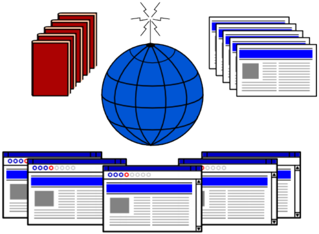
Unsupervised Learning of Relation Detection Patterns

Edgar Gonzàlez i Pellicer
Advisor: Dr. Jordi Turmo i Borràs

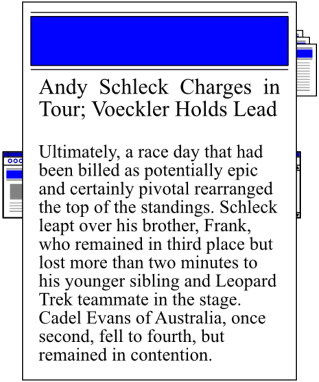
egonzalez@lsi.upc.edu
Universitat Politècnica de Catalunya

June 1st 2012

Information Extraction



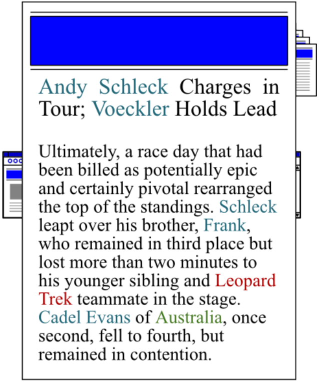
Information Extraction



Andy Schleck Charges in Tour; Voeckler Holds Lead

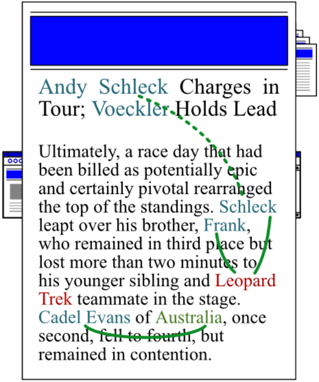
Ultimately, a race day that had been billed as potentially epic and certainly pivotal rearranged the top of the standings. Schleck leapt over his brother, Frank, who remained in third place but lost more than two minutes to his younger sibling and Leopard Trek teammate in the stage. Cadel Evans of Australia, once second, fell to fourth, but remained in contention.

Information Extraction

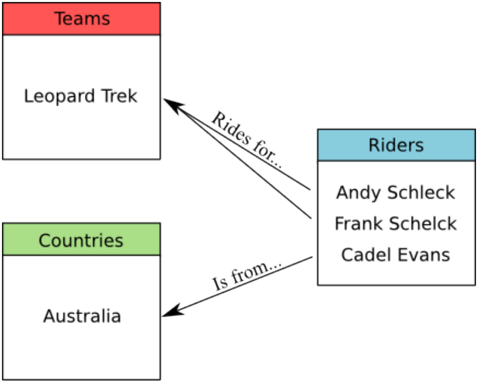

Andy Schleck Charges in Tour; Voeckler Holds Lead

Ultimately, a race day that had been billed as potentially epic and certainly pivotal rearranged the top of the standings. Schleck leapt over his brother, Frank, who remained in third place but lost more than two minutes to his younger sibling and Leopard Trek teammate in the stage. Cadel Evans of Australia, once second, fell to fourth, but remained in contention.

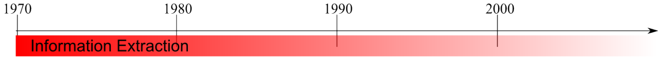
Information Extraction

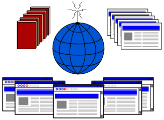


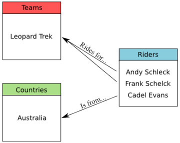
Information Extraction



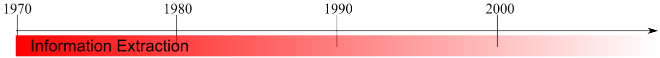
Reduction of Supervision

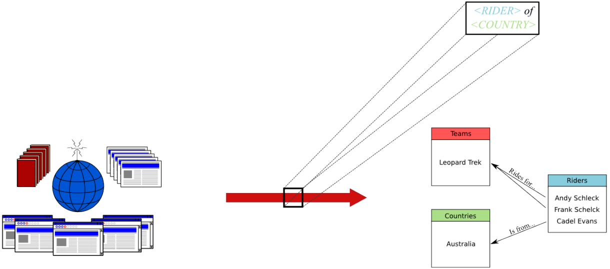




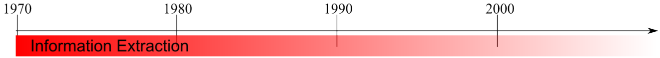


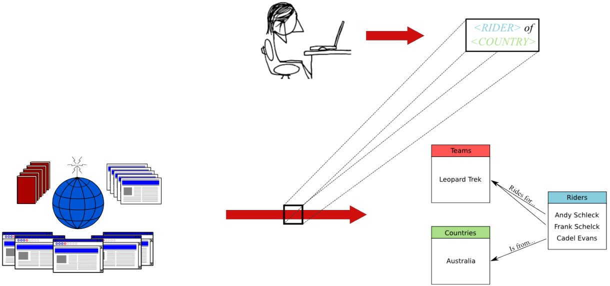
Reduction of Supervision



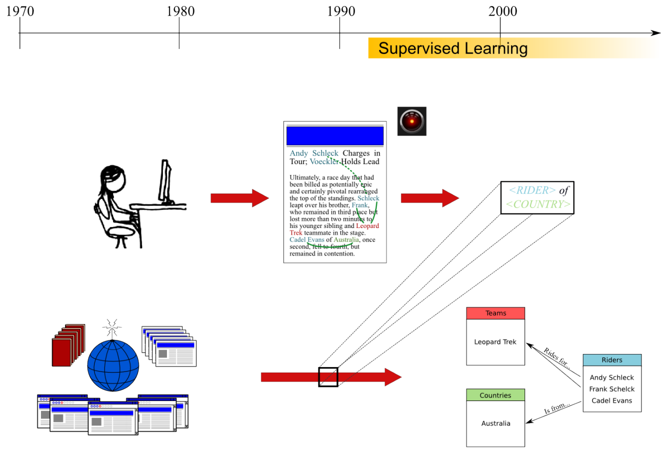


Reduction of Supervision

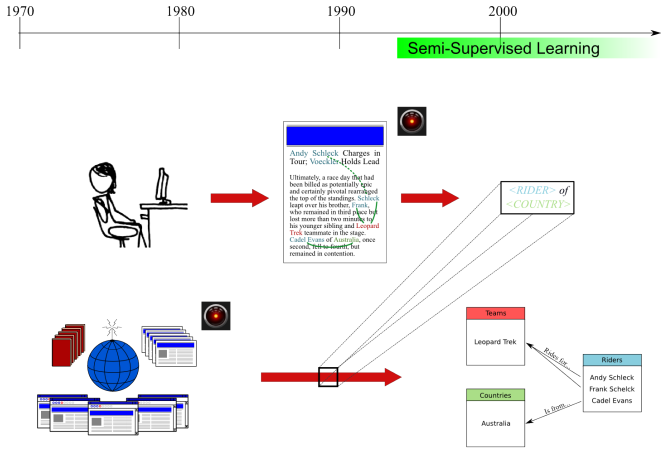




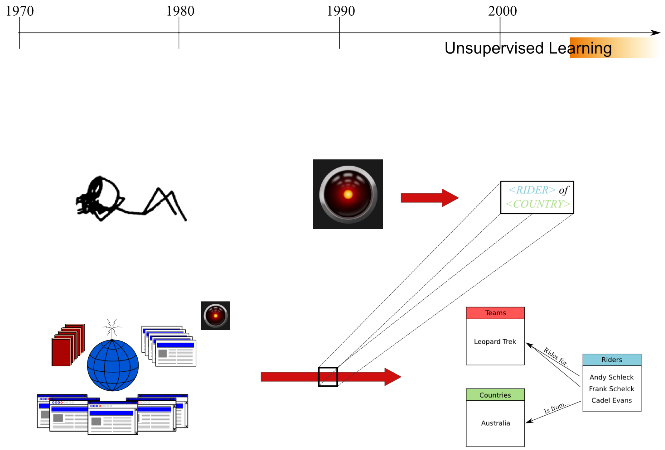
Reduction of Supervision



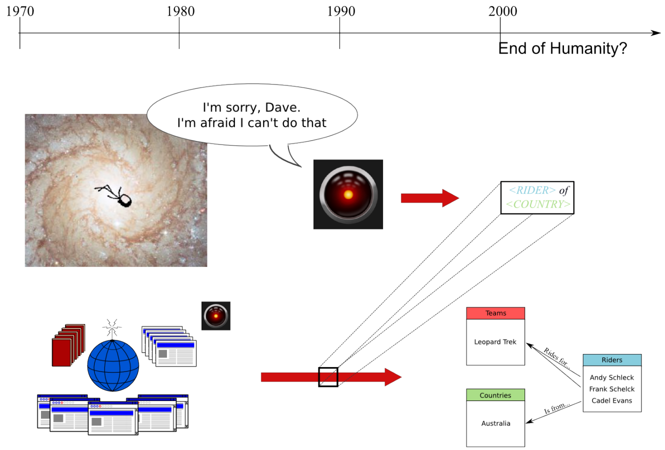
Reduction of Supervision



Reduction of Supervision



Reduction of Supervision



Weakly Supervised Approaches

- Seed documents
  - (Surdeanu et al., 2006)

