A Framework for Decreasing Textual Redundancy

Kapil Thadani, Kathleen McKeown
Columbia University

COLING 2008

Redundancy detection

- Identification of text which rephrases or restates information already present in input
- Needed when dataset consists of multiple documents on the same topic
  - eg: News articles, websites
- Common problem for summarization and QA systems
  - Redundant text can increase size of a valid answer or summary without improving information coverage

Common approaches

Clustering

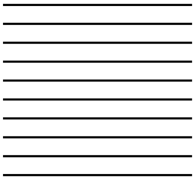
- Often used to detect and remove redundancy

Common approaches

Clustering

- Often used to detect and remove redundancy

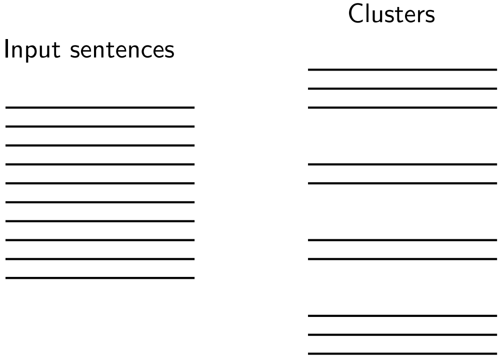
Input sentences



Common approaches

Clustering

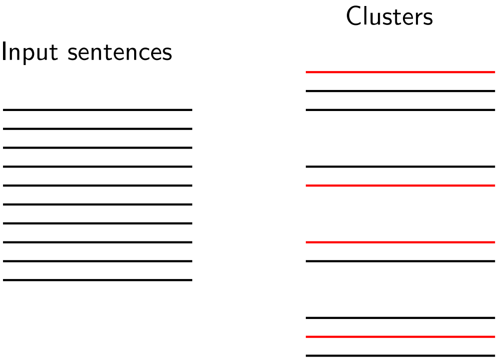
- Often used to detect and remove redundancy



Common approaches

Clustering

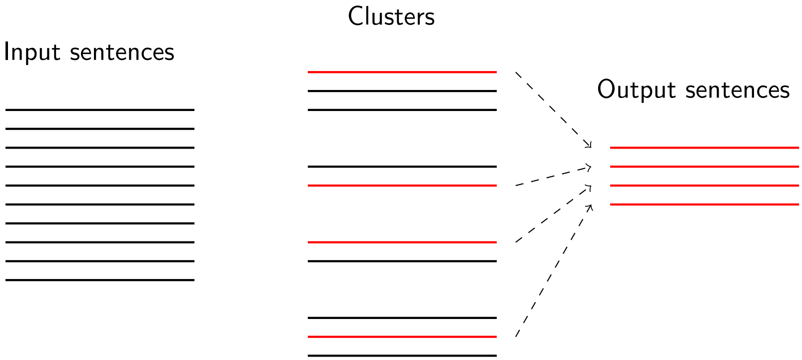
- Often used to detect and remove redundancy



Common approaches

Clustering

- Often used to detect and remove redundancy



Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm

Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm

Input sentences

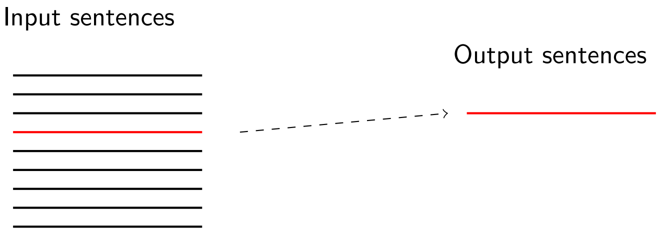
Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm



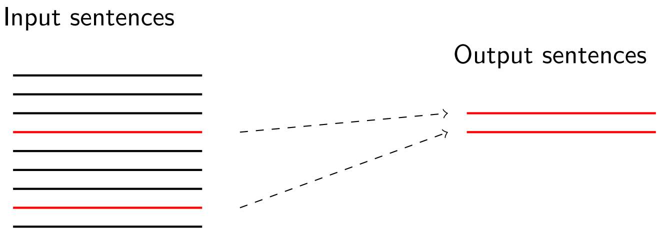
Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm



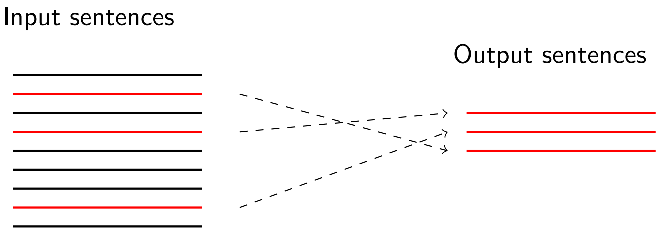
Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm



Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm

These methods:

1. Do not attempt to *preserve all information* in the document
2. Assume redundancy exists *at the sentence level*

Common approaches

Clustering

- Often used to detect and remove redundancy

MMR (Carbonell & Goldstein, 1998)

- Well-known diversity-based reranking algorithm

These methods:

1. Do not attempt to *preserve all information* in the document
2. Assume redundancy exists *at the sentence level*

This work

- Identifies redundancy below the sentence level through alignment
- Introduces bipartite graph representation for tracking repeated information
- Emphasis on preservation of information in a document

Related to:

- Sentence Fusion (Barzilay & McKeown, 2005), which avoids redundancy by fusion of aligned sentences
- Formal model for sentence selection (Filatova & Hatzivassiloglou, 2004), which introduces relationship between information summarization and set cover

Outline

Identifying redundancy

Reducing redundancy

Experiments

Outline

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Redundancy detection

- Sentences contain units of information or *concepts*
  - eg: Whittington, a lawyer, was shot in the chest
- Redundant information observed when other sentences have some similar information
  - eg: Whittington had been shot by Cheney during a quail hunt
- Need to efficiently remove the largest possible number of sentences from the document *without losing any concepts*
- Other considerations:
  - Minimize total number of words in answer (remove longer sentences)
  - Retain higher ranked sentences (given external ranks/weights)
  - Prefer more significant or more *central* sentences
  - Prefer sentences with greater coverage by concepts (more 'focused' sentences)

Terminology

Concept

- Unit of information (fact, opinion, idea)
- As small as it needs to be so that it appears whole in sentences
- No single textual realization; *only seen as a set of nuggets*

Nugget

- Textual realization of a concept in a sentence
- Nuggets for the same concept do not necessarily have the same text

An example: Sentences and concepts

Consider the following sentences:

1. Whittington is an attorney.
2. Cheney shot Whittington, the attorney.
3. Whittington, an attorney, was shot in Texas.
4. Whittington was shot by Cheney while hunting quail.
5. It was during a quail hunt in Texas.

An example: Sentences and concepts

Consider the following sentences:

1. Whittington is an attorney.
2. Cheney shot Whittington, the attorney.
3. Whittington, an attorney, was shot in Texas.
4. Whittington was shot by Cheney while hunting quail.
5. It was during a quail hunt in Texas.

An example: Sentences and concepts

Consider the following sentences:

Whittington was shot

1. Whittington is an attorney.
2. Cheney shot Whittington, the attorney.
3. Whittington, an attorney, was shot in Texas.
4. Whittington was shot by Cheney while hunting quail.
5. It was during a quail hunt in Texas.

