

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 | |
| <i>brother, mother, etc.</i> | | | | |
| Semantic Class | | <i>et al.</i> | | |
| | | | | S_0 S_1 S_2 |

2.2. Distance Metric

NP_i NP_j

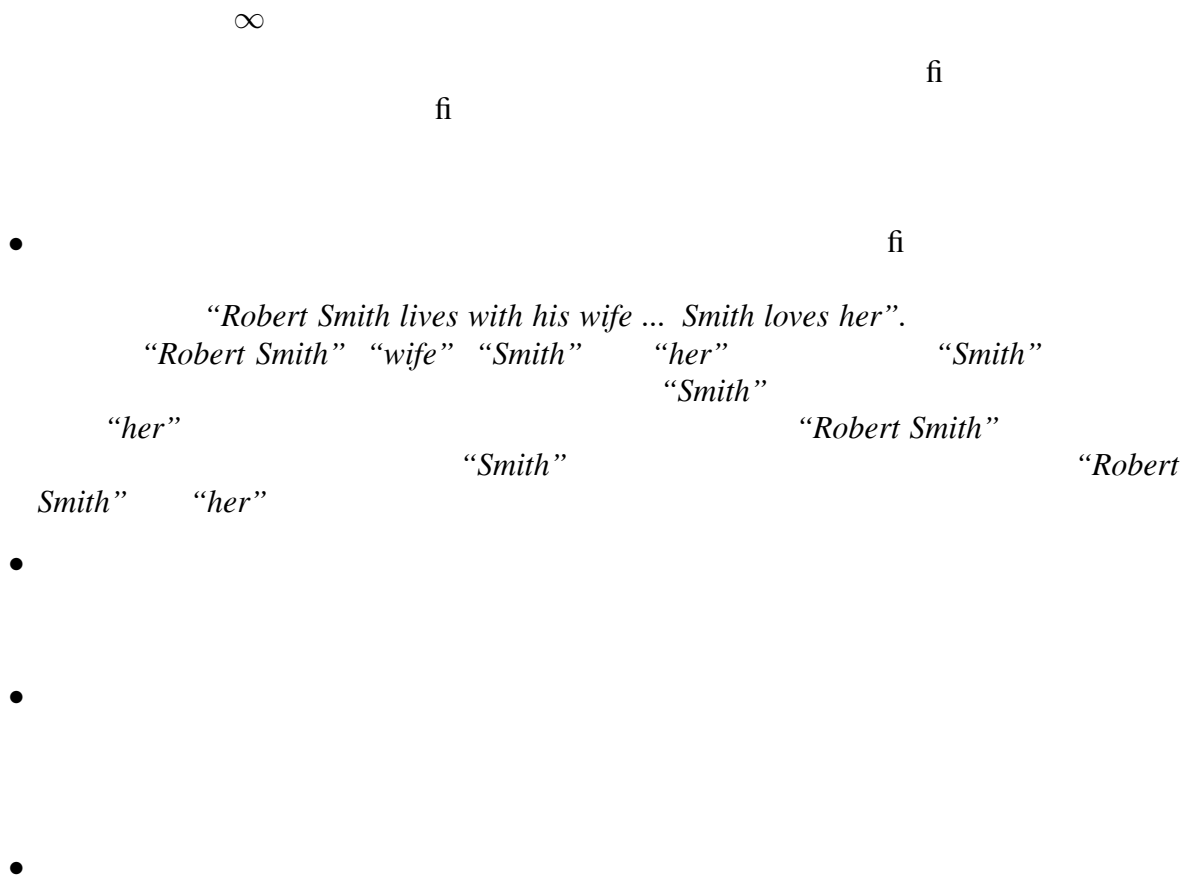
$$dist(NP_i, NP_j) = \sum_{f \in F} w_f * function_f(NP_i, NP_j)$$

F w_f $function_f$

∞ ∞ $-\infty$ ∞

3. The Clustering Methods

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



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



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.