Weakly Supervised Approaches

- Seed documents
  - (Surdeanu et al., 2006)
- Seed patterns
  - (Yangarber et al., 2000), (Yangarber, 2003), (Stevenson, 2004), (Greenwood and Stevenson, 2006)

Weakly Supervised Approaches

- Seed documents
  - (Surdeanu et al., 2006)
- Seed patterns
  - (Yangarber et al., 2000), (Yangarber, 2003), (Stevenson, 2004), (Greenwood and Stevenson, 2006)
- Seed contexts
  - (Zhang, 2004), (Chen et al., 2006), (Zhou et al., 2008)

Weakly Supervised Approaches

- Seed documents
  - (Surdeanu et al., 2006)
- Seed patterns
  - (Yangarber et al., 2000), (Yangarber, 2003), (Stevenson, 2004), (Greenwood and Stevenson, 2006)
- Seed contexts
  - (Zhang, 2004), (Chen et al., 2006), (Zhou et al., 2008)
- Seed tuples
  - (Brin, 1998), (Agichtein and Gravano, 2000), (Rosenfeld and Feldman, 2006)

The Problems of Supervision

- Introduces a strong bias in the process

The Problems of Supervision

- Introduces a strong bias in the process
    - Avoiding it becomes costly as the corpus size increases

The Problems of Supervision

- Introduces a strong bias in the process
  - Avoiding it becomes costly as the corpus size increases
- Requires a priori knowledge of the scenario

The Problems of Supervision

- Introduces a strong bias in the process
  - Avoiding it becomes costly as the corpus size increases
- Requires a priori knowledge of the scenario
  - Reduces utility of IE as an exploratory tool

Unsupervised Approaches

- IR query
  - (Sudo et al., 2003), (Sekine, 2006), (Shinyama and Sekine, 2006)

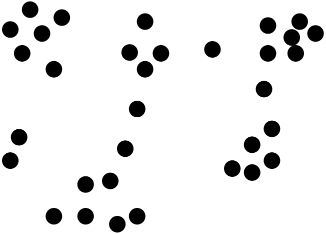
Unsupervised Approaches

- IR query
  - (Sudo et al., 2003), (Sekine, 2006), (Shinyama and Sekine, 2006)
- Trigger words
  - (Etzioni et al., 2004)

Unsupervised Approaches

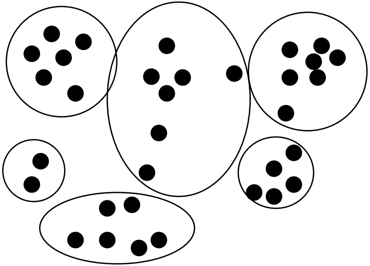
- IR query
  - (Sudo et al., 2003), (Sekine, 2006), (Shinyama and Sekine, 2006)
- Trigger words
  - (Etzioni et al., 2004)
- Entity types
  - (Chen et al., 2005), (Hassan et al., 2006)

Clustering



Tasoulis and Vrahatis, 2004: "The process of partitioning a set of patterns...

Clustering



. . . into disjoint and homogeneous meaningful groups, called clusters."

Clustering

- Clustering
  - is useful as an automatic tool for exploratory analysis of data collections (Hartigan, 1975)

Clustering

- Clustering
  - is useful as an automatic tool for exploratory analysis of data collections (Hartigan, 1975)
  - has been used across a wide spectrum of areas, such as:
    - image segmentation (Silverman and Cooper, 1988),
    - bridge building (Reich and Fenves, 1991),
    - or genetics (Eisen et al., 1998)

Clustering

- Clustering
  - is useful as an automatic tool for exploratory analysis of data collections (Hartigan, 1975)
  - has been used across a wide spectrum of areas, such as:
    - image segmentation (Silverman and Cooper, 1988),
    - bridge building (Reich and Fenves, 1991),
    - or genetics (Eisen et al., 1998)
  - has been applied to problems involving huge collections of data whose contents were mostly unknown

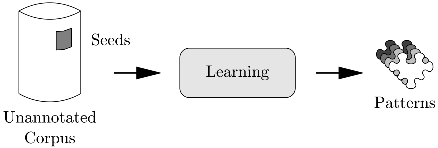
Our Proposal

- Incorporate clustering techniques into the process of IE pattern learning
  - Remove elements of human supervision

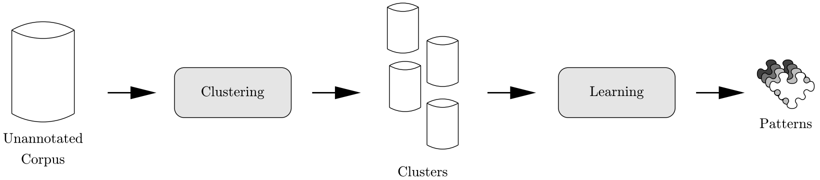
Our Proposal

- Incorporate clustering techniques into the process of IE pattern learning
  - Remove elements of human supervision
- Goal: develop a methodology that
  - from a completely unannotated collection of documents
  - without the need of expert-given seeds
  - produces good quality patterns
    - useful for IE and possibly other NLP tasks

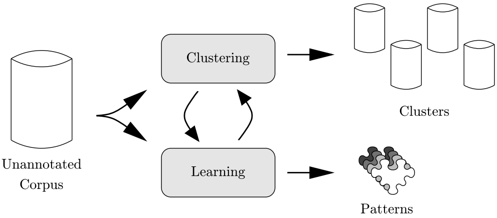
Manual seeding



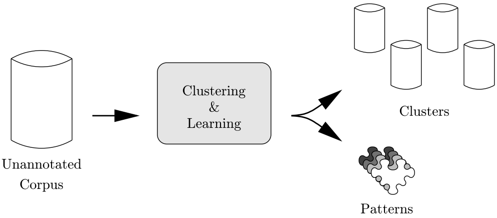
Sequential clustering and learning



Collaborative clustering and learning

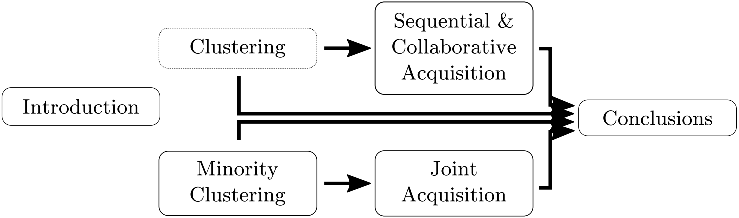


Joint clustering and learning



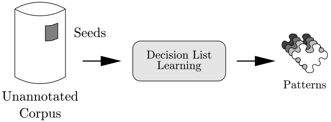
Relation Detection

Outline of the Thesis

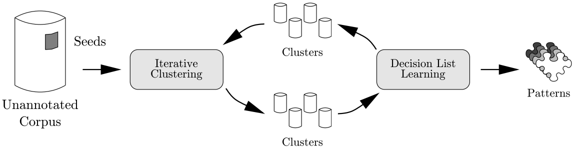


Outline

Decision-List-Based IE Pattern Learning

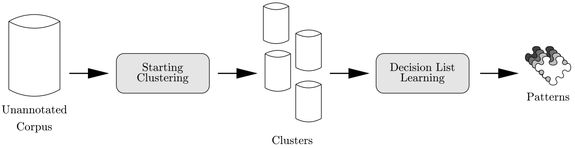


Sequential (Yarowsky, 1995)

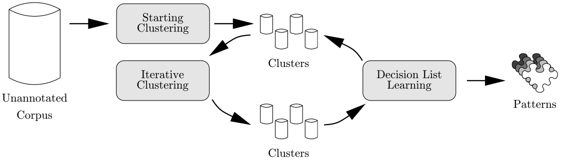


Collaborative (Surdeanu et al., 2005)

Decision-List-Based IE Pattern Learning



Sequential



Collaborative

Pattern Acquisition

- Find
  - Bag-of-words document representation
  - Bag-of-patterns document representation

Pattern Acquisition

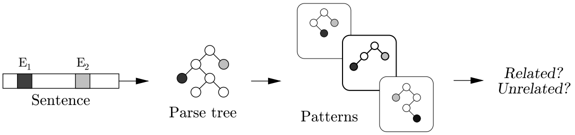
- Find
  - Bag-of-words document representation
  - Bag-of-patterns document representation
- Sequentially/collaboratively learn
  - Word clustering model
  - Pattern decision list classifier
    - $p_a \rightarrow \pi_c$
    - Collins/Riloff criteria

Pattern Acquisition

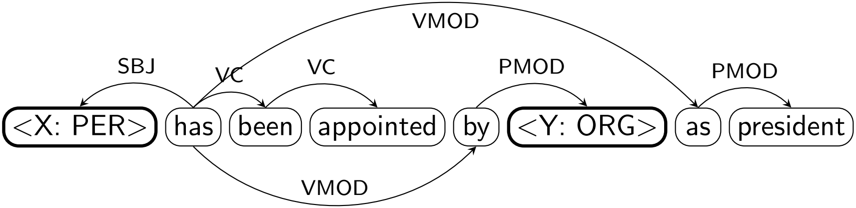
- Find
  - Bag-of-words document representation
  - Bag-of-patterns document representation
- Sequentially/collaboratively learn
  - Word clustering model
  - Pattern decision list classifier
    - $p_a \rightarrow \pi_c$
    - Collins/Riloff criteria
- Use selected antecedents as relation detection patterns

