Linguistically Aware Coreference Evaluation Metrics

Chen Chen and **Vincent Ng**

Human Language Technology Research Institute

The University of Texas at Dallas

Coreference Resolution

- Goal: Determine which mentions in a text or dialogue refer to the same real-world entity

Existing Scoring Metrics

- No consensus on which metric is the best

Existing Scoring Metrics

- No consensus on which metric is the best
- Therefore, CoNLL-2011 and CoNLL-2012 shared tasks take the average F-score of
  - MUC (Vilain et al., 1995)
  - B^3 (Bagga and Baldwin, 1988)
  - $CEAF_e$ (Luo, 2005)

Weakness

- However, all existing metrics are linguistically agnostic

Weakness

- However, all existing metrics are linguistically agnostic
  - Treat the mentions to be clustered as generic rather than linguistic objects

Illustrated Example

Gold Chains:

[(Hillary Clinton)-(she)-(she)]

Illustrated Example

Gold Chain:

[(Hillary Clinton)-(she)-(she)]

System Response A:

[(Hillary Clinton)-(she)]
[(she)]

System Response B:

[(Hillary Clinton)]
[(she)-(she)]

Illustrated Example

Gold Chain:

[(Hillary Clinton)-(she)-(she)]

System Response A:

[(Hillary Clinton)-(she)]

[(she)]

System Response B:

[(Hillary Clinton)]

[(she)-(she)]

All existing metrics assign same score to both responses

Illustrated Example

Gold Chain:

[(Hillary Clinton)-(she)-(she)]

System Response A:

[(Hillary Clinton)-(she)]
[(she)]

System Response B:

[(Hillary Clinton)]
[(she)-(she)]

However, intuitively, system response A should be better than B

Because we can infer what one mention of "she" refers to from response A

Goal

- Propose a framework for incorporating linguistic awareness into commonly-used coreference evaluation metrics to initiate further discussions

Plan for the Talk

- Existing Evaluation Metrics
- Formalizing Linguistic Awareness
- Evaluation
- Conclusion

Plan for the Talk

- Existing Evaluation Metrics
- Formalizing Linguistic Awareness
- Evaluation
- Conclusion

Notation

- For a coreference chain C
  - Define $|C|$ as the number of mentions in C

$$\text{Chain C: } \underbrace{m_1 \quad - \quad m_2 - \quad m_3 \quad \ldots\ldots \quad m_n}_{|C|}$$

Notation

- Define *d* as one document
- *K(d)* refers to key chains
  - $K(d)=\{K_i: i=1,2,\ldots,|K(d)|\}$

$$K_1 : m_a - m_b - m_c - \ldots$$

$$K_2 : m_d - m_e - m_f - \ldots$$

.....

$$K_{|K(d)|} : m_x - m_y - m_z - \ldots$$

Notation

- *S(d)* refers to system-generated chains
  - $S(d)=\{S_j: j=1,2,\ldots,|S(d)|\}$

$$S_1 : m_a - m_b - m_c - \ldots$$

$$S_2 : m_d - m_e - m_f - \ldots$$

$$\ldots\ldots$$

$$S_{|S(d)|} : m_x - m_y - m_z - \ldots$$

MUC (Vilain et al., 1995)

- Link-based metric, which counts links in one cluster

$$\text{Recall} = \frac{\text{number of common links}}{\text{number of key links}}$$

$$\text{Precision} = \frac{\text{number of common links}}{\text{number of system links}}$$

MUC (Vilain et al., 1995)

- To compute the number of common links, a partition $P(S_j)$ is created for system chain S_j

$$P(S_j) = \{C_j^i : i = 1,2,\ldots,|K(d)|\}$$

- Each C^i_j in the partition is formed by intersecting system chain S_j with one key chain K_i (C^i_j may be empty)

$$S_j : \underbrace{(m_a - m_b)}_{C_j^1} - \underbrace{(m_c - m_d)}_{C_j^2} - \underbrace{(m_e - m_f)}_{C_j^3} - \ldots$$

MUC (Vilain et al., 1995)

- The number of common links is defined as

$$c(K(d),S(d)) = \sum_{j=1}^{|S(d)|}\sum_{i=1}^{|K(d)|} w_c(C_j^i)$$

$$\text{where } w_c(C_j^i) = \begin{cases} 0 & \text{if } |C_j^i| = 0 \\ |C_j^i|\text{-}1 & \text{if } |C_j^i| > 0 \end{cases}$$

- If cluster *C* is non-empty, the minimum required number of links is *|C|-1*

MUC (Vilain et al., 1995)

- The number of key links is defined as

$$K_1 : m_a - m_b - m_c - \ldots$$

$$K_2 : m_d - m_e - m_f - \ldots$$

$$\ldots\ldots$$

$$K_{|K(d)|} : m_x - m_y - m_z - \ldots$$

$$k(K(d)) = \sum_{i=1}^{|K(d)|} w_k(K_i)$$

$$\text{where } w_k(K_i) = |K_i| - 1$$

MUC (Vilain et al., 1995)