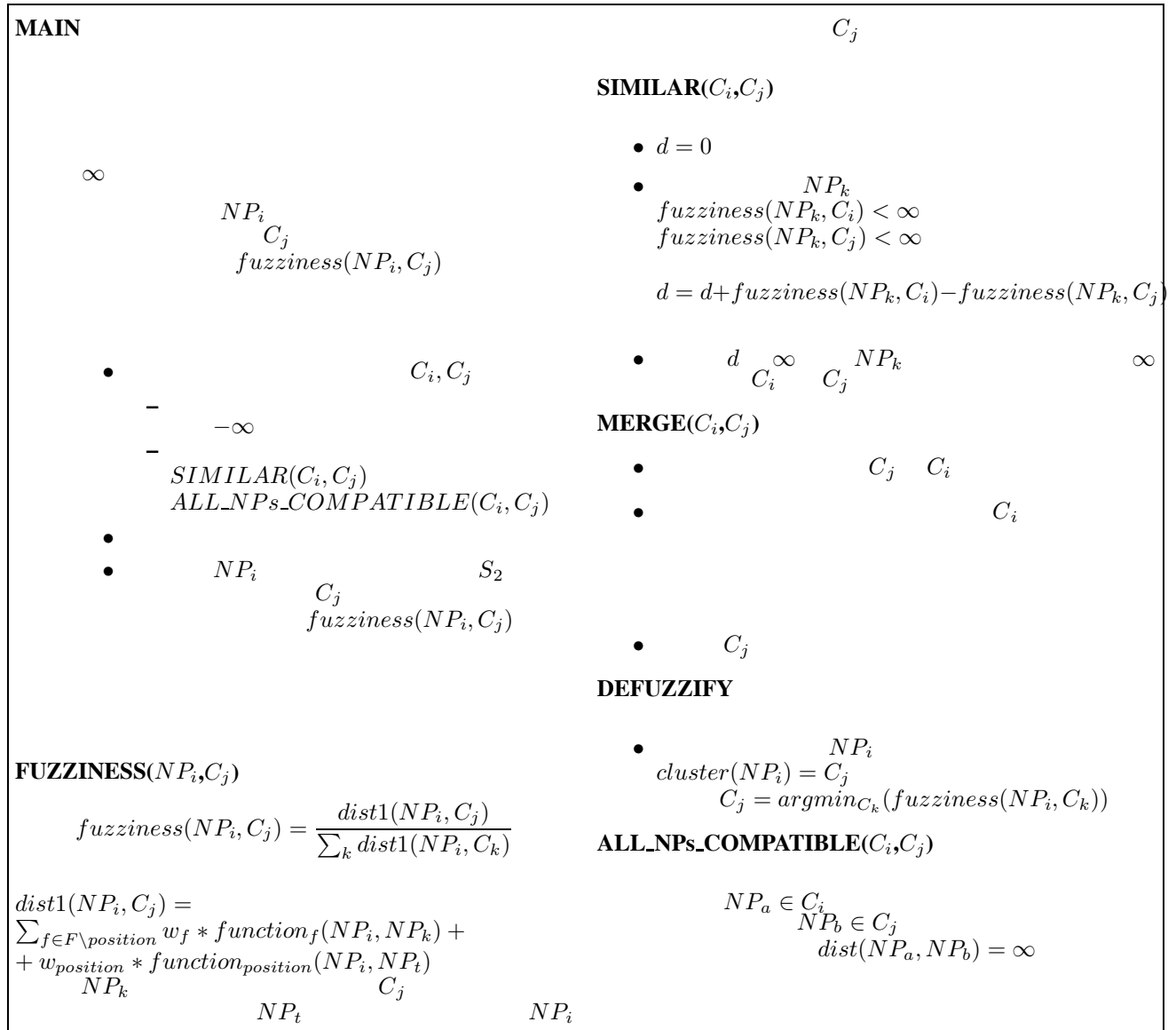


Figure 1. Progressive Fuzzy Clustering Algorithm

- Merging of clusters et al.

