Graph-Based Methods in Coreference Resolution

Presented by Huiqin Körkel-Qu
Department of Computer linguistic, Heidelberg University

December 9, 2010

Structure of the report

- Introduction to coreference
- Build anaphoricity in graph for coreference resolution
- BestCut-method based on min-cut algorithm
- An one step solution: hypergraph in coreference resolution
- Compare the methods

Coreference

What is coreference?

- **Entity:** an object or a set of objects in the real world.
- **Mention:** a textual reference to an entity
- **Coreference:** more mentions in language refer to the same entity
- **Eg:** Mary has a brother John, the boy is younger than the girl

entities: MARY, JOHN
mentions: Mary, a brother, John, the boy, the girl
coreference set: {Mary, the girl}, {a brother, John, the boy}

Anaphoricity

Eg: Mary has a brother John, the boy is younger than the girl

anaphoric: John, the boy, the girl
nonanaphoric: Mary, a brother

Eg: To repair the house will cost **a lot of money**

Two step methods

Eg: Mary has a brother John, the boy is younger than the girl

- **classification phase:** pair of mentions based on feature sets
  features: e.g the distance, string match feature. etc
  methods: decision trees, maximum entropy classifiers
  Eg: (Mary, a brother), (a brother, John). etc
- **clusterization phase**: to decide the entities(mentions that are coreferent)

Methods for clustering:

- locally optimized clustering
- globally optimized clustering.
  Bell tree, ILP, graph algorithms

ananphoricity classifier: the probability if a mention is anaphoric or not

the results of coreference classification/anaphoricity classification

- $P_C(m_i, m_j)$, the probability that mentions m_i and m_j are coreferent
- $P_A(m_i)$, the probability that mention m_i is anaphoric
- $1 - P_A(m_i)$, the probability that mention m_i is nonanaphoric

Construction of the graph– add anaphoricity

- create the source vertex s: ANAPHORIC
- create the sink vertex t: NON ANAPHORIC
- for each mention m_i create one vertex i
- two edges si and it
- the weight of si w_{si} is $P_A(m_i)$, i,e the probability that m_i is anaphoric
- the weight of it w_{it} is $1 - P_A(m_i)$, i,e the probability that m_i is nonanaphoric

add information of anaphoricity in the Graph

minimum s-t cut:

- assign any node(mention) i with $w_{si} > 0.5$,i.e $P_A(m_i) > 0.5$ to s
- assign any node i with $w_{it} > 0.5$,i.e $P_A(m_i) < 0.5$) to t
- remained nodes are assigned to one of s and t.

notice: 0.5 as threshold is too conservative,
i.e too fewer mentions are classified as anaphoric

In this system, P_A is rearranged so that the decision for anaphor is not too conservative

Incorporating coreference probability

- add one edge for every mention pair (m_i, m_j)(except s and t in the graph)
- the weight of edge between m_i and m_j, $w_{i,j}$ is $P_C(m_i, m_j)$

same as P_A, the weight here is also rearranged by learning to get the better result.

minimum cut of the graph

the costs of the minimum $s - t$ cut is:

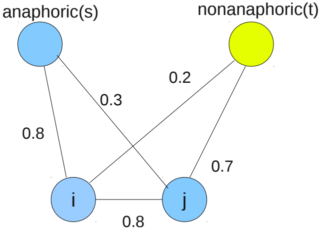
$$\min \sum_{m_i \in S-s, m_j \in T-t} w_{i,j} + \sum_{m \in S} w_{mt} + \sum_{n \in T} w_{sn}$$

Eg: two mentions m_i, m_j

$w_{ji} = w_{ij} = 0.8$

$w_{si} = 0.8, w_{it} = 0.2$

$w_{sj} = 0.3, w_{jt} = 0.7$



Add direction in the graph

Problem with the algorithm above: tend to classify **all mentions** of a coreference as anaphoric!
including **the first mention of an entity**.

Can we change the graph to be a directed graph?

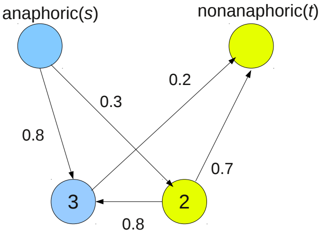
- from anaphoric node to mention node
- from mention node to nonanaphoric node
- from mention node with smaller index to mention node with bigger index node

Add direction in the graph

Mary$_1$ has **a brother**$_2$ **John**$_3$

consider mentions m_2, m_3

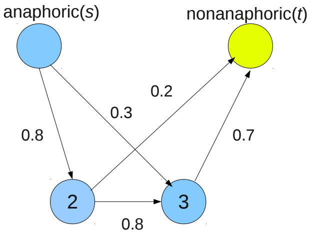
$$w_{23} = w_{32} = 0.8$$
$$w_{s3} = 0.8, w_{3t} = 0.2$$
$$w_{s2} = 0.3, w_{2t} = 0.7$$



the brother$_1$ **John**$_2$ is **a student**$_3$

consider mentions m_2, m_3

$$w_{23} = w_{32} = 0.8$$
$$w_{s3} = 0.3, w_{3t} = 0.7$$
$$w_{s2} = 0.8, w_{2t} = 0.2$$



evalution of this method

Baselines unter MUC-Score and CEAF-Score

- Berger(96), et al, no ananphoricity
- Ng(2004), where P_A and P_C are not coordinated
- Luo(2007), a heuristic search on Bell tree
- D&B(2007),Integer Linear Programming with ananphoricity as constraints(hard constraints)
- F&M(2008), ILP, with transitivity as hard constraints

Conclusion:

Large gain in precision
Small drop in recall
Improvement of F-score.

