A Unified Event Coreference Resolution by Integrating Multiple Resolvers

Bin Chen[1], Jian Su[2], Sinno Jialin Pan[3] and Chew Lim Tan[4]

[1,2,3]Institute for Infocomm Research, Singapore [4]National University of Singapore
[1]bchen,[2]sujian,[3]jspan@i2r.a-star.edu.sg [4]tancl@comp.nus.edu.sg

Abstract

Event coreference is an important and complicated task in cascaded event template extraction and other natural language processing tasks. Despite its importance, it was merely discussed in previous studies. In this paper, we present a globally optimized coreference resolution system dedicated to various sophisticated event coreference phenomena. Seven resolvers for both event and object coreference cases are utilized, which include three new resolvers for event coreference resolution. Three enhancements are further proposed at both mention pair detection and chain formation levels. First, the object coreference resolvers are used to effectively reduce the false positive cases for event coreference. Second, A revised instance selection scheme is proposed to improve link level mention-pair model performances. Last but not least, an efficient and globally optimized graph partitioning model is employed for coreference chain formation using spectral partitioning which allows the incorporation of pronoun coreference information. The three techniques contribute to a significant improvement of 8.54% in B^3 F-score for event coreference resolution on OntoNotes 2.0 corpus.

1 Introduction

Coreference resolution, the task of resolving and linking different mentions of the same object/event in a text, is important for an intelligent text processing system. The resolved coreferent mentions form a coreference chain representing a particular object/event. Following the natural order in the texts, any two consecutive mentions in a coreference chain form an anaphoric pair with the latter mention referring back to the prior one. The latter mention is called the anaphor while the prior one is named as the antecedent.

Most of previous works on coreference resolution such as (Soon et al, 2001; Yang et al, 2006), aimed at object coreference which both the anaphor and its antecedent are mentions of the same real world object such as person, location and organization. In contrast, an event coreference as defined in (Asher, 1993) is an anaphoric reference to an event, fact, and proposition which is representative of eventuality and abstract entities. In the following example:

*"Israel has **[fired]** missiles on the offices of the Palestinian Authority.*
***[It]** has caused 7 deaths with many injuries…*
*Israel helicopter gunships **[fired]** across the Gaza Strip for more than two hours.*
***[The attack]** in Gaza has been said to cause more violence in Gaza and West Bank and terminate the current round of mid-East peace talk in an unexpected way."*

The four mentions here, ***[fired]***, ***[it], [fired]*** and ***[the attack]*** are referring to the same event (an Israel attack in Gaza Strip on Palestinian Authority). The pronouns noun phrases and action verbs are taken as the representation of events which is also in line with OntoNotes 2.0 practices.

Event coreference resolution is an important task in natural language processing (NLP) research. According to our corpus study, 68.05% of articles in OntoNotes 2.0 corpus contain at least one event chain while 15.52% of all coreference chains are event chains. In addition to the significant proportion, event coreference resolution allows event extraction system to acquire necessary details. Considering the previous example, resolving the event chain ***[fired]-[it]-[fired]-[the attack]*** will provide us all necessary details about the "*air strike*" event mentioned in different sentences. Such details includes "Israel/Israel helicopter gunships" as the actuator, "offices of Palestinian Authority" as the target, "7 deaths and many injuries" as the consequence, "Gaza Strip" as the location and "more than two hours" as the duration. Without a successful event coreference resolution such separated pieces of information cannot be assembled properly.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 102–110,
Chiang Mai, Thailand, November 8 – 13, 2011.

On the other hand, event coreference resolution incurs more difficulties comparing to the traditional object coreference from two aspects. In a semantic view, an object (such as a person, location and etc.) is uniquely defined by its name (e.g. Barrack Obama) while an event requires its role[1] information to distinguish it from other events. For example, "the crash yesterday" – "crash in 1968" shares the same event head phrase "crash", but they are distinguished by the time arguments. In a syntactic view, object coreferences only involve mentions from noun category while event coreference involves mentions from different categories. The syntactic differences will cause the tradition coreference features crippled or malfunctioned as reported by (Chen et al, 2010a;b) for Verb-Pronoun/Verb-NP resolution. In addition to their findings, we further find that even the event NP-Pronoun/NP-NP resolution requires more sophisticated feature engineering than the traditional ones. For example, previous semantic compatibility features only focus on measuring the compatibility between object such as "person", "location" and etc. Event cases are generally falls in the "other" category which provides us no useful information in distinguishing different events. These extra syntactic and semantic difficulties make event coreference resolution a more complicated task comparing to object coreferences.