 - et al.

 - Search for the best match $-\infty$
et al.

 - Corpus-independent et al.
- Smith” “Smith” fi “Robert
“her” “Smith”

3.4. The Hard Variant (HC-V)

fi

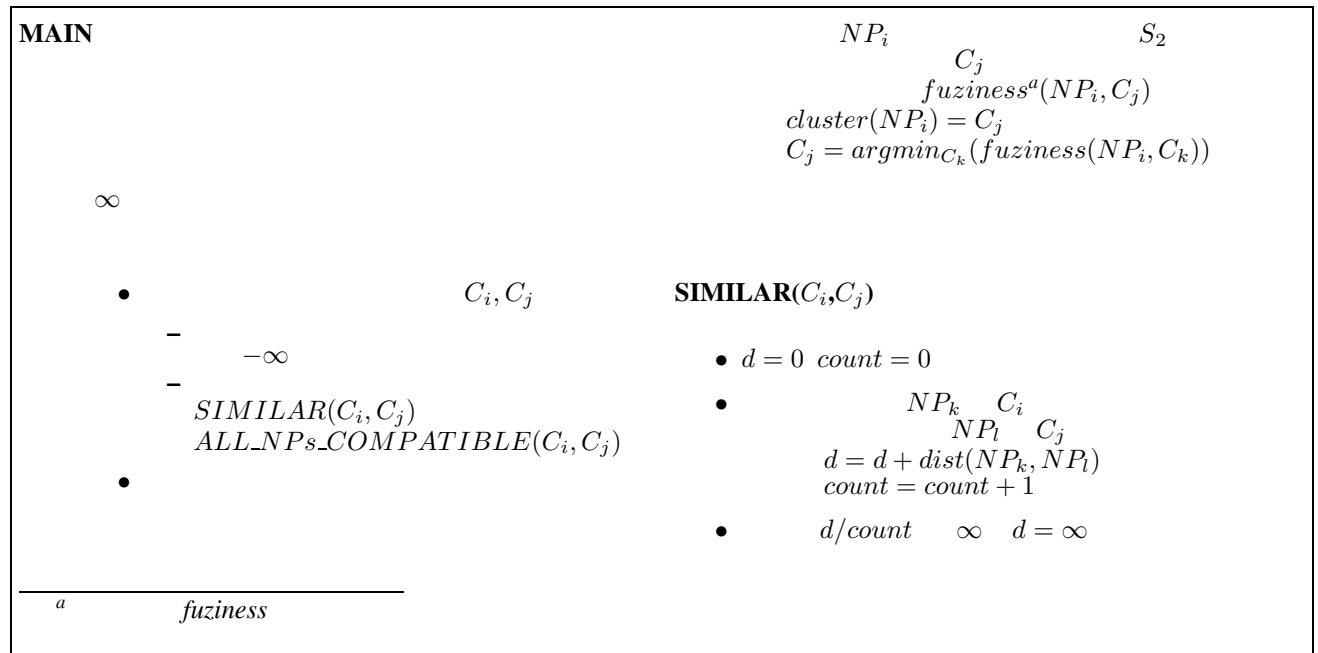


Figure 2. Hard Clustering Variant

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.*

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

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

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

R_i
 i **fi**

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

HC_i i CC_i

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

et al.

fi

fi

fi

fi

Experiment

fi

•

Bentonville Institute

•

“International Business Machines Corp.” and “IBM”

“the Congressional Black Caucus (CBC)”

•

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

Shinkle

“he” “his”

fi

•

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

fi

Experiment

fi

fi

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.

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

- P* *et al.*
-
- et al*
 Proceedings of Document Understanding Conference 2003
- Conference on Empirical Methods in NLP and Very Large Corpora* *Proceedings of the Joint*
-
- Technical Report CS-99-12*
-
- et al.* *International Journal*
of Lexicography (special issue)
-
- et al.* *Proceed-*
ings of the 4'th Dutch-Belgian Information Retrieval Workshop
-