- The number of system links is defined as

$$S_1 : m_a - m_b - m_c - \ldots$$

$$S_2 : m_d - m_e - m_f - \ldots$$

.....

$$S_{|S(d)|} : m_x - m_y - m_z - \ldots$$

$$s(S(d)) = \sum_{j=1}^{|S(d)|} w_s(S_j)$$

where $w_s(S_j) = |S_j| - 1$

B^3 (Bagga and Baldwin, 1998)

- B^3 is a mention-based metric, which counts the number of mentions. It computes:
  - Recall and precision for each mention
  - Average per-mention values to obtain the overall recall and precision

B^3 (Bagga and Baldwin, 1998)

- Define m_n as the nth mention in a document

B^3 (Bagga and Baldwin, 1998)

- Define m_n as the *n*th mention in a document
- K_i and S_j is the key chain and the system chain that contain m_n, respectively

$$K_i : m_a - m_b \ldots . m_m - \ldots - m_n$$
$$S_j : m_m - \ldots - m_n - \ldots - m_y - m_z$$

B^3 (Bagga and Baldwin, 1998)

- Define m_n as the nth mention in a document
- K_i and S_j is the key chain and the system chain that contain m_n, respectively
- C^i_j is the common subset between K_i and S_j

$$K_i : m_a - m_b \dots . m_m - \dots - m_n$$
$$S_j : m_m - \dots - m_n - \dots - m_y - m_z$$
$$C^i_j : m_m - \dots - m_n$$

B^3 (Bagga and Baldwin, 1998)

$$K_i : m_a - m_b \ldots . m_m - \ldots - m_n$$

$$S_j : m_m - \ldots - m_n - \ldots - m_y - m_z$$

$$C_j^i : m_m - \ldots - m_n$$

$$R(m_n) = \frac{w_c(C_j^i)}{w_k(K_i)}, P(m_n) = \frac{w_c(C_j^i)}{w_s(S_j)}$$

where $w_c(C_j^i) = |C_j^i|, w_k(K_i) = |K_i|$ and $w_s(S_j) = |S_j|$

CEAF (Luo, 2005)

- CEAF finds one-to-one alignment between chains in *K(d)* and *S(d)*

CEAF (Luo, 2005)

- Not all system chains and key chains are used
- Define $K_{min}(d)$ and $S_{min}(d)$ as the subset of key chains and system chains involved in the alignment

CEAF (Luo, 2005)

- Not all system chains and key chains are used
- Define $K_{min}(d)$ and $S_{min}(d)$ as the subset of key chains and system chains involved in the alignment
- Alignment function g which aligns one key chain K_i to system chain S_j is defined as

$$g(K_i) = S_j, K_i \in K_{\min}(d) \text{ and } S_j \in S_{\min}(d)$$

CEAF (Luo, 2005)

- $\varnothing(K_i, S_j)$ is to measure the similarity between two chains
- The score of alignment function *g* equals to the sum of similarity of all entries in alignment

$$\Phi(g) = \sum_{k_i \in K_{\min}(D)} \phi(K_{i,} g(K_i))$$

CEAF (Luo, 2005)

- Ø(K_i,S_j) is to measure the similarity between two chains
- The score of alignment function *g* equals to the sum of similarity of all entries in alignment

$$\Phi(g) = \sum_{k_i \in K_{\min}(D)} \phi(K_{i,} g(K_i))$$

- The optimal alignment *g** is the alignment whose Φ value is the largest among all possible alignments

CEAF (Luo, 2005)

- The recall (R) and precision (P) of a system partition can be computed as follows:

$$R = \frac{\Phi(g^*)}{\sum_{i=1}^{|K(d)|} \phi(K_i, K_i)}, P = \frac{\Phi(g^*)}{\sum_{j=1}^{|S(d)|} \phi(S_j, S_j)}$$

CEAF (Luo, 2005)

- The recall (R) and precision (P) of a system partition can be computed as follows:

$$R = \frac{\Phi(g^*)}{\sum_{i=1}^{|K(d)|} \phi(K_i, K_i)}, P = \frac{\Phi(g^*)}{\sum_{j=1}^{|S(d)|} \phi(S_j, S_j)}$$

- How to define Ø function?

CEAF (Luo, 2005)

$$\phi_3(K_i, S_j) = |K_i \cap S_j| = w_c(C_j^i) = |C_j^i|$$

- $\emptyset_3$ results in mention-based CEAF (a.k.a. CEAF_m)

CEAF (Luo, 2005)

$$\phi_4(K_i, S_j) = \frac{2|K_i \cap S_j|}{|K_i| + |S_j|} = \frac{2 * w_c(C_j^i)}{w_k(K_i) + w_s(S_j)} = \frac{2 * |C_j^i|}{|K_i| + |S_j|}$$

- $\emptyset_4$ results in entity-based CEAF (a.k.a. CEAF_e)

Common Functions

- Three functions common to MUC, B^3 and CEAF
  - $w_c(C^i_j)$, the **weight** of common subset of K_i and S_j
    - For MUC, its value is 0 if C^i_j is empty and $|C^i_j|$-1 otherwise; for B^3 and CEAF, its value is $|C^i_j|$

