A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution

Zheng Chen, Heng Ji, Robert Haralick

Department of Computer Science

The Graduate Center and Queens College

The City University of New York

September, 2009

Outline

- Task definition
- A pairwise event coreference resolution model
- Motivating examples
  - Add event attributes as features
  - Evaluation metrics
- Extraction of four event attributes
- Experiments and analysis
- Conclusions

Event Coreference Resolution Task

- Grouping all the event mentions into equivalent classes so that all the mentions in each class refer to a unified event (33 event types defined in US NIST *ACE* program)

1. An explosion in a cafe at one of the capital's busiest intersections killed one woman and injured another Tuesday

2. Police were investigating the cause of the explosion in the restroom of the multistory Crocodile Cafe in the commercial district of Kizilay during the morning rush hour

3. The blast shattered walls and windows in the building

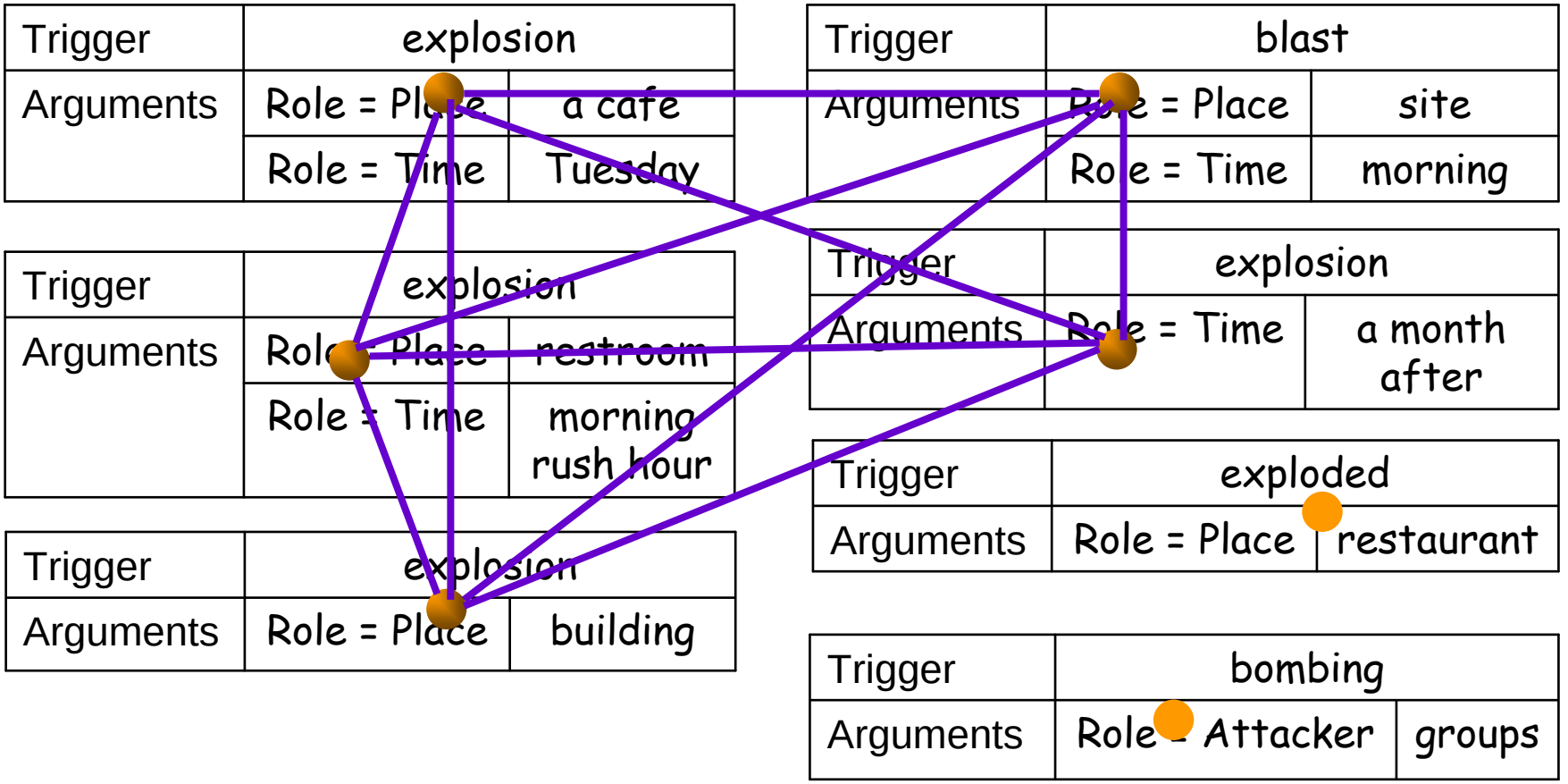
4. Ankara police chief Ercument Yilmaz visited the site of the morning blast