In this paper, we address the various different event coreference phenomena with seven distinct mention-pair resolvers designed with sophisticated features. We then propose three enhancements to boost up performance at both mention pair detection and chain formation level. First, for the mention-pair resolvers, we have proposed the technique to utilize competitive classifiers' results to further boost mention-pair resolvers' performances. Second, a revised instance selection strategy is proposed to avoid mention-pair resolvers from being misguided by locally preferred instances used previously. Last, on top of coreferent pairs identified by the mention-pair resolvers, we have incorporated the spectral partitioning approach to form the coreference chains in a globally optimized way. Especially, we proposed a technique to enhance the chain level performance by incorporating the pronoun information which the previous attempts did not utilized.

The rest of this paper will be organized in the following way. The next section (section 2) will introduce related works. A review on coreference resolution framework and its weaknesses is presented in section 3. After that we will move on to our proposed model to overcome the weaknesses in section 4. Section 5 will present the experiment results with discussions. Last section will wrap up with a conclusion and future research directions.

2 Previous Work

Although event coreference resolution is an important task, it has not attracted much attention. There is only a limited number of previous works related to this task.

In (Asher, 1993) chapter 6, a method to resolve references to abstract entities using discourse representation theory is discussed. However, no computational system was proposed.

Besides linguistic studies, there are only a few previous works attempting to tackle sub-problems of the event coreference resolution. (Byron, 2002; Müller, 2007; Chen et al, 2010a) attempted event pronoun resolution. (Chen et al, 2010b) attempted resolving noun phrases to verb mentions. All these works only focused on identifying pairs of coreferent event mentions in their targeted phenomena. The ultimate goal, which is extracting event chain, is lack of attention.

(Pradhan, et al, 2007) applied a conventional co-reference resolution system to OntoNotes1.0 corpus using the same set of features for object coreference resolution. However, there is no specific performance reported on event coreference. As (Chen et al, 2010b) pointed out, the conventional features do not function properly on event coreference problem. Thus, a thorough investigation on event coreference phenomena is required for a better understanding of the problem.

3 Resolution Framework

Before we introduce our proposed system to event coreference, we would like to revisit the two-step resolution framework to understand some of its weaknesses. Most of previous coreference resolution system employs a two-steps approach as in (Soon et al, 2001; Nicolae & Nicolae, 2006) and many others. The first step identifies all the pairs of coreferent mentions. The second step forms coreference chains using the coreferent pairs identified from the first step.

[1] Event roles refer to the arguments of the event such as actuator, patient, time, location and etc.

Although a handful of single-step frameworks were proposed recently such as (Cai & Strube, 2010), two-step framework is still widely in use because it has been well-studied. Conceptually, the two-step framework adopts a divide-and-conquer strategy which in turn, allows us to focus on different sub-problems at different stages. The mention-pair detection step allows us to employ many features associated with strong linguistic intuitions which have been proven useful in the previous linguistic study. The chain formation step allows us to leverage on efficient and robust graph partitioning algorithms such spectral partitioning used in this paper. Practically, the two-step framework is also more mature for practical uses and has been implemented as a number of standard coreference resolution toolkits widely available such as RECONCILE in (Stoyanov et al, 2010) and BART in (Versley et al, 2008). Performance-wise, two-step approaches also show comparable performance to single step approaches on some benchmark datasets[2].

In this paper, we are exploiting a brand new type of coreference phenomenon with merely previous attempts. Therefore, we employed the much matured two-step framework with innovative extensions to accommodate complicated event coreference phenomena. Such a divide-and-conquer strategy will provide us more insight for further advancements as well.

3.1 Mention-Pair Resolution Models

Most of mention-pair models adopt the well-known machine learning framework for object coreference as proposed in (Soon et al, 2001).

Instances Generation

In this learning framework, a training/testing instance has the form of $fv(cand_i, ana)$, where *ana* is the anaphor and $cand_i$ is the i^{th} candidate of the given anaphor. During training, we employed the widely used instance selection strategy described in (Ng & Cardie, 2002). In brief, only the closest antecedent of a given anaphor is used as positive instance while only candidates in between the anaphor and its closest antecedent are used as negative instances. During testing, an instance is generated in a similar manner with an additional constraint that the candidate must be within n sentences from the anaphor.