Common Functions

- Three functions common to MUC, B^3 and CEAF
  - $w_c(C^i_j)$, the **weight** of common subset of K_i and S_j
  - $w_k(K_i)$, the **weight** of key chain K_i
    - For MUC, its value is $|K_i|$-1; for B^3 and CEAF, its value is $|K_i|$

Common Functions

- Three functions common to MUC, B^3 and CEAF:
  - $w_c(C^i_j)$, the **weight** of common subset of K_i and S_j
  - $w_k(K_i)$, the **weight** of key chain K_i
  - $w_s(S_j)$, the **weight** of system chain S_j
    - For MUC, its value is $|S_j|-1$; for B^3 and CEAF, its value is $|S_j|$

Plan for the Talk

- Existing Evaluation Metrics
- Formalizing Linguistic Awareness
- Evaluation
- Conclusion

Formalizing Linguistic Awareness

- Existing metrics are linguistic agnostic, because
  - Three common functions are linguistic agnostic
- Modify above three common functions to encode linguistic awareness

What is Linguistic Awareness?

- Goal of (co)reference resolution
  - Facilitate automated text understanding by finding the referent for each referring expression

What is Linguistic Awareness?

- Goal of (co)reference resolution
  - Facilitate automated text understanding by finding the referent for each referring expression

- A resolver should be rewarded more if the selected antecedent allows the underlying entity to be easily inferred

What is Linguistic Awareness?

- Goal of (co)reference resolution
  - Facilitate automated text understanding by finding the referent for each referring expressions

- A resolver should be rewarded more if the selected antecedent allows the underlying entity to be easily inferred
  - NAME antecedents are preferable to NOMINAL antecedents
  - NOMINAL antecedents are preferable to PRONOUN antecedents

How to Encode Such Preference for NAME and NOMINAL Antecedents?

- Idea: assign different weights to different link types

- Given a link e_l, which connects two mentions, the weight of this link $w_l(e_l)$ is defined as,
  - If e_l involves a name, $w_l(e_l) = w_{nam}$
  - else if e_l involves a nominal, $w_l(e_l)=w_{nom}$
  - else $w_l(e_l)=w_{pro}$

How to Encode Such Preference for NAME and NOMINAL antecedents

- Idea: assign different weights to different link types

- Given a link e_l, which connects two mentions, the weight of this link $w_l(e_l)$ is defined as,
  - If e_l involves a name, $w_l(e_l) = w_{nam}$
  - else if e_l involves a nominal, $w_l(e_l)=w_{nom}$
  - else $w_l(e_l)=w_{pro}$

- w_{nam}, w_{nom}, w_{pro} are our model parameters. We want to set them so that $w_{nam} \geq w_{nom} \geq w_{pro}$

Scoring Singleton Cluster

- Singleton clusters have no link. How should they be scored?

Scoring Singleton Cluster

- Singleton clusters have no link. How should they be scored?
  - We create an additional parameter, w_{sing} , for any chain that only contains one mention
  - w_{sing} is the weight associated with singleton clusters

Incorporate Weights Variable

- $W=(w_{nam}, w_{nom}, w_{pro}, w_{sing})$
- Recall that we have three common functions
  - $w_c(C_j^i)$, the **weight** of common subset of key chain K_i and system chain S_j
  - $w_k(K_i)$, the **weight** of key chain K_i
  - $w_s(S_j)$, the **weight** of system chain S_j
- Below we show how to incorporate four weights into three weight functions

Linguistic Aware Weight Functions

- Weight of common subset of key&system chain
  - $w_c^L(C_j^i)$, the linguistically aware weight function of $w_c(C_j^i)$

- Weight of key chain
  - $w_k^L(K_i)$, the linguistically aware weight function of $w_k(K_i)$

- Weight of system chain
  - $w_s^L(S_j)$, the linguistically aware weight function of $w_s(S_j)$

Defining w_c^L

- Case 1: $|C_j^i| \geq 2$
- Case 2: $|C_j^i| = 0$
- Case 3: $|C_j^i| = 1$

Defining w_c^L

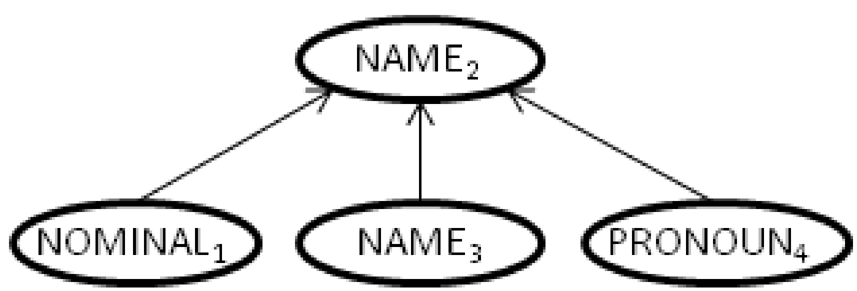
- Case 1: $|C_j^i| \geq 2$
  - Consider C_j^i contains four mentions: NOMINAL_1, NAME_2, NAME_3 and PRONOUN_4