5. The explosion comes a month after

6. a bomb exploded at a McDonald's restaurant in Istanbul, causing damage but no injuries

7. Radical leftist, Kurdish and Islamic groups are active in the country and have carried out the bombing in the past

Event Coreference Resolution



Problems on State-of-the-Art and Our Solutions

- All results were reported based on MUC program (single scenarios, e.g. management succession…)
  - → experiment with 33 event types in ACE program
- Useful linguistic event attributes were largely neglected due to their extremely low weights (~0) in ACE official scoring
  - → automatically label event attributes and incorporate them into coreference
- No formal comparison was conducted on various evaluation metrics
  - → compare three different metrics

Event Coreference Resolution as Agglomerative Clustering

Basic idea:

- Start with singleton event mentions, sort them according to the occurrence in the document
- Traverse through each event mention (from left to right), iteratively merge the active event mention into a prior event or start the event mention as a new event.

Event Coreference Resolution as Agglomerative Clustering

- $coref : E \times M \rightarrow (0,1)$ is a function to give a score to any (event, event mention) pair
- At each iteration, find $e_j \in E_k$ such that

$$e_j = \arg\max_{e_t \in E_k}(coref(e_t, em_k))$$

- If $coref(e_j, em_k) > \delta$, then merge em_k into event e_j otherwise create a new event and add it into E_k
- Train a Maximum-entropy model to learn the coreference function $coref(\cdot,\cdot)$

Standard Features (Base and Distance)

Category	Features	Feature Values (aem : the active event mention, e : a partially-established event, lem: the last event mention in e)
Base	type_subtype	pair of event type and subtype in aem
	nominal	1 if the trigger of aem is nominal
	nom_number	plural or singular if the trigger of aem is nominal
	pronominal	1 if the trigger of aem is pronominal
	exact_match	1 if the trigger spelling of aem matches the trigger spelling of an event mention in e
	stem_match	1 if the trigger stem in aem matches the trigger stem of an event mention in e
	trigger_sim	the maximum of quantized semantic similarity scores (0-5) using WordNet resource among the trigger pairs of aem and an event mention in e
	trigger_pair	trigger pair of aem and lem
	pos_pair	part-of-speech pair of triggers of aem and lem
Distance	token_dist	how many tokens between triggers of aem and lem (quantized)
	sentence_dist	how many sentences aem and lem are apart (quantized)
	event_dist	how many events in between aem and lem (quantized)

Standard Features (Arguments)

Arguments	overlap_num, overlap_roles	overlap number of arguments and their roles (role and id exactly match) between aem and e
	prior_num, prior_roles	the number of arguments that only appear in e and their roles
	act_num, act_roles	the number of arguments that only appear in aem and their roles
	coref_num	the number of arguments that corefer with each other but have different roles between aem and e
	time_conflict	1 if both aem and e have an argument with role "Time-Within" and their values conflict
	place_conflict	1 if both aem and e have an argument with role "Place" and their values conflict

Take a Close Look at Event Attributes

- Modality
  - Expressing *degrees of possibility*, *belief*, *evidentiality*, *expectation*, *attempting*, and *command* (Sauri et al., 2006); An Event is ASSERTED when the author or speaker makes reference to it as though it were a real occurrence; All other events are annotated as OTHER
- Polarity
  - Polarity has a value of NEGATIVE if an event did not occur, otherwise, it has a value of POSITIVE
- Genericity
  - Genericity has a value of SPECIFIC if an event is a singular occurrence at a particular place and time, otherwise, it has a value of GENERIC
- TENSE
  - It is determined with respect to the speaker or author. Possible values: PAST, FUTURE, PRESENT, and UNSPECIFIED

