Large Scale Cross-Document Coreference Resolution

Kambiz Ghoorchian
Sarunas Girdzijauskas

ghoorian@kth.se
02/06/2014
LCN - KTH

Overview

Problem

Cross-Document Coreference Resolution

Overview

Problem

Cross-Document Coreference Resolution

Current Solution

Vector Space Modeling

Limitations

Overview

Problem

Cross-Document Coreference Resolution

Current Solution

Vector Space Modeling

Limitations

Contribution

Graph Based Modeling

Diffusion Based Clustering

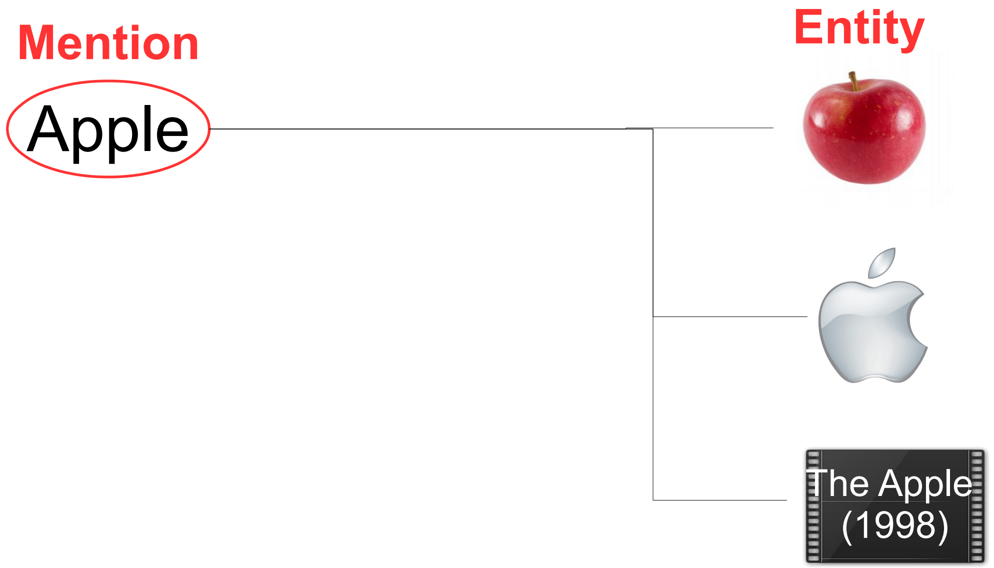
Definition

Entity

Entity

Distinctive and independent thing in the real world.

Definition



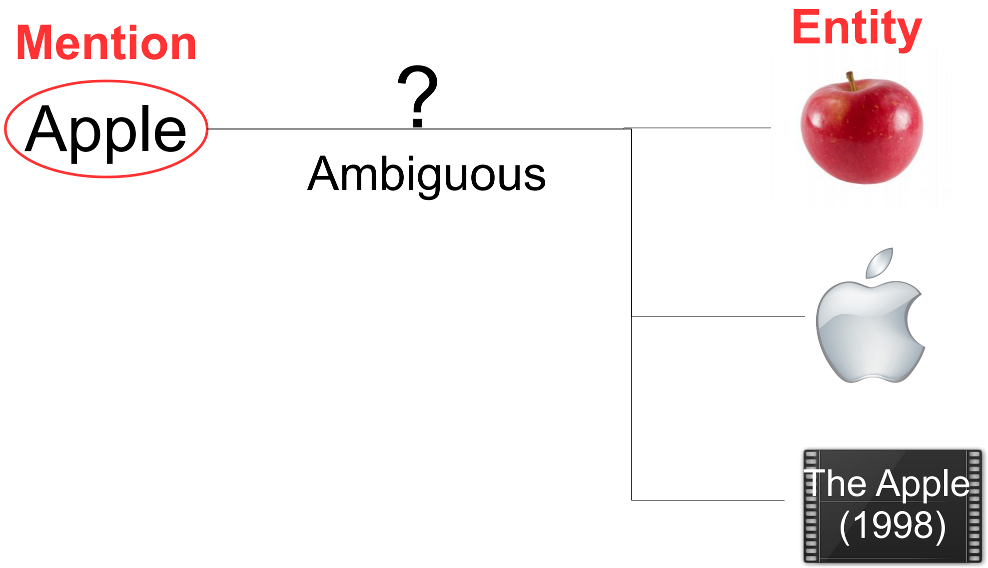
Entity

Distinctive and independent thing in the real world.

Mention

Linguistic phenomenon (word or phrase) refers to an entity.

Definition



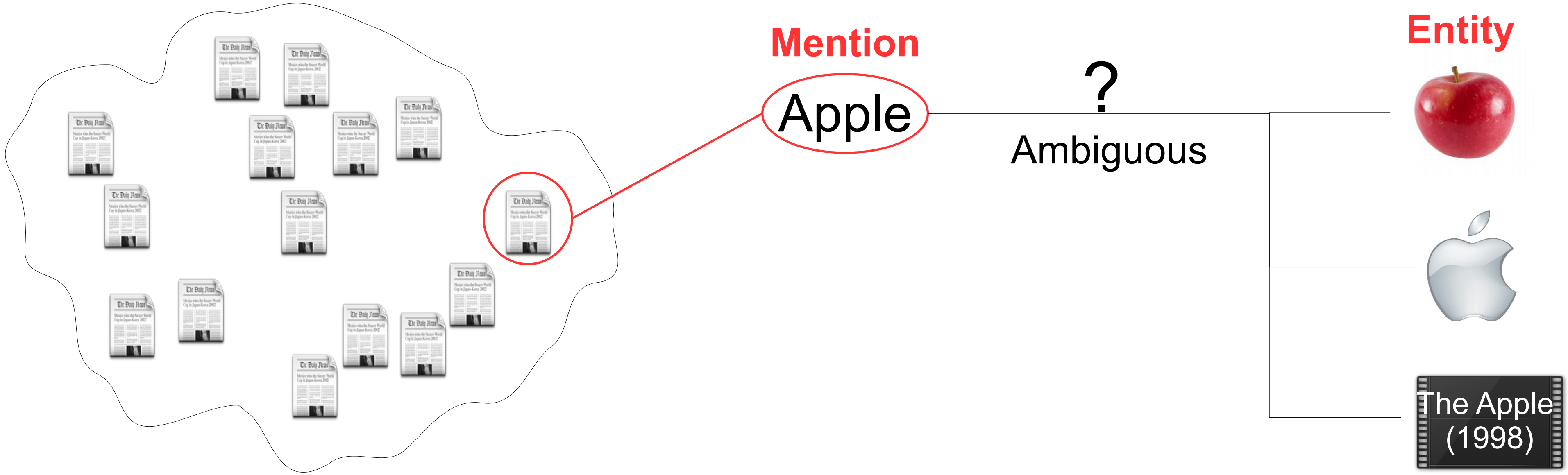
Entity

Distinctive and independent thing in the real world.

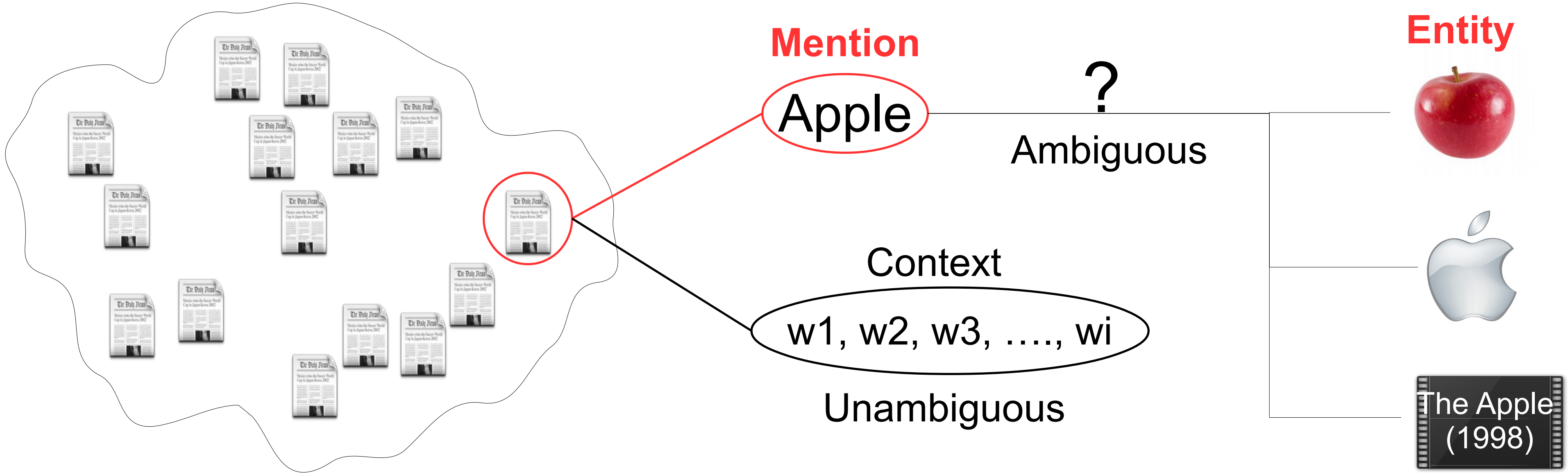
Mention

Linguistic phenomenon (word or phrase) refers to an entity.

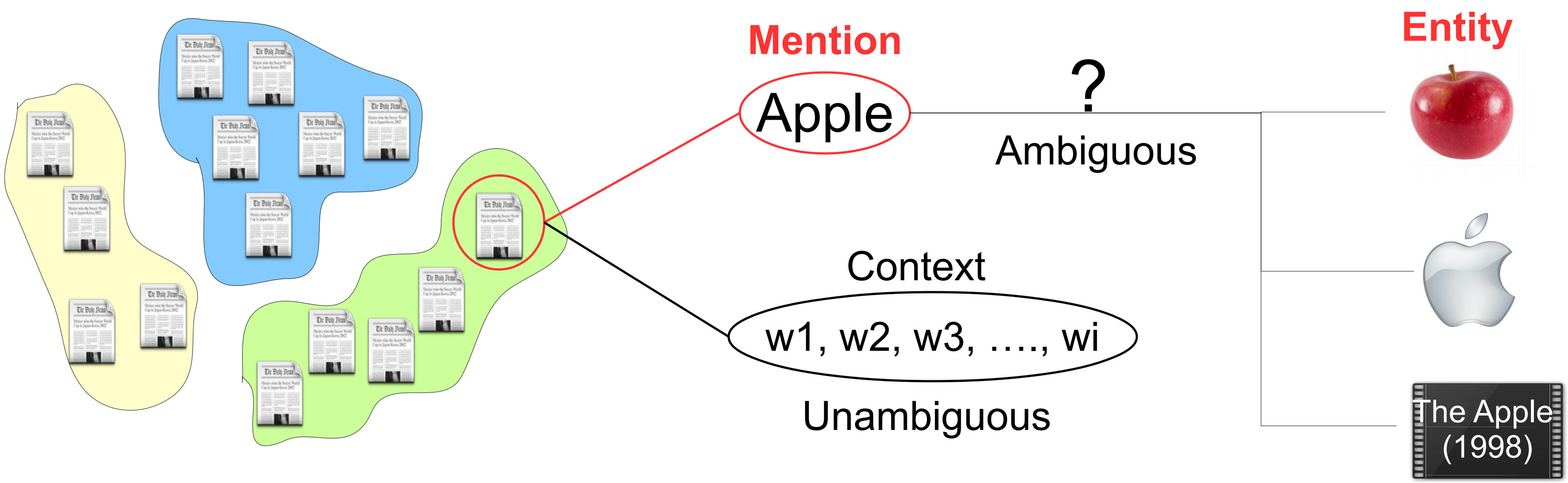
Problem



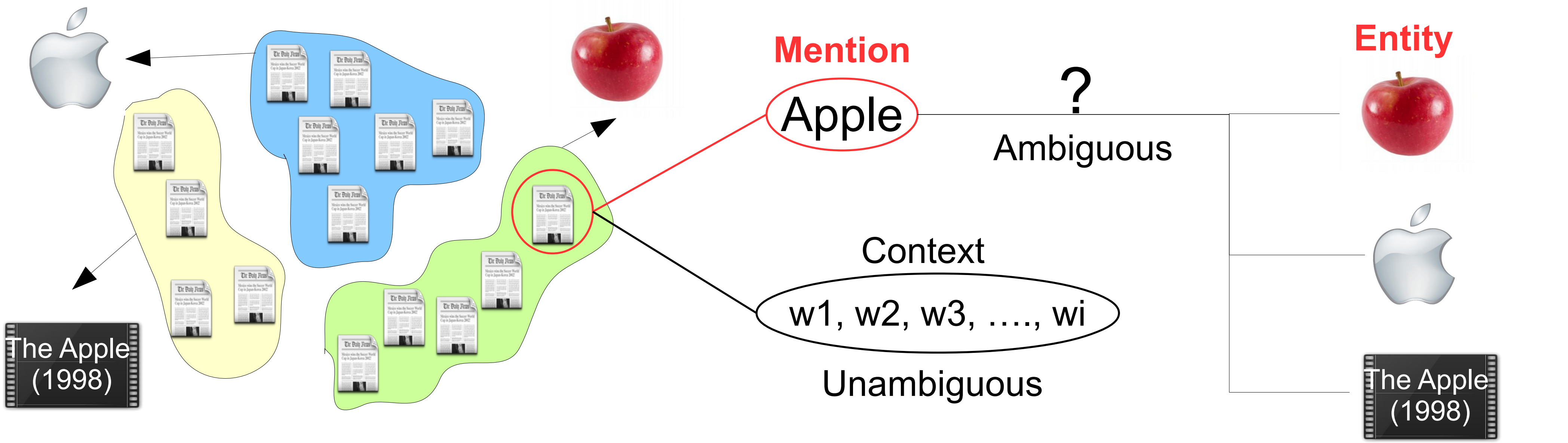
Problem


Mention
Apple
?
Ambiguous
Context
w1, w2, w3, …., wi
Unambiguous
Entity
The Apple
(1998)

Problem


Mention
Apple
?
Ambiguous
Context
w1, w2, w3, …., wi
Unambiguous
Entity
The Apple (1998)

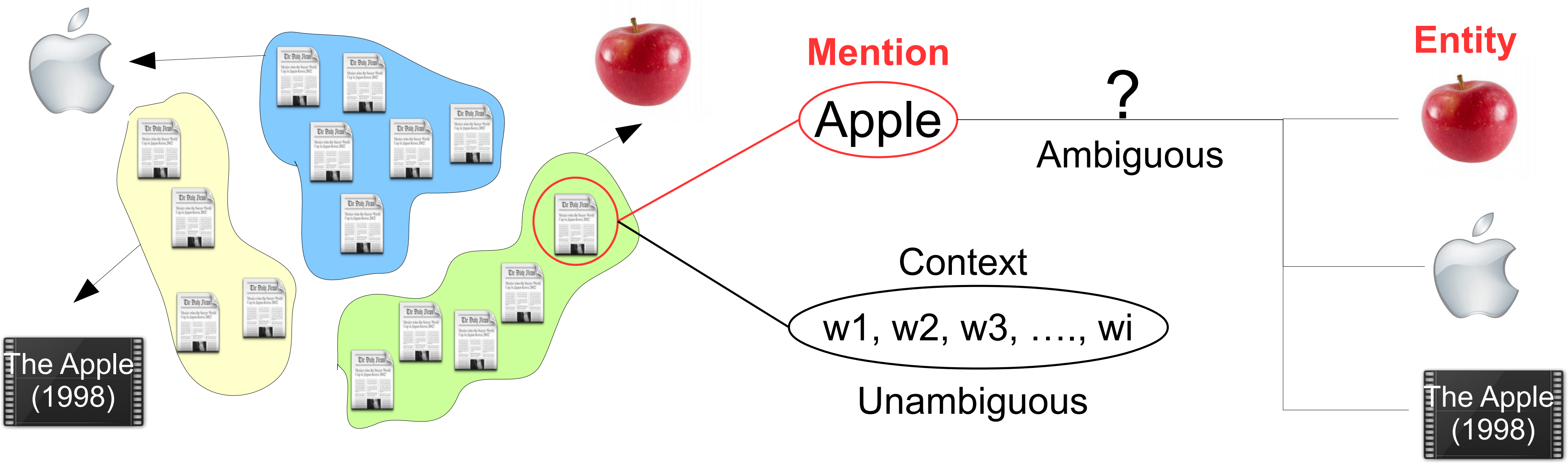
Problem



Questions:

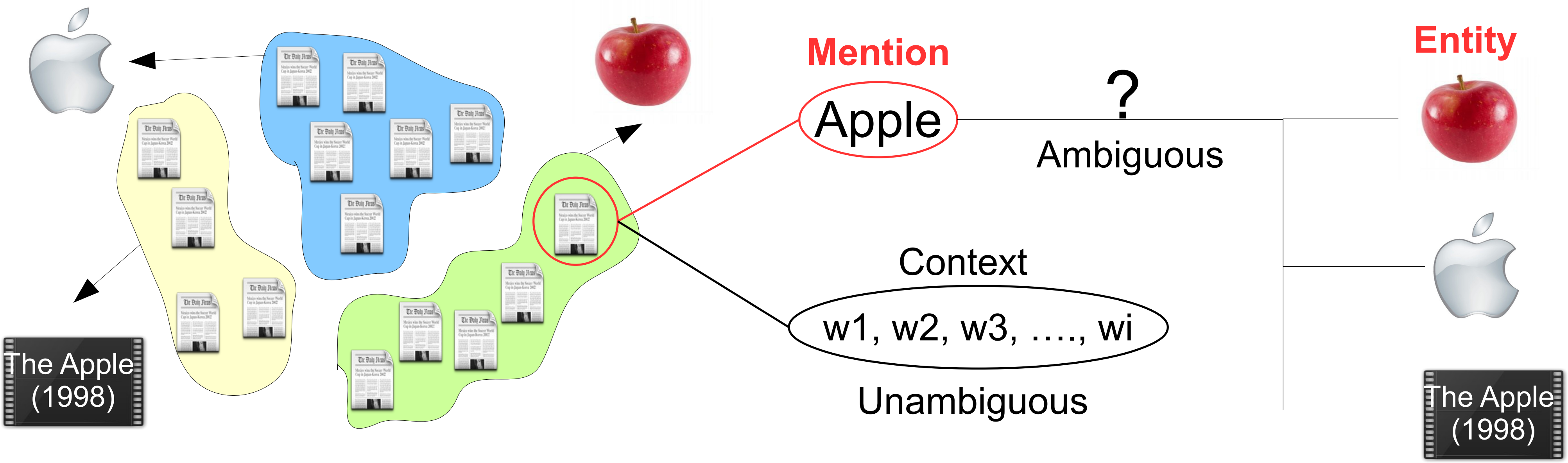
How to group these documents such that all the Mentions in each group refer to the same Entity in the real word?

Problem



Clustering

Problem

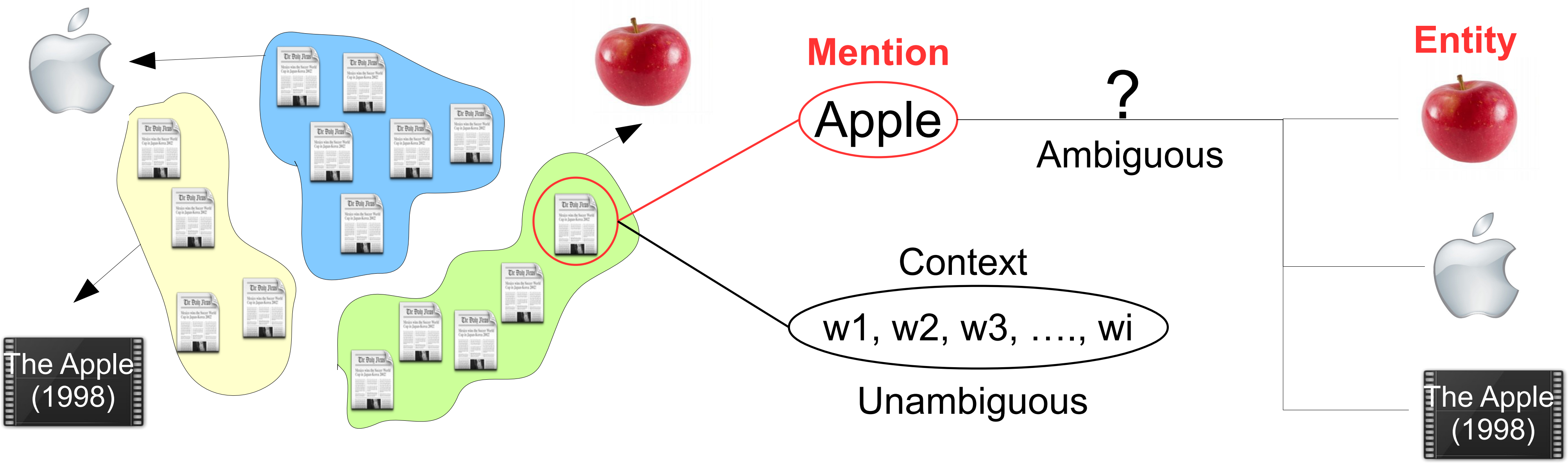


Clustering

Group documents based on their similarity

Similarity(D1,D2) $\propto$ Count(common context words)

Problem



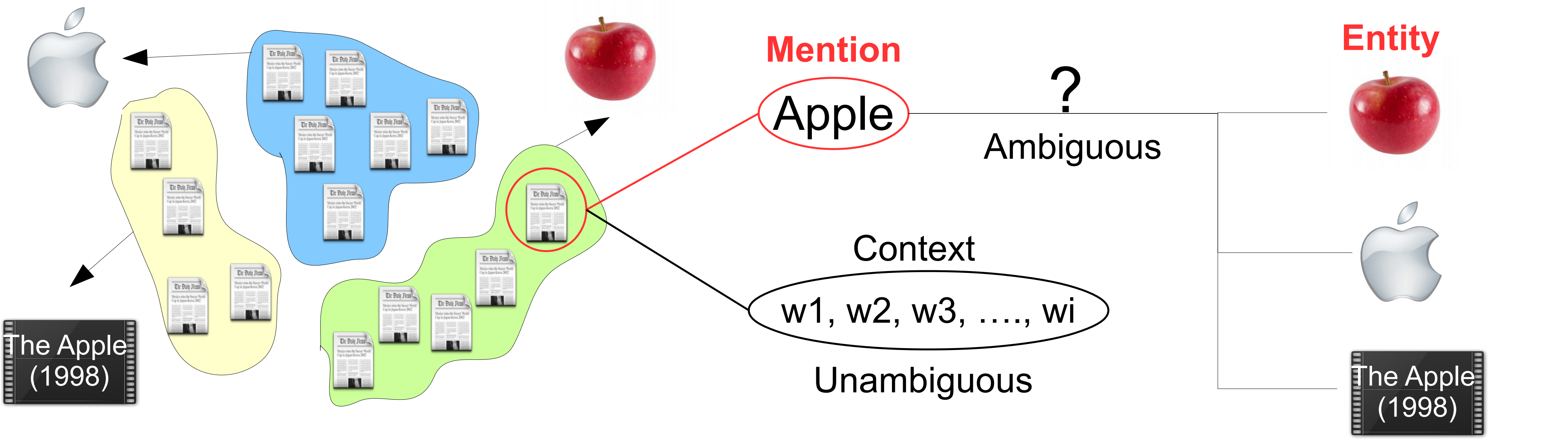
Clustering

Group documents based on their similarity

Similarity(D1,D2) $\propto$ Count(common context words)

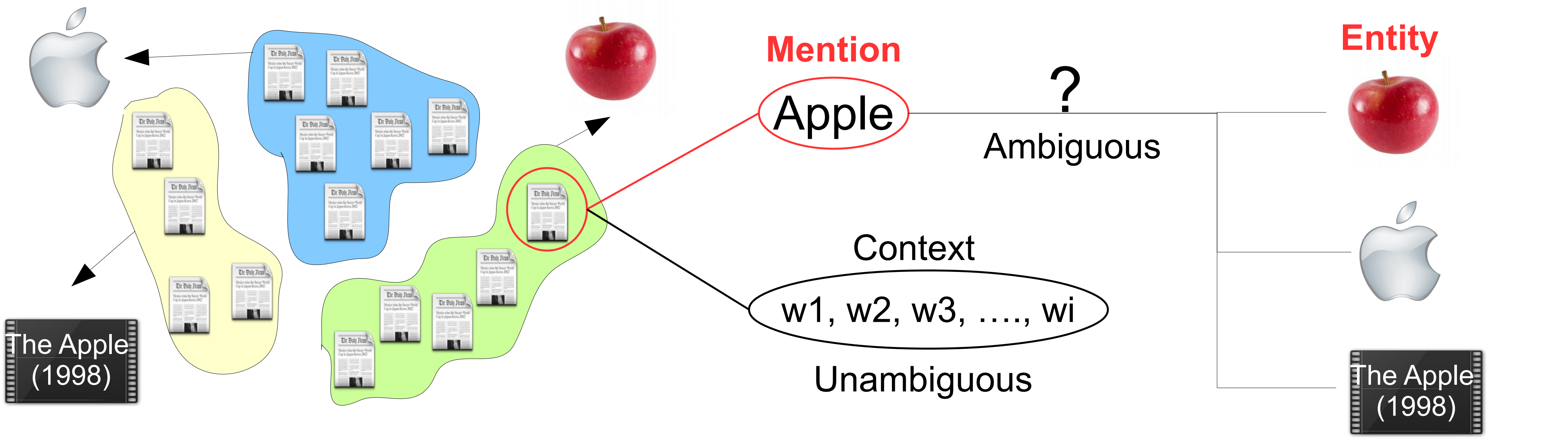
Mathematical model of the documents

Solution – VSM



Vector Space Model (VSM)