An obvious weakness of such an instance selection strategy is the representation power of the selected instances. Ideally, the selected instances should represent the coreferent status between any two mentions. However this strategy turns the selected set into a local preference representation. The positive instance is the closest preferred mention while the negatives are local non-preferable ones. Such an instance set may help in locally choosing a preferable candidate. But it may be harmful if we want to use the classifier's results in a global approach such as graph partitioning. In the section 4, we will propose a revised instance selection strategy to overcome such a weakness.

SVM with Tree-Kernel

In such a learning framework, many well-known learning models can be applied to the coreference resolution task. In this paper, support vector machine (SVM) is employed for its robust performance in high dimensional space.

In addition to the traditional SVM, we incorporate the syntactic structures through a convolution tree kernel. Tree kernel is used to capture the implicitly structural knowledge embedded in the syntax tree. Effectiveness of various structures was investigated in (Yang et al, 2006; Chen et al, 2010a;b). Based on their findings, we choose minimum-expansion for this paper. In brief, it contains only the path in the parse tree connecting an anaphor and its antecedent. The convolution tree kernel and traditional flat kernel are combined to form a composite kernel.

3.2 Coreference Chain Formation

After the coreferent mention pairs are identified, coreference chains are formed based on those coreferent pairs. There are two major ways to form coreference chains in the literature, best-link heuristic and graph partitioning.

Best-Link Heuristics Approach

The best-link heuristic selects the candidate with highest confidence for each anaphor and forms a "best-link" between them. After that, it simply joins all the mentions connected by "best-links" into the same coreference chain. The best-link heuristic approach is widely used as in (Soon et al, 2001; Yang et al, 2006) because of its simplicity and reasonably good performance.

[2] (Stoyanov et al, 2010) reported RECONCILE(two-steps) achieving 74.25% B^3 f-score on ACE 2005. (Haghighi & Klein, 2010) using single-step approach reported 75.10% B^3 f-score on the same dataset with same train/test-splitting. According to our experiences, such a 0.95% difference is not statistically significant. Other single-step works as (Rahman & Ng, 2009) and (Poon & Domingo, 2008) reported clearly lower B^3 f-score than RECONCILE using same datasets but different train/test-splitting.

The major critics of best-link heuristic fall on its lack of global consideration when forming the coreference chains. The mentions are only joined through locally selected "best-links". Thus the chain consistency is not enforced. Remedies to such a critic are proposed such as best-cut in the next subsection and our proposed method.

Graph Partitioning Approach

Graph partitioning approaches are proposed by various researchers to form coreference chains with global consideration. Here we take Best-Cut proposed in (Nicolae & Nicolae, 2006) as a representative of graph partitioning approaches. Best-Cut is a variant from the well-known minimum-cut algorithm. A graph is formed using all the mentions as vertices. An edge is added between two mentions if a positive output from the mention-pair model. Then the set of edges are iteratively cut to form the coreference chains.

According to (Nicolae & Nicolae, 2006), best-cut does not utilize coreferent pairs involving pronouns. However, event coreference chains contain a significant proportion of pronouns (18.8% of event coreference mentions in the OntoNotes2.0 corpus). Leaving them untouched is obviously not a preferable choice. In the next section, we will propose an alternative chain formation method to incorporate coreferent pronouns into the graph partitioning to accommodate its intensive occurrences in event chains.

4 Our Proposed Model

Our proposed resolution framework follows a similar system flow as the two-step framework which is illustrated in figure 1 for an overview of our resolution system. A brief discussion on various types of event coreference is given in the first subsection 4.1. Each type corresponds to a distinct mention-pair resolver. New features are proposed to capture 3 newly encountered phenomena. After that, we proposed two techniques to improve the mention-pair performance, namely a revised instance selection strategy and utilizing competing classifiers' results. At chain formation step, we also proposed the alternative method, spectral graph partitioning to utilizing pronoun coreferent information.