Predicate-Argument Structures p:a

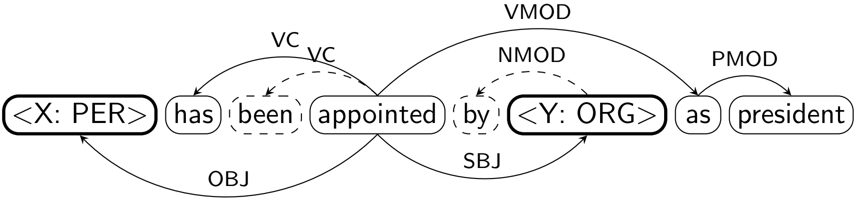
Predicate-Argument Structures p:a



Predicate-Argument Structures `p:a`



Predicate-Argument Structures p:a



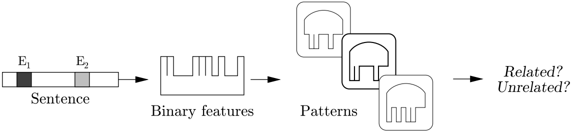
Predicate-Argument Structures `p:a`

- (Yangarber et al., 2000), (Surdeanu et al., 2006)

```
  sv:ORG:appoint
 svo:ORG:appoint:PER
svoio:ORG:appoint:PER:as:president
 svio:ORG:appoint    :as:president
  vo    :appoint:PER
 voio   :appoint:PER:as:president
  vio   :appoint    :as:president
```

Binary Feature Conjunctions

Binary Feature Conjunctions



Binary Feature Conjunctions

	w:t	c:l	w:t+c:l
Structure-based			
Distance between the pair is %d words	●	●	●
Distance between the pair is %d chunks	·	●	●
Left/rightmost entity is of type %t	●	●	●
Word-based			
Word %d positions before/after the left/rightmost entity...			
...has POS tag %t	●	·	●
Chunk-based			
Chunk %d positions before/after that containing the left/rightmost entity...			
...has type %t	·	●	●
...has a head with lemma %l	·	●	●

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	`left`	`after:left:1` `before:right:4`	`after:left:2` `before:right:3`	`after:left:3` `before:right:2`	`after:left:4` `before:right:1`	`right`
Tr.+Val.	`type=PER`	`tag=VBZ`	`tag=VBN`	`tag=VBN`	`tag=IN`	`type=ORG`
Features						

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5,					

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	`left`	`after:left:1` `before:right:4`	`after:left:2` `before:right:3`	`after:left:3` `before:right:2`	`after:left:4` `before:right:1`	`right`
Tr.+Val.	`type=PER`	`tag=VBZ`	`tag=VBN`	`tag=VBN`	`tag=IN`	`type=ORG`
Features	`dist=5`, `left/type=PER`,					

Binary Feature Conjunctions

Word	<X: PER>	has	been	appointed	by	<Y: ORG>
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5, left/type=PER, after:left:1/tag=VBZ, before:right:4/tag=VBZ,					

Binary Feature Conjunctions

Word	<**X: PER**>	has	been	appointed	by	<**Y: ORG**>
Position	`left`	`after:left:1` `before:right:4`	`after:left:2` `before:right:3`	`after:left:3` `before:right:2`	`after:left:4` `before:right:1`	`right`
Tr.+Val.	`type=PER`	`tag=VBZ`	`tag=VBN`	`tag=VBN`	`tag=IN`	`type=ORG`
Features	`dist=5`, `left/type=PER`, `after:left:1/tag=VBZ`, `before:right:4/tag=VBZ`, `after:left:2/tag=VBN`, `before:right:3/tag=VBN`,					

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5, left/type=PER, after:left:1/tag=VBZ, before:right:4/tag=VBZ, after:left:2/tag=VBN, before:right:3/tag=VBN, after:left:3/tag=VBN, before:right:2/tag=VBN,					

Binary Feature Conjunctions

Word	<**X: PER**>	has	been	appointed	by	<**Y: ORG**>
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5, left/type=PER, after:left:1/tag=VBZ, before:right:4/tag=VBZ, after:left:2/tag=VBN, before:right:3/tag=VBN, after:left:3/tag=VBN, before:right:2/tag=VBN, after:left:4/tag=IN, before:right:1/tag=IN,					

Binary Feature Conjunctions

Word	<**X: PER**>	has	been	appointed	by	<**Y: ORG**>
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5, left/type=PER, after:left:1/tag=VBZ, before:right:4/tag=VBZ, after:left:2/tag=VBN, before:right:3/tag=VBN, after:left:3/tag=VBN, before:right:2/tag=VBN, after:left:4/tag=IN, before:right:1/tag=IN, right/type=ORG					

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	left	after:left:1 before:right:4	after:left:2 before:right:3	after:left:3 before:right:2	after:left:4 before:right:1	right
Tr.+Val.	type=PER	tag=VBZ	tag=VBN	tag=VBN	tag=IN	type=ORG
Features	dist=5, left/type=PER, after:left:1/tag=VBZ, before:right:4/tag=VBZ, after:left:2/tag=VBN, before:right:3/tag=VBN, after:left:3/tag=VBN, before:right:2/tag=VBN, after:left:4/tag=IN, before:right:1/tag=IN, right/type=ORG					

- Patterns are conjunctions of these features

  left/type=PER $\wedge$ before:right:1/tag=IN $\wedge$ right/type=ORG

Binary Feature Conjunctions

Word	**<X: PER>**	has	been	appointed	by	**<Y: ORG>**
Position	`left`	`after:left:1` `before:right:4`	`after:left:2` `before:right:3`	`after:left:3` `before:right:2`	`after:left:4` `before:right:1`	`right`
Tr.+Val.	`type=PER`	`tag=VBZ`	`tag=VBN`	`tag=VBN`	`tag=IN`	`type=ORG`
Features	`dist=5`, `left/type=PER`, `after:left:1/tag=VBZ`, `before:right:4/tag=VBZ`, `after:left:2/tag=VBN`, `before:right:3/tag=VBN`, `after:left:3/tag=VBN`, `before:right:2/tag=VBN`, `after:left:4/tag=IN`, `before:right:1/tag=IN`, `right/type=ORG`					

- Patterns are conjunctions of these features

  `left/type=PER` ∧ `before:right:1/tag=IN` ∧ `right/type=ORG`

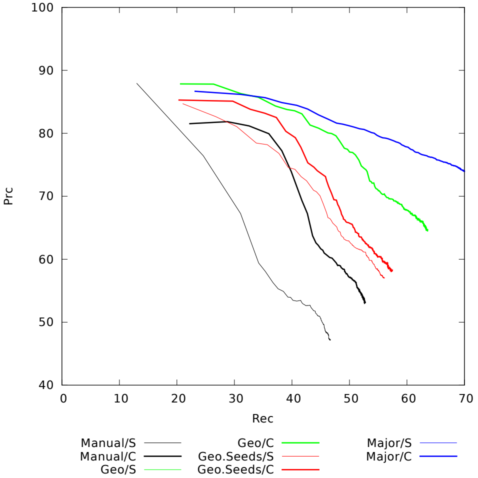
- Requires frequent-itemset mining algorithm (Bayardo, 1998)

Evaluation

- Comparison to (Surdeanu et al., 2006)
- Document corpora
  - APW, LAT, REU, SMT
- Seeds
  - Manual
  - Clustering-based
    - Individual: Geo (Surdeanu et al., 2005)
    - Ensemble: Major
- Approaches
  - Sequential/Collaborative
- Features
  - `p:a`

Results (Extract)

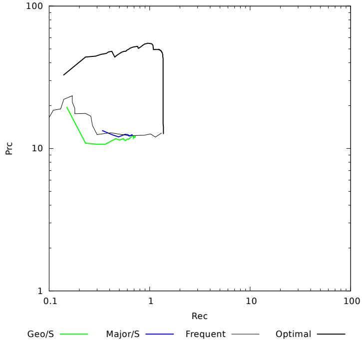
- APW corpus



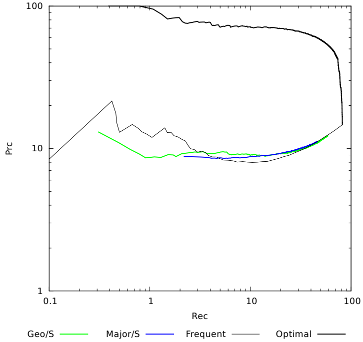
Evaluation

- IE corpus
  - ACE evaluations 2003–2005
- Seeds
  - Clustering-based: Geo, Major
- Approaches
  - Sequential
  - Frequent baseline / Optimal upper-bound
- Features
  - `p:a`
  - `w:t`, `c:l`, `w:t+c:l`