An example: Sentences and concepts

Consider the following sentences:

Whittington is an attorney

1. Whittington is an attorney.
2. Cheney shot Whittington, the attorney.
3. Whittington, an attorney, was shot in Texas.
4. Whittington was shot by Cheney while hunting quail.
5. It was during a quail hunt in Texas.

An example: Sentences and concepts

Consider the following sentences:

Cheney was the shooter

1. Whittington is an attorney.
2. Cheney shot Whittington, the attorney.
3. Whittington, an attorney, was shot in Texas.
4. Whittington was shot by Cheney while hunting quail.
5. It was during a quail hunt in Texas.

An example: Concepts

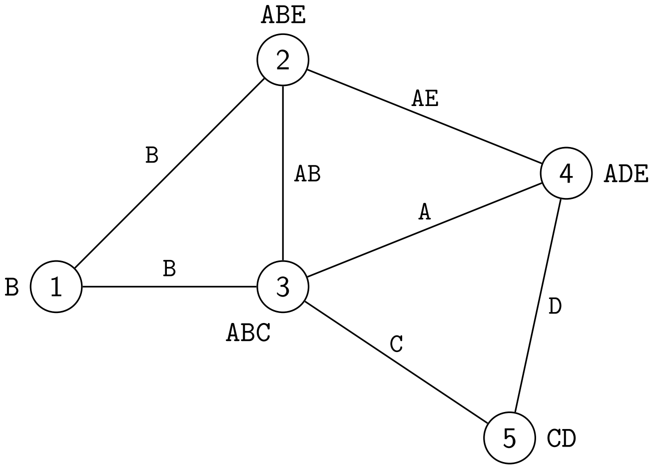
The sentences contain the following concepts:

A	Whittington was shot
B	Whittington is an attorney
C	The shooting occurred in Texas
D	It happened during a hunt for quail
E	Cheney was the shooter

Alignment between sentences

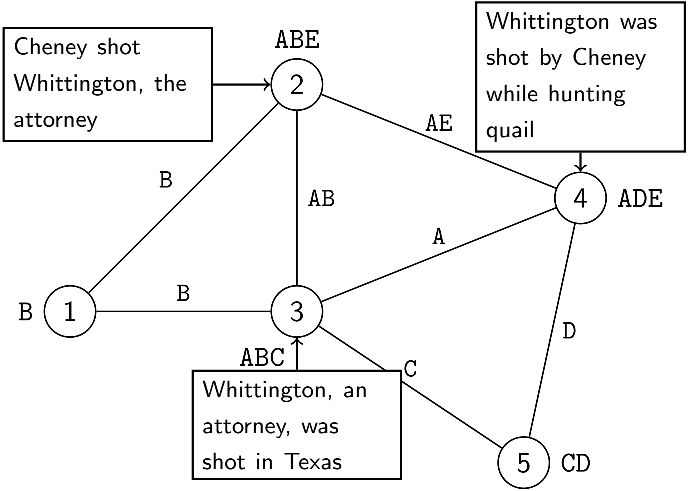
- Need approach that can find nuggets expressing the same concept in two sentences
  - Bag-of-words overlap
  - Substring matching
- Dependency tree alignment:
  - Useful for detecting overlap across non-contiguous segments within sentences
  - Increases overlap precision since syntactic dependencies maintained
  - Normalization techniques to capture further syntactic variation

Pairwise alignments



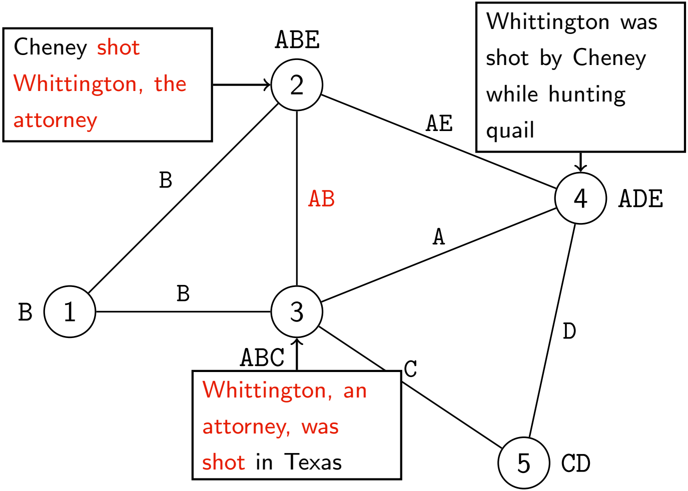
Graph showing pairwise alignments between sentences

Pairwise alignments



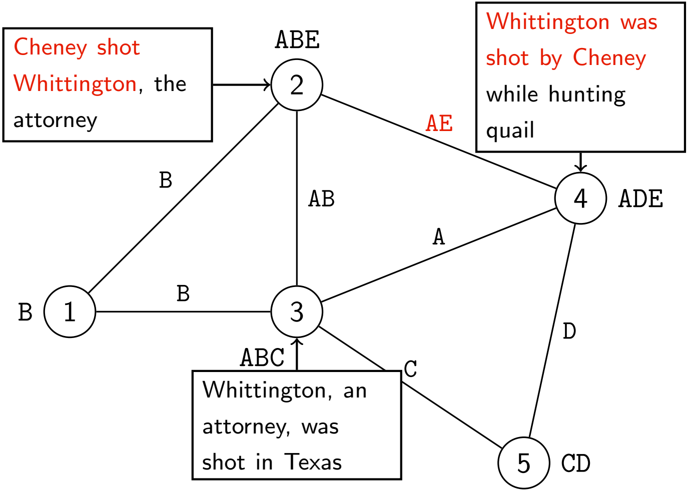
Graph showing pairwise alignments between sentences

Pairwise alignments



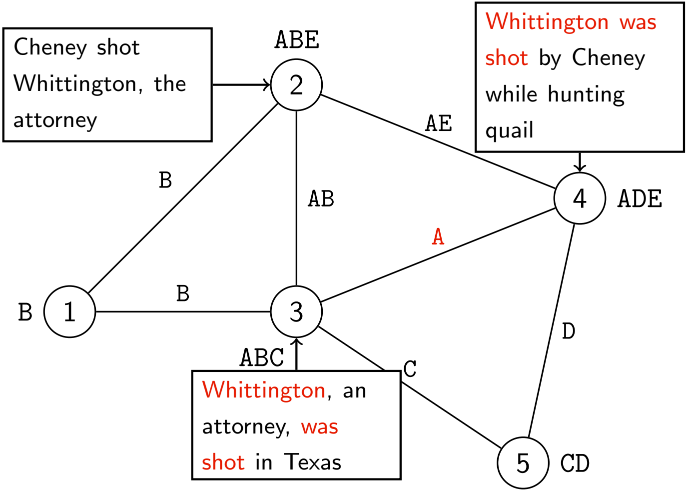
Graph showing pairwise alignments between sentences