4.1 Seven Distinct Mention-Pair Models

As we mentioned, one major difficulty of event coreference lies in the gap between different syntactic types of mentions (e.g. nouns, verbs and pronouns). As discussed in (Chen et al, 2010a;b), different syntactic types of coreferent mentions behave differently which requires different features to resolve them. Following this insight, we have built five distinct resolution models for event coreferences involving noun phrases (NP), pronouns and verbs. They are Verb-Pronoun, Verb-NP, Verb-Verb, NP-NP and NP-Pronoun

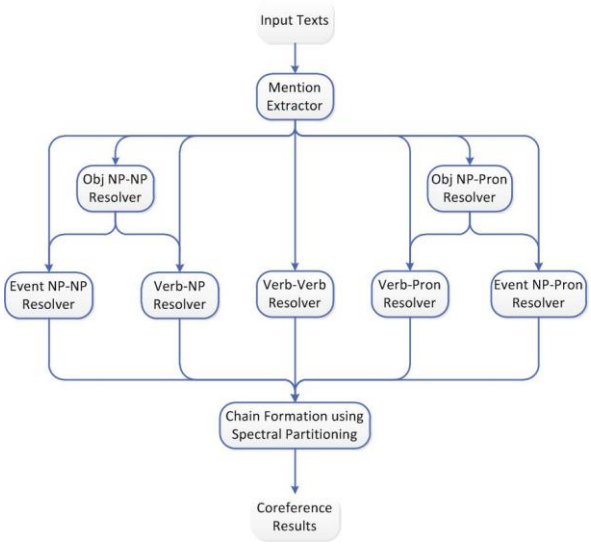

Figure 1: System Overview

resolver. Conventionally, pronouns can only appear as anaphor but not antecedent. Therefore we do not train Pronoun-Pronoun resolvers.

In addition to the syntactic difference, we find event NPs have different behaviors from the object NPs. Event NPs require the event roles to distinguish it from other events while the object NPs are quite self-explaining. The conventional features such as string-matching and head-matching will not work properly when handling cases like "confliction in Mid-East" vs. "confliction in Afghanistan". In our approach, a sophisticated argument matching feature is proposed to capture such information. The arguments information is extracted automatically from the premodifiers and propositional phrase attachments.

Similarly, conventional features try to match mentions into semantic categories like person, location and etc. Then it evaluates the semantic-matching features to pair-up mentions from the same semantic type. However, event NPs exhibit a very different hierarchy in WordNet from the object NPs. A dedicated event hierarchy matching feature is proposed to match event of the same type. With respect to the differences between object NPs and Event NPs, we train two distinct models to handle object NP-NP and event NP-NP resolution separately with distinct features. Similarly, we train separate resolvers

with distinct features for event/object NP-Pronoun. In total we have seven distinct mention-pair resolvers for different syntactic and semantic types of mentions. Five of them focus on event coreference while the other two aim at object coreference. Object coreference results are used to enhance event coreference performance by rule out in appropriate anaphors. All the features we incorporated are tabulated below.

Features	Detail	Used In
Distance	Sentence Distance, Word Distance, Phrase Distance and etc.	All
String-Matching	Full-Match, Partial-Match, Head-Match, Contained-In, Similarity Measures and etc.	eNN oNN VV
Argument-Matching	Event arguments from pre-modifiers and PP-attachments	VN VV eNN
Contexts-Matching	Non-stop-words near the anaphor and antecedent	eNN VN
NP Type	Definite / Indefinite / Proper Name	oNN eNN VN NP
Verb Type	Predicative / Model / Passive / Common	VN VP VV
Pronoun Type	Possessive/Reflexive/Common	oNP VP NP
NE-Semantic	Named entity semantic type	oNN
WN-Event-Semantic	WordNet semantic types of event	eNN eNP
WN-Object-Semantic	WordNet semantic types of object	oNN oNP
Grammatical Roles	Subject/Object in main/sub clauses	All
Synonymic Relation	If anaphor and antecedent share synonym list	eNN VV VN
Morphological Relation	If anaphor and antecedent are morphological	VN
Structural Information	Minimum-Expansion	Except o/eNN

Table 1: Feature List (e:Event; o:object; N:NP; P:Pronoun; V:Verb)

Besides the new features we proposed above (e.g. Event-Semantic and Argument-Matching), the other features we used in the seven mention pair resolvers are employed from a number of previous works such as (Soon et al, 2001; Yang et al, 2008) for object coreference feature, (Chen et al, 2010a;b) for features involving verbs.