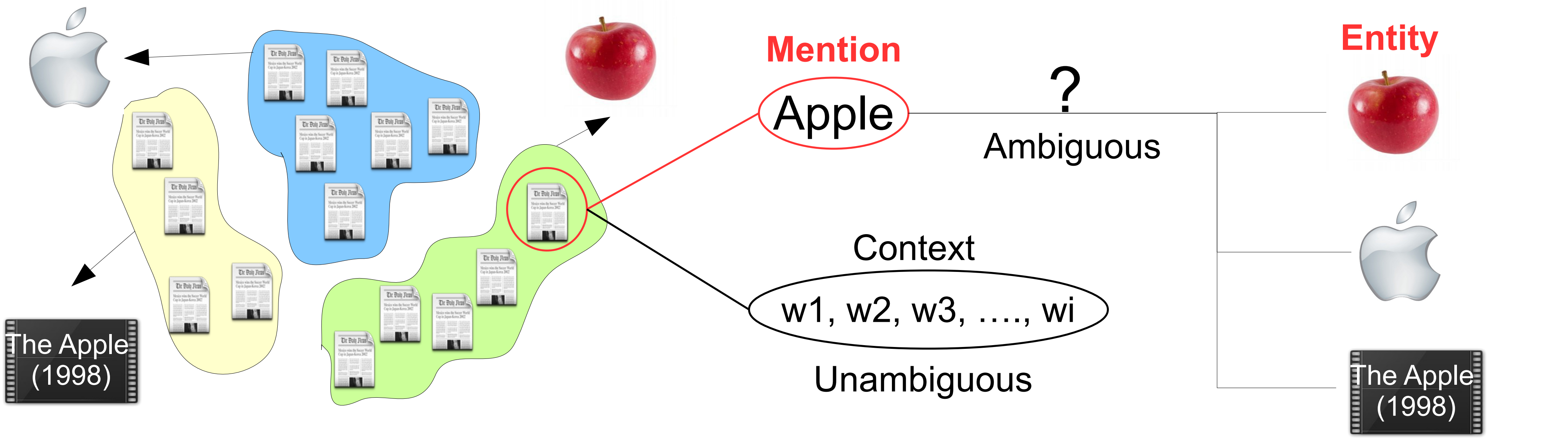
Solution – VSM



Unique Context Words

$$\{ w_1, w_2, w_3, \ldots, w_m \}$$

Solution – VSM

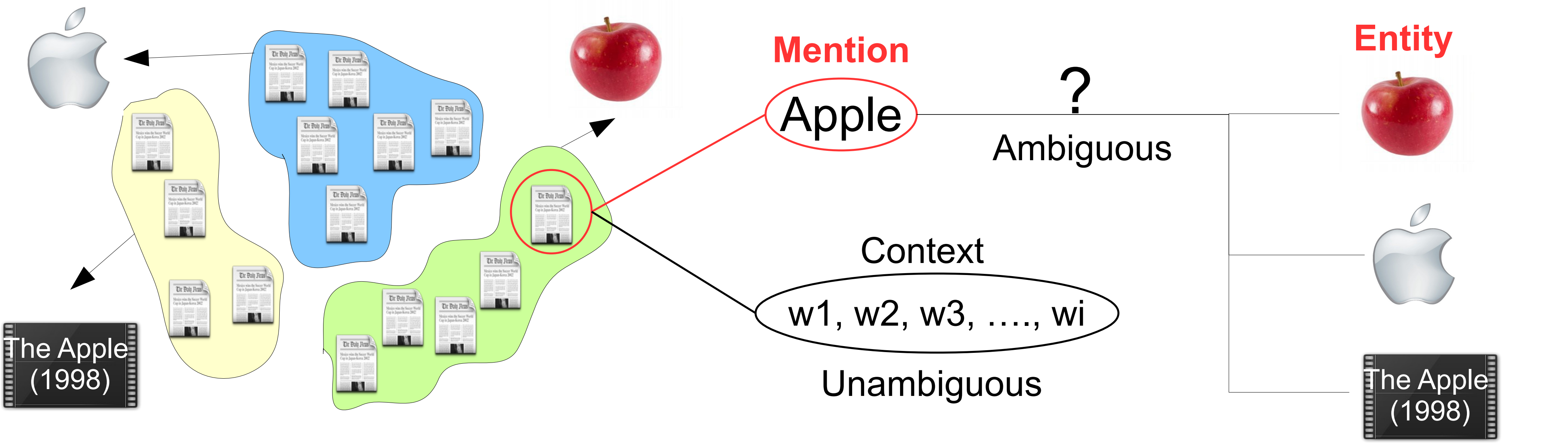


Unique Context Words

$$\left\{ w_1, w_2, w_3, \ldots, w_m \right\}$$

$$D_1 \longrightarrow \vec{V}_1 = \left\{ 1, 0, 1, \ldots, 0 \right\}$$

Solution – VSM



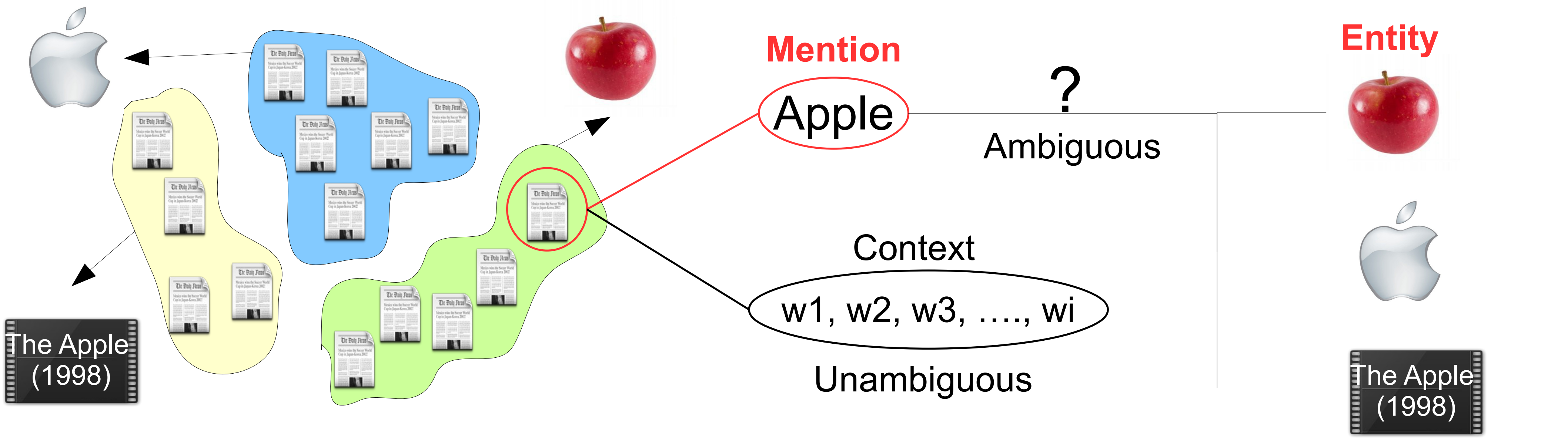
Unique Context Words

$$\{w_1, w_2, w_3, \ldots, w_m\}$$

$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

Solution – VSM



Unique Context Words

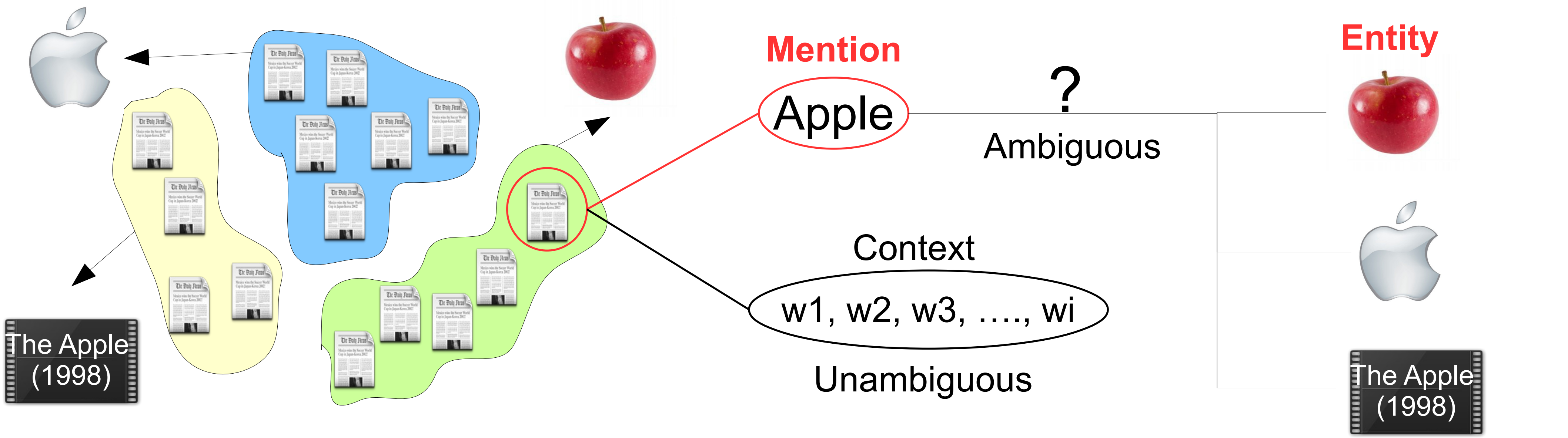
$$\{w_1, w_2, w_3, \ldots, w_m\}$$

$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

$$D_n \longrightarrow \vec{V}_n = \{1, 1, 1, \ldots, 0\}$$

Solution – VSM



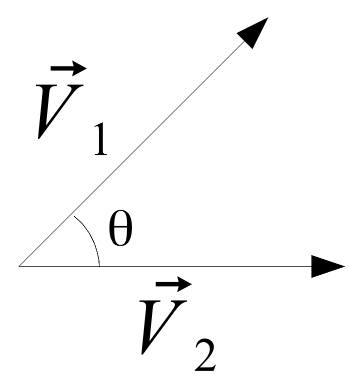
Unique Context Words

$$\{w_1, w_2, w_3, \ldots, w_m\}$$

$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

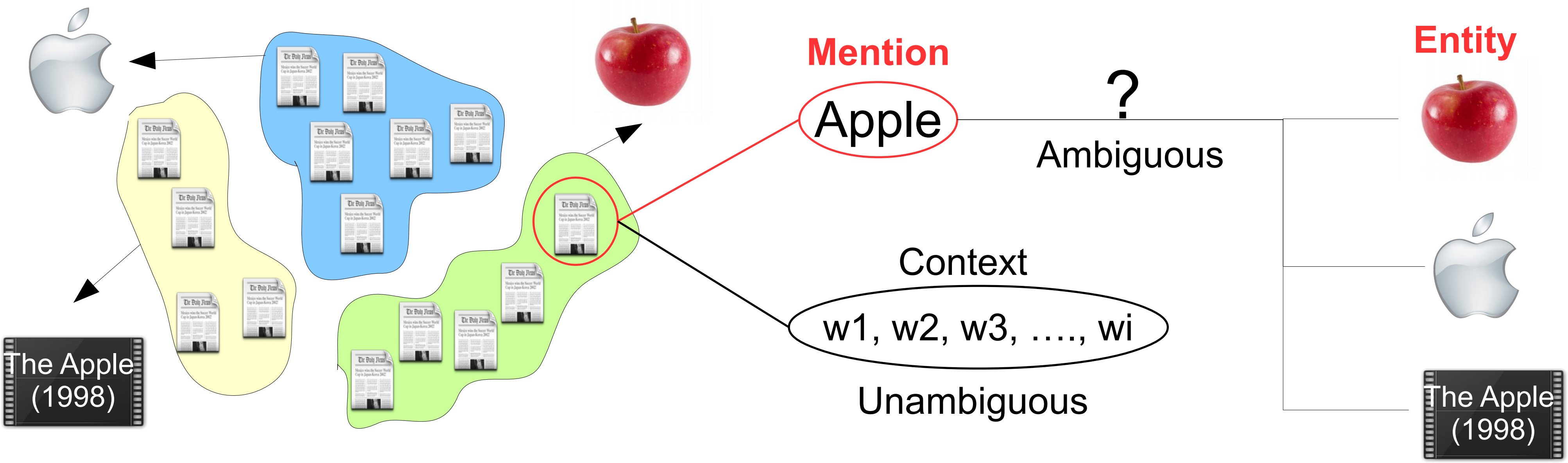
$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

$$D_n \longrightarrow \vec{V}_n = \{1, 1, 1, \ldots, 0\}$$



$$Similarity(D_1, D_2) \propto \cos(\theta)$$

Solution – VSM



Unique Context Words

$$\{w_1, w_2, w_3, \ldots, w_m\}$$

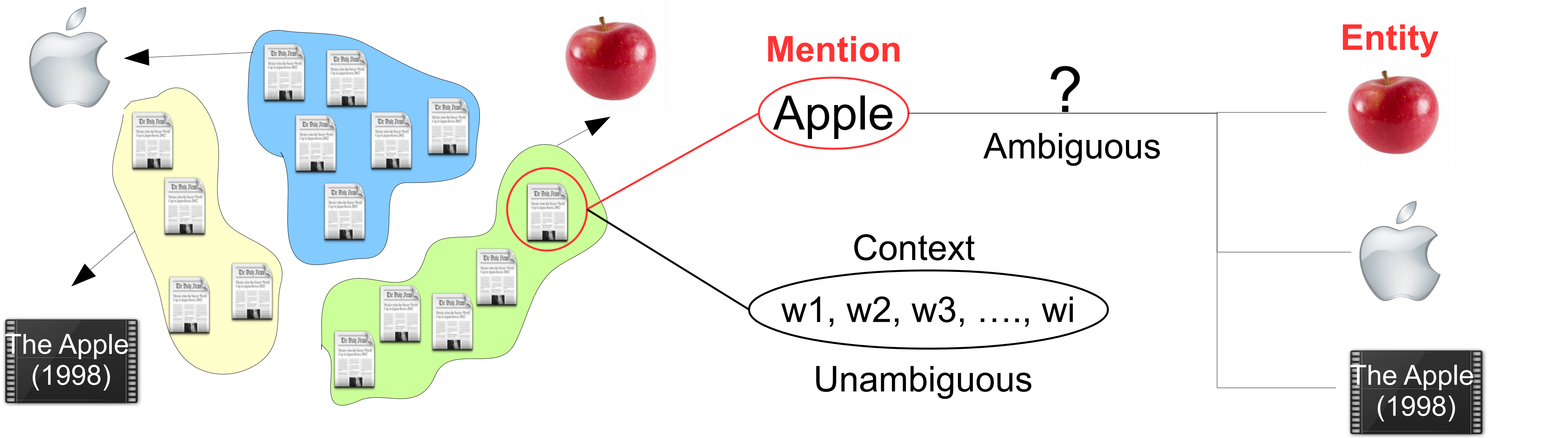
$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

$$D_n \longrightarrow \vec{V}_n = \{1, 1, 1, \ldots, 0\}$$

Limitations

Solution – VSM



Unique Context Words

$$\{w_1, w_2, w_3, \ldots, w_m\}$$

$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

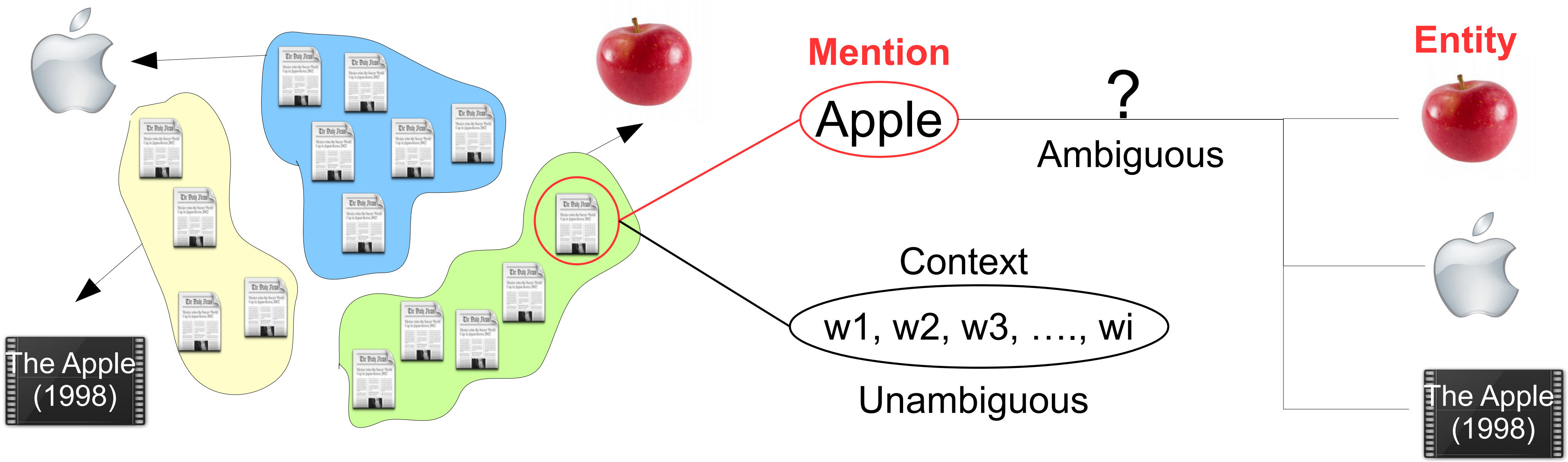
$$D_n \longrightarrow \vec{V}_n = \{1, 1, 1, \ldots, 0\}$$

Limitations

Number of Comparisons

Not every two documents need to be compared

Solution – VSM



Unique Context Words

$$\{w_1, w_2, w_3, \ldots, w_m\}$$

$$D_1 \longrightarrow \vec{V}_1 = \{1, 0, 1, \ldots, 0\}$$

$$D_2 \longrightarrow \vec{V}_2 = \{0, 0, 1, \ldots, 1\}$$

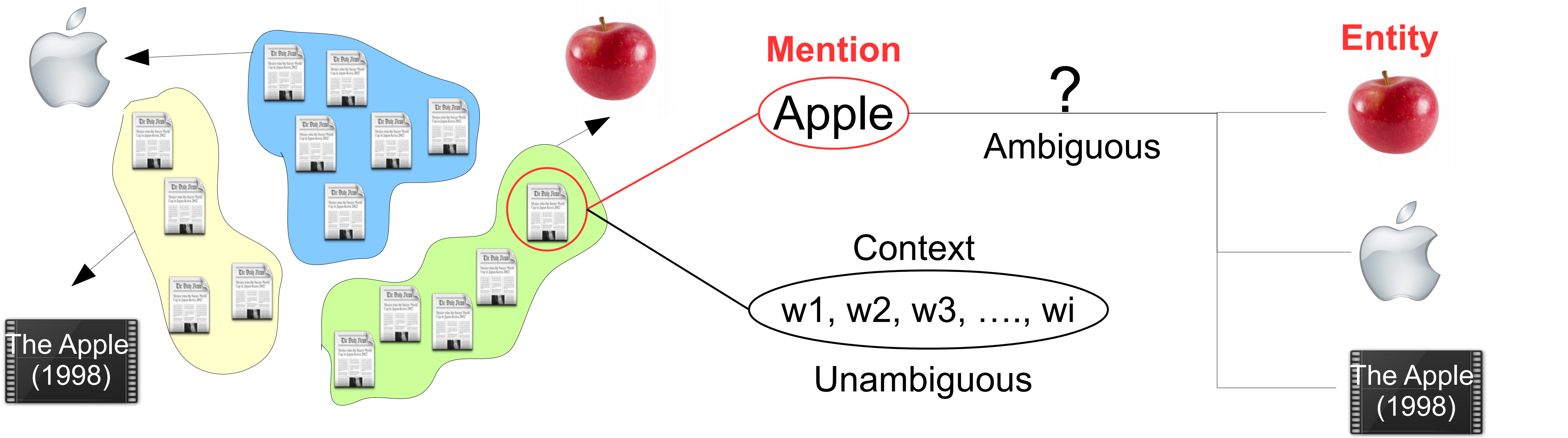
$$D_n \longrightarrow \vec{V}_n = \{1, 1, 1, \ldots, 0\}$$

Limitations

Number of Comparisons
Not every two documents need to be compared

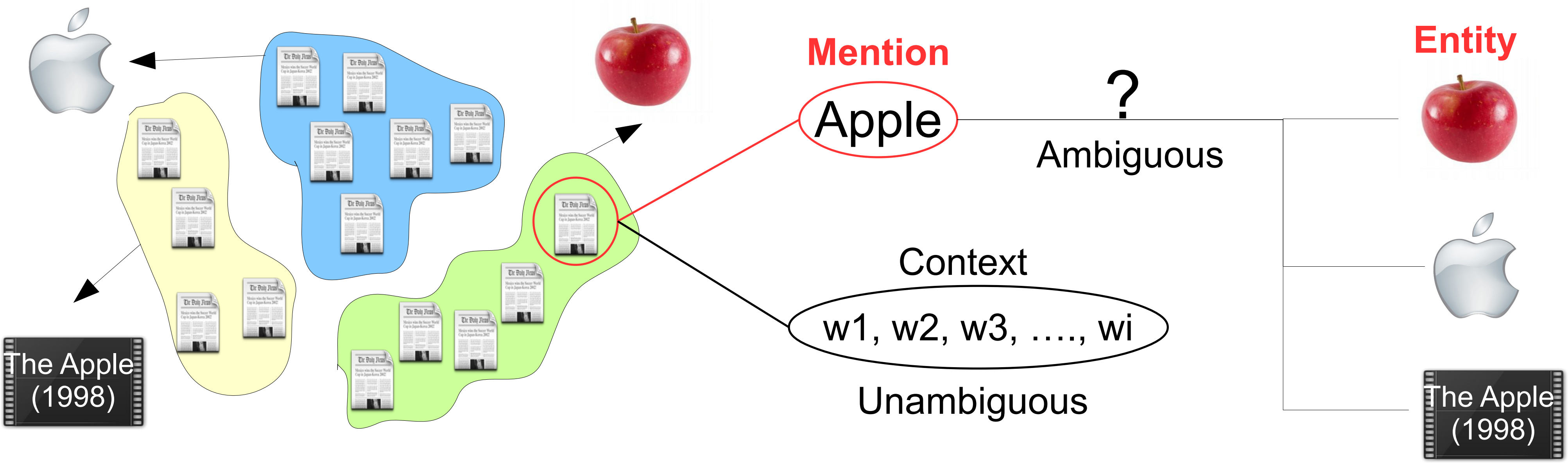
Size of each Comparison
Total number of unique context words (n)