Defining w_c^L

- Case 1: $|C_j^i| \geq 2$
  - Consider C_j^i contains four mentions: NOMINAL$_1$, NAME$_2$, NAME$_3$ and PRONOUN$_4$
  - Generate maximum spanning tree in terms of total weights of links

Defining w_c^L

- Case 1: $|C_j^i| \geq 2$
  - Consider C_j^i contains four mentions: $NOMINAL_1$, $NAME_2$, $NAME_3$ and $PRONOUN_4$
  - Generate maximum spanning tree in terms of total weights of links
  - One possible maximum spanning tree :



Defining w_c^L

- Case 1: $|C_j^i| \geq 2$. Let E be the edge set of the maximum spanning tree

$$w_c^L(C_j^i) = \sum_{e_l \in E} w_l(e_l)$$

Defining w_c^L

- Case 2: $|C_j^i|=0$

Defining w_c^L

- Case 2: $|C_j^i|=0$

$$w_c^L(C_j^i) = 0$$

Defining w_c^L

- Case 3: $|C_j^i|=1$

Defining w_c^L

- Case 3: $|C_j^i|=1$
  - If C_j^i, K_i *and* S_j are all singleton clusters, which means this system chain is a correctly resolved singleton cluster, w_{sing}
  - 0, otherwise

Defining w_c^L

- The linguistically aware weight function of common subset between K_i and S_j is defined as

$$w_c^L(C_j^i) = \begin{cases} \sum_{e_l \in E} w_l(e_l) & \text{if } |C_j^i| > 1 \\ w_{\text{sing}} & \text{if } |C_j^i|, |K_i|, |S_j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

Linguistic Aware Weight Functions

- Weight of common subset of key&system chain
  - $w_c^L(C_j^i)$, the linguistically aware weight function of $w_c(C_j^i)$
- Weight of key chain
  - $w_k^L(K_i)$, the linguistically aware weight function of $w_k(K_i)$
- Weight of system chain
  - $w_s^L(S_j)$, the linguistically aware weight function of $w_s(S_j)$

Defining w_k^L

- Case 1: $|K_i| \geq 1$
- Case 2: $|K_i| = 1$

Defining w_k^L

- Case 1: $|K_i| \geq 1$
  - Generate maximum spanning tree over K_i, let E be the edges in the tree

$$w_k^L(K_i) = \sum_{e_l \in E} w_l(e_l)$$

Defining w_k^L

- Case 2: $|K_i|=1$

$$w_k^L(K_i) = w_{\text{sing}}$$

Defining w_k^L

- The linguistically aware weight function of key chain k_i is defined as

$$w_k^L(K_i) = \begin{cases} \sum_{e_l \in E} w_l(e_l) \text{ if } |K_i| > 1 \\ w_{\text{sing}} \qquad\qquad \text{if } |K_{\mathrm{i}}| = 1 \end{cases}$$

Linguistic Aware Weight Functions

- Weight of common subset of key&system chain
  - $w_c^L(C_j^i)$, the linguistically aware weight function of $w_c(C_j^i)$
- Weight of key chain
  - $w_k^L(K_i)$, the linguistically aware weight function of $w_k(K_i)$
- Weight of system chain
  - $w_s^L(S_j)$, the linguistically aware weight function of $w_s(S_j)$

Defining w_s^L

- Case 1: $|S_j|=1$
- Case 2: $|S_j|\geq 1$

Defining $w_s{}^L$

- Case 1: $|S_j|=1$

$$w_S^L(S_j) = w_{\mathrm{sing}}$$

Defining w_s^L

- Case 2: $|S_j|>1$

Defining w_s^L

- Recall that we can create a partition $P(S_j)$ for each system chain S_j

$$P(S_j) = \{C_j^i : i = 1, 2, \ldots, |K(d)|\}$$

- Each C_j^i in $P(S_j)$ is formed by intersecting S_j with K_i

$$S_j : \underbrace{(m_a - m_b)}_{C_j^1} - \underbrace{(m_c - m_d)}_{C_j^2} - \underbrace{(m_e - m_f)}_{C_j^3} - \ldots$$

Defining w_s^L

- Recall that we can create a partition $P(S_j)$ for each system chain S_j

$$P(S_j) = \{C_j^i : i = 1,2,\ldots,|K(d)|\}$$

- Each C^i_j in $P(S_j)$ is formed by intersecting S_j with K_i

Spurious links

$$S_j : \underbrace{(m_a - m_b)}_{C_j^1} - \underbrace{(m_c - m_d)}_{C_j^2} - \underbrace{(m_e - m_f)}_{C_j^3} - \ldots$$

Defining w_s^L

- Only spurious links should be penalized as precision error

Spurious links

$$S_j : \underbrace{(m_a - m_b)}_{C_j^1} - \underbrace{(m_c - m_d)}_{C_j^2} - \underbrace{(m_e - m_f)}_{C_j^3} - \ldots$$