Pairwise alignments



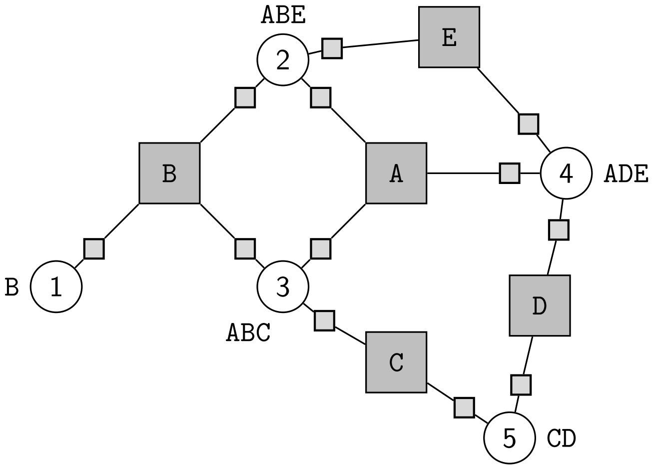
Graph showing pairwise alignments between sentences

Pairwise alignments



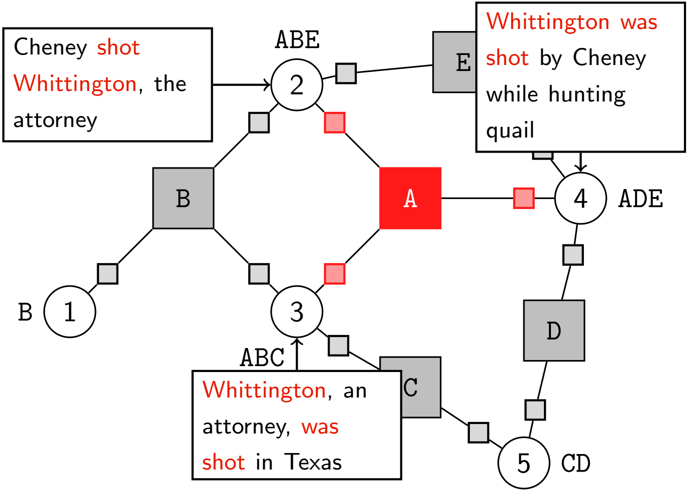
Graph showing pairwise alignments between sentences

Concept graph representation



Structure of the equivalent concept graph

Concept graph representation

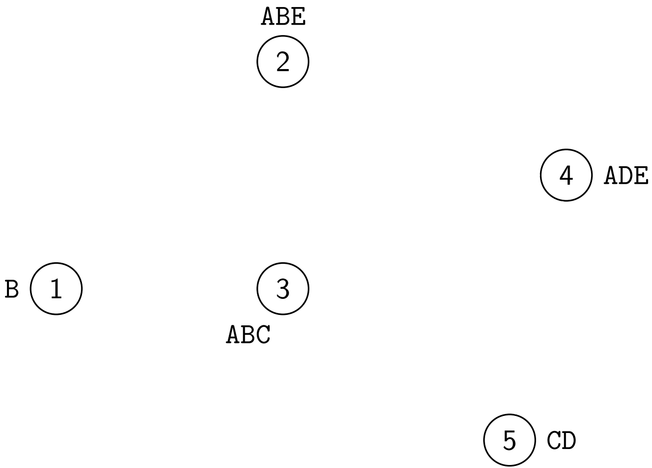


Structure of the equivalent concept graph

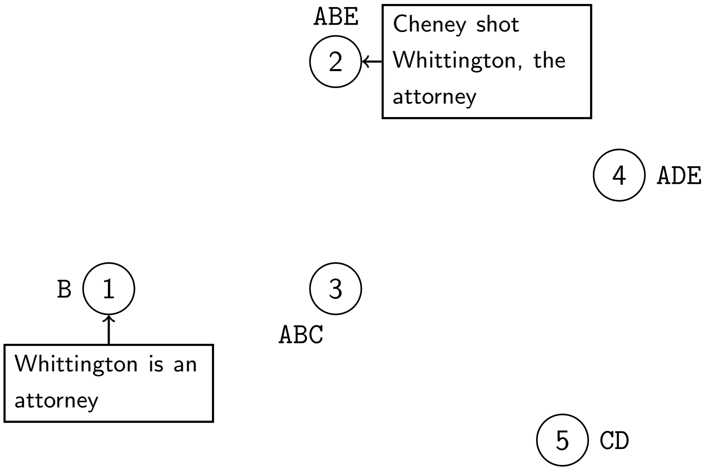
Constructing the graph

- Requires all pairwise alignments between sentences
- Pairwise alignments assumed to be symmetric and transitive
- Exploits graph structure to make construction process efficient
- At every alignment step between a pair of sentences:
  - A pair of newly aligned *fragments* of text may be generated
  - The fragment from one of the sentences must be compared with *all its other nuggets*
  - Comparison determines whether the aligned fragments belong to an existing concept or a new concept

