Joint Modeling for Chinese Event Extraction with Rich Linguistic Features

Chen Chen and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Goal

- Advance the state of the art in Chinese event extraction
  - This work's focus: ACE Chinese event extraction

ACE Event Extraction

Task: Extract instances of predefined event type from documents

([Resnef] **<u>acquired</u>** [Yugansk], paying only [9.35 billion dollars])
[Resneft]<u>收购</u>[尤甘斯克]付出了仅[93.5亿美元]

- 4 subtasks:
  - Trigger identification
    - **<u>收购(acquired)</u>**
  - Trigger type determination
    - ***Transfer-Money***
  - Argument identification
    - **Resneft(Resneft) ,尤甘斯克(Yugansk) and 93.5亿美元(9.35 billion dollars)**
  - Argument role determination
    - Resneft[***BUYER***],尤甘斯克[***ARTIFACT***] and 93.5亿美元[***PRICE***]

Evaluation Dataset

- All 633 Chinese document in Automatic Content Extraction (ACE) Evaluation 2005 training corpus
  - 33 trigger types
    - E.g., DIE, Transport, Attack and so on.

Baseline System

- Our implementation of Li et al.'s (2012) system
  - State of the art ACE Chinese event extraction system
  - Pipeline architecture
  - Provides us with a Baseline feature set

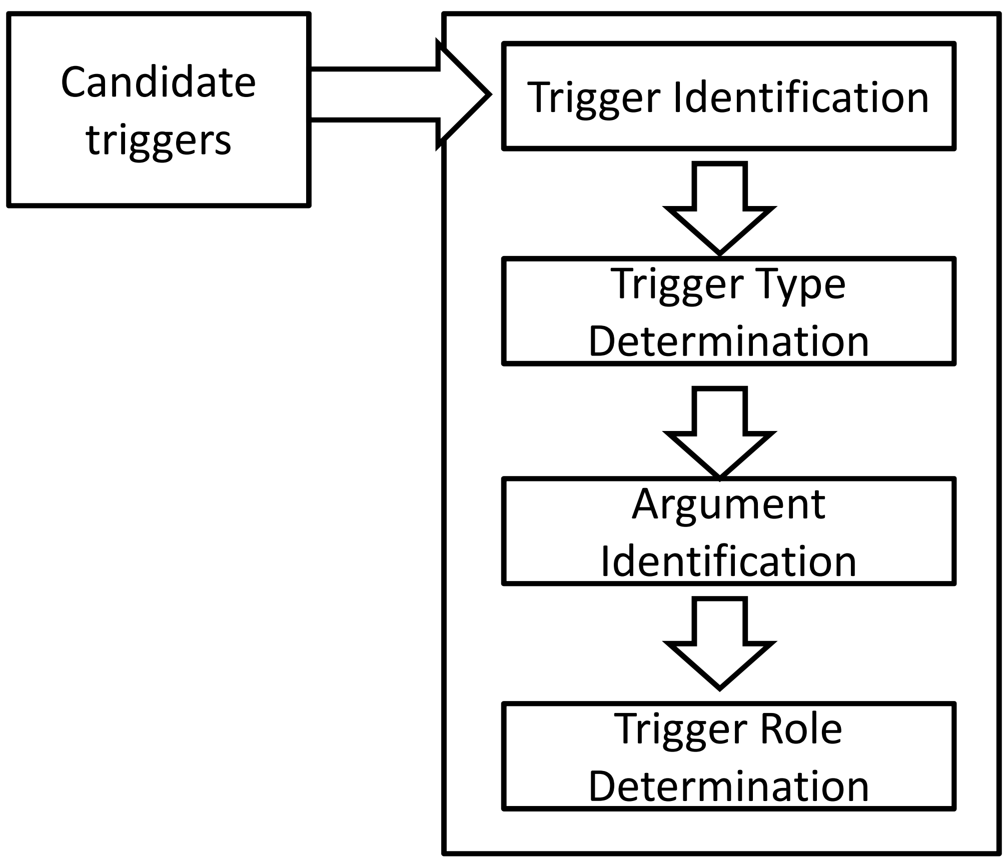
Two Extensions to Li et al.'s System

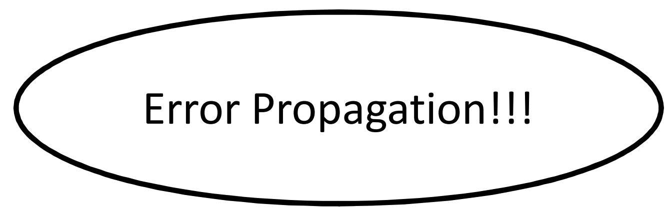
- Joint Learning architecture
  - Goal: to reduce error propagation in pipeline architecture

- Rich linguistic features
  - employ features that capture linguistic information ranging from the character level to the discourse level
  - Goal: use these features to augment the Baseline feature set