Utilizing Competing Classifiers' Results

For the same mention, different mention-pair resolvers will resolve it to different antecedents. Some of these resolution results contradict each other. In the following example:

"USA Today reports ***{some evidence}*** *that has been uncovered shows Bin Laden financed* ***[the attack]*** *and assigned one of his top assistants to supervise* ***[it]****."*

For the anaphor ***[it]***, event NP-Pronoun resolver may pick ***[the attack]*** as antecedent while object NP-Pronoun resolver may pick ***{some evidence}*** as antecedent. Instead of choosing one as the final resolution result from these contradicting outputs, we feed the object resolver results into the event resolvers as a feature and re-train the event resolvers. The idea behind is to provide the learning models with a confidence on how likely the anaphor refers to an object.

Revised Training Instances Selection Strategy

As we mentioned previously, the traditional training instance selection strategy as in (Ng & Cardie, 2002) has a significant weakness. The original purpose of mention pair resolvers is to identify any two coreferent mentions (not restricted to the closest one). By using the previous training instance selection strategy, the selected training instances actually represent a sample space of locally closest preferable mention vs. locally non-preferable mentions. In most of previous works, it shows a reasonably good performance when using with "best-link" chain formation technique. Our investigation shows it actually misguided the graph partitioning methods. Therefore, we propose a revised training instance selection strategy which reflects the true sample space of the original coreferent/non-coreferent status between mentions. In brief, our revised strategy exhaustively selects all the coreferent mention-pairs as positive instances and non-coreferent pairs as negative instances regardless of their closeness to the anaphor. Considering the following example,

"…linking ***{Saudi terrorist Osama Bin Laden}*** *to* ***[the bombing]****.* ***{USA Today}*** *reports* ***{some evidence}*** *that has been uncovered shows* ***{Bin Laden}*** *financed* ***[the attack]*** *and assigned one of his* ***{top assistants}*** *to supervise* ***[it]****."*

The traditional instance selection scheme will only select ***[the attack]–[it]*** as positive instance and ***{top assistants}–[it]*** as negative instance. Our revised instance selection scheme will select an additional positive instance ***[the bombing]–[it]*** and additional negative instance as ***{Bin Laden}–[it]***, ***{USA Today}–[it]*** and other curly brackets NP mentions. Thus the full sample

space is represented using our training instances selection strategy.

4.2 Spectral Graph Partitioning

After deriving the potential coreferent mention pairs, we further use spectral graph partitioning as described in (Ng et al, 2002) to form the globally optimized coreference chains. As we mentioned previously, traditional chain formation technique suffers from a local decision (as in best-link approaches) or failure to incorporate pronoun information (as in best-cut approaches). Spectral graph partitioning shows its advantages over previous approaches. Spectral graph partitioning (aka. Spectral clustering) has made its success in a number of fields such as image segmentation in (Shi & Malik, 2000) and gene expression clustering in (Shamir & Sharan, 2002).

Compared to the "traditional algorithms" such as k-means or minimum-cut, spectral clustering has many fundamental advantages. Results obtained by spectral clustering often outperform the traditional approaches, spectral clustering is very simple to implement and can be solved efficiently by standard linear algebra methods. More attractively, according to (Luxburg, 2006), spectral clustering does not intrinsically suffer from local optima problem. In this paper, the similarity graph is formed in similar way as in (Nicolae & Nicolae, 2006) using SVM confidence[3] outputs.

Utilizing Pronoun Information

Besides the simplicity and efficiency of spectral graph partitioning, one particular reason to employ spectral partitioning is that the previous best-cut approach failed to incorporate pronoun information in their similarity graph. It may not be an issue in object coreference scenario as pronouns are only a relatively small proportion (9.78% of object mentions in OntoNotes). However, in event cases, pronouns contribute 18.8% of the event mentions. As we further demonstrated in our corpus study, event chains are relatively more sparse and shorter than object chains. Removing pronouns from the similarity graph will break a significant proportion of the event chains. Thus we propose this spectral graph partitioning approach to overcome this weakness from the previous models.