Results

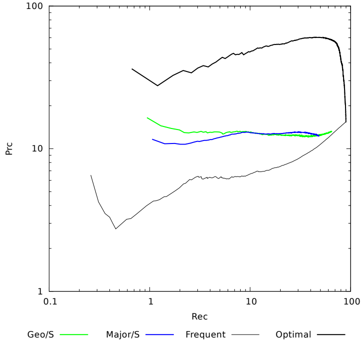


p:a

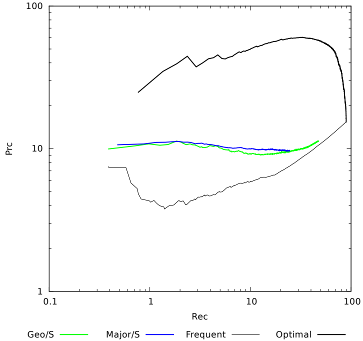


w:t

Results



c:l



w:t+c:l

Analysis

- For each pair of related entities
  - Consider random variables
    - Π_+: assigned document cluster
    - T_+: ACE relation type

Analysis

- For each pair of related entities
  - Consider random variables
    - Π_+: assigned document cluster
    - T_+: ACE relation type
- Their normalized mutual information, $\bar{\mathrm{I}}(\Pi_+ ; T_+)$, is:
  - 0.144 for GEO
  - 0.206 for MAJOR

Analysis

- For each pair of related entities
  - Consider random variables
    - Π_+: assigned document cluster
    - T_+: ACE relation type
- Their normalized mutual information, $\bar{\mathrm{I}}(\Pi_+ ; T_+)$, is:
  - 0.144 for GEO
  - 0.206 for MAJOR

Analysis

- For each pair of related entities
  - Consider random variables
    - Π_+: assigned document cluster
    - T_+: ACE relation type
- Their normalized mutual information, $\bar{\mathrm{I}}(\Pi_+ ; T_+)$, is:
  - 0.144 for GEO
  - 0.206 for MAJOR
- Mismatch!
  - Clustering → Domain-specific relations
    - *one-domain-per-pattern* assumption
  - ACE → Generic relations

Conclusions

- Clustering methods

Conclusions

- Clustering methods
  - . . . can successfully replace manual seeding in sequential and collaborative approaches

Conclusions

- Clustering methods
  - . . . can successfully replace manual seeding in sequential and collaborative approaches
  - . . . but perform poorly on relation detection of generic relations!

Conclusions

- Clustering methods
  - . . . can successfully replace manual seeding in sequential and collaborative approaches
  - . . . but perform poorly on relation detection of generic relations!
    - The *one-domain-per-pattern* assumption is violated

Conclusions

- Clustering methods
  - . . . can successfully replace manual seeding in sequential and collaborative approaches
  - . . . but perform poorly on relation detection of generic relations!
    - The *one-domain-per-pattern* assumption is violated
- Will joint acquisition overcome these limitations?

Joint acquisition

- For two given entity types

Joint acquisition

- For two given entity types
  - Cluster the contexts of co-occurring entities

Joint acquisition

- For two given entity types
  - Cluster the contexts of co-occurring entities
  - (Hasegawa et al., 2004), (Chen et al., 2005), (Eichler et al., 2008)...

Joint acquisition

- For two given entity types
  - Cluster the contexts of co-occurring entities
  - (Hasegawa et al., 2004), (Chen et al., 2005), (Eichler et al., 2008)...
- Two types of objects

Joint acquisition

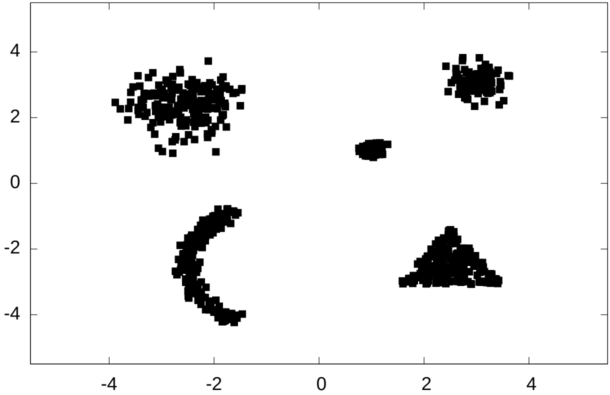
- For two given entity types
  - Cluster the contexts of co-occurring entities
  - (Hasegawa et al., 2004), (Chen et al., 2005), (Eichler et al., 2008)...
- Two types of objects
  - Unrelated pairs → disparate, spread

Joint acquisition

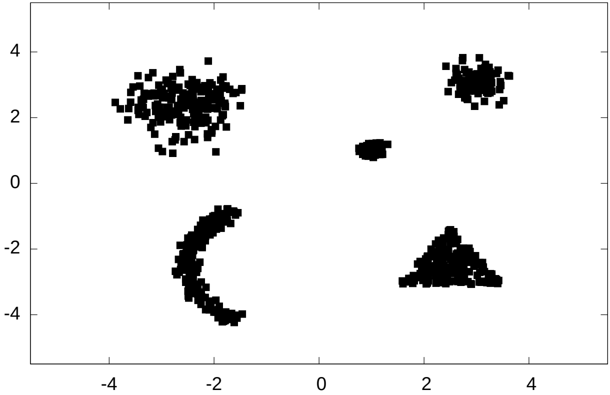
- For two given entity types
  - Cluster the contexts of co-occurring entities
  - (Hasegawa et al., 2004), (Chen et al., 2005), (Eichler et al., 2008)...
- Two types of objects
  - Unrelated pairs → disparate, spread
  - Related pairs → similar, clustered

Outline

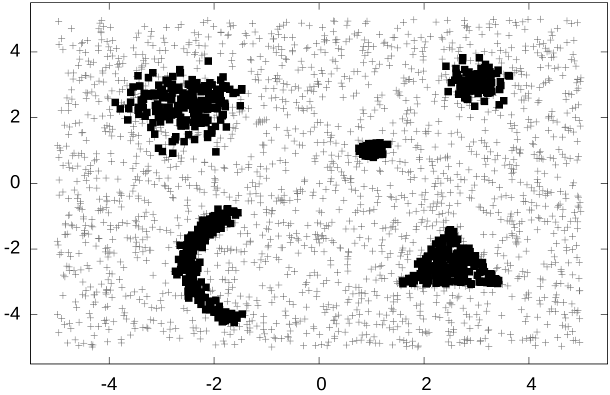
Clustering



All-in Clustering



Minority Clustering



Minority Clustering

- Has also been named
  - One-class clustering (Crammer and Chechik, 2004)
  - Density-based clustering (Gupta and Ghosh, 2006)
  - Minority detection (Ando and Suzuki, 2006)

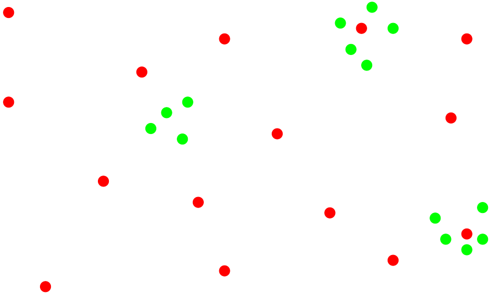
Minority Clustering

- Has also been named
  - One-class clustering (Crammer and Chechik, 2004)
  - Density-based clustering (Gupta and Ghosh, 2006)
  - Minority detection (Ando and Suzuki, 2006)
- Has been compared to
  - "*Clustering needles in a haystack*" (Ando and Suzuki, 2006)