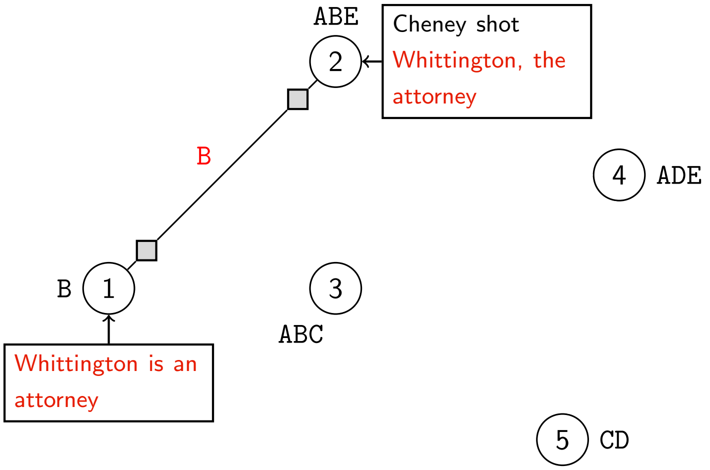
Constructing the graph



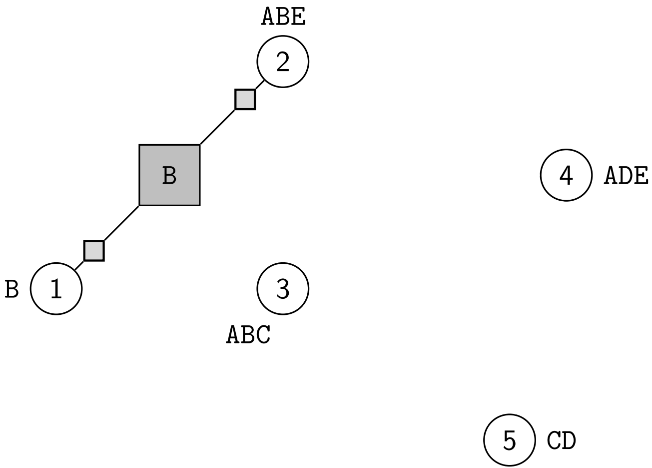
Constructing the graph



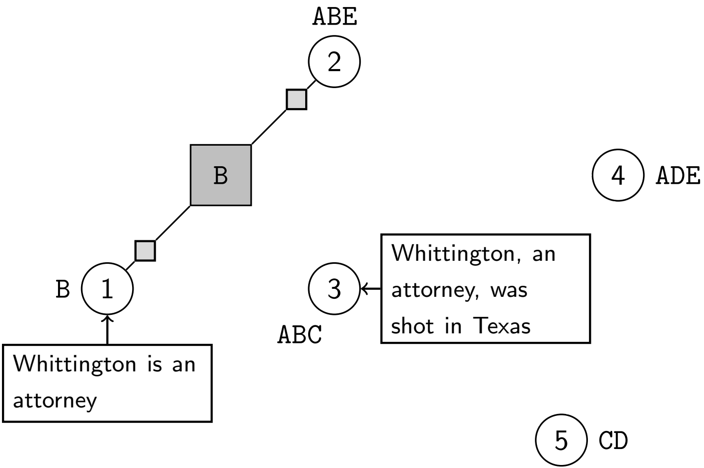
Constructing the graph



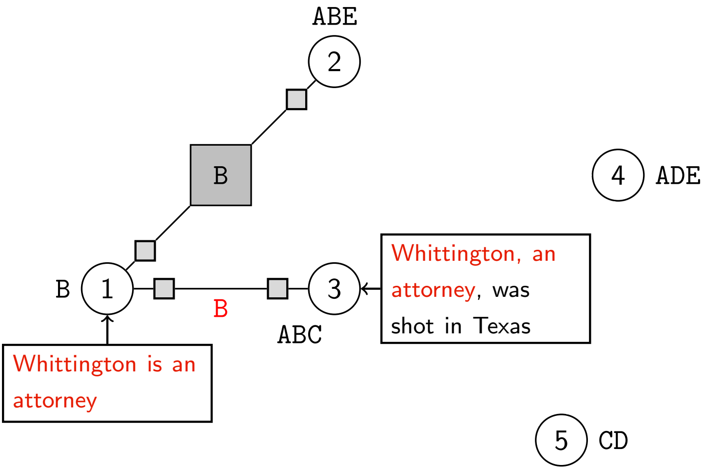
Constructing the graph



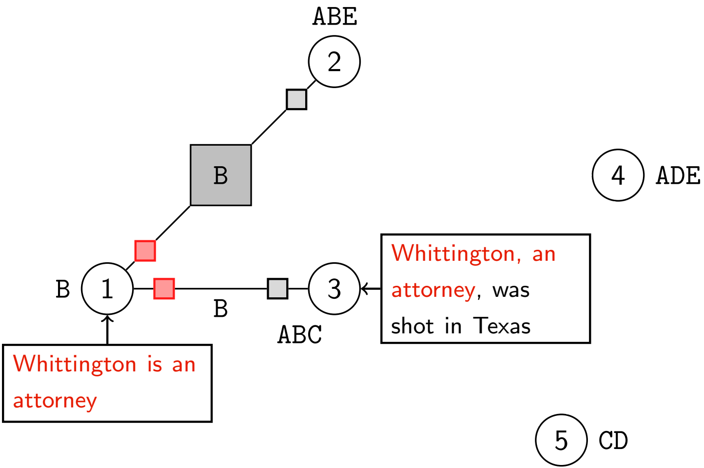
Constructing the graph



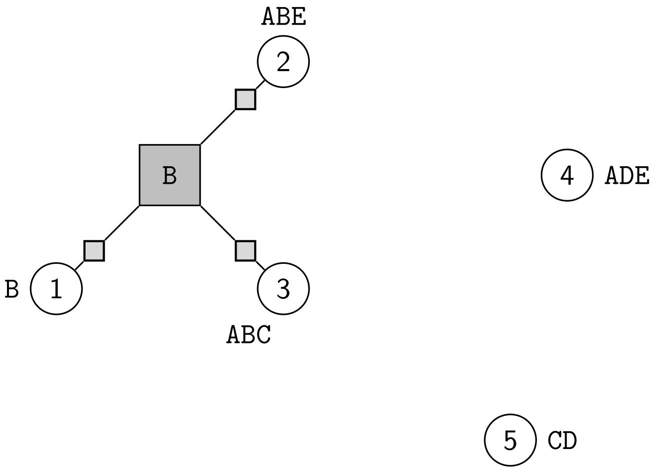
Constructing the graph



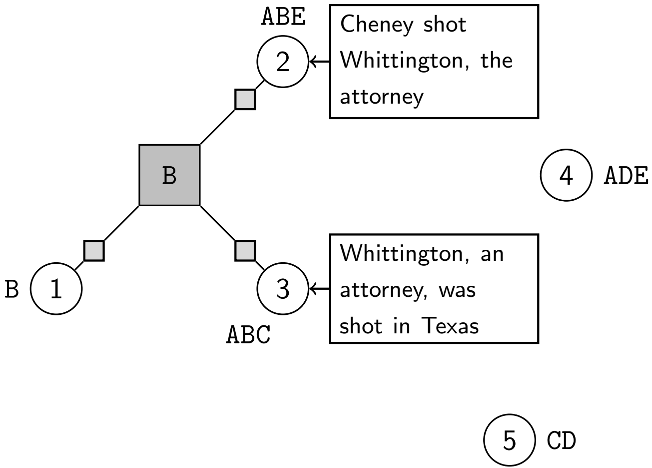
Constructing the graph



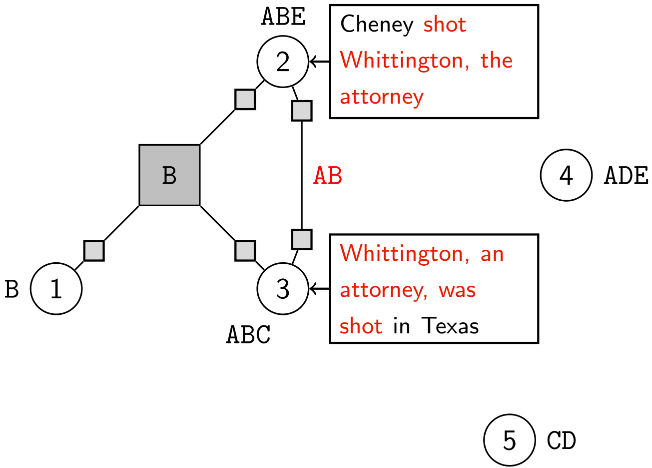
Constructing the graph



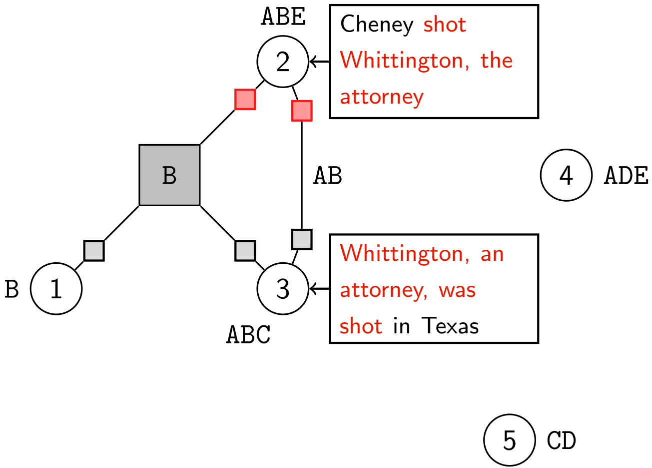
Constructing the graph



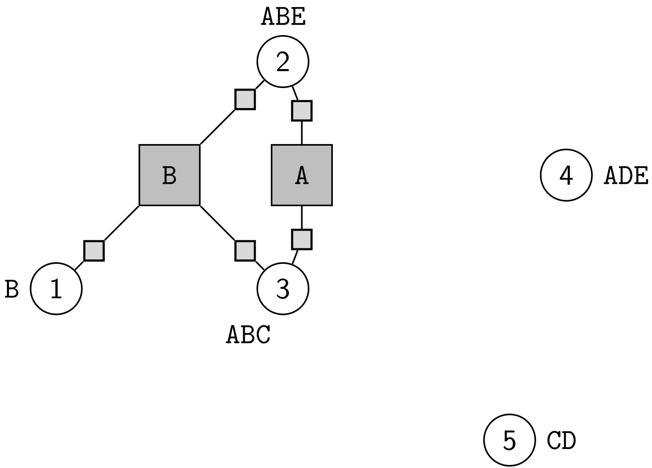
Constructing the graph



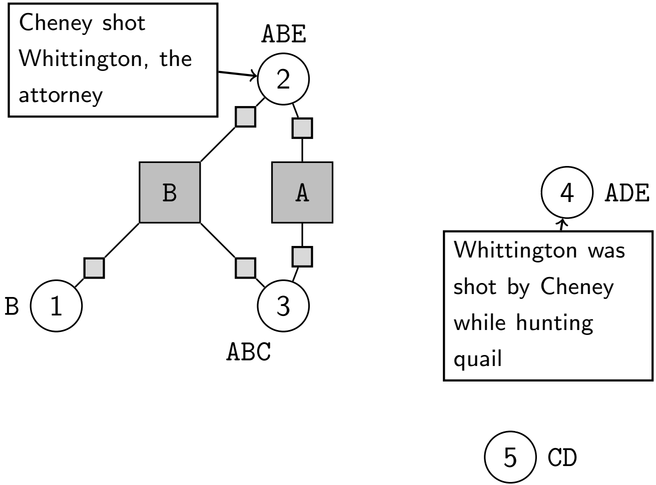