Instead of re-implementing the minimum-cut algorithm, we apply the spectral partitioning to a similarity graph without pronoun information. This setting is based on two considerations. Firstly, spectral partitioning is theoretically equivalent to minimum-cut partitioning which means they can handle the same problem set. Secondly, by using the same model, we can eliminate any empirical difference in these two partitioning algorithms and show the true contribution from incorporating pronoun information.

[3] Confidence is computed from kernel outputs using sigmoid function.

5 Experiment Settings and Results

In this section, we present various sets of experiment results to verify the effectiveness of our proposed methods individually and collectively.

5.1 Corpus Study

The corpus we used is OntoNotes2.0 which contains 300K of English news wire data from Wall Street Journal and 200K of English broadcasting news from various sources including (ABC, CNN and etc.). OntoNotes2.0 provides gold annotation for parsing, named entity, and coreference. The distribution of event coreference is tabulated below.

	# of Articles	# of Chains	# of Mentions
Event	1033	2693	7314
Object	1511	14655	54753
Total	1518	17348	62067

Table 2: Event Coreference Distribution in OntoNotes2.0

The distribution of event chains is quite sparse. In average, an article contains only 2.6 event chains comparing to 9.7 object chains. Furthermore, event chains are generally shorter than object chains. Each event chain contains 2.72 mentions comparing to 3.74 in the object chains.

5.2 Performance Metrics & Experiment Settings

In this work, we employ two performance metrics for evaluation purposes. At mention-pair level, we used the standard pair-wise precision/recall/f-score to evaluate the seven mention-pair resolvers. At coreference chain level, we use B-Cube (B^3) measure as proposed in (Bagga & Baldwin, 1998). B^3 provides an overall evaluation of coreference chains instead of coreferent links. Thus it is widely used in previous works.

For each experiment conducted, we use the following data splitting. 400 articles are reserved to train the object NP-Pronoun and NP-NP resolvers. (400 news articles are sufficient for object coreference training, comparing with other data sets used for both training and testing such as 519 articles in ACE-02, 60 articles in MUC-6 and 50 articles in MUC-7.) Among the remaining 1118 articles, we random selected 894 (80%) for

training the 5 event resolvers while the other 224 articles are used for testing.

In order to separate the propagated errors from preprocessing procedures such as parsing and NE tagging, we used OntoNotes 2.0 gold annotation for Parsing and Named Entities only. Coreferent mentions are generated by our system instead of using the gold annotations.

In order to test the significance in performance differences, we perform paired t-test at 5% level of significance. We conduct the experiments 20 times through a random sampling method to perform meaningful statistical significance test.

5.3 Experiment Results

In this section, we will present the experiment results to verify each of the improvements we proposed in previous sections.

Mention-Pair Models Performances

The first set of experiment results presented here is the seven mention-pair resolvers using all conventional settings without any proposed methods.

The Verb-Verb resolver performance is particularly low due to lack of training instances where only 48 positive instances available from the corpus. Our Mention-pair models are not directly comparable with (Chen et al, 2010a;b) which used gold annotation for object coreference information while we resolve such coreferent pairs using our trained resolvers. There are also a number of differences in the preprocessing stage which makes the direct comparison impractical.

Mention-Pair Score	Precision	Recall	F-Score
Event Resolvers			
Verb-Pronoun	32.34	68.32	43.90
Verb-NP	54.22	68.56	60.55
Verb-Verb	22.47	83.33	35.40
NP-Pronoun	46.62	70.47	56.12
NP-NP	48.83	60.08	53.88
Object Resolvers			
NP-NP	58.89	66.04	62.26
NP-Pronoun	61.37	84.33	71.04
Event Chain B^3	**Precision**	**Recall**	**F-Score**
BL	26.67	68.09	38.33

Table 3: Mention-Pair Performance in %

The coreference chains formed using spectral partitioning without any proposed improvements yields a B^3 f-score of 38.33% which serves as our initial baseline (BL) for further comparisons.

Utilizing Competing Classifiers' Results

Since object resolver results are in general better than event resolver, we propose to utilize competing object classifiers' results to improve event resolvers' performance. The experiment results are tabulated below. The "BL+CC" row presents the performance when utilized competing classifiers' results into the baseline system.