Graph Based Modeling

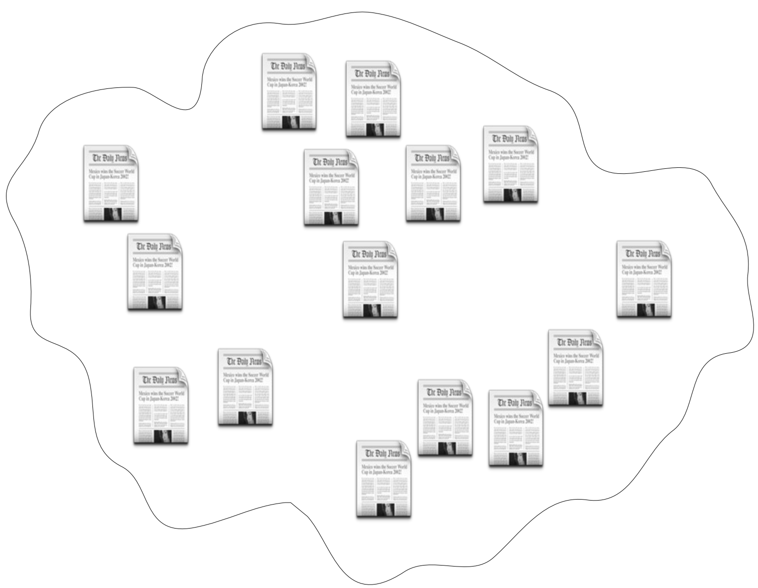


Graph Based Modeling

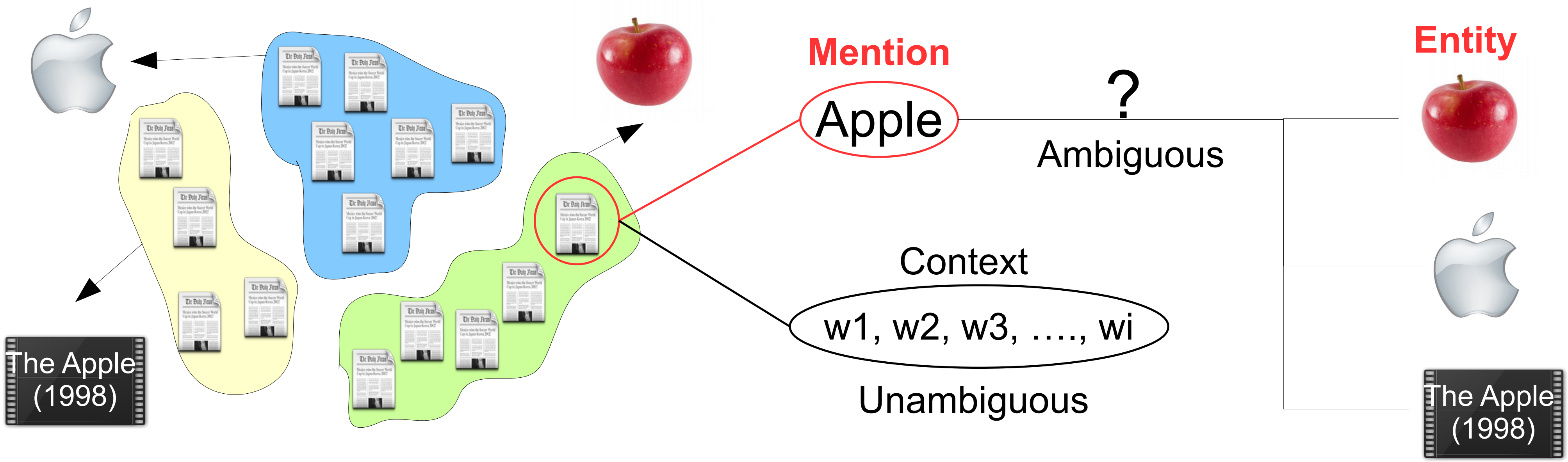
Graph Based Modeling

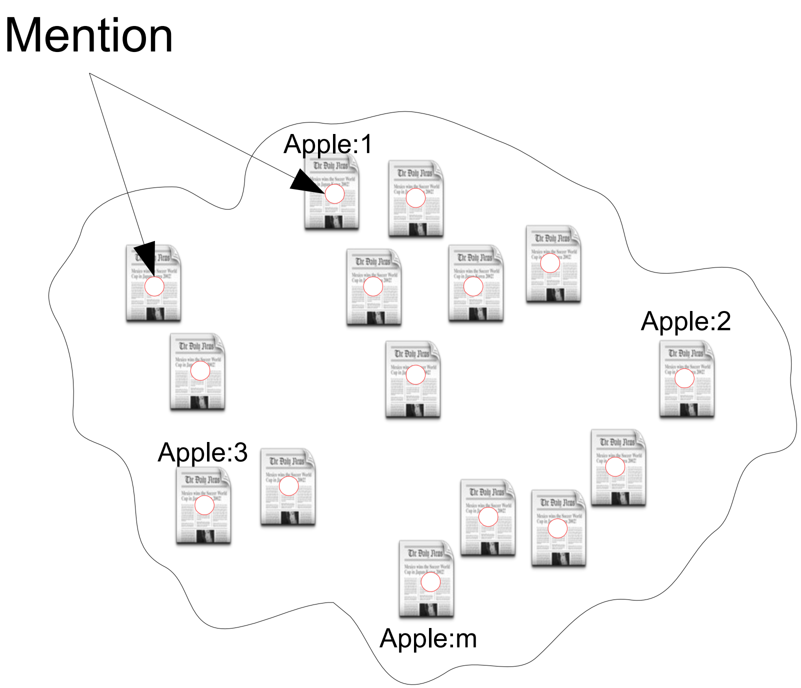


Graph based modeling



Graph Based Modeling

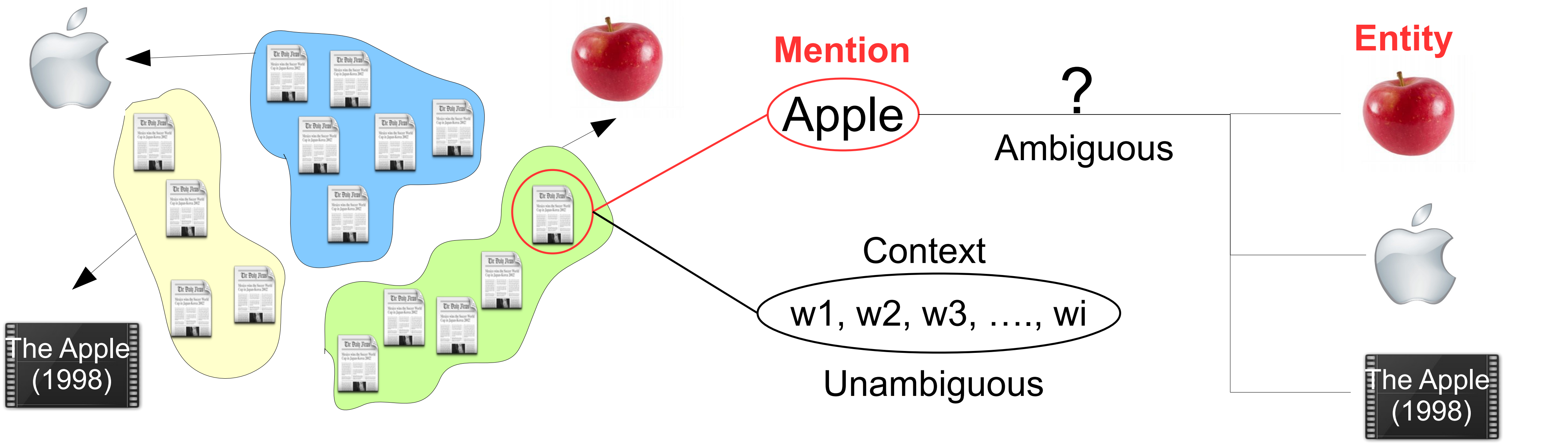


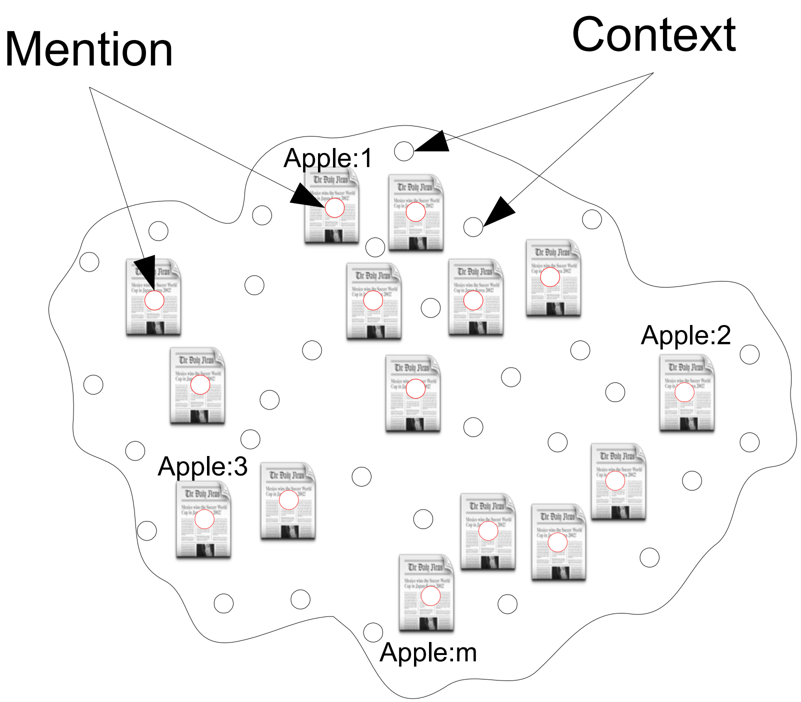


Graph based modeling

- Assign a unique vertex to each Mention

Graph Based Modeling

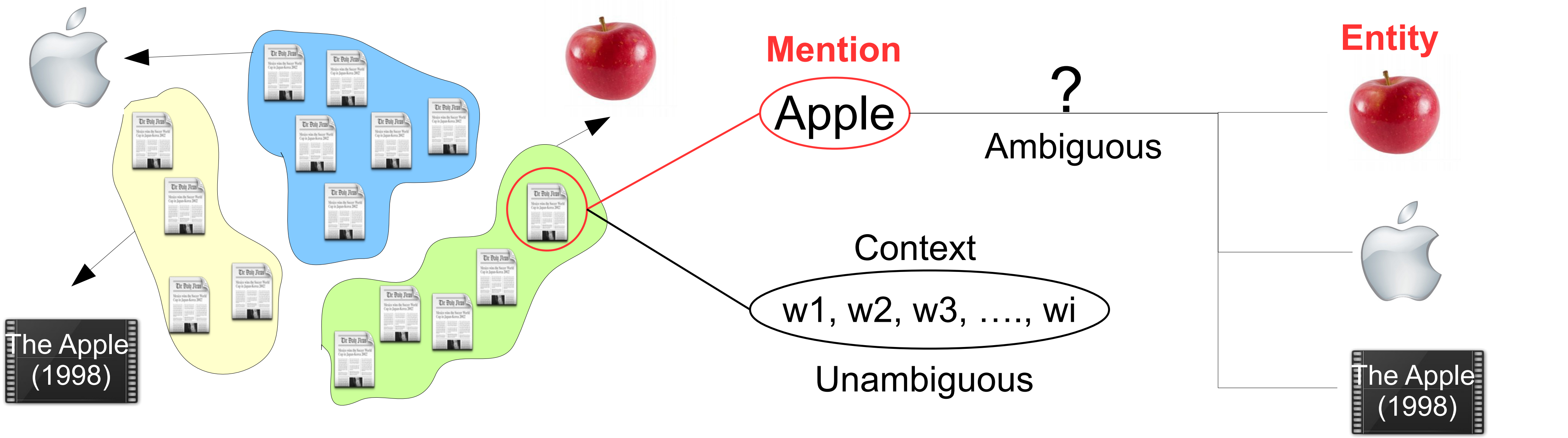


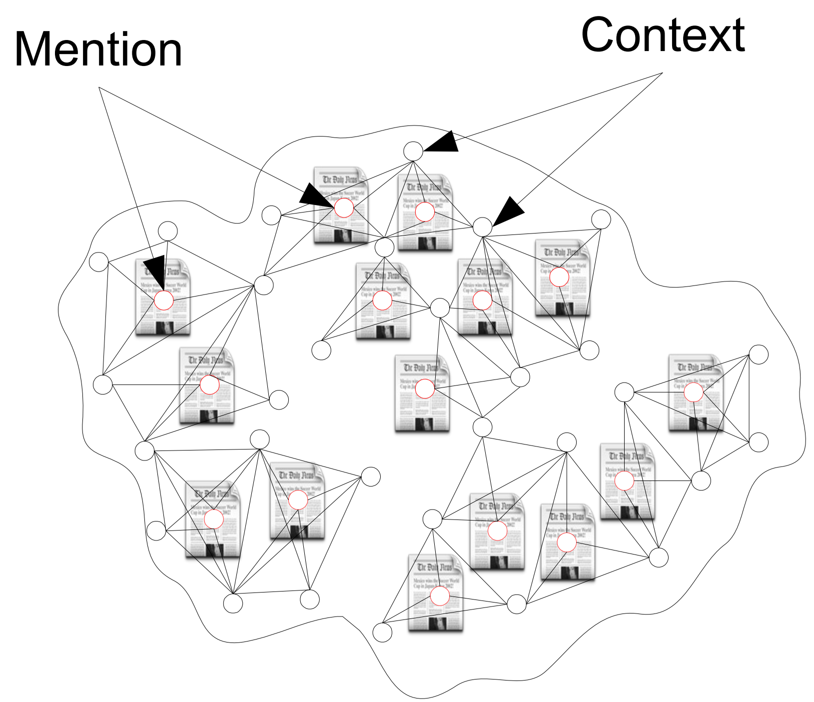


Graph based modeling

- Assign a unique vertex to each Mention
- Assign a unique vertex to each context word

Graph Based Modeling

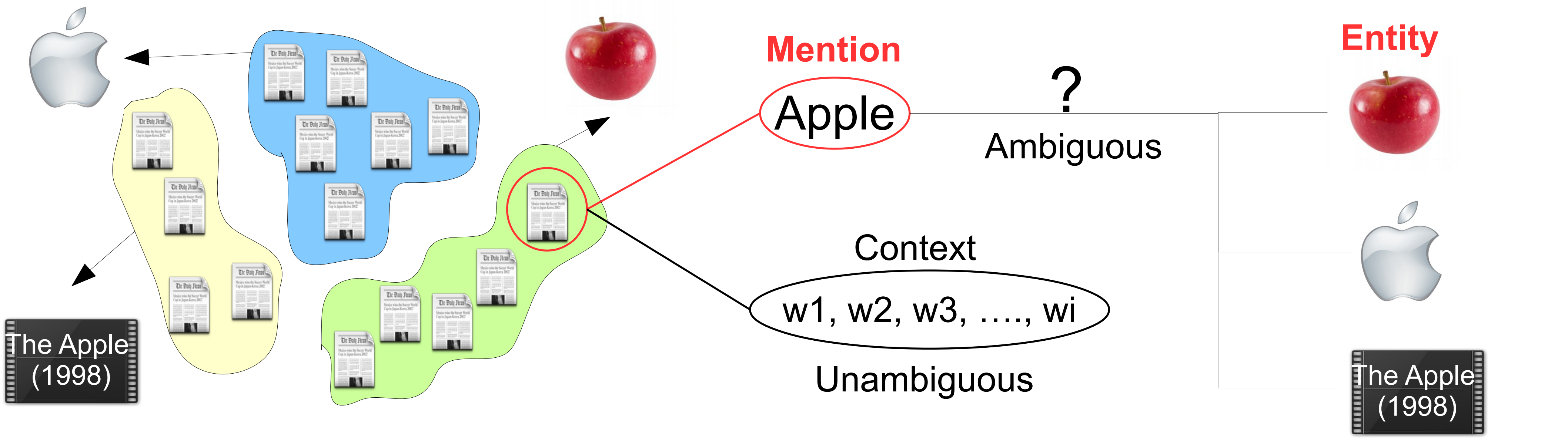


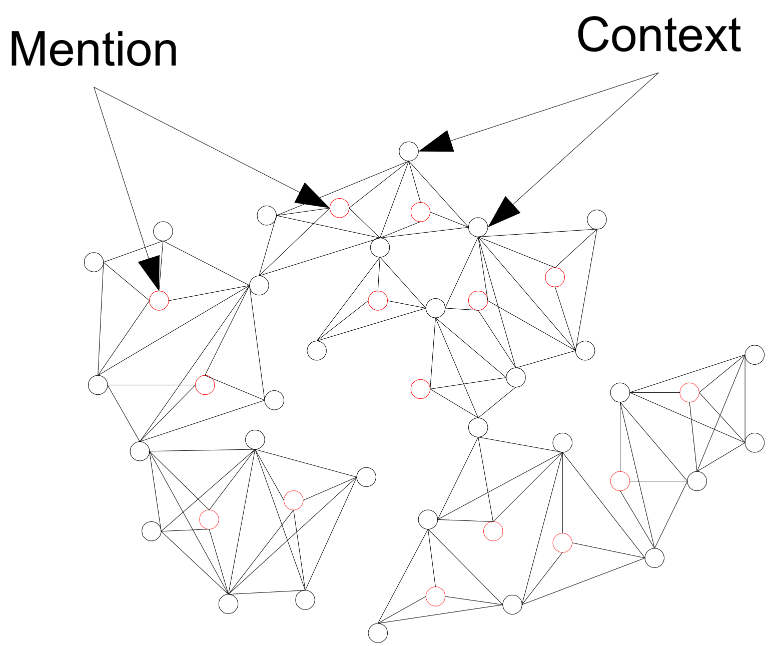


Graph based modeling

- Assign a unique vertex to each Mention
- Assign a unique vertex to each context word
- Connect vertices from same document