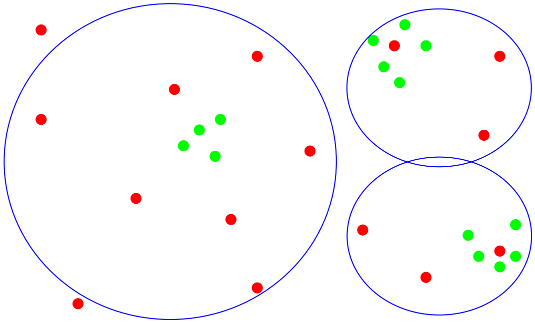
Minority Clustering

- Foreground objects are
  - Dense
  - Sparse

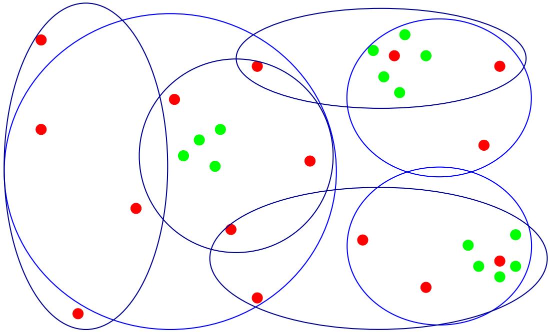
Ensemble Minority Clustering



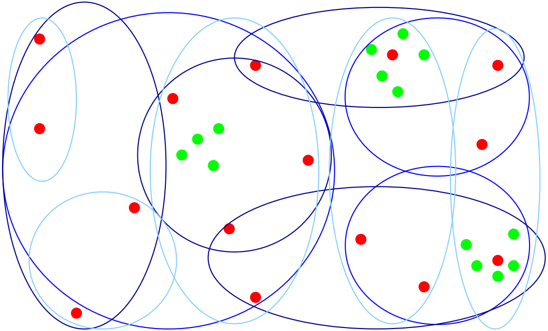
Ensemble Minority Clustering



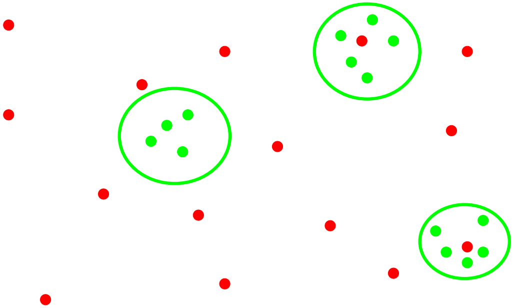
Ensemble Minority Clustering



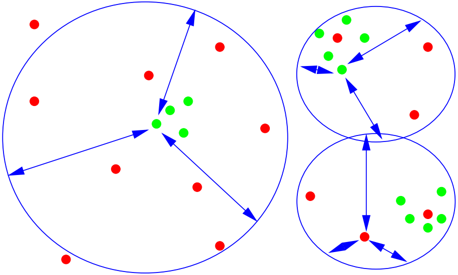
Ensemble Minority Clustering



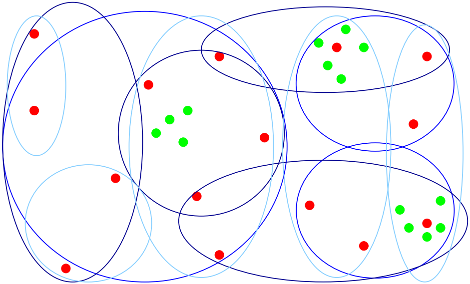
Ensemble Minority Clustering



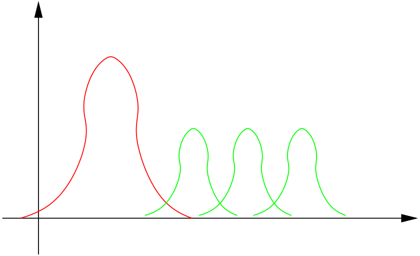
Ensemble Minority Clustering



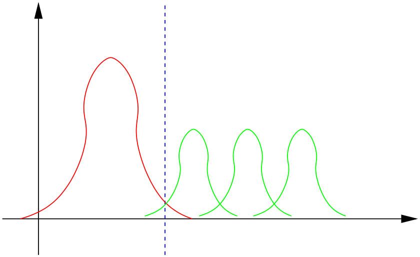
Ensemble Minority Clustering



Ensemble Minority Clustering



Ensemble Minority Clustering



Ensemble Minority Clustering

EWOCS

Ensemble Minority Clustering

Ensemble Weak minOrity Cluster Scoring

Ensemble Minority Clustering

Ensemble **W**eak min**O**rity **C**luster **S**coring

(Yes, pun intended)

Individual clustering

- Random Splitting (RSPLIT; Topchy et al., 2003)
- Random Bregman Clustering (RBC)
  - Family of divergence functions suitable for clustering (Banerjee et al., 2005)
  - Includes:
    - Euclidean distance
    - Mahalanobis distance
    - Kernel-generated distances
    - . . . Gaussian kernel

Individual clustering

- Random Splitting (RSPLIT; Topchy et al., 2003)
- Random Bregman Clustering (RBC)
  - Family of divergence functions suitable for clustering (Banerjee et al., 2005)
  - Includes:
    - Euclidean distance
    - Mahalanobis distance (MAH)
    - Kernel-generated distances
    - . . . Gaussian kernel (G(α,γ), G(AUTO))

Threshold detection

- Threshold detection
  - Best
  - Size
  - Dist
  - nGauss+Best
  - nGauss+Var
  - 2Gauss

Threshold detection

- Threshold detection
  - BEST
  - SIZE
  - DIST
  - NGAUSS+BEST
  - NGAUSS+VAR
  - 2GAUSS

Evaluation

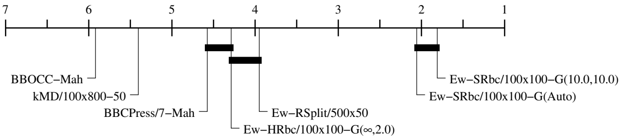
- 160 synthetic datasets
- 2-, 3-, 5- and 8-dimensional data
- Gaussians over uniform distributions

Evaluation

- 160 synthetic datasets
- 2-, 3-, 5- and 8-dimensional data
- Gaussians over uniform distributions
- Baselines
  - Random
  - AllFG
- Alternatives
  - BBOCC (Gupta and Ghosh, 2005)
  - BBCPress (Gupta and Ghosh, 2006)
  - kMD (Ando, 2007)
- Upper-bound
  - Convex

Bergmann-Hommel Tests

F1 (Best threshold)



AUC

Results (Extract)

			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

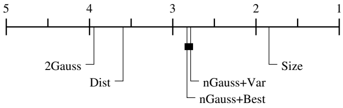
			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Results (Extract)

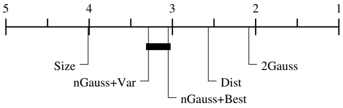
			2 Dimensions				3 Dimensions			
				Best				Best		
			AUC	Prc	Rec	F1	AUC	Prc	Rec	F1
Random	-		**0.500**	14.5	14.5	**14.5**	**0.500**	14.5	14.5	**14.5**
AllFG	-		**0.500**	14.5	100.0	**24.9**	**0.500**	14.5	100.0	**24.9**
BBOCC	-	Mah	**0.752**	40.9	69.4	**44.5**	**0.841**	61.6	62.3	**56.9**
BBCPress	7	Mah	**0.849**	55.6	68.1	**60.7**	**0.934**	79.0	76.8	**77.4**
kMD	100×800–50		**0.808**	82.0	63.7	**68.5**	**0.945**	93.6	90.0	**91.6**
Ew-RSplit	500×2	-	**0.843**	41.1	78.0	**52.1**	**0.911**	55.9	77.2	**63.6**
	500×50	-	**0.862**	45.0	75.9	**54.5**	**0.950**	66.0	83.9	**73.1**
Ew-HRbc	100×100	Mah	**0.896**	59.1	71.5	**63.6**	**0.971**	76.1	85.0	**79.9**
		G(∞,2)	**0.896**	58.9	71.7	**63.5**	**0.971**	76.3	84.9	**79.9**
Ew-SRbc	100×100	Mah	**0.799**	37.1	73.3	**47.5**	**0.901**	53.6	78.2	**62.5**
		G(10,10)	**0.958**	66.4	85.9	**74.6**	**0.991**	85.3	94.9	**89.7**
		G(Auto)	**0.937**	64.5	83.7	**72.5**	**0.986**	83.7	93.2	**88.1**
Convex	-		**0.957**	67.6	100.0	**79.3**	**0.996**	95.4	100.0	**97.6**

Bergmann-Hommel Tests

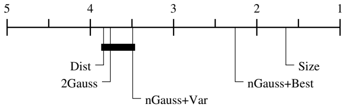
- EwSrbc/100 × 100-G(10,10)



Precision



Recall



F1

Results (2 Dimensions, Extract)

			Best			Size		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	59.1	71.5	**63.6**	61.2	61.2	**61.2**
		G(∞,2)	58.9	71.7	**63.5**	61.2	61.3	**61.2**
Ew-SRbc	100x100	Mah	37.1	73.3	**47.5**	38.1	38.1	**38.1**
		G(10,10)	66.4	85.9	**74.6**	71.2	71.2	**71.2**
		G(Auto)	64.5	83.7	**72.5**	68.7	68.7	**68.7**

			nGauss+Best			nGauss+Var		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	56.0	72.2	**60.5**	41.2	80.0	**45.6**
		G(∞,2)	56.6	72.0	**60.4**	39.6	81.6	**45.5**
Ew-SRbc	100x100	Mah	35.7	74.1	**46.3**	24.3	88.8	**31.7**
		G(10,10)	62.0	88.0	**70.9**	50.3	94.2	**64.2**
		G(Auto)	58.5	87.2	**68.7**	63.7	60.1	**51.4**

Results (2 Dimensions, Extract)

			Best			Size		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	59.1	71.5	**63.6**	61.2	61.2	**61.2**
		G(∞,2)	58.9	71.7	**63.5**	61.2	61.3	**61.2**
Ew-SRbc	100x100	Mah	37.1	73.3	**47.5**	38.1	38.1	**38.1**
		G(10,10)	66.4	85.9	**74.6**	71.2	71.2	**71.2**
		G(Auto)	64.5	83.7	**72.5**	68.7	68.7	**68.7**

			nGauss+Best			nGauss+Var		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	56.0	72.2	**60.5**	41.2	80.0	**45.6**
		G(∞,2)	56.6	72.0	**60.4**	39.6	81.6	**45.5**
Ew-SRbc	100x100	Mah	35.7	74.1	**46.3**	24.3	88.8	**31.7**
		G(10,10)	62.0	88.0	**70.9**	50.3	94.2	**64.2**
		G(Auto)	58.5	87.2	**68.7**	63.7	60.1	**51.4**

Results (2 Dimensions, Extract)