Mention-Pair	Precision	Recall	F-Score
Event Verb-Pronoun Resolver			
w/o object info	32.34	68.32	43.90
with object info	45.09	64.73	53.00
Event Verb-NP Resolver			
w/o object info	54.22	68.56	60.55
with object info	56.67	67.61	61.66
Event NP-Pronoun Resolver			
w/o object info	46.62	70.47	56.12
with object info	57.83	69.15	62.99
Event NP-NP Resolver			
w/o object info	48.83	60.08	53.88
with object info	51.35	59.20	55.00
Event Chain B^3	**Precision**	**Recall**	**F-Score**
BL	26.67	68.09	38.33
BL + CC	32.33	67.08	43.61

Table 4: Performance in % using competing classifiers' results

By incorporating the object coreference information, we manage to improve the event coreference resolution significantly, more than 9% F-score for Verb-Pronoun resolver and about 7% F-score for event NP-Pronoun resolver. Object coreference information improves pronoun resolution more than NP resolution. This is mainly because pronouns contain much less information than NP. Such additional information will helps greatly in preventing object pronouns to be resolved by event resolvers mistakenly. Although object coreference is incorporated at mention-pair level, we still measures its contribution to B^3 score at chain level. It improves the B^3 f-score from 38.33% to 43.61% which is a 5.28% improvement. This observation also shows the importance of collective decision of multiple classifiers.

Revised Instance Selection

The second technique we proposed is a revised training instances selection strategy. Table 5 shows improvement using revised instance selection strategy. We refer the traditional instance selection strategy as "BL+CC" and our proposed instance selection strategy as "BL+CC+RIS" (Revised Instance Selection). At mention-pair level we take event NP-Pronoun resolver for demonstration. Similar behaviors are observed in all the mention-pair models. In order to demonstrate power of revised instance selection scheme, we evaluate the mention-pair results in two different ways. The best-candidate evaluation follows the traditional mention pair evaluation. It firstly groups mention-pair predictions by

anaphor. Then an anaphor is correctly resolved as long as the candidate-anaphora pair with highest resolver's score is the true antecedent-anaphor pair. The correct/wrong of other candidates' resolution outputs are not counted at all. The coreferent link evaluation counts each candidate-anaphor pair resolution separately. Intuitively, best-candidate evaluation measures how good a resolver can rank the candidates while the coreferent link evaluation measures the how good a resolver identifies coreferent pairs.

An interesting phenomenon here is the performance evaluation using the best candidate actually drops 3.26% in f-measure when employing the revised instance selection scheme. But when we look at the coreferent link results, the revised instance selection scheme improves the performance by 2.84% f-measure. As a result, our revised instance selection scheme trains better classifier with higher coreferent link prediction results. Since this coreferent link information is further used in the final chain formation step. Our revised scheme contributes an improvement on the final event chain formation by 2.02% F-Score in B^3 measure.

Mention-Pair Score	**Precision**	**Recall**	**F-Score**
Event NP-Pronoun using **Best Candidate** Evaluation			
BL+CC	57.83	69.15	62.99
BL+CC+RIS	52.05	67.11	58.63
Event NP-Pronoun using **Coreferent Link** Evaluation			
BL+CC	39.96	64.03	49.21
BL+CC+RIS	43.33	65.47	52.15
Event Chain B^3	**Precision**	**Recall**	**F-Score**
BL+CC	32.33	67.08	43.61
BL+CC+RIS	35.21	64.74	45.63[4]

Table 5: Performance in % using revised instance selection

This observation shows that the traditional mention-pair model should be revised to maximize the coreferent link performance instead of the traditional best-candidate performance. Because the coreferent link performance is more influential to the final chain formation process using graph partitioning approach.

Spectral Partitioning Utilizing Pronoun Information

The third improvement we proposed is the spectral partitioning with pronoun information. The performance improvement is demonstrated in table 6. In order to separate the contribution from incorporating pronouns and revising instance selection, we conducted the experiment using traditional training instance selection.

B^3 Performance	**Precision**	**Recall**	**F-Score**
BL	26.67	68.09	38.33
BL+CC	32.33	67.08	43.61
BL+CC+Pron	34.14	69.65	45.82[5]
BL+CC+RIS+Pron	35.27	70.02	46.91

Table 6: Performance in % using pronoun information

By incorporating the coreferent pronoun information, the performance is significantly improved by 2.19% in f-measure. By further incorporating the revised instance selection scheme, we achieve B^3-Precision/Recall/F-Score as 35.27 / 70.02 / 46.91% respectively which is an 8.54% F-score improvement from the initial resolution system. 46.91% F-score is the highest performance we achieved in this event coreference resolution work.