Constructing the graph



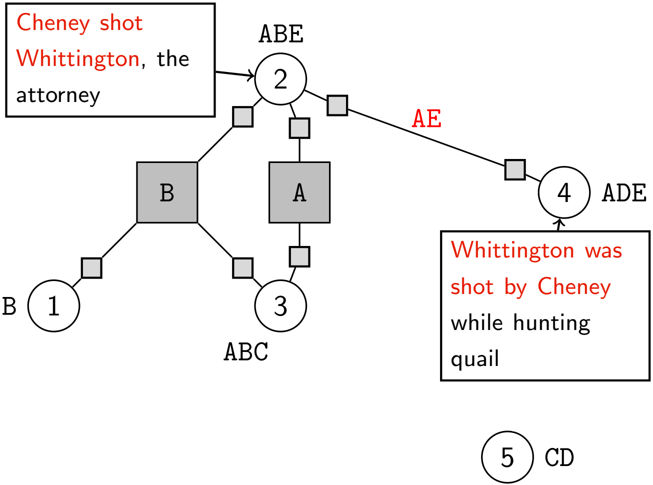
Constructing the graph



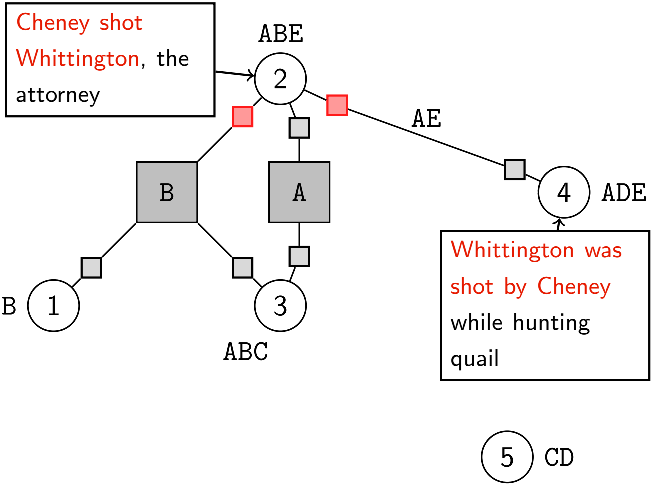
Constructing the graph



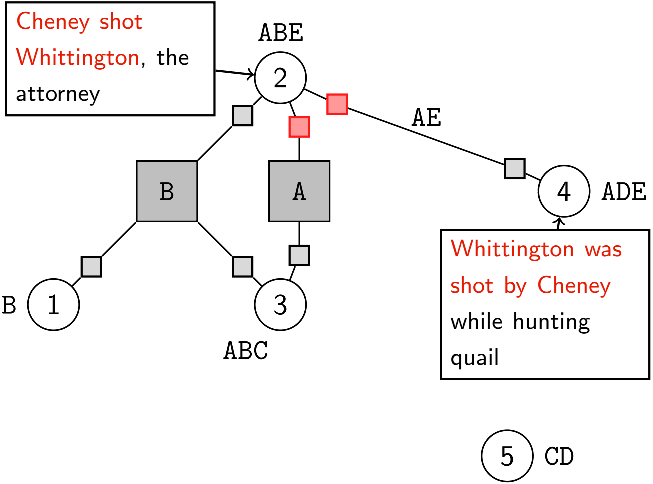
Constructing the graph



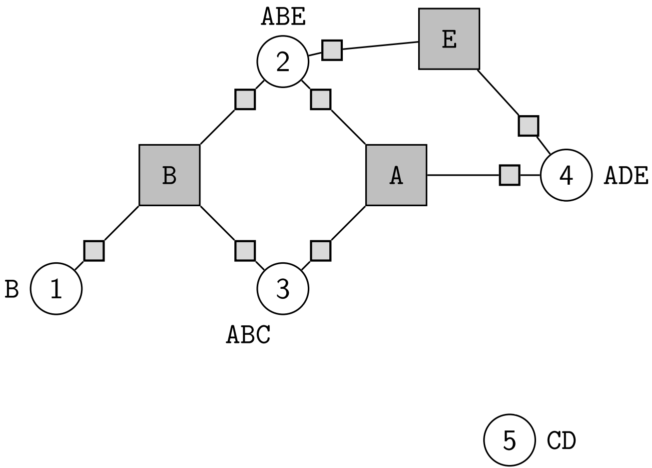
Constructing the graph



Constructing the graph



Constructing the graph



Outline

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

`ABC` represents a sentence where `A`,`B` and `C` represent units of information (concepts)

- `ABC` and `BC`
- `AB`, `BC` and `AC`
- `ABC`, `A` and `BCD`

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

`ABC` represents a sentence where `A`,`B` and `C` represent units of information (concepts)