Event Attribute Disagreement Examples

Event Attributes	Event Mentions	Attribute Value
Modality	*Toyota Motor Corp. said Tuesday it will promote Akio Toyoda, a grandson of the company's founder who is widely viewed as a candidate to **some day head** Japan's largest automaker.*	Other
	*Managing director Toyoda, 46, grandson of Kiichiro Toyoda and the eldest son of Toyota honorary chairman Shoichiro Toyoda, **became** one of 14 senior managing directors under a streamlined management system set to be…*	Asserted
Polarity	*At least 19 people were **killed** in the first blast*	Positive
	*There were **no** reports of **deaths** in the blast*	Negative
Genericity	*An **explosion** in a cafe at one of the capital's busiest intersections killed one woman and injured another Tuesday*	Specific
	*Roh has said **any** pre-emptive **strike** against the North's nuclear facilities could prove disastrous*	Generic
Tense	*Israel holds the Palestinian leader responsible for the latest violence, even though the recent **attacks** **were** carried out by Islamic militants*	Past
	*We are warning Israel not to exploit this war against Iraq **to** carry out more **attacks** against the Palestinian people in the Gaza Strip and destroy the Palestinian Authority and the peace process.*	Future

Incorporate Event Attributes into Coreference

- Two event mentions cannot be coreferential if any of the attributes conflict with each other
- State-of-the-art ACE systems all ignored event attributes because of their zero weights in the evaluation scoring
- Our solution: train automatic (MaxEnt) classifiers to predict four event attributes, and then use them as additional features in event coreference resolution
  - Attribute values as features
  - Whether the attributes of an event mention and its candidate antecedent event conflict or not

Features for Event Attribute Classification

Attribute	Features for classification
Common	the trigger and its part-of-speech
	event type and subtype
	the left two words of the trigger (lower case) and their POS tags
	the right two words of the trigger (lower case) and their POS tags
Polarity	the embedding verb of the trigger if any
	a boolean feature indicating whether a negative word exists (not, no, cannot or a word ending with n't) ahead of the trigger and within the clause containing the trigger
Modality	a boolean feature indicating whether a modal auxiliary (may, can, etc.) or modal adverbs (possibly, certainly, etc.) exists ahead of the trigger and within the clause containing the trigger
Genericity	a boolean feature indicating whether the event mention has a "PLACE" argument
	a boolean feature indicating whether the event mention has a "TIME-WITHIN" argument
	the number of arguments that the event mention has except "PLACE" and "TIME-WITHIN"
Tense	the first verb within the clause containing the trigger and its POS tag
	the head words of the "TIME-WITHIN" argument if the event mention has one

Performance of Event Attribute Models

- Most event mentions are POSITIVE, ASSERTED, SPECIFIC and PAST
- Improvements for Polarity, Modality and Genericity over the baselines (majority) are quite limited
- Improvements for Tense are significant, either using perfect event mentions or using system generated event mentions.

	Polarity			Modality			Genericity			Tense		
	P	R	F	P	R	F	P	R	F	P	R	F
Perfect (majority)	0.966	1.0	0.983	0.748	1.0	0.856	0.777	1.0	0.874	0.510	1.0	0.675
Perfect (model)	0.968	1.0	0.984	0.784	1.0	0.879	0.795	1.0	0.885	0.644	1.0	**0.783**
System (majority)	0.969	0.573	0.720	0.779	0.519	0.622	0.792	0.523	0.629	0.550	0.432	0.483
System (model)	0.974	0.574	0.722	0.805	0.527	0.637	0.799	0.525	0.633	0.677	0.484	**0.564**

