Clustering Algorithms for Noun Phrase Coreference Resolution

{ }

Abstract

fi

Keywords:

1. Introduction

fi

fi

fi

fi

fi

et al.

fi

f	w_f	$function_f$
		NP_i NP_j
		NP_j fi
	∞	NP_i NP_j
	∞	NP_j NP_i
	∞	
	∞	
	∞	
	∞	

Table 1. Set of features and their weights used in coreference resolution

2. Methods

2.1. *Feature Selection*

fi

Individual Words

Head Noun fi
Position n
Pronoun
Article fi fi

Appositive

Number
Proper Name fi
Gender fi
brother, mother, etc.

Semantic Class *et al.*
S_0 S_1 S_2

2.2. *Distance Metric*

NP_i NP_j

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f * function_f(NP_i, NP_j)$$

F w_f $function_f$
∞ $-\infty$
∞ ∞

3. The Clustering Methods

3.1. *Hard Clustering Cardie et al. (HC-C)*

∞

fi

fi

- fi

  “Robert Smith lives with his wife ... Smith loves her”.
  “Robert Smith” “wife” “Smith” “her” “Smith”
  “Smith”
  “her” “Robert Smith”
  “Smith” “Robert
  Smith” “her”

-

-

-

3.2. *Fuzzy Clustering Bergler et al. (FC-B)*

et al.

fi

et al.

et al.

fi

fi

3.3. *Progressive Fuzzy Clustering (FC-P)*

et al.

et al. fi

$\in S_0$ $\in S_1$ $\in (S_0 \cup S_1)$ $\in S_2$

Appositive merging

Restriction on pronoun coreferencing

"After he saw the danger, Quayle got scared"

- *Progressive nature* *et al.*

MAIN

∞

NP_i
C_j
$fuzziness(NP_i, C_j)$

- C_i, C_j
  - $-\infty$
  - $SIMILAR(C_i, C_j)$
    $ALL\_NPs\_COMPATIBLE(C_i, C_j)$
- NP_i S_2
  C_j
  $fuzziness(NP_i, C_j)$

FUZZINESS(NP_i,C_j)

$$fuzziness(NP_i, C_j) = \frac{dist1(NP_i, C_j)}{\sum_k dist1(NP_i, C_k)}$$

$dist1(NP_i, C_j) =$
$\sum_{f \in F \setminus position} w_f * function_f(NP_i, NP_k) +$
$+ w_{position} * function_{position}(NP_i, NP_t)$
NP_k C_j
NP_t NP_i

C_j

SIMILAR(C_i,C_j)

- $d = 0$
- NP_k
  $fuzziness(NP_k, C_i) < \infty$
  $fuzziness(NP_k, C_j) < \infty$

  $d = d + fuzziness(NP_k, C_i) - fuzziness(NP_k, C_j)$
- d ∞ NP_k ∞
  C_i C_j

MERGE(C_i,C_j)

- C_j C_i
- C_i

- C_j

DEFUZZIFY

- NP_i
  $cluster(NP_i) = C_j$
  $C_j = argmin_{C_k}(fuzziness(NP_i, C_k))$

ALL_NPs_COMPATIBLE(C_i,C_j)

$NP_a \in C_i$
$NP_b \in C_j$
$dist(NP_a, NP_b) = \infty$

Figure 1. Progressive Fuzzy Clustering Algorithm

- *Merging of clusters* *et al.*

- *et al.*

  $-\infty$

- *Search for the best match* *et al.*

  fi *"Robert Smith"* *"Smith"* *"her"* *"Smith"*

- *Corpus-independent* *et al.*

3.4. *The Hard Variant (HC-V)*

fi

MAIN

NP_i S_2

C_j

$fuziness^a(NP_i, C_j)$

$cluster(NP_i) = C_j$

$C_j = argmin_{C_k}(fuziness(NP_i, C_k))$

∞

- C_i, C_j
  - $-\infty$
  - $SIMILAR(C_i, C_j)$
    $ALL\_NPs\_COMPATIBLE(C_i, C_j)$
- **SIMILAR(C_i,C_j)**
  - $d = 0$ $count = 0$
  - NP_k C_i
    NP_l C_j
    $d = d + dist(NP_k, NP_l)$
    $count = count + 1$
  - $d/count$ ∞ $d = \infty$

[a] *fuziness*

Figure 2. Hard Clustering Variant

et al. S_2

et al.

it I me our fi

4. Corpora and Evaluation

“he”, “she”, “him”, “her”, “they”, “them”

et al.

et al.

et al.

$$R = \frac{\sum_i (|C_i| - |p(C_i)|)}{\sum (|C_i| - 1)}$$

C_i $p(C_i)$ C_i
$|p(C_i)|$ C_i

$$R = \frac{\sum_i R_i}{n}$$

R_i i n
i fi

$$R_i = \frac{|HC_i \cap CC_i|}{|HC_i|}$$

HC_i i CC_i
i

$F = \frac{2*P*R}{P+R}$

et al.

fi

fl

fi

fi fi

Experiment

- *Bentonville Institute*

- *“International Business Machines Corp.” and “IBM”*

“the Congressional Black Caucus (CBC)”

- fi *“He spends most of his time talking to associates and customers,” Shinkle said, “and he always comes back with many ideas from them” “he” “his”*
Shinkle
fi

- *“Coca-Cola”, “Coke” “million”, “cents”*
“the U.S.”, “the country” fi

Experiment

fi

fl

5.2. *Pronouns*

Table 3. Precision, recall and F-measure obtained considering all entities, using the different algorithms, starting with correct semantic classes for a) dryrun subcorpus of the MUC-6 corpora, b) formal-test subcorpus of the MUC-6 corpora and c) DUC subcorpus.

fi

6. Conclusion and Future Work

Table 4. Precision, recall and F-measure obtained considering just the pronouns, using the different algorithms for a) dryrun subcorpus of the MUC-6 corpora, b) formaltest subcorpus of the MUC-6 corpora and c) DUC subcorpus.

References

et al.

P

et al

Proceedings of Document Understanding Conference 2003

Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora

Technical Report CS-99-12

et al. *International Journal of Lexicography (special issue)*

et al. *Proceedings of the 4'th Dutch-Belgian Information Retrieval Workshop*