55.5**	63.5	57.0	**61.0**	65.6	56.3	**60.6**
Luo (2007)	62.7	51.1	**56.3**	64.6	55.4	**59.6**	67.0	56.8	**61.5**
Denis & Baldridge (2007)	63.8	51.4	**56.9**	62.6	53.6	**57.8**	67.0	56.8	**61.5**
Kleener (2007)	63.2	51.3	**56.7**	62.6	53.6	**57.8**	66.7	56.7	**61.3**
Graph Minimum Cut	61.4	57.6	**59.4**	64.1	59.4	**61.7**	65.7	61.9	**63.8**

- 1/3 of training data for joint tuning of t and t_2; 2/3 for training $\mathbf{P_A}$ and $\mathbf{P_C}$
- significant improvement over “No Anaphoricity” baseline
  - large gains in precision and smaller drops in recall
- significant improvement over best baseline (D&B)
- best F-measure score achieved for each dataset

Summary

- Proposed a graph-cut-based approach to anaphoricity determination that
  - directly optimizes the desired coreference evaluation metric
  - exploits the probabilities provided by the coreference model
  - achieves the best results on all three ACE datasets according to both the MUC scorer and the CEAF scorer
  - provides a flexible mechanism for co-ordinating anaphoricity and coreference decisions