Semi-Supervised Multiple Cross-Document Coreference Resolution

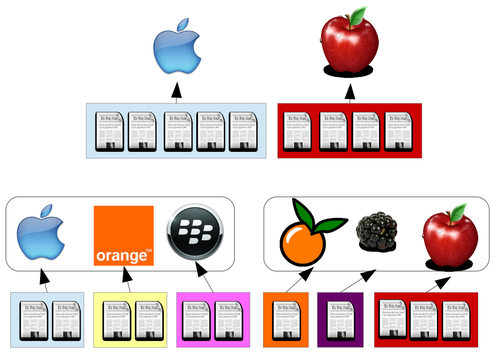
Kambiz Ghoorchian, Fatemeh Rahimian, Sarunas Girdzijauskas

ghoorian@kth.se, rahimian@kth.se, sarunasg@kth.se

Problem

- Cross Document Coreference resolution (CDCR): categorization of documents based on co-reference relation.
- Single Disambiguation [1]: Single Mention referring to multiple Entities.
  - Lack of documents containing a specific mention.
  - Inefficiency - time and processing
- Multiple Disambiguation: Multiple mentions referring to multiple entities (News feeds, Blog posts).
  - Semantic overlap due to hierarchical structure of entities.
  - Increases coherency and noise in aggregated documents.
  - Lower clustering resolution compensated by supervised sampling.