			Best			Size		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	59.1	71.5	**63.6**	61.2	61.2	**61.2**
		G(∞,2)	58.9	71.7	**63.5**	61.2	61.3	**61.2**
Ew-SRbc	100x100	Mah	37.1	73.3	**47.5**	38.1	38.1	**38.1**
		G(10,10)	66.4	85.9	**74.6**	71.2	71.2	**71.2**
		G(Auto)	64.5	83.7	**72.5**	68.7	68.7	**68.7**

			nGauss+Best			nGauss+Var		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	56.0	72.2	**60.5**	41.2	80.0	**45.6**
		G(∞,2)	56.6	72.0	**60.4**	39.6	81.6	**45.5**
Ew-SRbc	100x100	Mah	35.7	74.1	**46.3**	24.3	88.8	**31.7**
		G(10,10)	62.0	88.0	**70.9**	50.3	94.2	**64.2**
		G(Auto)	58.5	87.2	**68.7**	63.7	60.1	**51.4**

Results (2 Dimensions, Extract)

			Best			Size		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	59.1	71.5	**63.6**	61.2	61.2	**61.2**
		G(∞,2)	58.9	71.7	**63.5**	61.2	61.3	**61.2**
Ew-SRbc	100x100	Mah	37.1	73.3	**47.5**	38.1	38.1	**38.1**
		G(10,10)	66.4	85.9	**74.6**	71.2	71.2	**71.2**
		G(Auto)	64.5	83.7	**72.5**	68.7	68.7	**68.7**

			nGauss+Best			nGauss+Var		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	56.0	72.2	**60.5**	41.2	80.0	**45.6**
		G(∞,2)	56.6	72.0	**60.4**	39.6	81.6	**45.5**
Ew-SRbc	100x100	Mah	35.7	74.1	**46.3**	24.3	88.8	**31.7**
		G(10,10)	62.0	88.0	**70.9**	50.3	94.2	**64.2**
		G(Auto)	58.5	87.2	**68.7**	63.7	60.1	**51.4**

Results (2 Dimensions, Extract)

			Best			Size		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	59.1	71.5	**63.6**	61.2	61.2	**61.2**
		G(∞,2)	58.9	71.7	**63.5**	61.2	61.3	**61.2**
Ew-SRbc	100x100	Mah	37.1	73.3	**47.5**	38.1	38.1	**38.1**
		G(10,10)	66.4	85.9	**74.6**	71.2	71.2	**71.2**
		G(Auto)	64.5	83.7	**72.5**	68.7	68.7	**68.7**

			nGauss+Best			nGauss+Var		
			Prc	Rec	F1	Prc	Rec	F1
Ew-HRbc	100×100	Mah	56.0	72.2	**60.5**	41.2	80.0	**45.6**
		G(∞,2)	56.6	72.0	**60.4**	39.6	81.6	**45.5**
Ew-SRbc	100x100	Mah	35.7	74.1	**46.3**	24.3	88.8	**31.7**
		G(10,10)	62.0	88.0	**70.9**	50.3	94.2	**64.2**
		G(Auto)	58.5	87.2	**68.7**	63.7	60.1	**51.4**

Conclusions

- EWOCS is an effective method for ensemble minority clustering

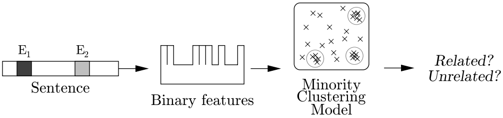
Conclusions

- EWOCS is an effective method for ensemble minority clustering
- In particular, we can build competitive approaches using only unsupervised components

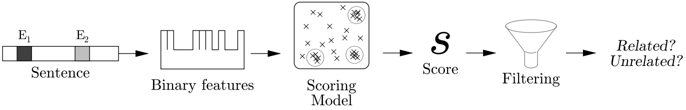
Outline

Relation Detection (Revisited)

Relation Detection (Revisited)



Relation Detection (Revisited)



Advantages

- General and domain-specific relations
- Generation of a model
- Reduced computational cost
- Feature flexibility

Feature sets

	w:t	w:tl	w:tw	c:t	c:tl	c:tw	w:t+c:t	w:t+c:tl
Structure-based								
Distance between the pair is %d words	●	●	●	●	●	●	●	●
Distance between the pair is %d chunks	·	·	·	●	●	●	●	●
Left/rightmost entity is of type %t	●	●	●	●	●	●	●	●
Word-based								
Word %d positions before/after the left/rightmost entity...								
...has POS tag %t	●	●	●	·	·	·	●	●
...has lemma %l	·	●	·	·	·	·	·	·
...can have synset %w	·	·	●	·	·	·	·	·
Chunk-based								
Chunk %d positions before/after that containing the left/rightmost entity...								
...has type %t	·	·	·	●	●	●	●	●
...has a head with POS tag %t	·	·	·	●	●	●	●	●
...has a head with lemma %l	·	·	·	·	●	·	·	●
...has a head which can have synset %w	·	·	·	·	·	●	·	·

EWOCS

- Individual clustering
  - Probabilistic EM (PROB)
  - Random Bregman Clustering (RBC)
  - Random Support Vector Clustering (RSVC)
- Threshold detection
  - DIST

Evaluation

- Training: Unannotated corpus

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus
  - ACE evaluations 2003–2005

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus
  - ACE evaluations 2003–2005
- 11 most frequent entity type pairs

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus
  - ACE evaluations 2003–2005
- 11 most frequent entity type pairs
- Baselines

Evaluation

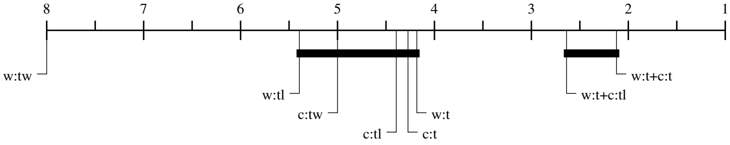
- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus
  - ACE evaluations 2003–2005
- 11 most frequent entity type pairs
- Baselines
  - Base

Evaluation

- Training: Unannotated corpus
  - APW-2000 subset of AQUAINT (29Mw)
- Test: IE corpus
  - ACE evaluations 2003–2005
- 11 most frequent entity type pairs
- Baselines
  - Base
  - Grams (Hassan et al., 2006)

Bergmann-Hommel Tests (Features)

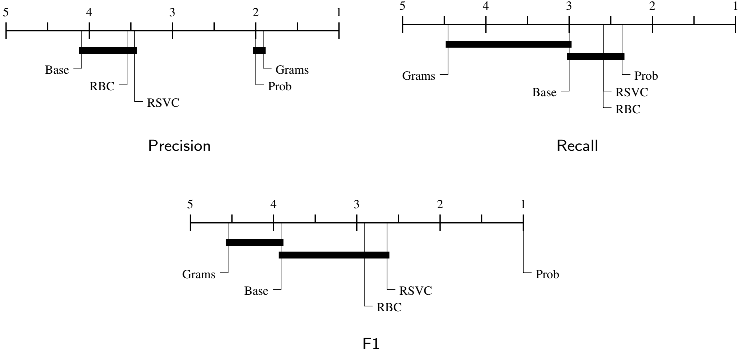
- BEST threshold



F1

Bergmann-Hommel Tests (Clusterer)

- w:t+c:t features, BEST threshold



Results (Extract)

		GRAMS			PROB			RBC			RSVC		
		Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1
GPE-LOC	DIST	—	—	—	60.8	67.6	**64.0**	71.9	47.7	**57.3**	71.3	48.0	**57.2**
	BEST	67.8	58.6	**62.8**	69.4	65.6	**67.5**	61.6	71.9	**66.3**	61.4	72.0	**66.3**
GPE-ORG	DIST	—	—	—	70.1	64.9	**67.2**	67.4	63.9	**65.6**	66.2	65.8	**66.0**
	BEST	68.3	73.8	**70.9**	65.5	78.5	**71.3**	58.6	77.8	**66.8**	61.8	73.6	**66.9**
GPE-PER	DIST	—	—	—	62.5	57.7	**60.0**	71.2	51.3	**59.6**	70.8	51.6	**59.7**
	BEST	55.1	62.6	**58.6**	69.0	55.5	**61.5**	64.7	58.0	**61.2**	63.8	59.0	**61.3**
GPE-VEH	DIST	—	—	—	54.5	73.8	**62.7**	64.1	67.8	**65.9**	63.3	69.4	**66.2**
	BEST	71.1	50.6	**59.1**	63.5	70.5	**66.8**	63.9	69.8	**66.7**	65.1	69.3	**67.1**
ORG-PER	DIST	—	—	—	66.9	59.3	**62.8**	68.1	52.6	**59.3**	67.3	54.5	**60.2**
	BEST	52.1	71.4	**60.2**	67.1	60.3	**63.5**	56.5	70.0	**62.5**	56.5	71.5	**63.1**
ORG-VEH	DIST	—	—	—	73.1	65.0	**68.8**	70.2	66.3	**68.2**	69.4	66.3	**67.8**
	BEST	91.1	50.5	**65.0**	70.1	70.3	**70.2**	78.2	65.3	**71.2**	78.7	64.6	**70.9**

Results (Extract)