Graph Based Modeling

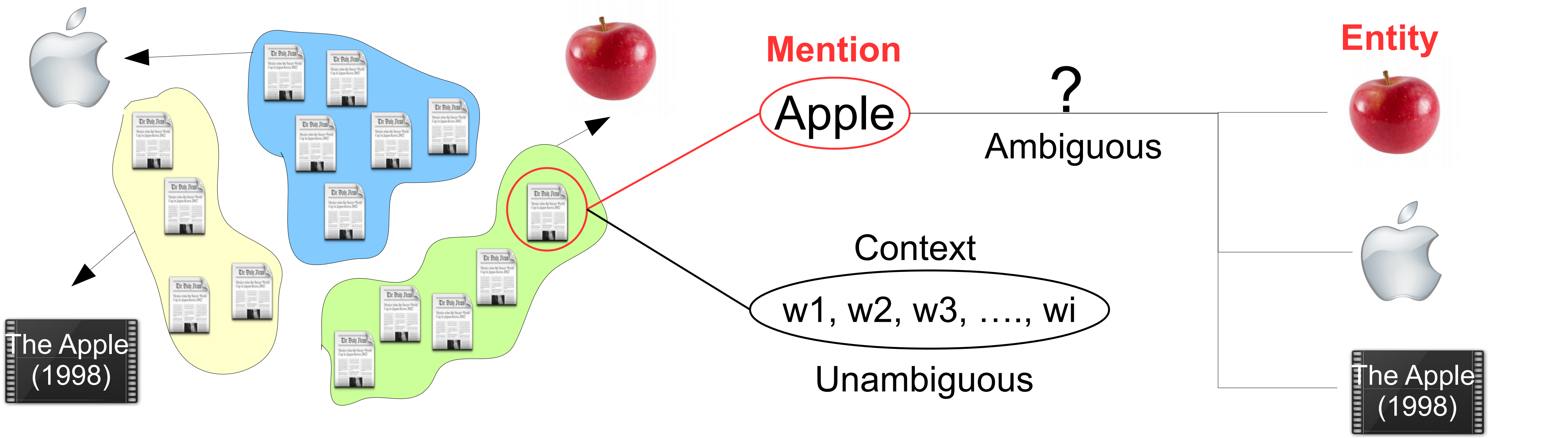


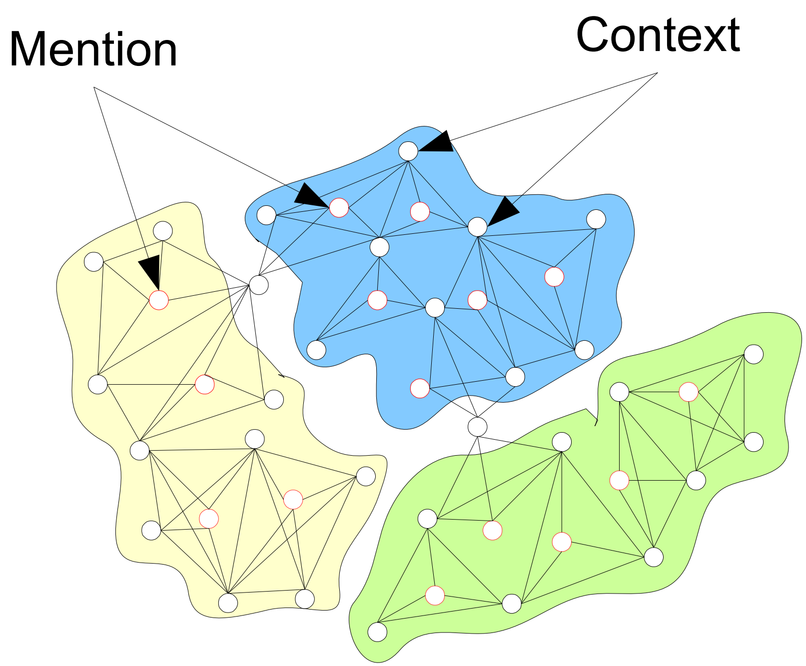


Graph based modeling

- Assign a unique vertex to each Mention
- Assign a unique vertex to each context word
- Connect vertices from same document
- Graph Representation

Graph Based Modeling

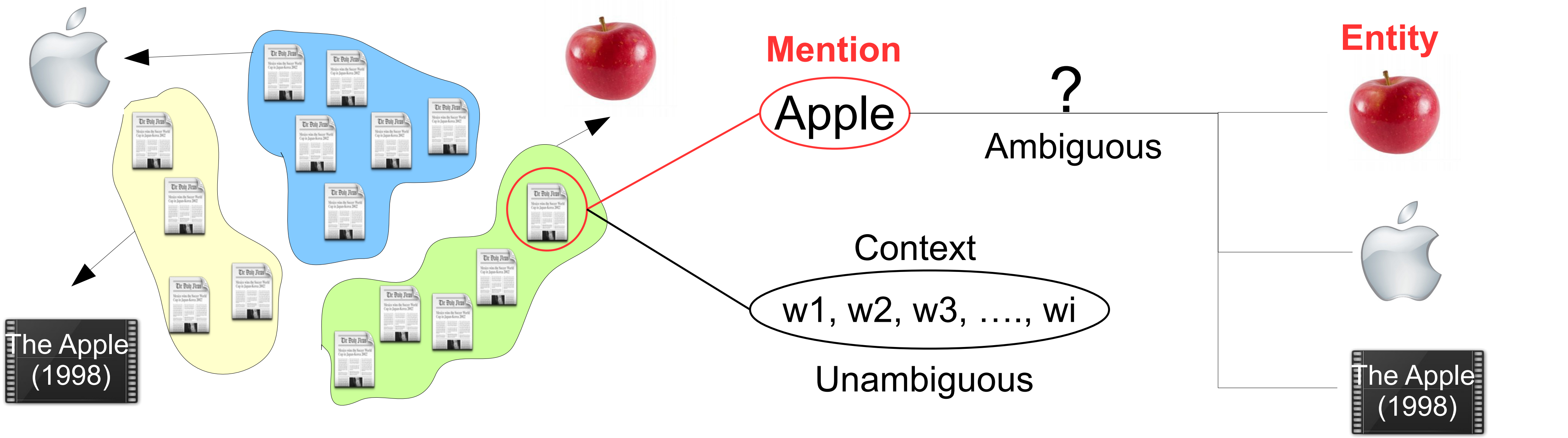




Graph based modeling

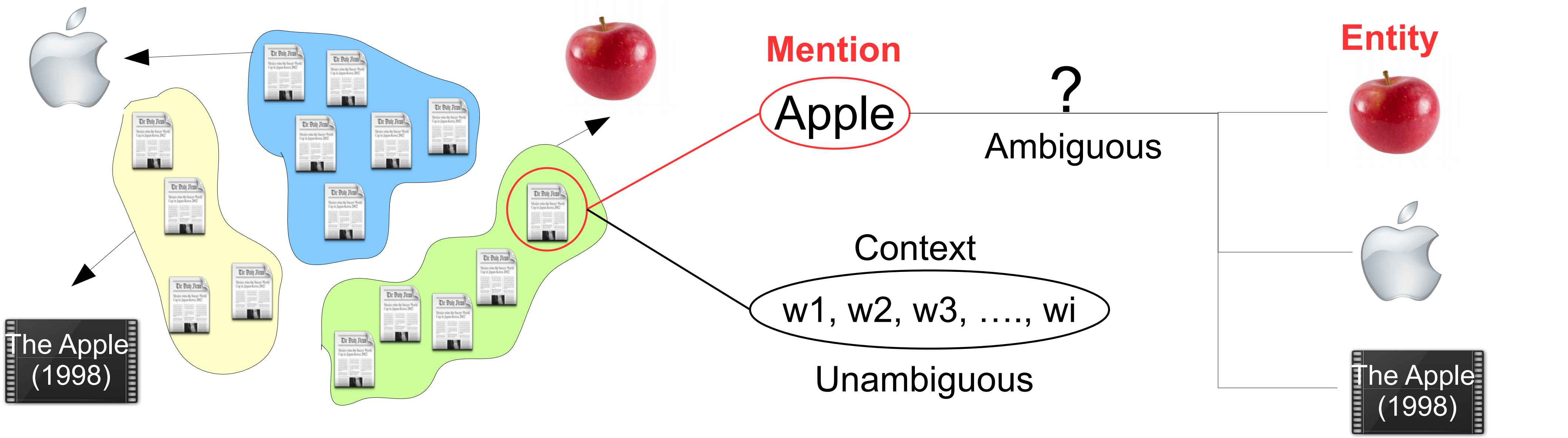
- Assign a unique vertex to each Mention
- Assign a unique vertex to each context word
- Connect vertices from same document
- Graph Representation