- `ABC` and `BC`
- `AB`, `BC` and `AC`
- `ABC`, `A` and `BCD`

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

`ABC` represents a sentence where `A`,`B` and `C` represent units of information (concepts)

- `ABC` and `BC`
- `AB`, `BC` and `AC`
- `ABC`, `A` and `BCD`

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Some cases

ABC represents a sentence where A,B and C represent units of information (concepts)

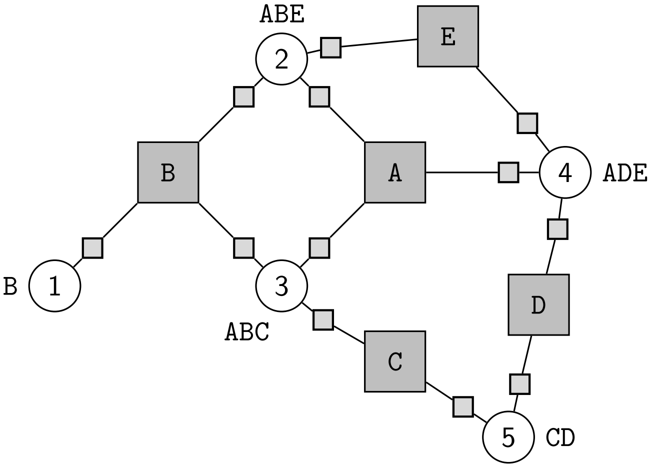
- ABC and BC
- AB, BC and AC
- ABC, A and BCD

Set cover

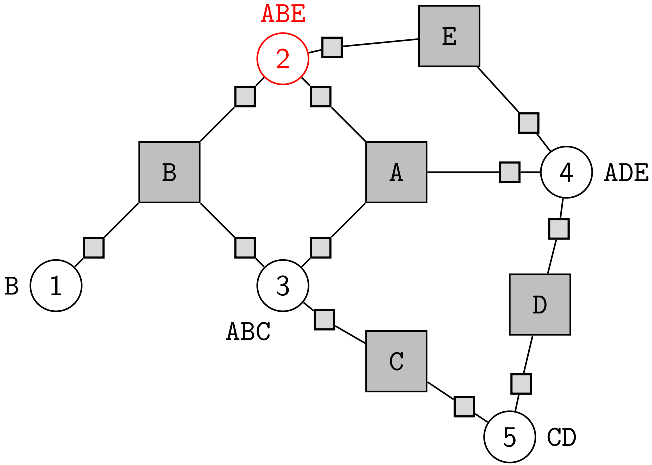
Want to find the smallest set of sentences that cover all concepts

- Reduces to *minimum set cover* which is NP-hard (Filatova & Hatzivassiloglou, 2004)
- Other considerations such as sentence length, ranking can be accounted for by assigning weights
- Greedy approximation algorithm exists for weighted set cover (Hochbaum, 1997)
- Best known polynomial time approximation algorithm; can be used with our representation