		Grams			Prob			Rbc			Rsvc		
		Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1
Gpe-Loc	Dist	—	—	—	60.8	67.6	**64.0**	71.9	47.7	**57.3**	71.3	48.0	**57.2**
	Best	67.8	58.6	**62.8**	69.4	65.6	**67.5**	61.6	71.9	**66.3**	61.4	72.0	**66.3**
Gpe-Org	Dist	—	—	—	70.1	64.9	**67.2**	67.4	63.9	**65.6**	66.2	65.8	**66.0**
	Best	68.3	73.8	**70.9**	65.5	78.5	**71.3**	58.6	77.8	**66.8**	61.8	73.6	**66.9**
Gpe-Per	Dist	—	—	—	62.5	57.7	**60.0**	71.2	51.3	**59.6**	70.8	51.6	**59.7**
	Best	55.1	62.6	**58.6**	69.0	55.5	**61.5**	64.7	58.0	**61.2**	63.8	59.0	**61.3**
Gpe-Veh	Dist	—	—	—	54.5	73.8	**62.7**	64.1	67.8	**65.9**	63.3	69.4	**66.2**
	Best	71.1	50.6	**59.1**	63.5	70.5	**66.8**	63.9	69.8	**66.7**	65.1	69.3	**67.1**
Org-Per	Dist	—	—	—	66.9	59.3	**62.8**	68.1	52.6	**59.3**	67.3	54.5	**60.2**
	Best	52.1	71.4	**60.2**	67.1	60.3	**63.5**	56.5	70.0	**62.5**	56.5	71.5	**63.1**
Org-Veh	Dist	—	—	—	73.1	65.0	**68.8**	70.2	66.3	**68.2**	69.4	66.3	**67.8**
	Best	91.1	50.5	**65.0**	70.1	70.3	**70.2**	78.2	65.3	**71.2**	78.7	64.6	**70.9**

Results (Extract)

		Grams			Prob			Rbc			Rsvc		
		Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1
Gpe-Loc	Dist	—	—	—	60.8	67.6	**64.0**	71.9	47.7	**57.3**	71.3	48.0	**57.2**
	Best	67.8	58.6	**62.8**	69.4	65.6	**67.5**	61.6	71.9	**66.3**	61.4	72.0	**66.3**
Gpe-Org	Dist	—	—	—	70.1	64.9	**67.2**	67.4	63.9	**65.6**	66.2	65.8	**66.0**
	Best	68.3	73.8	**70.9**	65.5	78.5	**71.3**	58.6	77.8	**66.8**	61.8	73.6	**66.9**
Gpe-Per	Dist	—	—	—	62.5	57.7	**60.0**	71.2	51.3	**59.6**	70.8	51.6	**59.7**
	Best	55.1	62.6	**58.6**	69.0	55.5	**61.5**	64.7	58.0	**61.2**	63.8	59.0	**61.3**
Gpe-Veh	Dist	—	—	—	54.5	73.8	**62.7**	64.1	67.8	**65.9**	63.3	69.4	**66.2**
	Best	71.1	50.6	**59.1**	63.5	70.5	**66.8**	63.9	69.8	**66.7**	65.1	69.3	**67.1**
Org-Per	Dist	—	—	—	66.9	59.3	**62.8**	68.1	52.6	**59.3**	67.3	54.5	**60.2**
	Best	52.1	71.4	**60.2**	67.1	60.3	**63.5**	56.5	70.0	**62.5**	56.5	71.5	**63.1**
Org-Veh	Dist	—	—	—	73.1	65.0	**68.8**	70.2	66.3	**68.2**	69.4	66.3	**67.8**
	Best	91.1	50.5	**65.0**	70.1	70.3	**70.2**	78.2	65.3	**71.2**	78.7	64.6	**70.9**

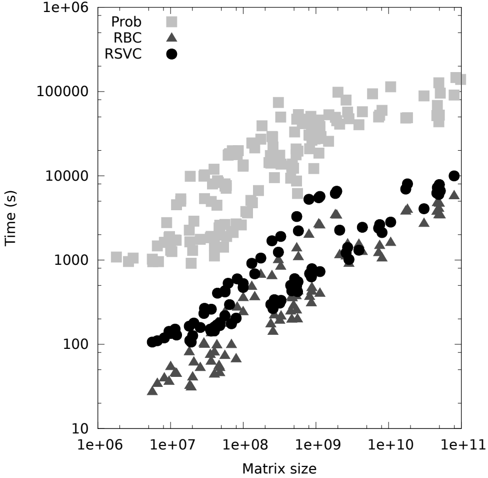
Results (Extract)

		Grams			Prob			Rbc			Rsvc		
		Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1
Gpe-Loc	Dist	—	—	—	60.8	67.6	**64.0**	71.9	47.7	**57.3**	71.3	48.0	**57.2**
	Best	67.8	58.6	**62.8**	69.4	65.6	**67.5**	61.6	71.9	**66.3**	61.4	72.0	**66.3**
Gpe-Org	Dist	—	—	—	70.1	64.9	**67.2**	67.4	63.9	**65.6**	66.2	65.8	**66.0**
	Best	68.3	73.8	**70.9**	65.5	78.5	**71.3**	58.6	77.8	**66.8**	61.8	73.6	**66.9**
Gpe-Per	Dist	—	—	—	62.5	57.7	**60.0**	71.2	51.3	**59.6**	70.8	51.6	**59.7**
	Best	55.1	62.6	**58.6**	69.0	55.5	**61.5**	64.7	58.0	**61.2**	63.8	59.0	**61.3**
Gpe-Veh	Dist	—	—	—	54.5	73.8	**62.7**	64.1	67.8	**65.9**	63.3	69.4	**66.2**
	Best	71.1	50.6	**59.1**	63.5	70.5	**66.8**	63.9	69.8	**66.7**	65.1	69.3	**67.1**
Org-Per	Dist	—	—	—	66.9	59.3	**62.8**	68.1	52.6	**59.3**	67.3	54.5	**60.2**
	Best	52.1	71.4	**60.2**	67.1	60.3	**63.5**	56.5	70.0	**62.5**	56.5	71.5	**63.1**
Org-Veh	Dist	—	—	—	73.1	65.0	**68.8**	70.2	66.3	**68.2**	69.4	66.3	**67.8**
	Best	91.1	50.5	**65.0**	70.1	70.3	**70.2**	78.2	65.3	**71.2**	78.7	64.6	**70.9**

Results (Extract)

		Grams			Prob			Rbc			Rsvc		
		Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1
Gpe-Loc	Dist	—	—	—	60.8	67.6	**64.0**	71.9	47.7	**57.3**	71.3	48.0	**57.2**
	Best	67.8	58.6	**62.8**	69.4	65.6	**67.5**	61.6	71.9	**66.3**	61.4	72.0	**66.3**
Gpe-Org	Dist	—	—	—	70.1	64.9	**67.2**	67.4	63.9	**65.6**	66.2	65.8	**66.0**
	Best	68.3	73.8	**70.9**	65.5	78.5	**71.3**	58.6	77.8	**66.8**	61.8	73.6	**66.9**
Gpe-Per	Dist	—	—	—	62.5	57.7	**60.0**	71.2	51.3	**59.6**	70.8	51.6	**59.7**
	Best	55.1	62.6	**58.6**	69.0	55.5	**61.5**	64.7	58.0	**61.2**	63.8	59.0	**61.3**
Gpe-Veh	Dist	—	—	—	54.5	73.8	**62.7**	64.1	67.8	**65.9**	63.3	69.4	**66.2**
	Best	71.1	50.6	**59.1**	63.5	70.5	**66.8**	63.9	69.8	**66.7**	65.1	69.3	**67.1**
Org-Per	Dist	—	—	—	66.9	59.3	**62.8**	68.1	52.6	**59.3**	67.3	54.5	**60.2**
	Best	52.1	71.4	**60.2**	67.1	60.3	**63.5**	56.5	70.0	**62.5**	56.5	71.5	**63.1**
Org-Veh	Dist	—	—	—	73.1	65.0	**68.8**	70.2	66.3	**68.2**	69.4	66.3	**67.8**
	Best	91.1	50.5	**65.0**	70.1	70.3	**70.2**	78.2	65.3	**71.2**	78.7	64.6	**70.9**

Runtime


Prob
RBC
RSVC
Time (s)
Matrix size
1e+06
100000
10000
1000
100
10
1e+06
1e+07
1e+08
1e+09
1e+10
1e+11

Sample acquisition

- FAC-GPE entities, PROB method, `w:t+c:tl` features

	Feature
+	ch/common/type=NP
+	dist=0ch
+	ch/common/head-tag=FAC
+	ch/before:right:1/head-lemma=in
+	ch/before:right:1/head-tag=IN
+	ch/before:right:1/type=PP
+	dist=1tk
+	dist=2tk
+	ch/common/head-lemma=hospital
+	ch/common/head-lemma=home
+	ch/common/head-lemma=plant
	...
-	ch/left/type=NP
-	ch/before:right:9/head-lemma=mile
-	ch/after:left:4/head-lemma=mile
-	ch/before:right:9/head-tag=NNS
	...

Sample acquisition

- FAC-GPE entities, PROB method, `w:t+c:tl` features

	Feature
+	ch/common/type=NP
+	dist=0ch
+	ch/common/head-tag=FAC
+	ch/before:right:1/head-lemma=in
+	ch/before:right:1/head-tag=IN
+	ch/before:right:1/type=PP
+	dist=1tk
+	dist=2tk
+	ch/common/head-lemma=hospital
+	ch/common/head-lemma=home
+	ch/common/head-lemma=plant
	...
-	ch/left/type=NP
-	ch/before:right:9/head-lemma=mile
-	ch/after:left:4/head-lemma=mile
-	ch/before:right:9/head-tag=NNS
	...