6 Conclusion and Future Works

This paper presents a unified event coreference resolution system by integrating multiple mention-pair classifiers including 3 new mention-pair resolvers. Furthermore, we proposed three techniques to enhance the resolution performance. First, we utilize the competing classifiers' results to enhance mention-pair model. Then we propose the revised training instance selection scheme to provide better coreferent link information to graph partitioning model. Lastly, we employ spectral partitioning method with pronoun information to improve chain formation performance. All the three techniques contribute to a significant improvement of 8.54% over the initial 38.33% in B^3 F-score. In future, we plan to incorporate more semantic knowledge for mention-pair models such as semantic roles and word senses. For chain formation, we plan to incorporate domain knowledge to enforce chain consistency.

[4] The B^3-F-Score difference between RIS and Baseline is statistically significant using paired t-test at 5% level of significance

[5] The B^3-F-Score difference between Baseline and Baseline+Pronoun is statistically significant using paired t-test at 5% level of significance

References

Asher,N. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publisher.

Bagga,A. & Baldwin,B. 1998. Algorithms for scoring coreference chains. *In Proceedings of the Linguistic Coreference Workshop at Conference on Language Resources and Evaluation (LREC-1998).*

Byron,D. 2002. Resolving Pronominal Reference to Abstract Entities, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, USA.

Cai,J. & Strube,M. 2010. End-to-End Coreference Resolution via Hypergraph Partitioning. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, China,

Chen,B.; Su,J. & Tan,C.L. 2010a. A Twin-Candidate Based Approach for Event Pronoun Resolution using Composite Kernel. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, China,

Chen,B.; Su,J. & Tan,C.L. 2010b. Resolving Noun Phrases to their Verbal Mentions. In *Proceeding of conference on Empirical Methods in Natural Language Processing (EMNLP)*, USA.

Hovy,E.; Marcus,M.; Palmer,M.; Ramshaw,L. & Weischedel,R. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, USA

Luxburg,U. 2006. A tutorial on spectral clustering. *In (MPITechnical Reports No. 149).* Tubingen: Max Planck Institute for Biological Cybernetic

Müller,C. 2007. Resolving it, this, and that in unrestricted multi-party dialog. *In Proceedings of ACL-2007*, Czech Republic.

Ng,A.; Jordan,M. & Weiss,Y. 2002. On spectral clustering: analysis and an algorithm. *In Advances in Neural Information Processing Systems* 14 (pp. 849 – 856). MIT Press

Ng,V. & Cardie,C. 2002. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of conference on Empirical Methods in Natural Language Processing (EMNLP)*, USA.

Nicolae,C & Nicolae,G. 2006. BESTCUT: A Graph Algorithm for Coreference Resolution. In *Proceedings of conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austrilia.

Poon,H. & Domingos.P. 2008. Joint Unsupervised Coreference Resolution with Markov Logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008*.

Pradhan,S.; Ramshaw,L.; Weischedel,R.; MacBride,J. & Micciulla,L. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, USA.

Shamir,R.; and Sharan,R. 2002. Algorithmic approaches to clustering gene expression data. *Current Topics in Computational Molecular Biology*, 269–300.

Shi,J. & Malik,J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI).*

Sidner,C.L. 1983. Focusing in the comprehension of definite anaphora. In *Computational Models of Discourse*. MIT Press.

Soon,W.; Ng,H. & Lim, D. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521– 544.

Stoyanov,V.; Cardie,C.; Gilbert,N.; Riloff,E.; Buttler,D. & Hysom,D. 2010 Coreference Resolution with Reconcile. In *Proceedings of the Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*

Versley,Y.; Ponzetto,S.P.; Poesio,M.; Eidelman,V.; Jern,A.; Smith,J.; Yang,X. & Moschitti,A. 2008. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Yang,X.; Su,J. & Tan,C.L. 2006. Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge. *In Proceedings of the Conference of the 46th Annual Meeting of the Association for Computational Linguistics*. Australia.

Yang,X.; Su,J. & Tan,C.L. 2008. A Twin-Candidates Model for Learning-Based Coreference Resolution. In *Computational Linguistics*, 34(3):327-356.