Defining w_s^L

- Only spurious links should be penalized as precision error
- Thus, intuitively, w_s^L should be defined as the sum of weights of all spurious links and weights of all subset C_j^i

Spurious links

$$S_j : \underbrace{(m_a - m_b)}_{C_j^1} - \underbrace{(m_c - m_d)}_{C_j^2} - \underbrace{(m_e - m_f)}_{C_j^3} - \ldots$$

Weights of Spurious Links

- Given *n* non-empty clusters in partition $P(S_j)$, there are different sets of $(n\text{-}1)$ spurious links that can connect non-empty clusters together
- We define $E_t(S_j)$ as the set which contains the largest sum of weights of links

Weights of Spurious Links

- Given *n* non-empty clusters in partition $P(S_j)$, there are different sets of (n-1) spurious links that can connect them together
- We define $E_t(S_j)$ as the set which contains the largest sum of weights of links

$$w_s^L(S_j) = \sum_{C_j^i \in P(S_j)} w_c^L(C_j^i) + \sum_{e \in E_t(S_j)} w_l(e)$$

Weights of common subsets

Weights of spurious links

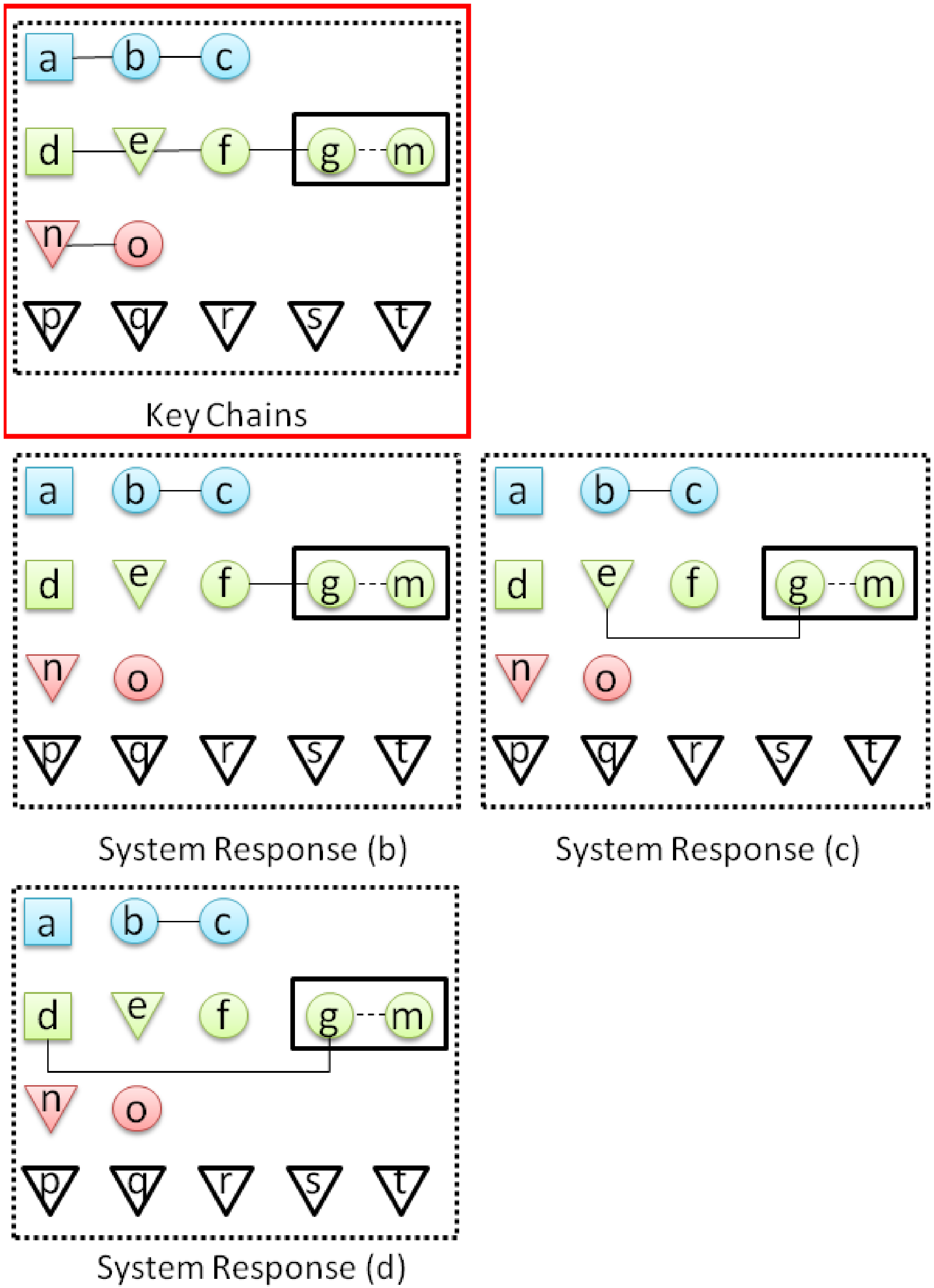
Defining w_s^L

- The linguistically aware weight function of key chain k_i is defined as

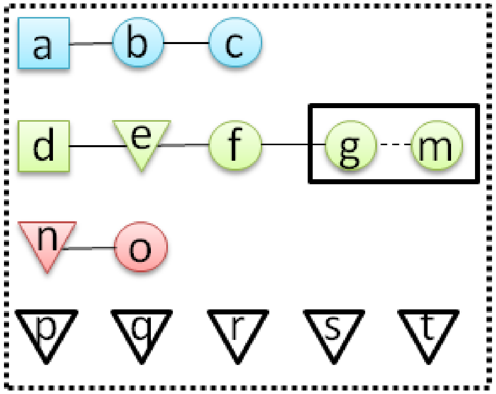
$$w_s^L(S_j) = \begin{cases} \sum_{C_j^i \in P(S_j)} w_c^L(C_j^i) + \sum_{e \in E_t(S_j)} w_l(e) & \text{if } |S_j| > 1 \\ w_{\text{sing}} & \text{if } |S_j| = 0 \end{cases}$$

Plan for the Talk

- Existing Evaluation Metrics
- Formalizing Linguistic Awareness
- Evaluation
- Conclusion

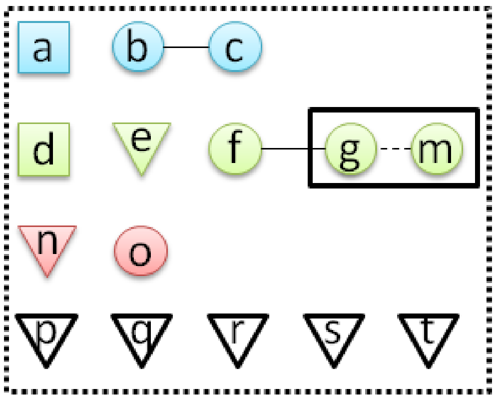


A square denotes a NAME mention
A triangle denotes a NOMINAL mention
A circle denotes PRONOUN mention

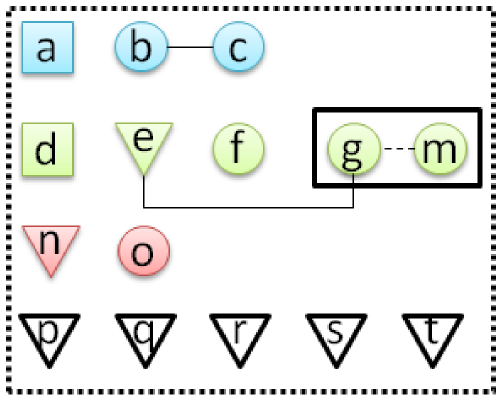


Key Chains

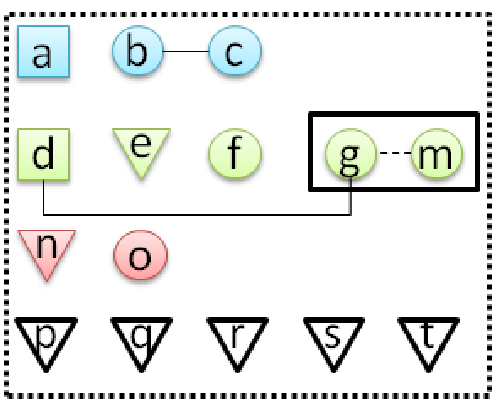
A square denotes a NAME mention
A triangle denotes a NOMINAL mention
A circle denotes PRONOUN mention



System Response (b)