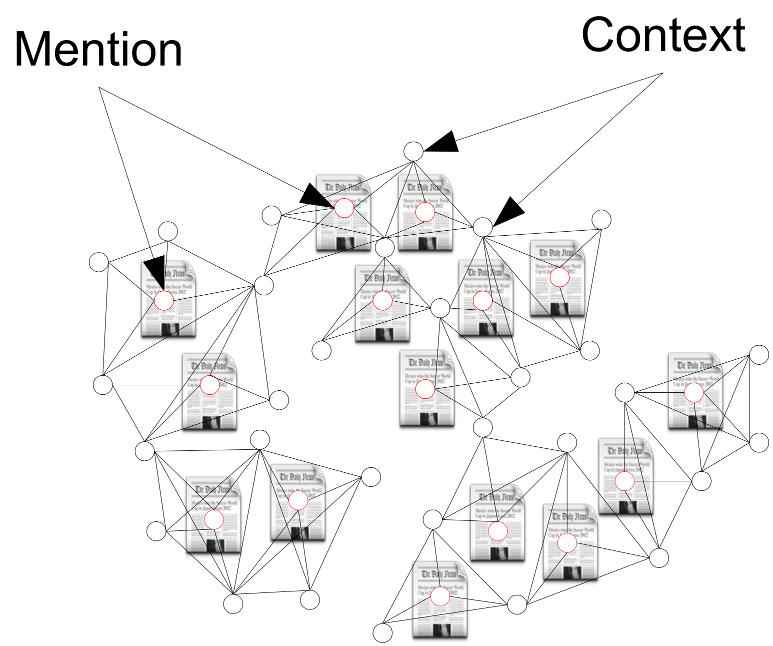
Diffusion based Clustering



Diffusion based Clustering

Diffusion based Clustering

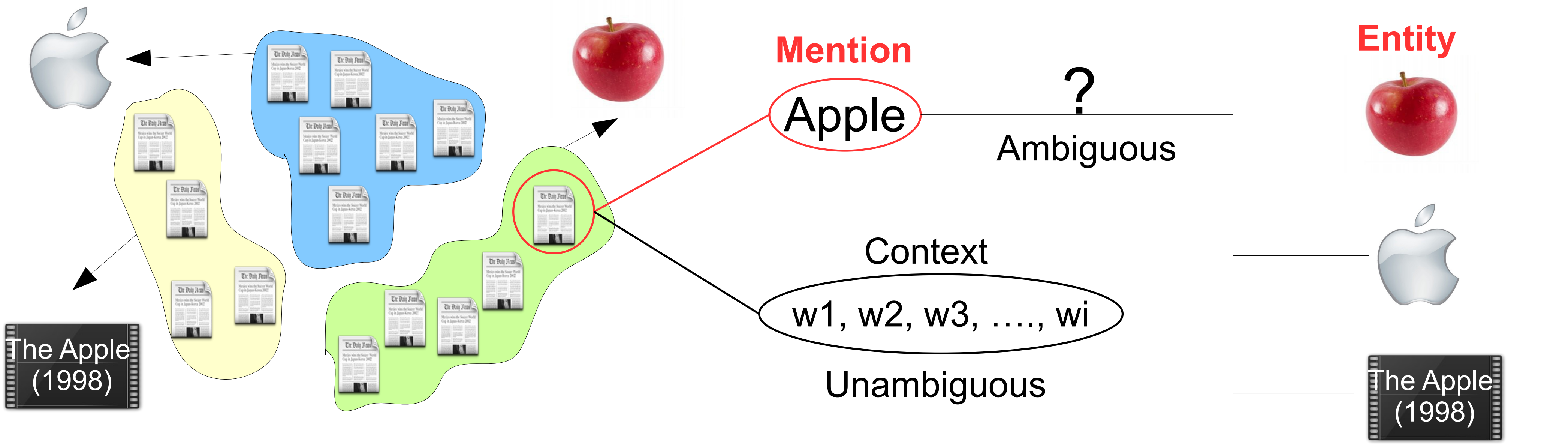


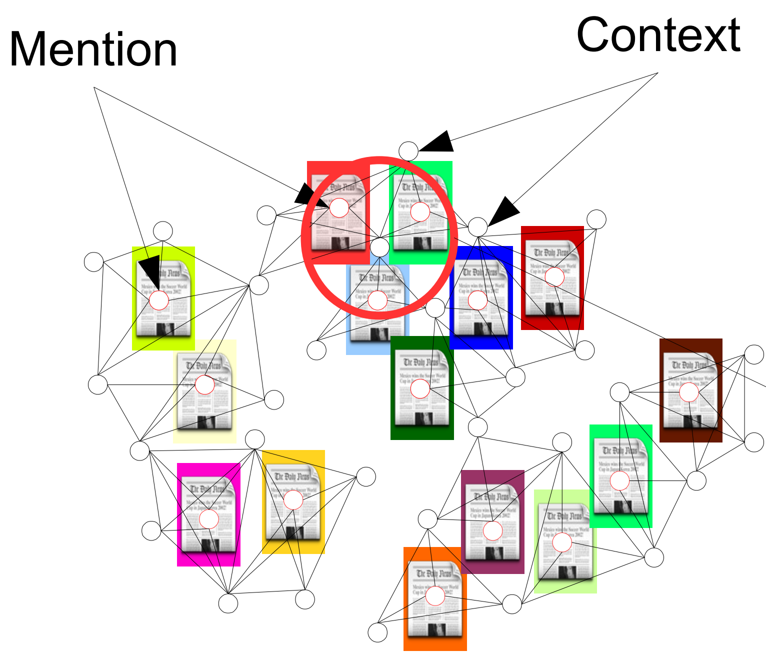


Diffusion Clustering

- Initialization
- Diffusion

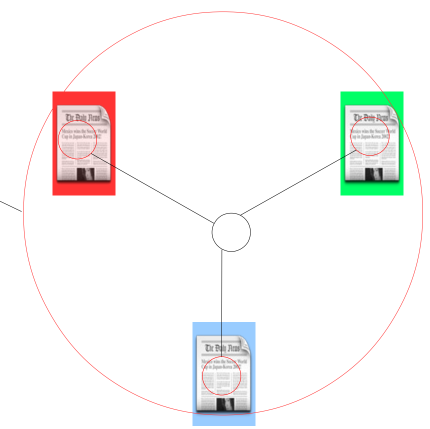
Diffusion based Clustering



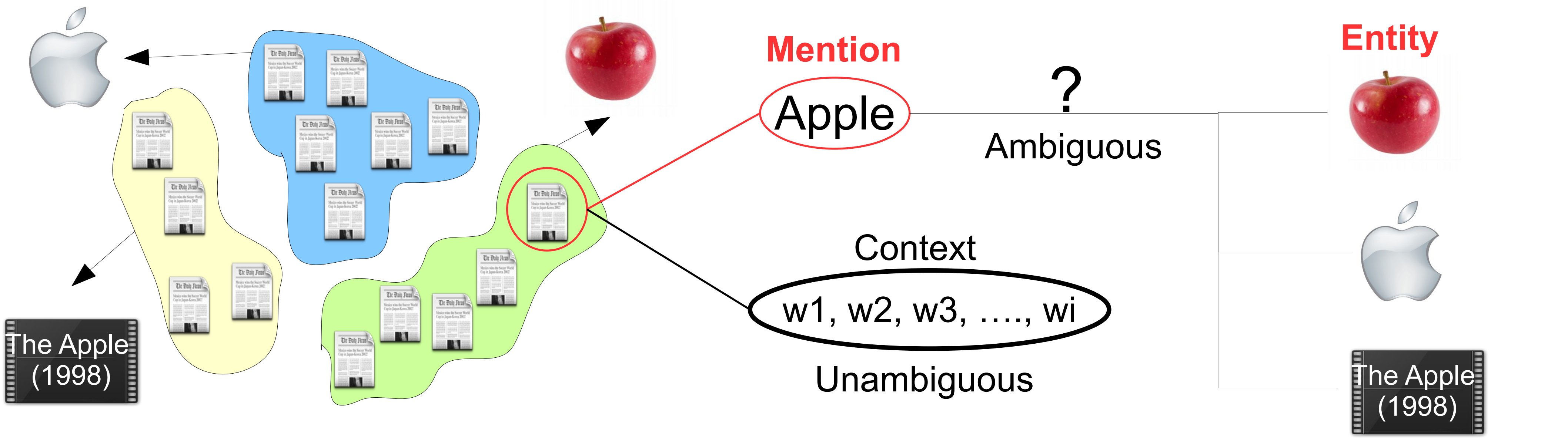


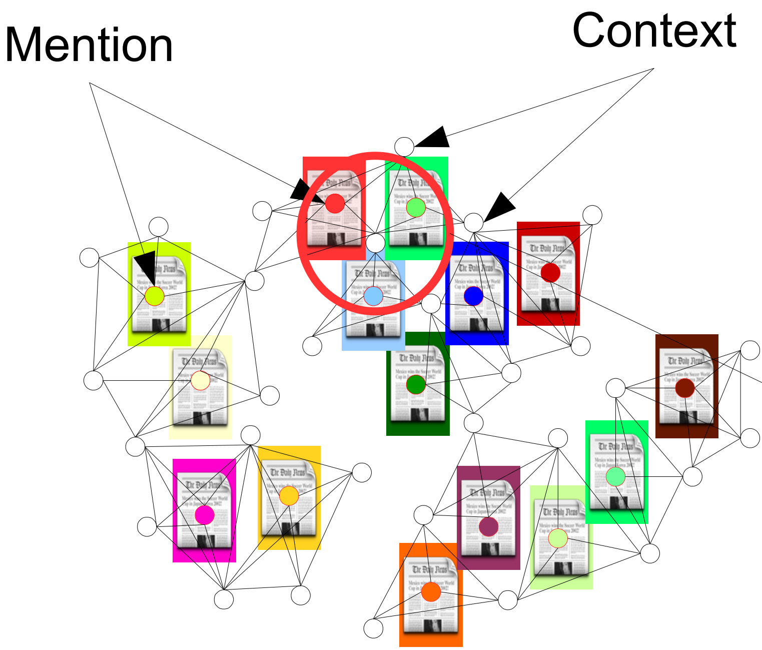
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion



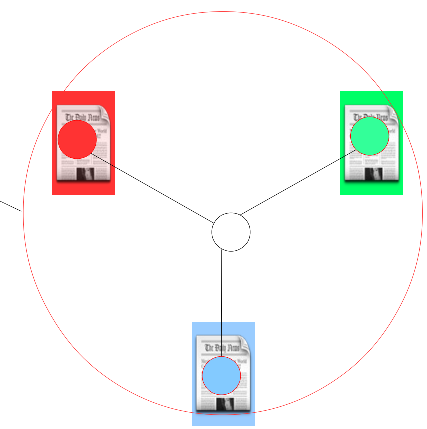
Diffusion based Clustering



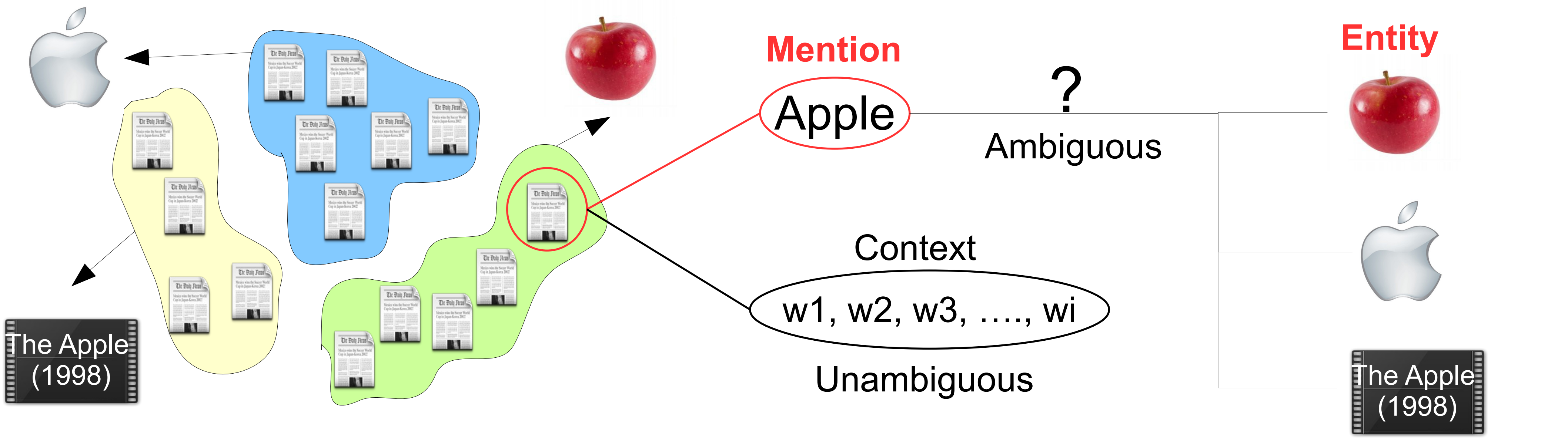


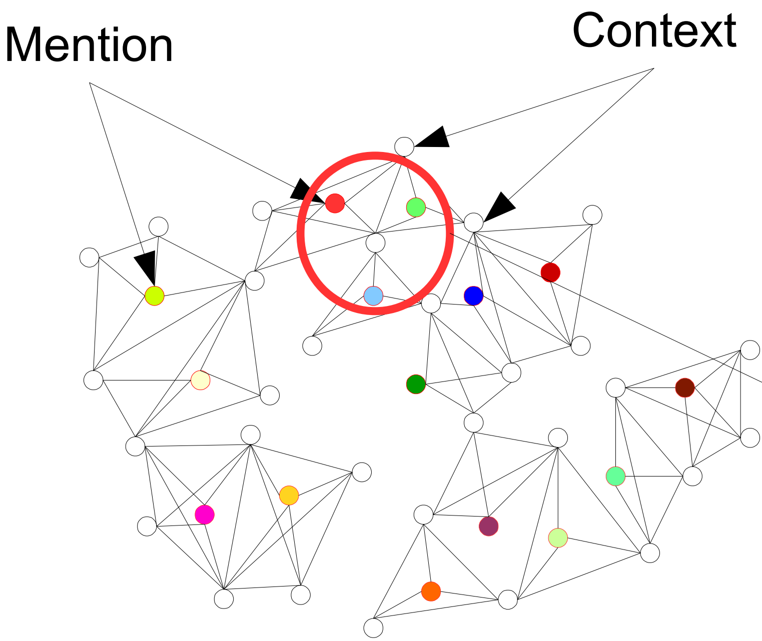
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion



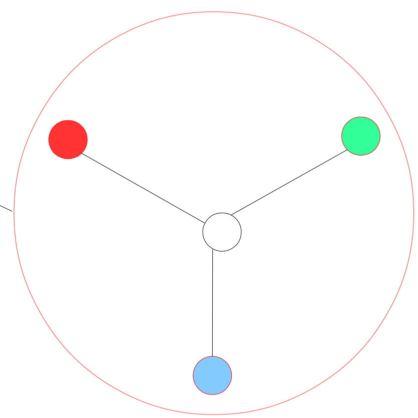
Diffusion based Clustering



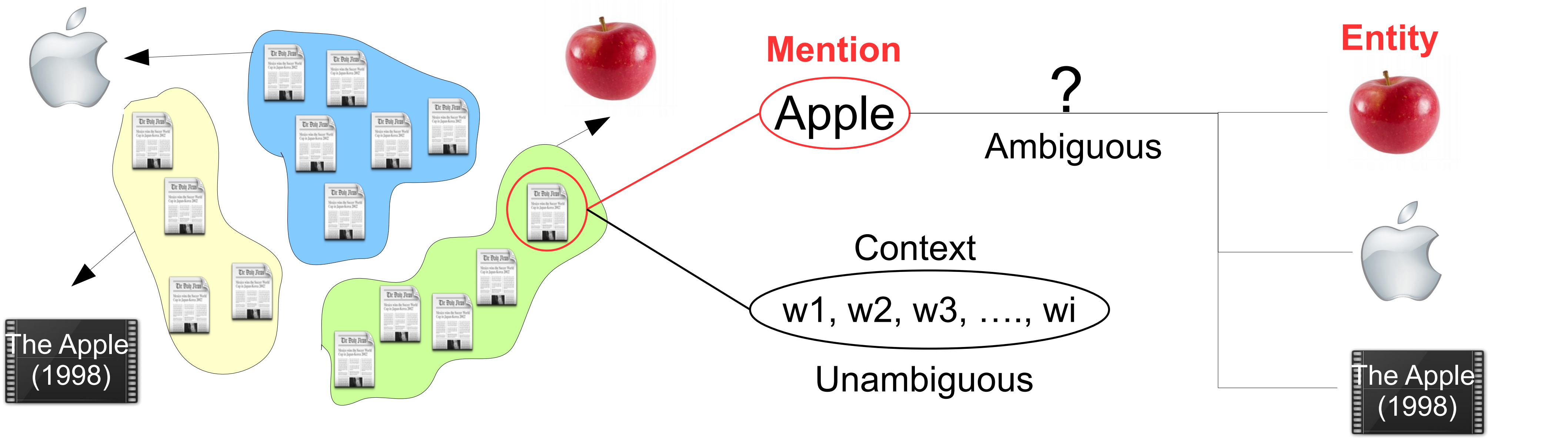


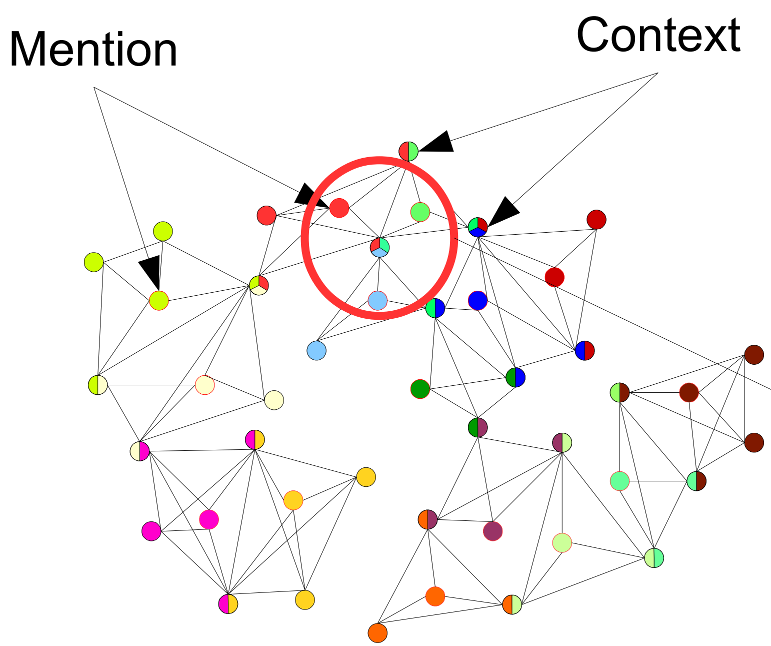
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion



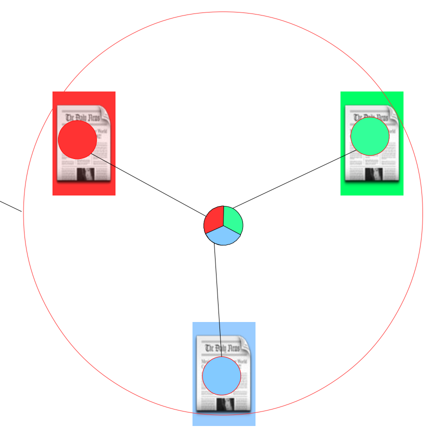
Diffusion based Clustering



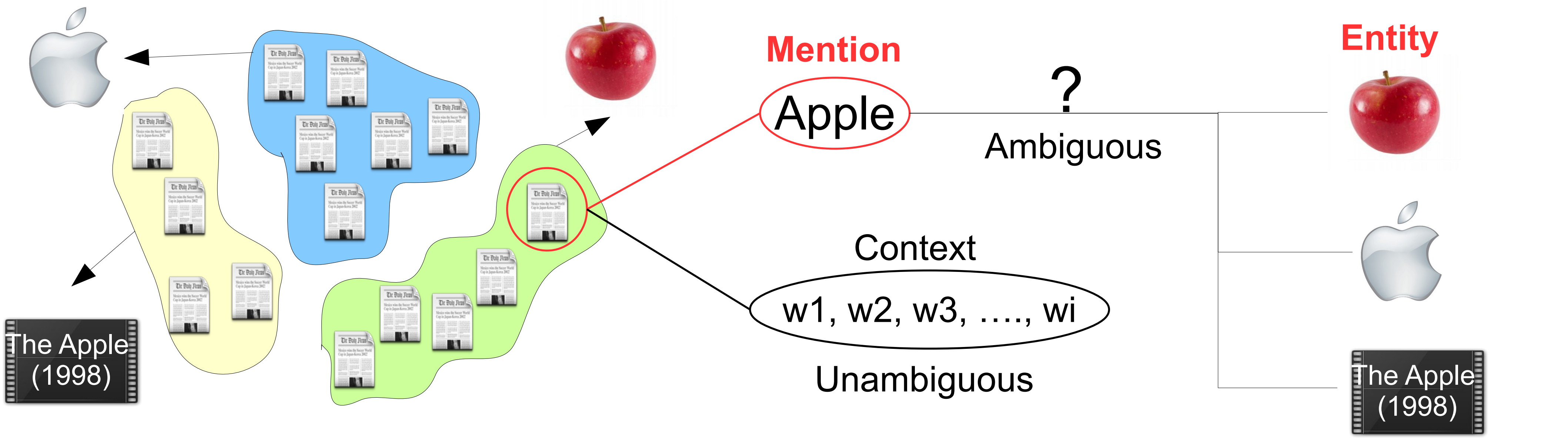


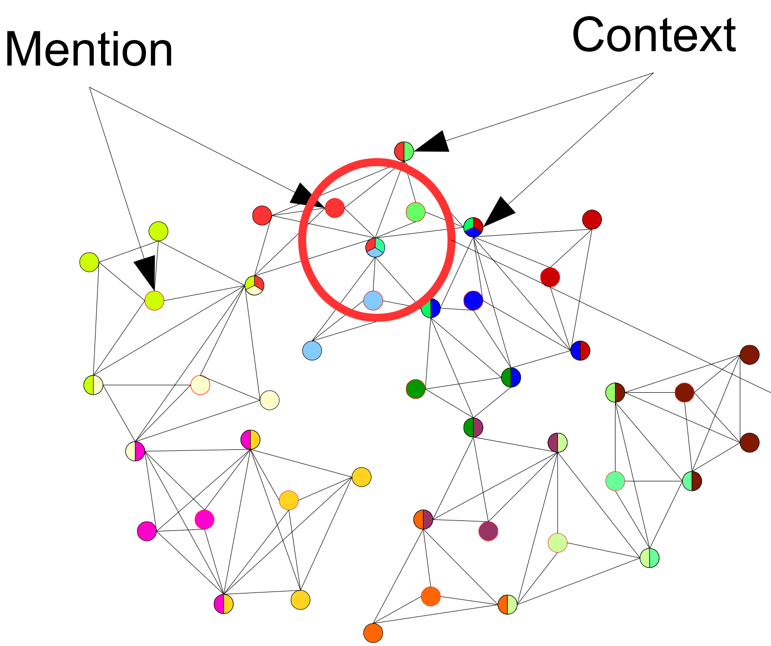
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion



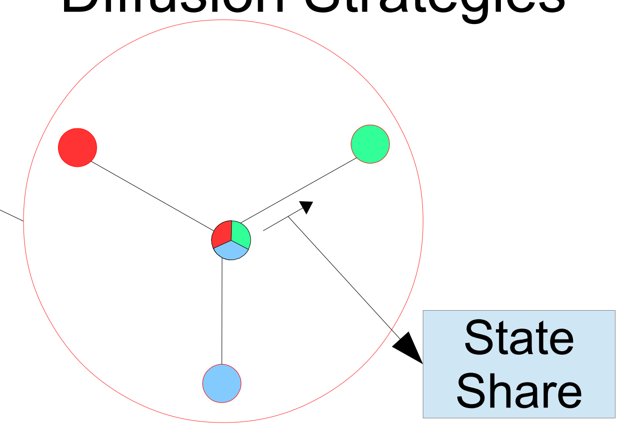
Diffusion based Clustering



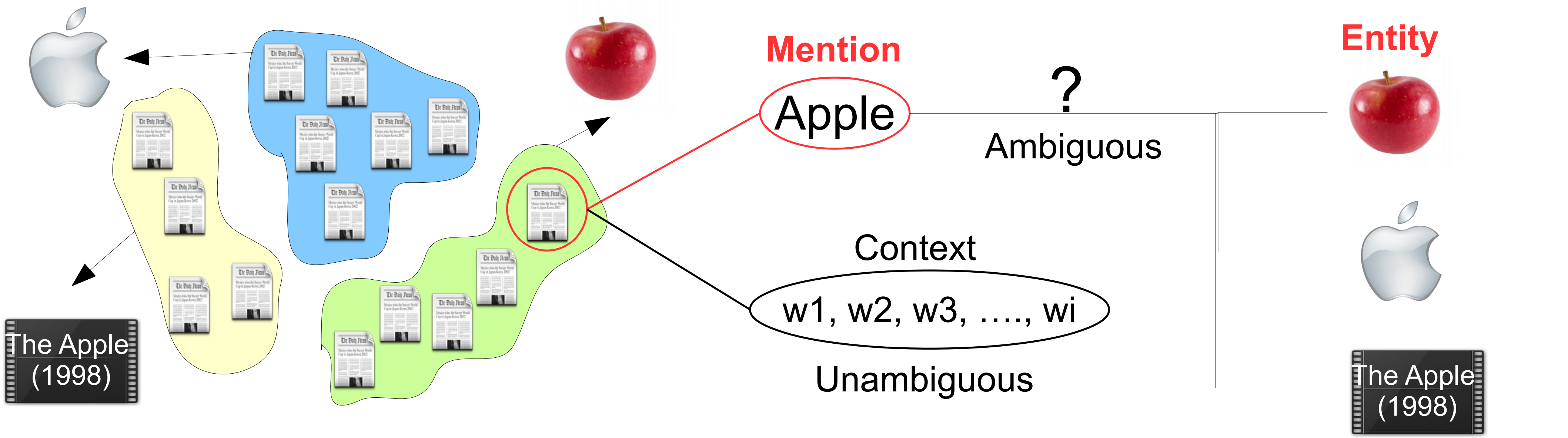


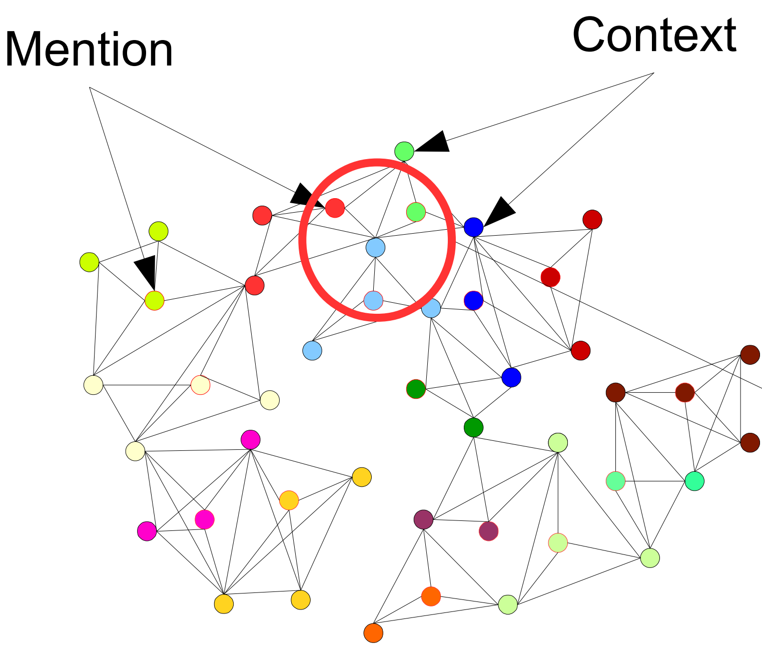
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion
  - Diffusion Strategies



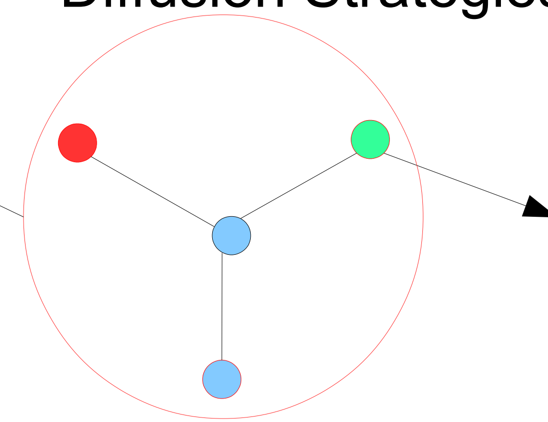
Diffusion based Clustering





Diffusion Clustering

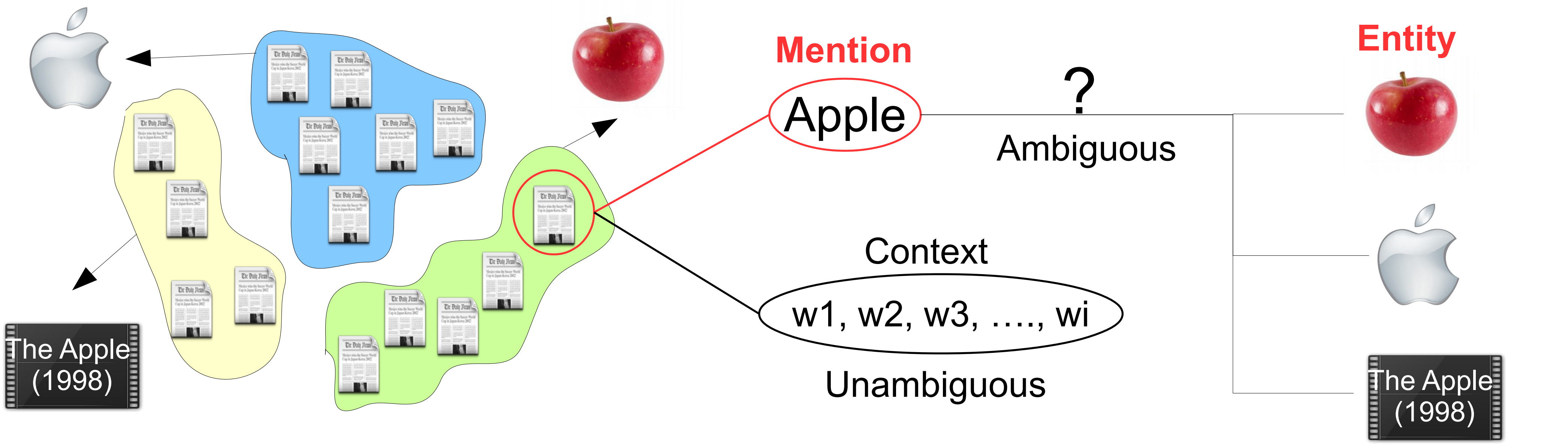
- Initialization
  - Color assignment
- Diffusion
  - Diffusion Strategies

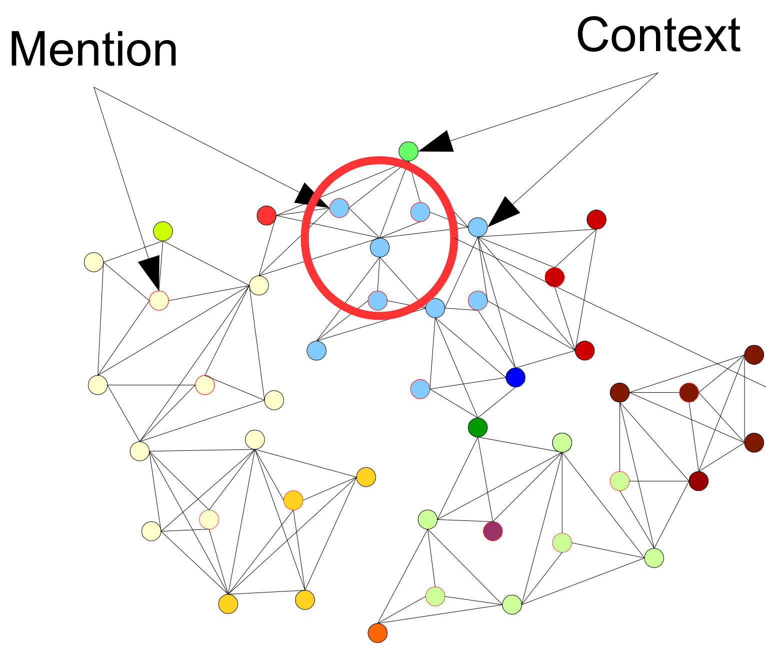


Information:

Own state
Neighbor state
Incoming colors

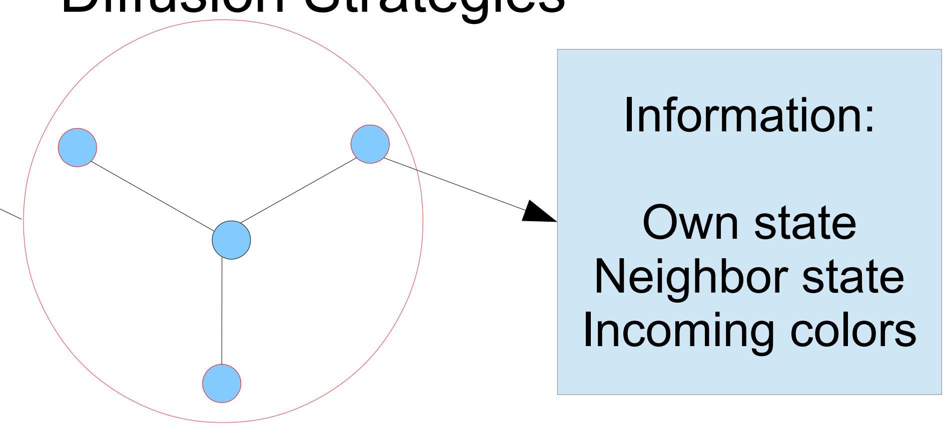
Diffusion based Clustering



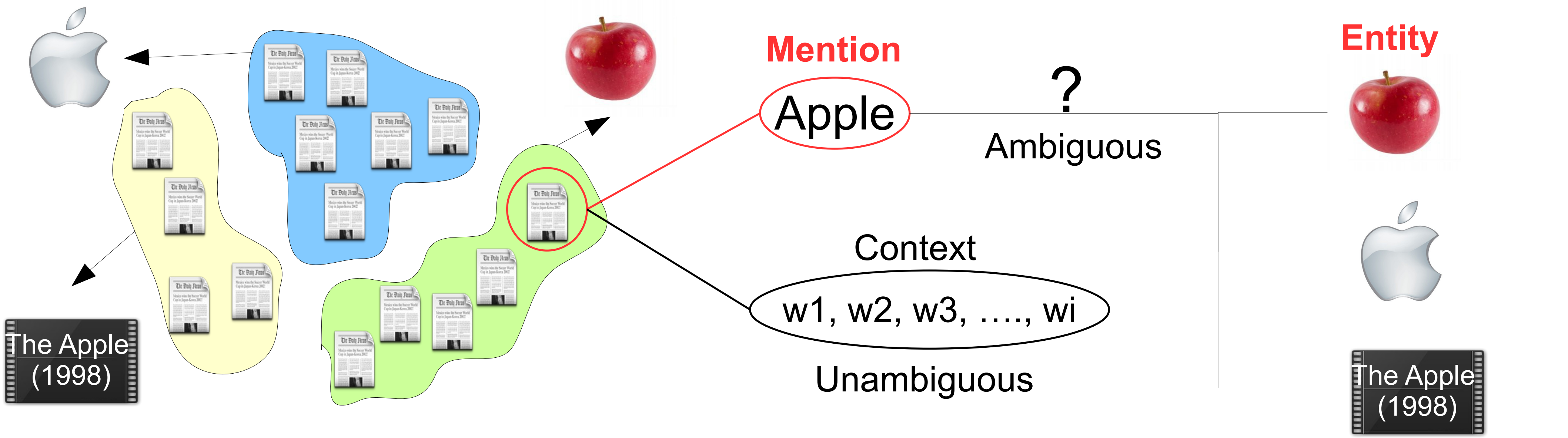


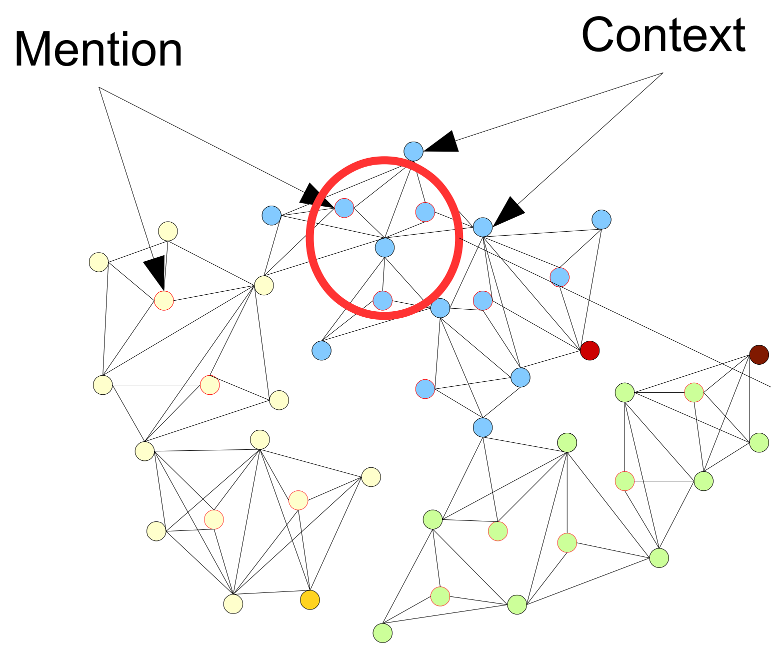
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion
  - Diffusion Strategies