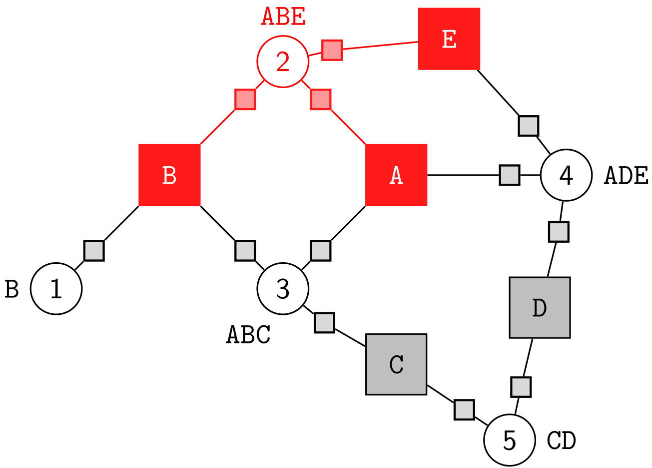
Identifying redundant sentences



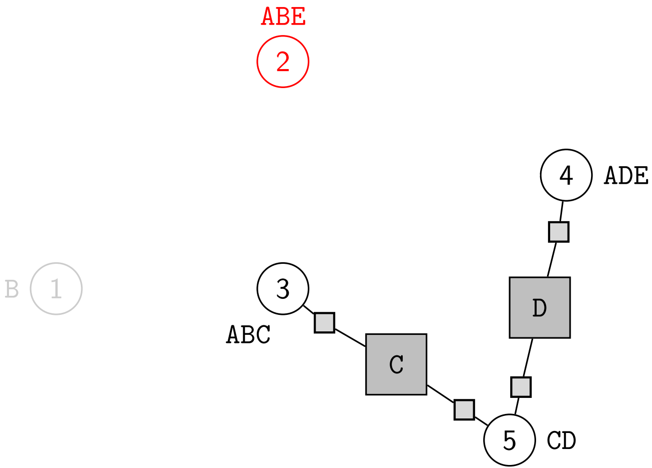
Identifying redundant sentences



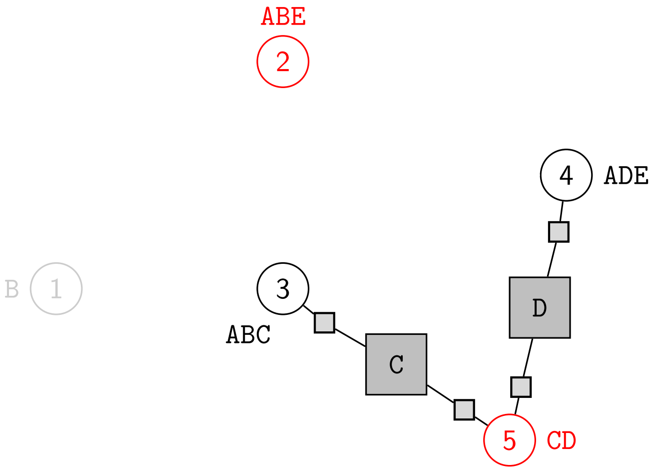
Identifying redundant sentences



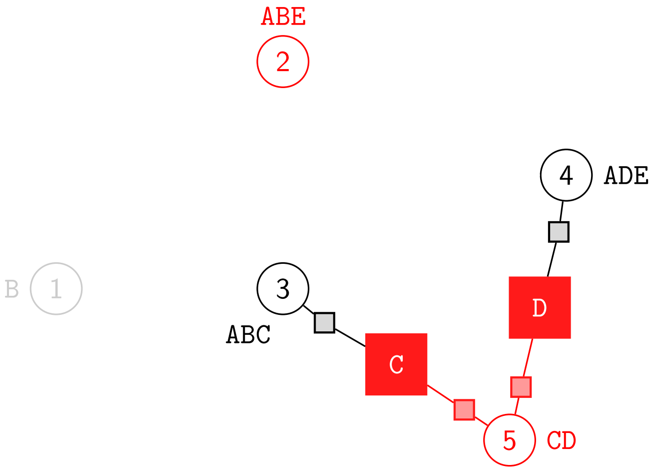
Identifying redundant sentences



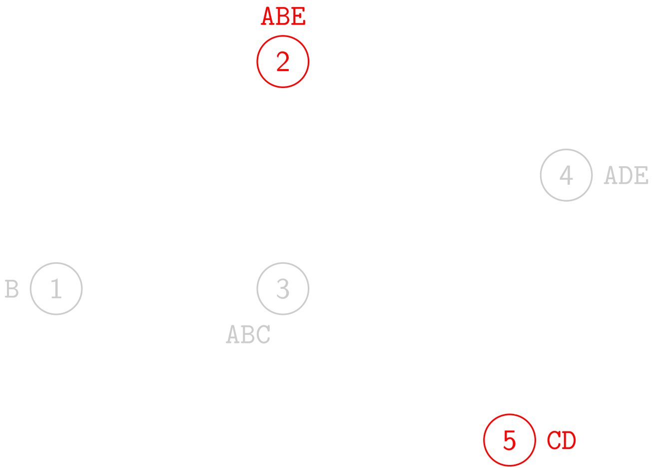
Identifying redundant sentences



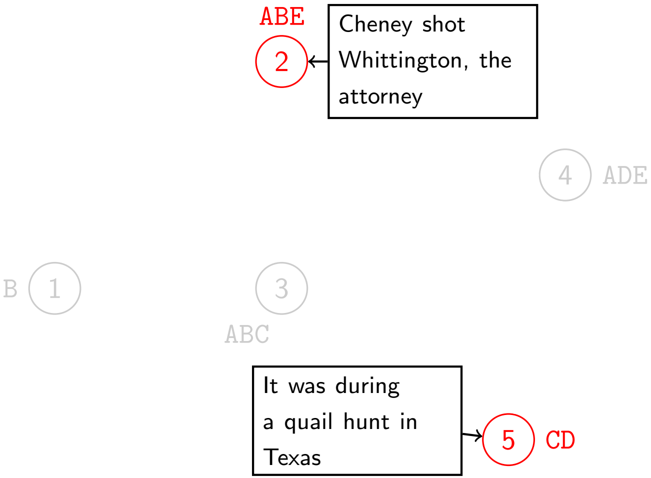
Identifying redundant sentences



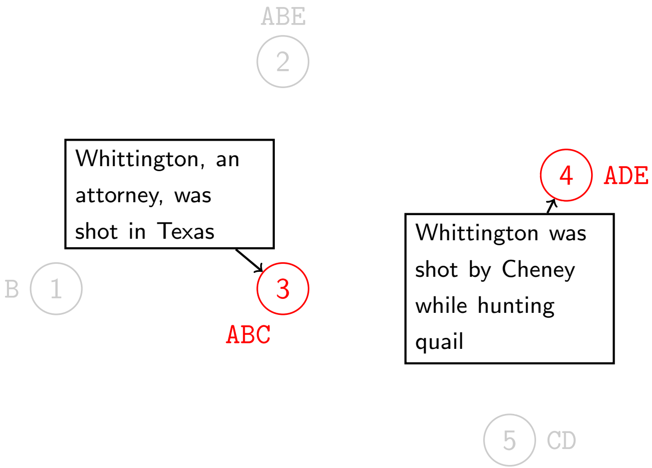
Identifying redundant sentences



Identifying redundant sentences



Identifying redundant sentences



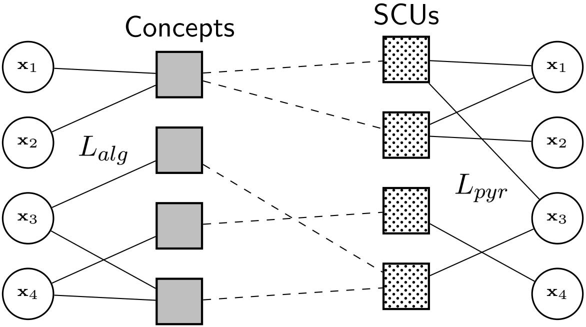
Outline

Dataset

- Experiments test quality of graph construction algorithm
- Pyramid data from DUC 2005 (Nenkova et al, 2007)
  - 20 documents (1941 sentences)
  - Each has 7 human-generated summaries of the same news article (lots of redundancy)
  - Human-annotated *semantic content units* or SCUs → concepts
  - Contributors for each SCU from the summaries → nuggets

Evaluation metrics

- Concepts are mapped to SCUs by calculating the *longest common subsequence* between nuggets (from the concept) and contributors (from the SCU)



Evaluation metrics

- Metrics draw on well-known IR measures of precision, recall, F-measure

$$\begin{aligned} \text{Precision} &= \frac{L_{alg} \cap L_{pyr}}{L_{alg}} \\ \text{Recall} &= \frac{L_{alg} \cap L_{pyr}}{L_{pyr}} \end{aligned}$$

- F_1 score is their unweighted harmonic mean

Evaluation results

Focused random baseline

- Statistics drawn from distributions of corresponding gold-standard concept graphs (number of concepts, number of concepts per sentence)
- Best scores from 100 runs per document considered

Measure	Random
Precision	0.0510
Recall	0.0515
F_1 score	0.0512

Evaluation results

Clustering approach

- Based on spectral partitioning (Shi & Malik, 2000)
- Each cluster forms a concept
- Parameter to control recursion depth swept over; clustering configuration with maximum F_1 score considered

Measure	Random	Clustering
Precision	0.0510	0.2961
Recall	0.0515	0.1162
F_1 score	0.0512	0.1669

Evaluation results

Concept graph approach

- Dependency tree alignment used; trees generated by MINIPAR (Lin, 1998)

Measure	Random	Clustering	Concepts
Precision	0.0510	0.2961	**0.4496**
Recall	0.0515	0.1162	**0.3266**
F_1 score	0.0512	0.1669	**0.3783**

Per-document results

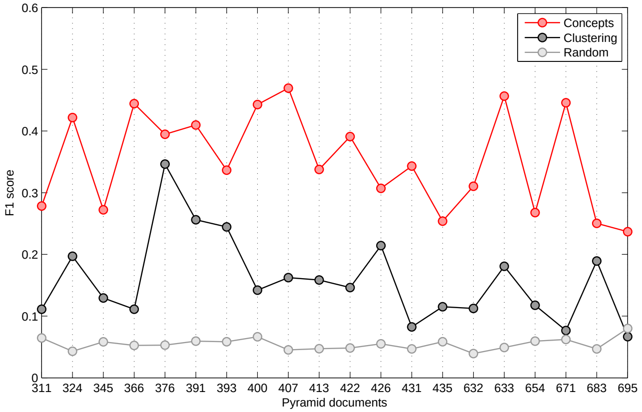


Figure: F_1 scores over each document

Conclusion

- Common information in sentences uncovered through pairwise alignment
- Concept graph representation tracks repeated information in document
- Set cover approximation algorithm used to reduce redundancy

Further directions

- Synthesis of new non-redundant sentences along the lines of sentence fusion (Barzilay & McKeown, 2005)
- Support for unidirectional redundancy to be identified through entailment approaches

Questions?