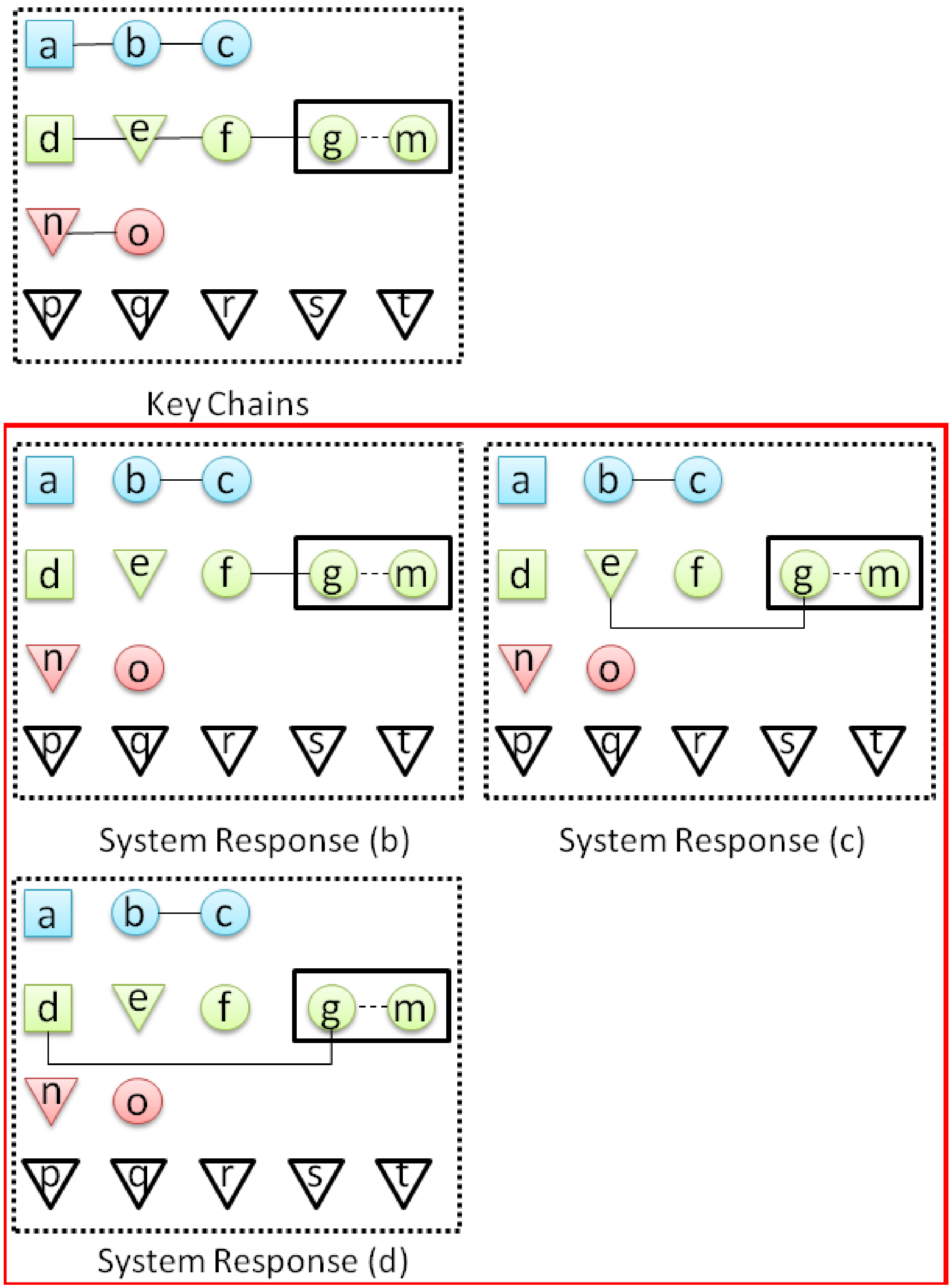
System Response (c)



System Response (d)

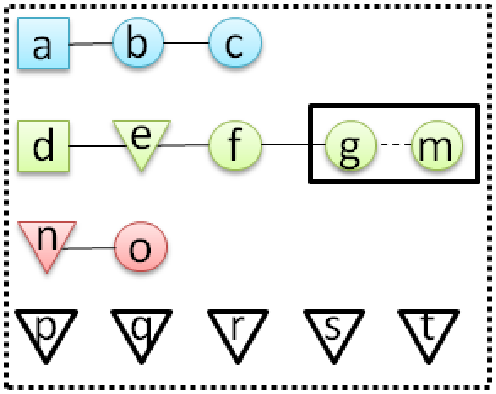
System Response (b) (c) and (d) differ in resolving mentions g to m, to a PRONOUN mention, a NOMINAL mention and a NAME mention respectively. Intuitively, response (d) is better than (c), while response (c) is better than (b)

A square denotes a NAME mention
A triangle denotes a NOMINAL mention
A circle denotes PRONOUN mention



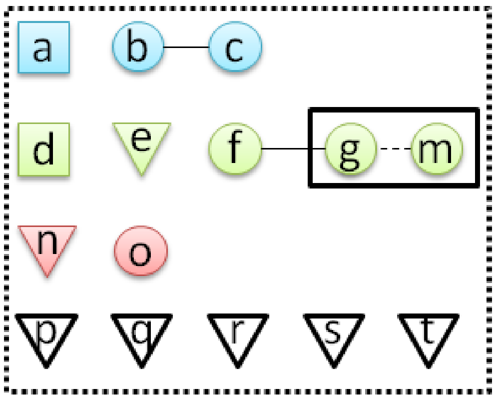
System Response (b) (c) and (d) differ in resolving mentions g to m, to a PRONOUN mention, a NOMINAL mention and a NAME mention respectively. Intuitively, response (d) is better than (c), while response (c) is better than (b)

Original metrics assign identical scores to system response (b), (c) and (d)