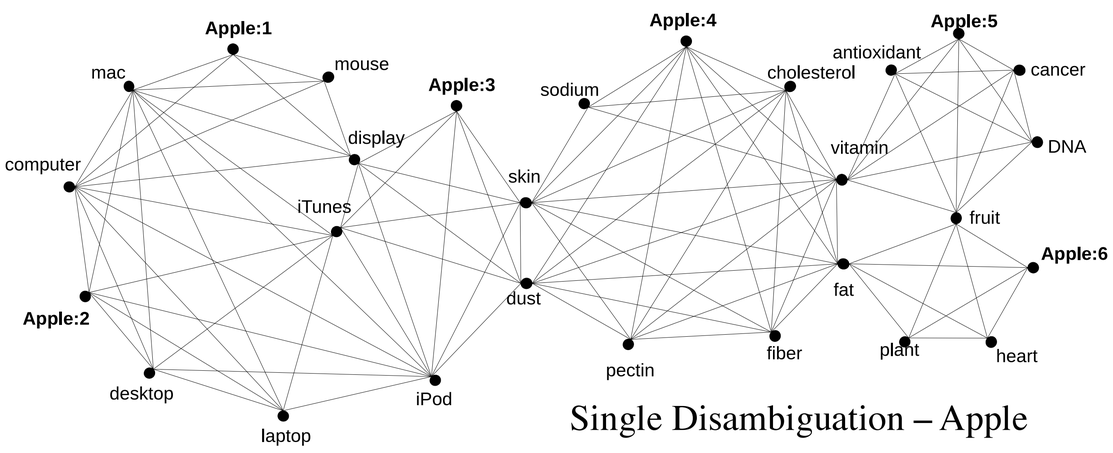
Single vs Multiple Disambiguation



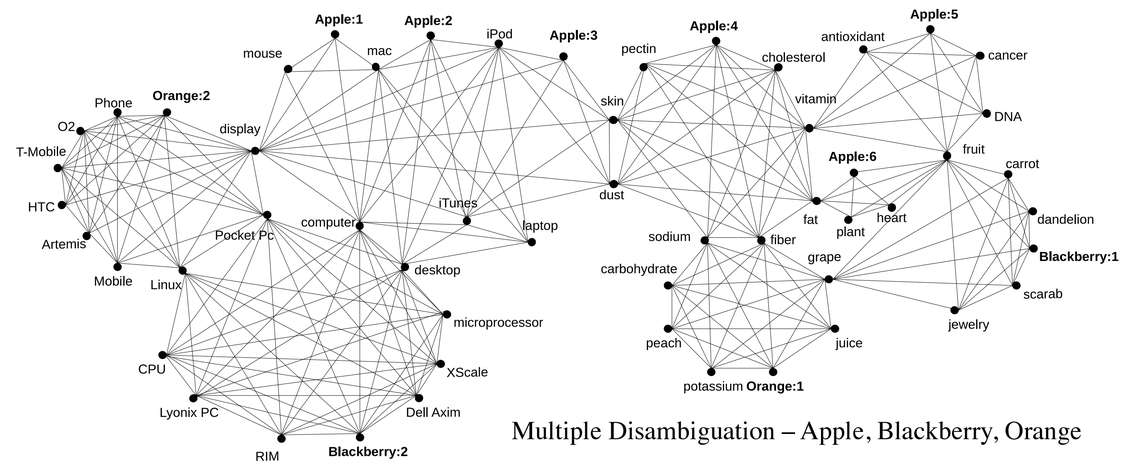
Graph Construction

1. Assign a unique vertex to each ambiguous word (Mention - "Apple").
2. Assign a single vertex to each unambitious word (Context Words - "iPod", "Fiber", ...).
3. Connect vertices based on their co-occurrence in the same document.
4. Multiple disambiguation is similar except that in this case all the ambiguous mentions ("Apple", "Blackberry" and "Orange") are included.

- Apple sells a variety of computer accessories for Macs, including Thunderbolt display and Magic mouse.
- Apple designs and creates iPod, iTunes, Mac laptop and Desktop computers.
- Apple skin protects your iPod. It also fits your iTunes and is enhanced with anti dust treatment. It gives sharper look to the display.
- Apple contains no fat, sodium or cholesterol and is a good source of fiber. Its skin has the highest concentration of vitamins and is mainly covered with pectin and dust.
- Like many fruits, Apple contains vitamins as well as other antioxidants, which may reduce the risk of cancer by preventing DNA damage.
- The Apple has the most diverse fruit plants in the world. It was found to increases the burn of fat and reduces the risk of heart attack.
- Grapes and peaches contain carbohydrates and dietary fiber in moderate amounts. Sodium levels in their juice are high compared to an Orange, and potassium content is moderate.
- Motifs such as wild carrots, dandelions, and fruits like grape and Blackberry were quite common on Tiffany's designs. The scarab theme was also used quite frequently as a decorative motive in his jewelry items.
- XScale microprocessors can be found in products such as the popular RIM BlackBerry hand-held, the Dell Axim family of Pocket PCs. It is used as the main CPU in the Iyonix PC desktop computer running Linux.
- Artemis is a Linux Mobile 5.0 Pocket PC Phone edition, manufactured by HTC. The device is sold by mobile phone operators Orange, O2 and T-Mobile. The latter version lacks touch display.



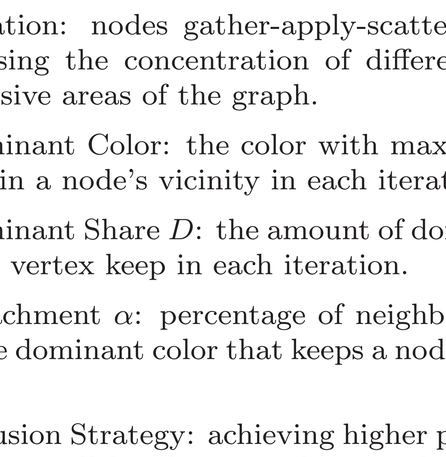
Single Disambiguation – Apple



Multiple Disambiguation – Apple, Blackberry, Orange

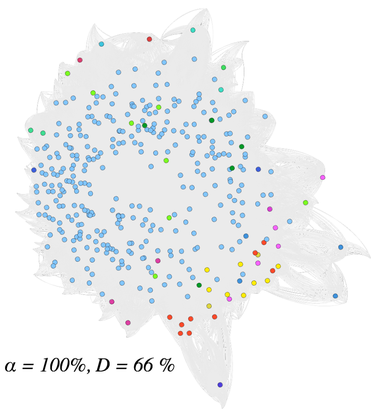
Diffusion Algorithm

- Initialization: each document and its context words receive a unique color.
- Iteration: nodes gather-apply-scatter colors, increasing the concentration of different colors in cohesive areas of the graph.
- Dominant Color: the color with maximum quantity in a node's vicinity in each iteration.
- Dominant Share D: the amount of dominant color each vertex keep in each iteration.
- Attachment α: percentage of neighbors with the same dominant color that keeps a node in the cluster.
- Diffusion Strategy: achieving higher precision and lower recall by changing the Attachment α from 100% to 75%. Lower overall quality of the results compared to the original model that is compensated using supervised sampling of a small portion of the documents (4%).