Unique information

- In real-world documents, sentences can have unique information that never aligns with other sentences
- These can't be selected as redundant (unless assumed irrelevant)
- Set cover algorithm should select these first
  - Covers information that would end up in output anyway
- Need a more principled approach to minimizing effect of unique information; perhaps along the lines of fusion (Barzilay & McKeown, 2005)

Concept membership

At every alignment step, for the first sentence in the alignment:

- Need to compare the fragment uncovered in the alignment with existing nuggets
- Comparison based on word-indices; efficient
- Every comparison yields three sets of words
  - Words that are common between fragment and nugget: $w_{F \cap N}$
  - Words that occur only in the fragment: w_F
  - Words that occur only in the nugget: w_N
- If w_N is *significant*, it becomes a new concept
  - First recursively compared with other nuggets
- If $w_{F \cap N}$ is significant, and w_F is not $\longrightarrow$ fragments belong to concept of that nugget
- If both $w_{F \cap N}$ and w_F are significant $\longrightarrow$ existing concept contains multiple units of information; should be split up

Splitting up concepts

- Concepts are effectively a collection of mappings between participating nuggets
- Must be able to split them up
  - Only maintain mappings of meaningful words (higher-*idf*)
  - Non-meaningful words (auxiliaries, determiners, etc) accompany their parents in dependency structure
  - Meaningful words can appear in *both* new nuggets
    - eg: subjects/objects of propositions, nouns for adjectives
  - Approaches can vary depending on linguistic information available

Nugget restriction examples

Fragments that consist of only the following:

- Proper names (from NER)
- Propositions with unresolved pronouns, demonstratives
- Strings of stem words (low-*idf*)
- Solitary words (excluding numbers)

are not *significant* nuggets and cannot form concepts by themselves.

More per-document results

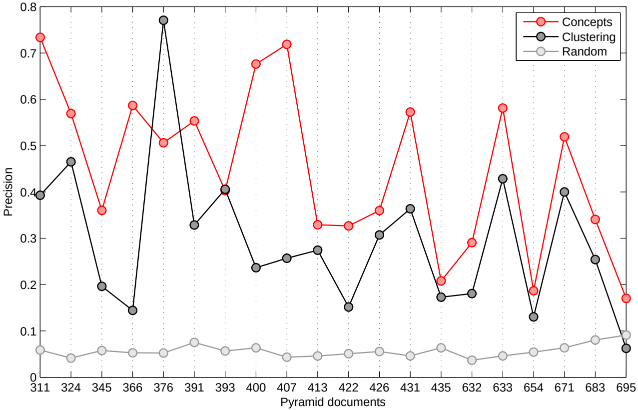


Figure: Precision over each document

More per-document results

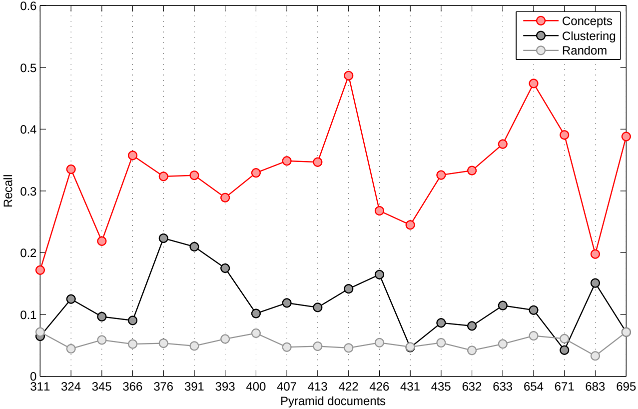


Figure: Recall over each document

Analysis

Example: Concepts from D376

“Albanian-laid mines”, “international tribunal”, “Libya brought case”, “stop acts of genocide”, “decisions carry diplomatic weight”
“It ordered”, “decisions”, “two”, “Court”, “1989”, “enforcement powers”

Table: Variation in quality of concepts detected

Analysis

Example: Partially-redundant sentences from D376

{*The Court does not have the powers to enforce its decisions,*} but they usually {*carry diplomatic weight*} {*The court also considered*} reciprocal Bosnian-Serbian {*accusations*} of genocide
Military disputes {*are*} very common cases {*It heard*} US appeals for release of hostages held by Iran Sixteen {*permanent judges*} preside in the Peace Palace

Table: Variation in quality of whole nuggets detected

Chunks & citations

SENTENCE: Whittington, the attorney and political figure, was shot by the Vice President.	CHUNK: Whittington, the attorney ... CITATION: Harry Whittington is an American lawyer.
	CHUNK: Whittington ... was shot ... CITATION 1: Whittington was shot in the chest during a quail-hunting trip. CITATION 2: Whittington was shot in the chest by Dick Cheney.
SENTENCE: {*Whittington was shot*} in the chest during a quail-hunting trip.	CHUNK: {*Whittington was shot*} in the chest ... CITATION: {*Whittington was shot*} in the chest by Dick Cheney.
	CHUNK: ... a quail-hunting ... CITATION: The incident occurred during a quail hunt.

Figure: An example of the representation

Real examples

- On February 11, 2006, Whittington, a Bush-Cheney campaign contributor, was accidentally shot and injured by U.S. Vice President Dick Cheney during a quail hunting trip, at a ranch in south Texas owned by Katharine Armstrong. (Wikipedia on *Harry Whittington*)
- On February 11, 2006, U.S. Vice President Dick Cheney accidentally shot Harry Whittington, a 78-year-old Texas attorney, while participating in a quail hunt on a ranch in Kenedy County, Texas. (Wikipedia on *Dick Cheney shooting incident*)
- It's never a good thing to be a punch line in politics, and the vice president had the field to himself after accidentally shooting his hunting companion, Austin lawyer Harry Whittington, at a Texas ranch late Saturday. (Washington Post)
- A Texas attorney remains in intensive care after being shot during a weekend hunting trip with Vice President Dick Cheney. (Time Magazine)

Spectral partitioning

- Create affinity matrix $\mathbf{A}$ through pairwise comparisons; each element $\mathbf{a}_{ij} = \mathbf{a}_{ji}$ is the IDF-weighted cosine similarity of overlapping stems from sentence i and sentence j
- Build degree matrix $\mathbf{D}$ such that $\mathbf{d}_{ii} = \sum_j \mathbf{a}_{ij}$ and $\mathbf{d}_{ij} = 0, i \neq j$
- Compute stochastic matrix $\mathbf{D}^{-1}\mathbf{A}$ or Laplacian $\mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-\frac{1}{2}}$
- Take second eigenvector of this matrix and sort it (eigengap) to get an ordering of sentences
- Compute normalized cut between every split of this ordering and partition at point of minimum normalized cut

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

where $assoc(A,V) = \sum_{a \in A, v \in V} w(a,v)$
- Use *cluster depth* parameter to control recursion depth

Assumptions

- Redundancy not necessarily bidirectional
  - Units of information may be more general or specific variants (eg: gun vs shotgun)
  - Specific details may be irrelevant
- Require full knowledge of relevance + entailment recognition
- Constrain problem with two assumptions:
  1. All information in the document is relevant and must be preserved
  2. General information (at a lower level of granularity) cannot be inferred from more specific information