BestCut: a globally clustering based on Min-cut on graph

Construction of weighted undirected graphs

- **Mention Detection:** 6 entity types
  PERSON, ORGANIZATION, LOCATION, FACILITY, GPE, UNK
- **classification:** $P_C(m_i, m_j)$, the probability that mention m_i and m_j are coreferent

- **Number of Graphs:** for every entity type, a graph will be constructed(6 graphs)
  the mentions from different type will not be coreferent
-
  - **vertex:** every mention in the type is a vertex
  - **edge:** the weight between two vertexes m_i, m_j(two mentions) is $P_C(m_i, m_j)$

BestCut: a globally clustering based on Min-cut on graph

Construction of weighted undirected graphs

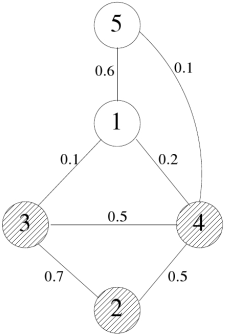
- **Mention Detection:** 6 entity types
  PERSON, ORGANIZATION, LOCATION, FACILITY, GPE, UNK
- **classification:** $P_C(m_i, m_j)$, the probability that mention m_i and m_j are coreferent

- **Number of Graphs:** for every entity type, a graph will be constructed(6 graphs)
  the mentions from different type will not be coreferent
-
  - **vertex:** every mention in the type is a vertex
  - **edge:** the weight between two vertexes m_i, m_j(two mentions) is $P_C(m_i, m_j)$

Construction of the graphs: an example

Mary$_1$ has **a brother**$_2$ **John**$_3$. **The boy**$_4$ is older than **the girl**$_5$

5 mentions
entity type: PERSON
Only one graph will be constructed.



When to stop the cut?

G: current graph.
S, T: the two parts after cut
$S.V, T.V$: the vertexes in the two parts. $|S.V| \geq |T.V|$
$S.E, T.E$: the edges in the two parts
$C.E$: the edges crossing the cut

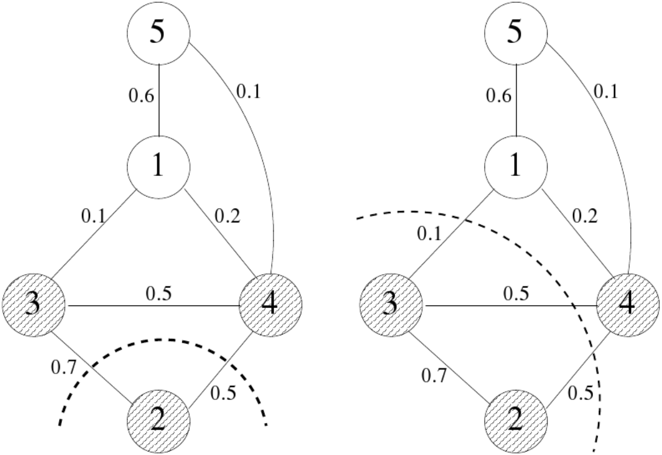
Stop the cut or continue?
Features for stopping the cut

- $|S.V|/|T.V|$
- $|C.E|/|G.E|$
- $max(C.E), min(C.E), avg(C.E)$
- etc

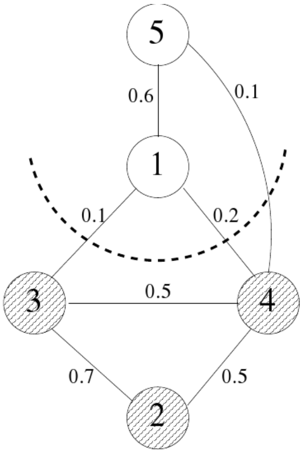
Procedure of a Cut

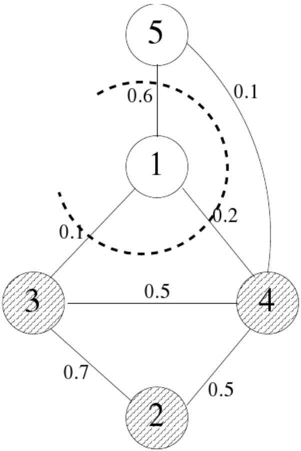
a sequence of s-t cuts, and the BestCut is one of them

Eg:



Procedure of a cut



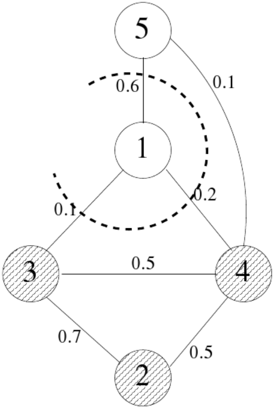


How to choose the BestCut

Scoring a cut (s-t cut from the above procedure)

- average weight to decide if a vertex belongs to its group.
- maximum weight to decide if a vertex belongs to its group.