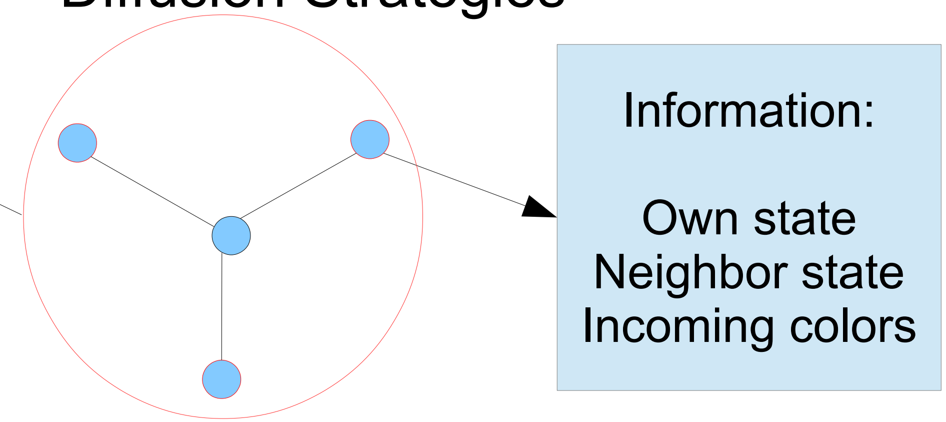
Diffusion based Clustering



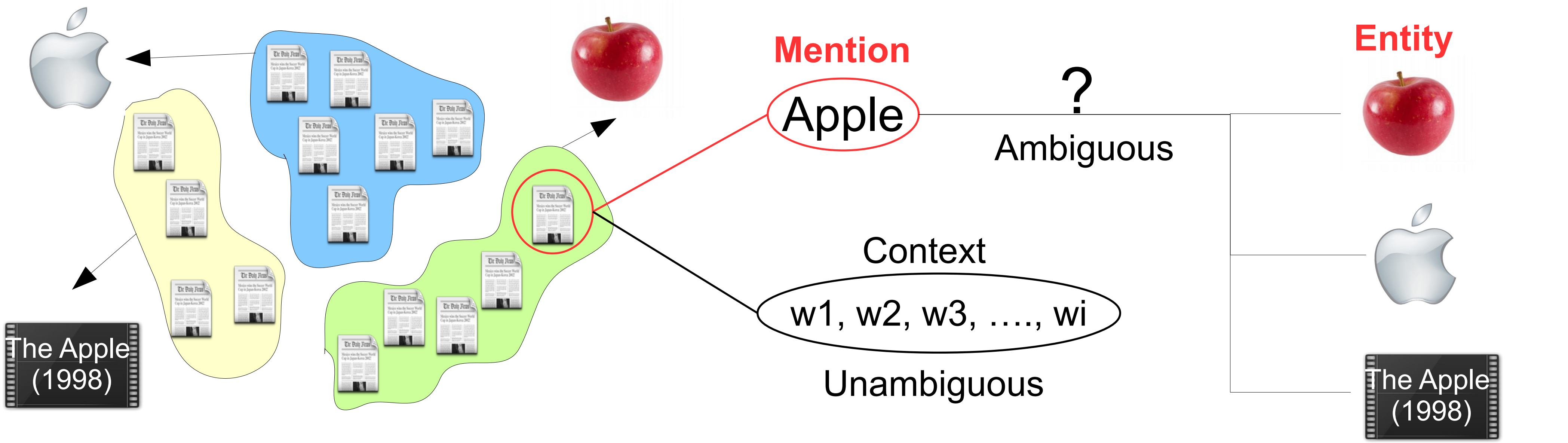


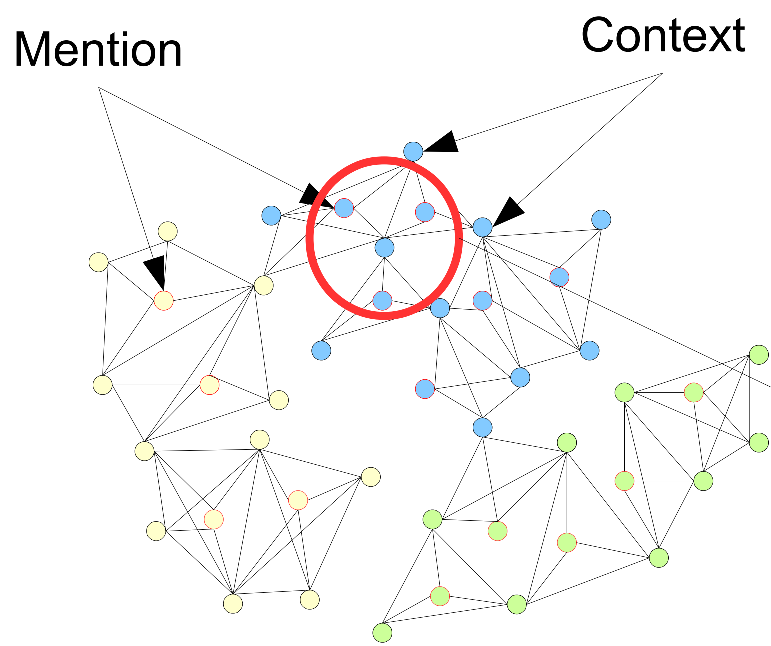
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion
  - Diffusion Strategies



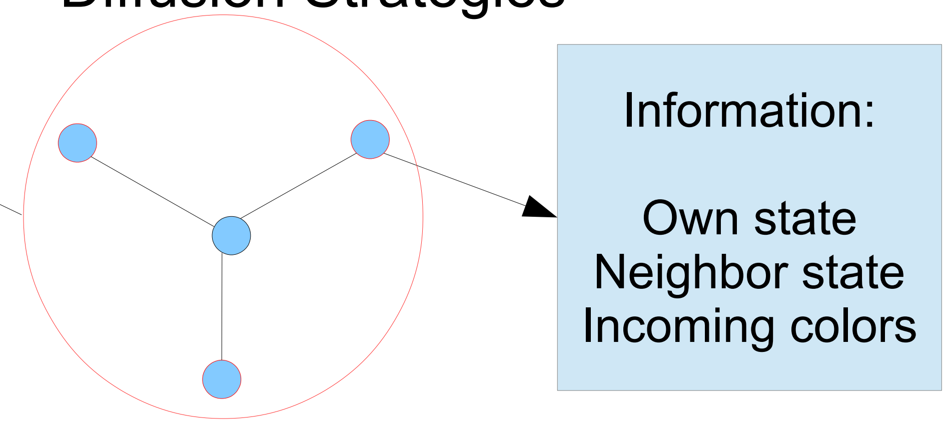
Diffusion based Clustering



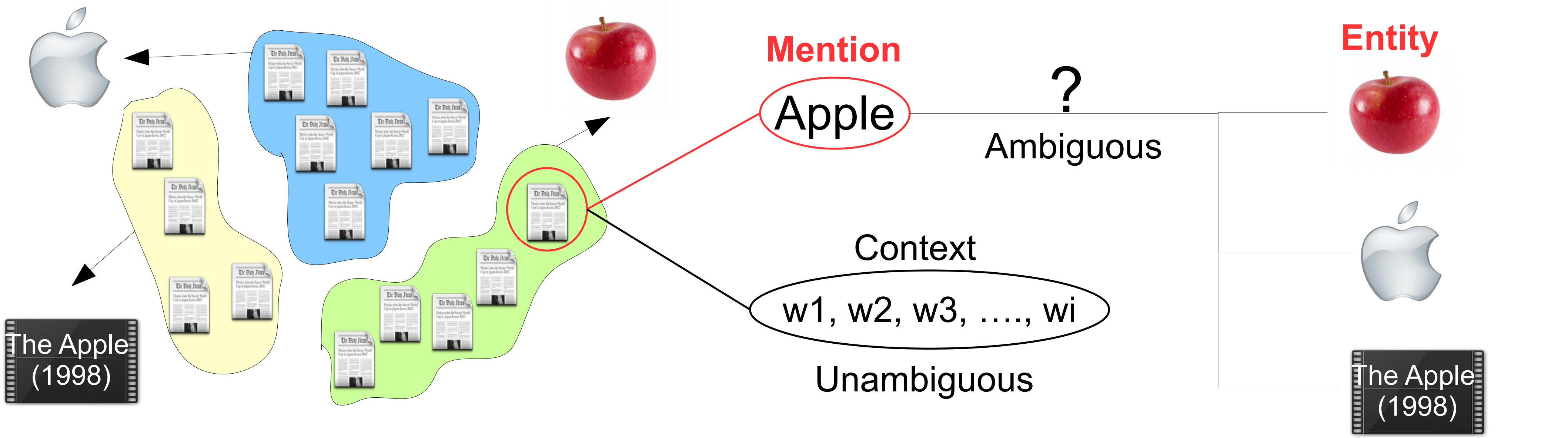


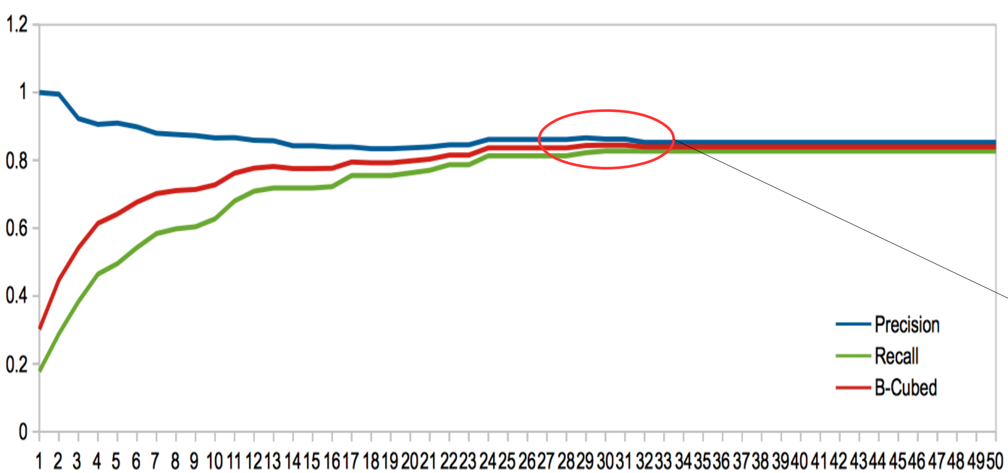
Diffusion Clustering

- Initialization
  - Color assignment
- Diffusion
  - Diffusion Strategies



Results

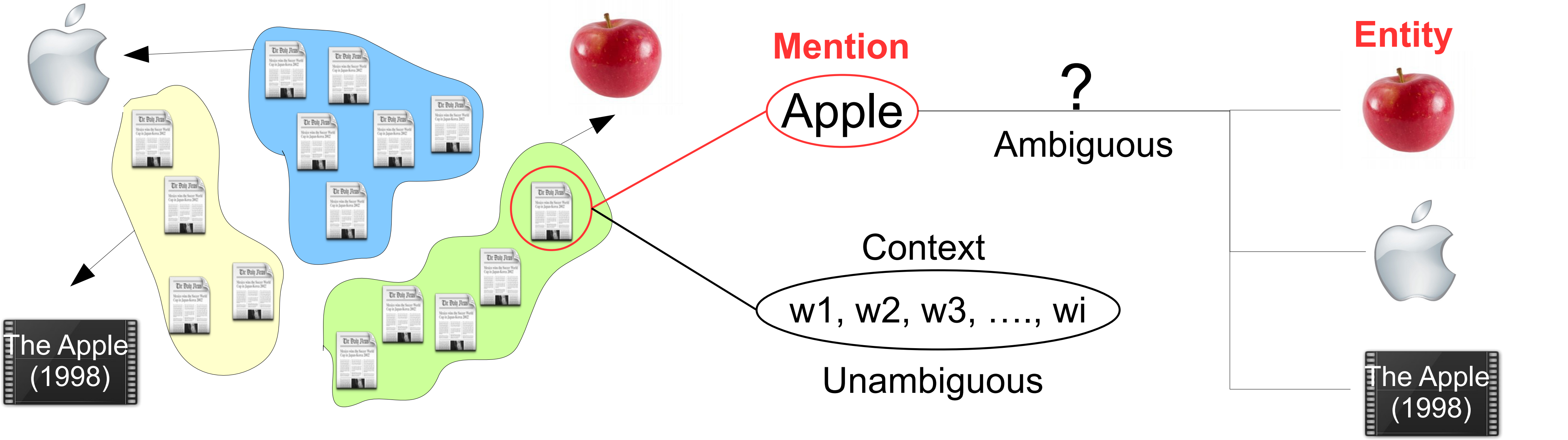


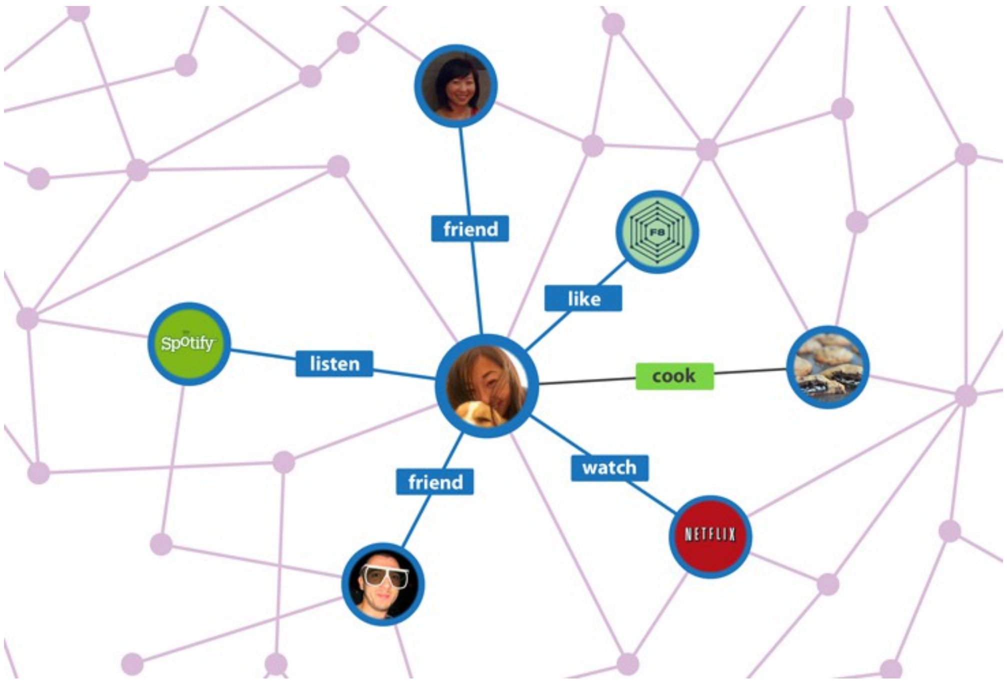


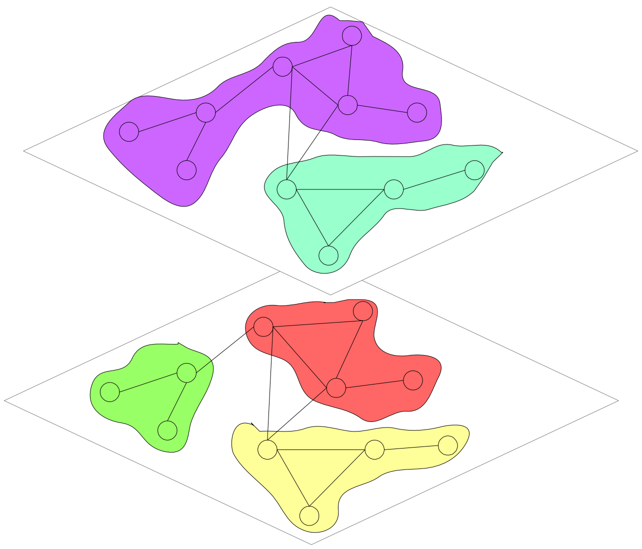
$\alpha = 0, \beta = 0.7, \gamma = 0.96, \delta = 0.05$

Model	Accuracy(F1%)
Bagga & Baldwin	84.6
Google	66.4
Basic	84.5
Extended	**88.9**

What Next?


The Apple
(1998)
Mention
Apple
?
Ambiguous
Context
w1, w2, w3, …., wi
Unambiguous
Entity
The Apple
(1998)


friend
like
listen
cook
friend
watch
Spotify
NETFLIX


Activity
Profile

Thank You!