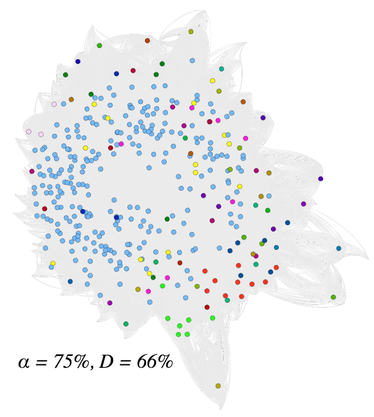


Ground Truth

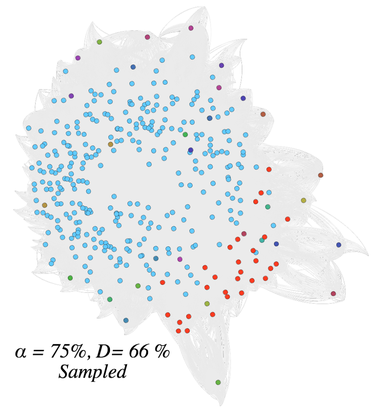
Experiments



α = 100%, D = 66 %

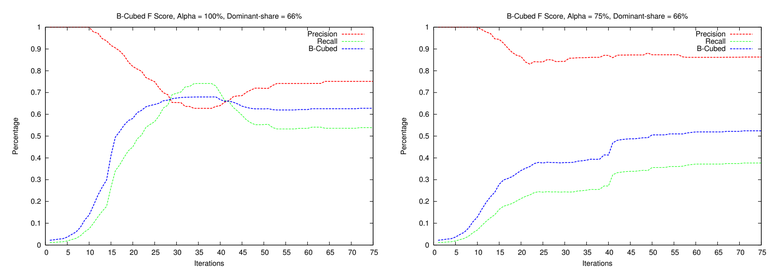


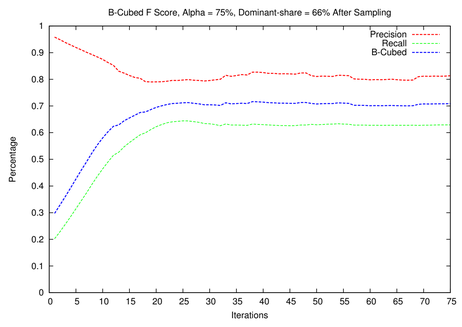
α = 75%, D = 66%



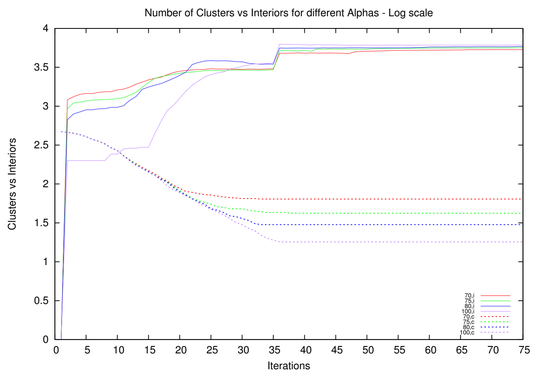
α = 75%, D= 66 %
Sampled

Results





α	B-Cubed	V-Measure
100%	62.7%	29.0%
75%	52.7%	28.8%
75% + sampling	**71.5%**	**40.0%**



References

[1] Fatemeh Rahimian, Sarunas Girdzijauskas, and Seif Haridi,: Parallel Community Detection For Cross-Document Coreference. Presented at the 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WI'14), Warsaw, Poland, August 2014.