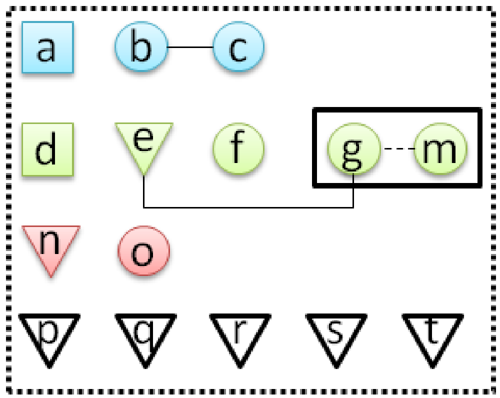


Key Chains

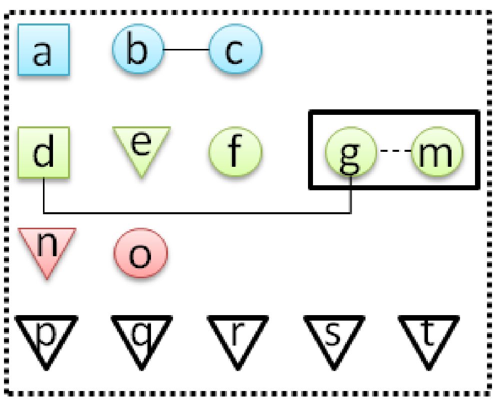
A square denotes a NAME mention
A triangle denotes a NOMINAL mention
A circle denotes PRONOUN mention



System Response (b)



System Response (c)



System Response (d)

System Response (b) (c) and (d) differ in resolving mentions g to m, to a PRONOUN mention, a NOMINAL mention and a NAME mention respectively. Intuitively, response (d) is better than (c), while response (c) is better than (b)

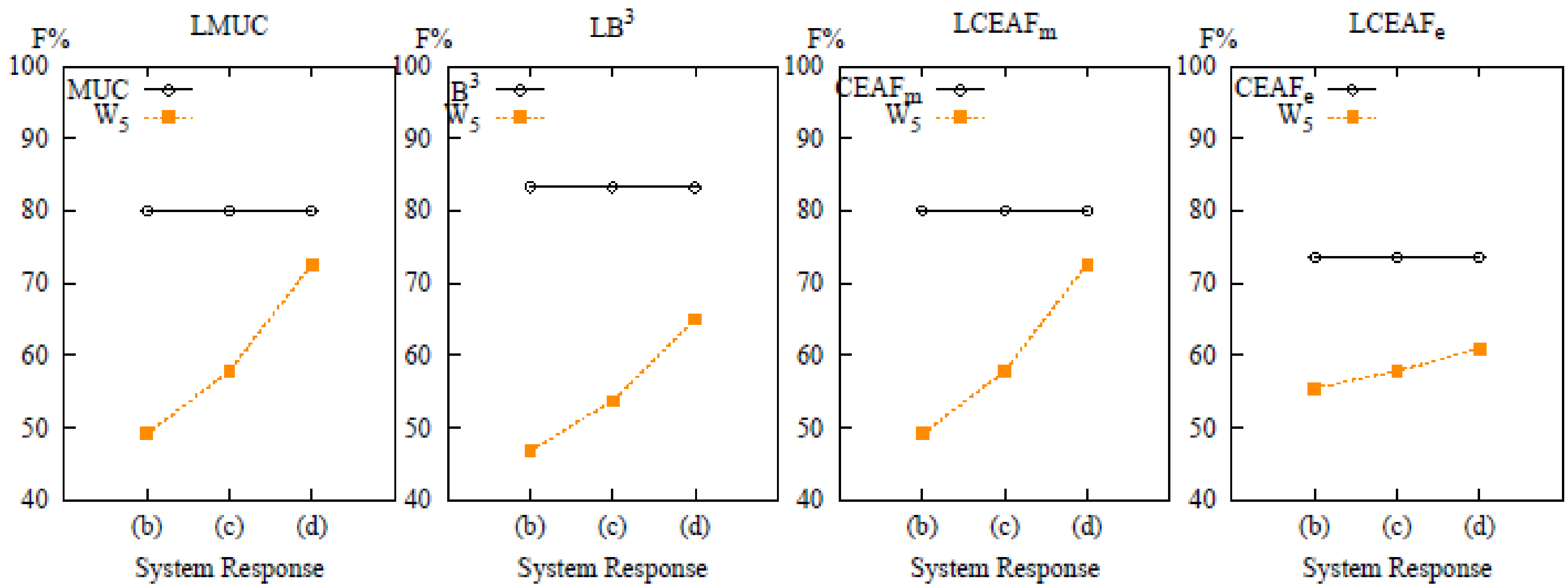
Original metrics assign identical scores to system response (b), (c) and (d)

Goal:
Show how linguistically aware metrics behave on response (b), (c) and (d)

Weight Variable

- $W=(w_{nam}, w_{nom}, w_{pro}, w_{sing})$
- $W_5=(1.0, 0.5, 0.25, 1.0)$

Evaluation Result



- Under linguistically aware metrics, response (d) has higher score than (c); response (c) has higher score than (b), as expected

Plan for the Talk

- Existing Evaluation Metrics
- Formalizing Linguistic Awareness
- Evaluation
- Conclusion

Conclusion

- We addressed the problem of linguistic agnosticity by proposing a framework that enables linguistic awareness to be incorporated into existing metrics
- See the paper for extensive experimentation and analysis of the differences between the linguistically agnostic and linguistically aware evaluation metrics