Eg:



Evaluation and conclusion

evaluated under EMC-F Score and MUC P, R, F Scores
compared with two baselines: Belltree(Luo 04) and Link-Best

- outperform the baselines with true mentions and detected mentions(If the entity types are known).
- for undetected mentions it works not so well

Conclusion: the mention detections and the decision of entity types are important for this algorithm

A one step solution using hypergraph

The procedure of this system:

- learn the hyperedge weights
- create a hypergraph
- partition the hypergraph in to subhypergraphs so that each subhypergraphs represents an entity

Features used to construct the hypergraph

- StrMatch_Npron, StrMatch_Pron
- Alias
- Appositive
- distance, etc

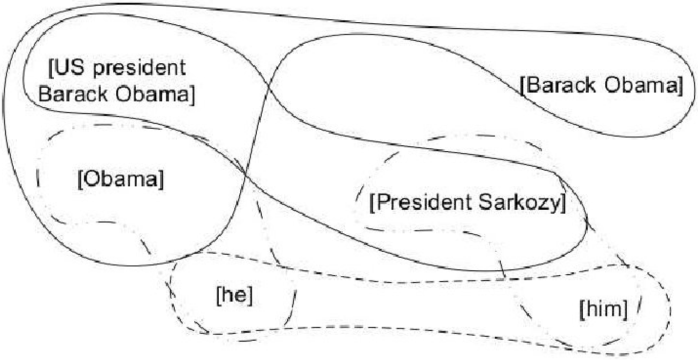
Training result for hyperedge weight:

Edge Name	Weight
Alias	0.777
StrMatch_Pron	0.702
Appositive	0.568
StrMatch_Npron	0.657
ContinuousDistAgree	0.403

An example

US President Barack Obama came to Toronto today.
Obama discussed the nancial crisis with **President Sarkozy**
He talked to **him** about the recent downturn of the European markets.
Barack Obama will leave Toronto tomorrow.

2 entities: BALACK OBAMA
NICOLAS SARKOZY



partial string match:

{US President Barack Obama, Obama, Barack Obama}

{US President Barack Obama, President Sarkozy}

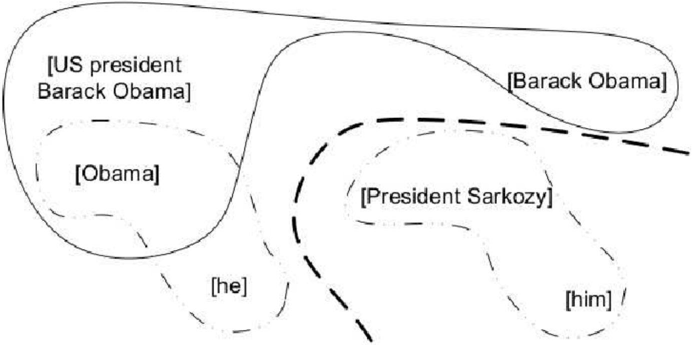
pronoun match: {he, him}

all speak: {Obama, he}

{President Sarkozy, him}

partition of the hypergraph

- *recursive 2-way partitioning*
- flat-K partitioning



evalution

Using MUC, B^3_{sys}, $CEAF_{sys}$ Scores
Baselines: SOON(2001)-BART(2008)
B&R(2008), 'the best performance on ACE2004'

Conclusion:

gain in recall
drop in precision
better F-Measure than B& R in most of the cases.

compare the 3 methods

- how many steps?
  Bestcut: 2 steps, classification, and clustering
  s-t cut: 3 steps, classification (P_A, P_C), anaphoricity determination(s-t cut), coreference clustering
  Hypergraph: 1 step, construct hypergraph based on features, and cut(spectral clustering)
- the emphasis(new idea) of the 3 methods
  Bestcut: mincut in coreference clustering
  s-t cut: combine anaphoricity and coreference in graph
  hypergraph: all in one, no separation between 'classification' and clustering

Reference

- Cristina Nicolae, Gabriel Nicolae: BESTCUT: A Graph Algorithm for Coreference Resolution. EMNLP 2006:275-283
- Vincent Ng: Graph-Cut-Based Anaphoricity Determination for Coreference Resolution. HLT-NAACL 2009: 575-583
- Cai, Jie; Strube, Michael (2010). End-to-End Coreference Resolution via Hypergraph Partitioning In: COLING '10, pp.143-151

The end ;-)