Sample extraction

$(\text{ 13/CC })_{\mathrm{NP}}$ $(\text{ were/VBD released/VBN })_{\mathrm{VP}}$ $(\text{ from/IN })_{\mathrm{PP}}$
$(\text{ a/DT } \underline{\textbf{Virginia/GPE}}_1 \ \underline{\textbf{hospital/FAC}}_2)_{\mathrm{NP}}$
$(\text{ by/IN })_{\mathrm{PP}}$ $(\text{ Monday/NN afternoon/NN })_{\mathrm{NP}}$ ./.

Sample extraction

$(\text{ 13/CC })_{\text{NP}}$ $(\text{ were/VBD released/VBN })_{\text{VP}}$ $(\text{ from/IN })_{\text{PP}}$
$(\text{ a/DT } \underline{\textbf{Virginia/GPE}}_1\ \underline{\textbf{hospital/FAC}}_2)_{\text{NP}}$
$(\text{ by/IN })_{\text{PP}}$ $(\text{ Monday/NN afternoon/NN })_{\text{NP}}$ ./.

- `dist=1tk`
- `dist=0ch`, `ch/common/type=NP`
- `ch/common/head-tag=FAC`, `ch/common/head-lemma=hospital`

Conclusions

- Minority clustering approaches to pattern acquisition are a powerful alternative to existing methods
- EWOCS-based approaches allow the incorporation of new types of information
  - But feature sets cannot be arbitrarily extended

Outline

Our Proposal

- Incorporate clustering techniques into the process of IE pattern learning
  - Remove elements of human supervision

Our Proposal

- Incorporate clustering techniques into the process of IE pattern learning
  - Remove elements of human supervision
- Goal: develop a methodology that
  - from a completely unannotated collection of documents
  - without the need of expert-given seeds
  - produces good quality patterns
    - useful for IE and possibly other NLP tasks

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision
    - Reduction to minority clustering

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision
  - Reduction to minority clustering
  - Flexibility in features

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision
  - Reduction to minority clustering
  - Flexibility in features
  - Compared positively to methods in the state of the art in IE evaluations

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision
  - Reduction to minority clustering
  - Flexibility in features
  - Compared positively to methods in the state of the art in IE evaluations
- Ensemble methods are often superior to individual ones

Conclusions

- It is indeed possible to enhance pattern acquisition with clustering to reduce elements of supervision
  - Reduction to minority clustering
  - Flexibility in features
  - Compared positively to methods in the state of the art in IE evaluations
- Ensemble methods are often superior to individual ones
- Prevalence of Ugly Duckling theorem (Watanabe, 1969)

Contributions

- Joint approach for the acquisition of patterns using a minority clustering algorithm

Contributions

- Joint approach for the acquisition of patterns using a minority clustering algorithm
- EWOCS, a novel minority clustering algorithm based on ensembles

Contributions

- Joint approach for the acquisition of patterns using a minority clustering algorithm
- EWOCS, a novel minority clustering algorithm based on ensembles
- Sequential and collaborative approaches for pattern acquisition and document clustering

Future Work

- Include relation classification

Future Work

- Include relation classification
- Explore extended feature sets

Future Work

- Include relation classification
- Explore extended feature sets
    - Kernels

Future Work

- Include relation classification
- Explore extended feature sets
  - Kernels
- Improve collaborative approaches

Clustering

- Edgar Gonzàlez, Jordi Turmo
  "*Unsupervised Clustering of Spontaneous Speech Documents*"
  InterSpeech 2005
- Edgar Gonzàlez, Jordi Turmo
  "*Non-Parametric Document Clustering by Ensemble Methods*"
  Procesamiento del Lenguaje Natural, vol. 40
- Edgar Gonzàlez, Jordi Turmo
  "*Comparing Non-Parametric Ensemble Methods for Document Clustering*"
  NLDB 2008

Minority Clustering

- Edgar Gonzàlez, Jordi Turmo
  "*Unsupervised Ensemble Minority Clustering*"
  Machine Learning (Submitted)

Relation Detection

- Edgar Gonzàlez, Jordi Turmo
  "*Unsupervised Relation Extraction by Massive Clustering*"
  ICDM 2009

Thank you!

Thank you!

Outline

Bibliography I

Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *ACM Conference on Digital Libraries (DL)*, pages 85–94, 2000.

Shin Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *7th IEEE International Conference on Data Mining (ICDM)*, pages 13–22, 2007.

Shin Ando and Einoshin Suzuki. An information theoretic approach to detection of minority subsets in database. In *6th IEEE International Conference on Data Mining (ICDM)*, pages 11–20, 2006.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Bibliography II

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

Roberto J. Bayardo, Jr. Efficiently mining long patterns from databases. In *ACM SIGMOD International Conference on Management of Data*, pages 85–93, 1998.

Sergei Brin. Extracting patterns and relations from the World-Wide Web. In *International Workshop on the Web and Databases (WebDB)*, 1998.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Unsupervised feature selection for relation extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 262–267, 2005.

Bibliography III

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Relation extraction using label propagation based semi-supervised learning. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 2006.

Koby Crammer and Gal Chechik. A needle in a haystack: Local one-class optimization. In *21st International Conference on Machine Learning (ICML)*, pages 26–33, 2004.

Kathrin Eichler, Holmer Hemsen, and Günter Neumann. Unsupervised relation extraction from web documents. In *International Conference on Language Resources and Evaluation (LREC)*, 2008.

Bibliography IV

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–14868, 1998.

Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in KnowItAll. In *International World Wide Web Conference (WWW)*, 2004.

Mark A. Greenwood and Mark Stevenson. Improving semi-supervised acquisition of relation extraction patterns. In *Workshop on Information Extraction Beyond the Document*, pages 29–35, 2006.

Bibliography V

Gunjan Gupta and Joydeep Ghosh. Robust one-class clustering using hybrid global and local search. In *22nd International Conference on Machine Learning (ICML)*, pages 273–280, 2005.

Gunjan Gupta and Joydeep Ghosh. Bregman bubble clustering: A robust, scalable framework for locating multiple, dense regions in data. In *6th IEEE International Conference on Data Mining (ICDM)*, pages 232–243, 2006.

John A. Hartigan. *Clustering Algorithms*. John Wiley and Sons, 1975.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

Bibliography VI

Hany Hassan, Ahmed Hassan, and Ossama Emam. Unsupervised information extraction approach using graph mutual reinforcement. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

Dekang Lin and Patrick Pantel. DIRT - discovery of inference rules from text. In *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2001.

Yoram Reich and Steven J. Fenves. The formation and use of abstract concepts in design. In Douglas H. Fisher, Jr, Michael J. Pazzani, and Pat Langley, editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 323–354. Morgan-Kaufmann, 1991.

Ellen Riloff. Automatically generating extraction patterns from untagged text. In *13th National Conference on Artificial Intelligence (AAAI)*, pages 1044–1049, 1996.

Bibliography VII

Benjamin Rosenfeld and Ronen Feldman. Ures: an unsupervised web relation extraction system. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 2006.

Satoshi Sekine. On-demand information extraction. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 2006.

Yusuke Shinyama and Satoshi Sekine. Preemptive information extraction using unrestricted relation discovery. In *International Conference on Human Language Technology Research and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2006.

Bibliography VIII

Judith F. Silverman and David B. Cooper. Bayesian clustering for unsupervised estimation of surface and texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10 (4):482–495, 1988.

Mark Stevenson. An unsupervised WordNet-based algorithm for relation extraction. In *Beyond Named Entity: Semantic Labelling for NLP Tasks Workshop at LREC*, pages 37–42, 2004.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. An improved extraction pattern representation model for automatic IE pattern acquisition. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 224–231, 2003.

Mihai Surdeanu, Jordi Turmo, and Alicia Ageno. A hybrid unsupervised approach for document clustering. In *11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

Bibliography IX

Mihai Surdeanu, Jordi Turmo, and Alicia Ageno. A hybrid approach for the acquisition of information extraction patterns. In *EACL Workshop on Adaptive Text Extraction and Mining (ATEM)*, pages 49–56, 2006.

Dimitris K. Tasoulis and Michael N. Vrahatis. Unsupervised distributed clustering. In *IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN)*, 2004.

Alexander Topchy, Anil Kumar Jain, and William Punch. Combining multiple weak clusterings. In *3rd IEEE International Conference on Data Mining (ICDM)*, pages 331–338, 2003.

Satosi Watanabe. *Knowing and Guessing: A Quantitative Study of Inference and Information*. John Wiley and Sons, 1969.

Bibliography X

Roman Yangarber. Counter-training in discovery of semantic patterns. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 343–350, 2003.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Automatic acquisition of domain knowledge for information extraction. In *Conference on Applied Natural Language Processing (ANLP-NAACL)*, pages 940–946, 2000.

David Yarowsky. Unsupervised word sense disambiguation rivaling unsupervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1995.

Zhu Zhang. Weakly-supervised relation classification for information extraction. In *Conference on Information and Knowledge Management (CIKM)*, pages 581–588, 2004.

Bibliography XI

Guodong Zhou, Junhui Li, Longhua Qian, and Qiaoming Zhu. Semi-supervised learning for relation extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 32–39, 2008.