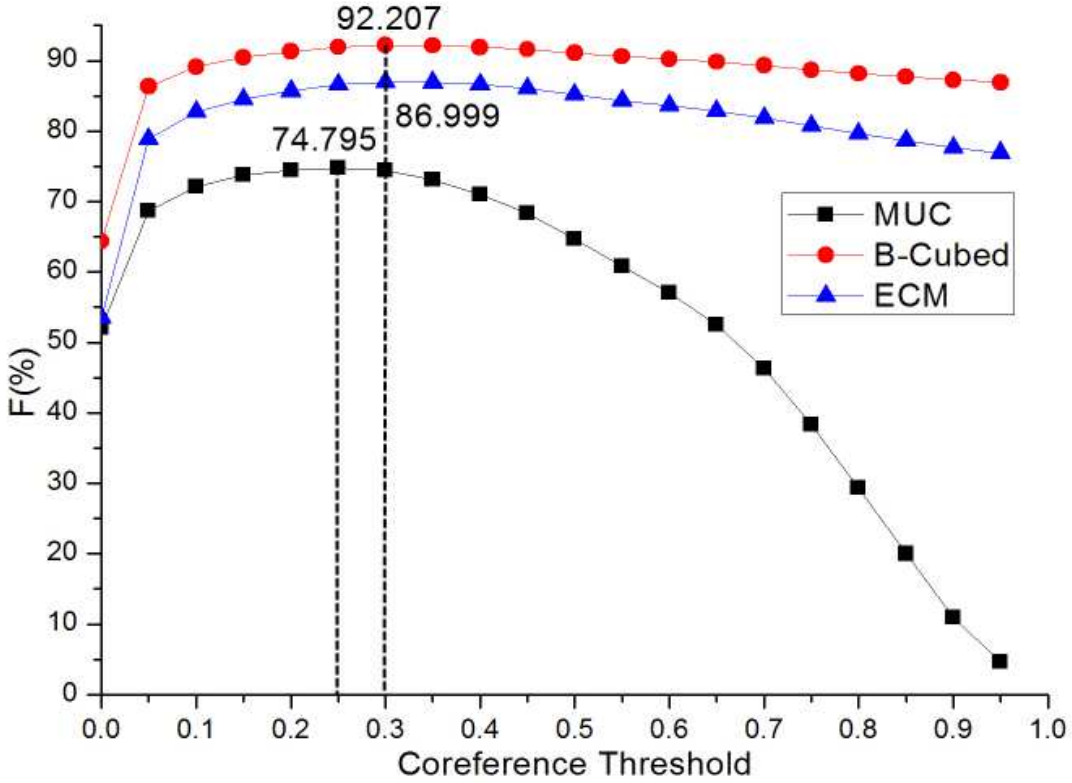
Experiments: Data and Evaluation Metrics

- ACE 2005 English corpus which contains 559 documents
- Use ground truth and system generated event mentions
- Ten times ten-fold cross validation and measured significance with the Wilcoxon signed rank test
- Evaluate the results by three conventional metrics used in entity coreference resolution:
  - MUC F-Measure (Vilain et al., 1995)
  - B-Cubed F-Measure (Bagga and Baldwin, 1998)
  - ECM F-Measure (Luo, 2005)

Evaluation Metrics (Cont')

	Remarks
MUC (Link based) Vilain et al., 1995	Compare the number of links from the hypothesis chains to the number of links from the reference chains. – Precision: the percentage of links out of the total number of links from the hypothesis chains that are correctly identified. – Recall: the percentage of links out of the total number of links from the reference chains that are correctly identified. Drawbacks: – Chains with a single mention are ignored. – Bias towards systems that return longer chains.
B-Cubed (Mention-based) Bagga and Baldwin., 1998	$\text{Precision}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the output chain containing entity}_i}$ $\text{Recall}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the truth chain containing entity}_i}$ $\text{Precision} = \sum_{i=1}^{N} w_i \times \text{Precision}_i$ $\text{Recall} = \sum_{i=1}^{N} w_i \times \text{Recall}_i$ Drawbacks: –An entity is still possible to be used multiple times and thus get double credits
ECM Luo, 2005	Based on the best one-to-one map between reference entities and system entities. Finding the best one-to-one map is a maximum bipartite matching problem and can be solved by the Kuhn-Munkres algorithm.

Determining Coreference Threshold δ



- The best MUC F-score, B-Cubed F-score and ECM F-score are obtained at $\delta = 0.25, \delta = 0.3, \delta = 0.3$ respectively
- MUC metric does not prefer results with many singleton event mentions

Feature Impact Using Ground Truth Event Mentions

	MUC F	B-Cubed F	ECM F
Baseline	38.55%	86.81%	77.67%
+Distance	44.45%	86.60%	78.10%
+Arguments	53.03%	87.90%	80.42%
+Attributes	72.25%	91.86%	86.50%

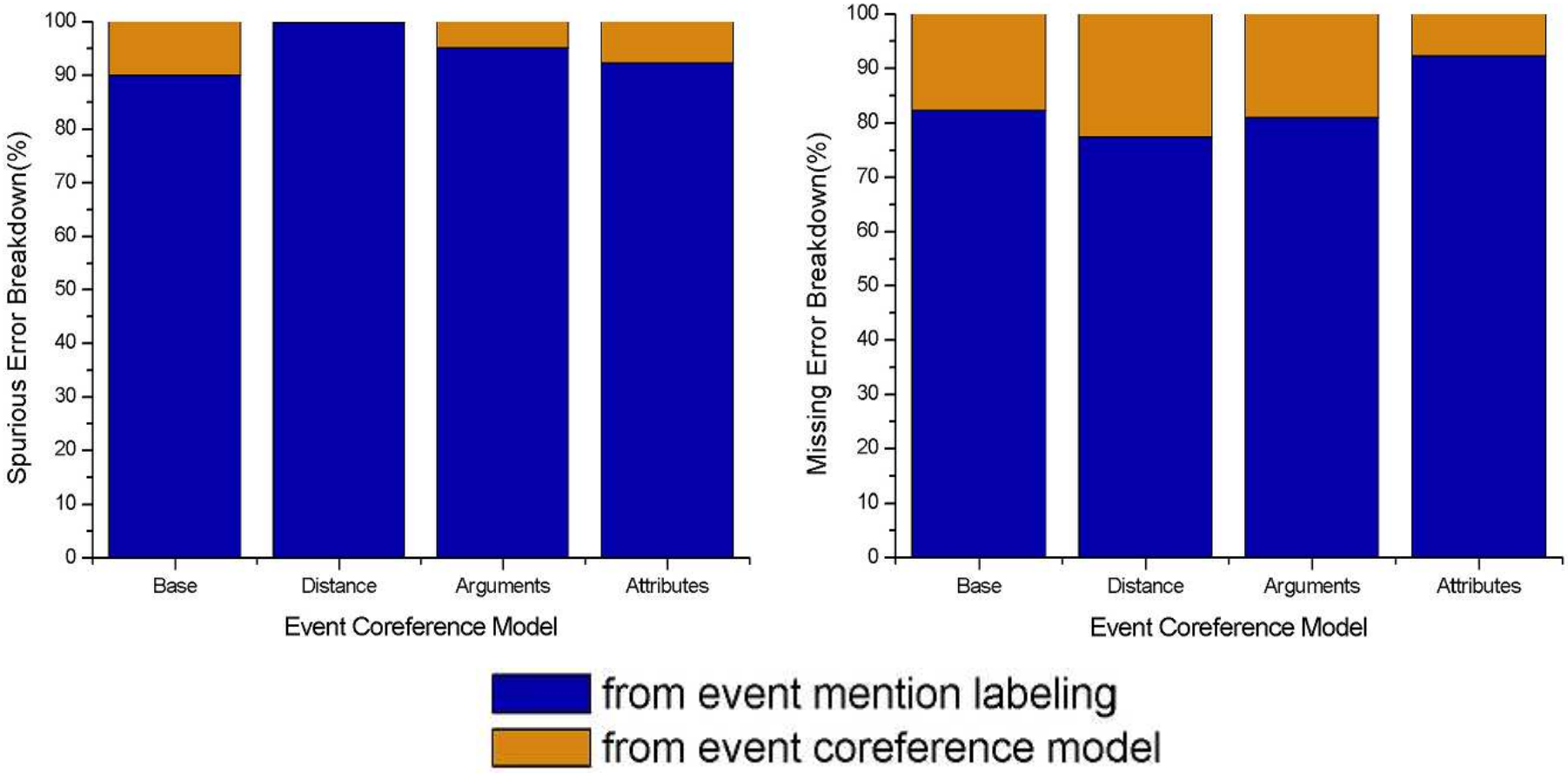
- The distance feature set contributes about 0.4% F-score improvement
- The arguments feature set contributes nearly 2.4% F-score improvement
- The attributes feature set contributes the most significant contribution (6.08% absolute improvement)

Feature Impact Using System Event Mentions

	MUC F	B-Cubed F	ECM F
Base	0.265	0.558	0.489
+Distance	0.254	0.548	0.483
+Arguments	0.274	0.552	0.490
+Attributes	0.28	0.554	0.492

- The aggregated features do not provide significant improvements

Error Analysis on Event Coreference with System Generated Event Mentions



- The performance bottleneck of event coreference resolution comes from the poor performance of event mention labeling

Conclusions and Future Work

- A formal statement of event coreference resolution and an algorithm for the task
- A close study of feature impact on the performance of the pairwise event coreference model
- A new set of features based on event attribute annotations
- A comparison of three evaluation metrics that were previously adopted in entity coreference resolution
- Error analysis proved that event mention labeling is the dominant bottleneck of event coreference resolution

- Future Work
  - Improve attribute labeling using FactBank (Sauri and Pustejovsky, 2009)
  - Incorporate located-in relation extraction results (“egypt”-“mideast”)
  - Event coreference as feedback to improve event mention extraction

Thank you

This work is supported by Google Research, CUNY Research Enhancement Program and GRTI Program