Li et al.'s Pipeline Architecture

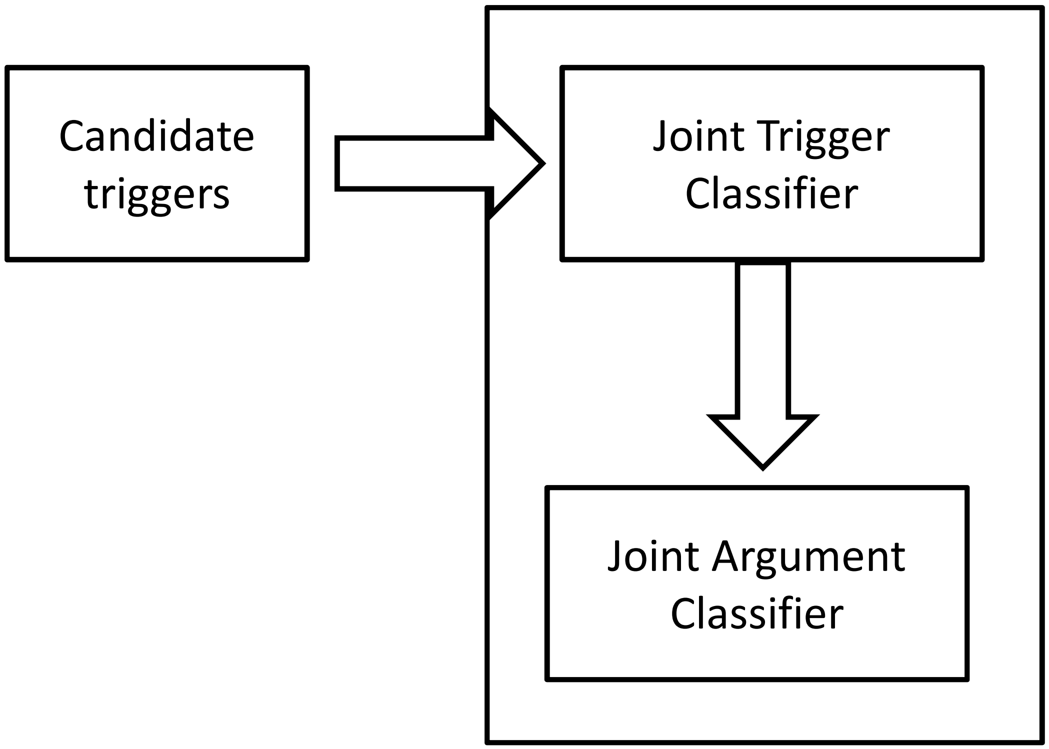
- Extract candidate triggers, then apply 4 classifiers, one for each subtask





Our Joint Learning Architecture

- After extracting candidate triggers, apply only two classifiers



Jointly learn the first two tasks in the pipeline

Jointly learn the last two tasks in the pipeline

Joint Trigger Classifier

- Jointly identify triggers and determine triggers' type
- To generate training data,
  – Create one instance for each word in training doc
    - If the word is not a trigger, the class label is ***NONE.***
    - Otherwise, the class label is trigger's type.
  – Train the model using *SVM_multiclass.*

Joint Trigger Classifier

- Testing
  - Create one instance for each heuristically extracted candidate trigger in test document
  - Apply SVM classifier on the test instances.
    - If the test instance is assigned the class ***NONE***, the corresponding trigger candidate is a non-trigger
    - Otherwise, the instance is classified as an identified trigger, and the trigger type is its assigned label.

Joint Argument Classifier

- Jointly identify arguments and determine arguments' role
- To generate training data,
  - Create one instance by pairing each trigger with each of its candidate arguments.
    - If the candidate argument is indeed a true argument of the trigger, the class label is the argument's role.
    - Otherwise, the class label is ***NONE***.
  - Train the model using *SVM_multiclass.*

Joint Argument Classifier

- Testing
  - Create one instance by pairing each predicted trigger with each of its candidate arguments.
  - Apply SVM classifier on the test instances.
    - If the test instance is assigned the class ***NONE***, the corresponding argument candidate is classified as not an argument of the trigger.
    - Otherwise, the argument is a true argument of the trigger, and the role is the class value assigned.

Linguistic Extensions

- 6 types of features
  - Character-Based Features
  - Semantic Role Labeling
  - Trigger Probability Feature
  - Zero Pronoun Features
  - Trigger Type Consistency Features
  - Argument Consistency Features

Character-Based Features

An example:

刺伤[injury by stabbing] is a known trigger, which appears in training set, while 撞伤[injury by hitting] is an unknown trigger in test set.

Character-based features can be included to exploit similarity between candidate trigger撞伤 and 刺伤, which have the common character伤.

Character-Based Features

- Create 4 character-based features for trigger-related classifiers.
  - The first/last character of the word
  - The synonym entry of the first/last character in a synonym dictionary from Harbin Institute of Technology NLP Group.

Linguistic Extensions

- 6 types of features
  - Character-Based Features
  - Semantic Role Labeling
  - Trigger Probability Feature
  - Zero Pronoun Features
  - Trigger Type Consistency Features
  - Argument Consistency Features

Semantic Role Labeling

- Goal: detect the arguments of a predicate and their semantic roles

Why is SRL useful for event extraction?

- A large portion of the triggers defined in the event extraction task are predicates
- If a predicate happens to be a trigger, the predicate's arguments are essentially the event's arguments
  - Helpful for argument identification
- There is a close correspondence between the PropBank-style roles (e.g., Arg0, Arg1) provided by a SRL and the FrameNet-Style event argument roles
  - Helpful for argument role determination

Semantic Role Labeling Features

- Run SRL tool on all documents
  - (Björkelund et al.,2009).
- Encode 5 features for trigger related classifiers
  - Whether the word under consideration is a predicate.
  - The semantic type/subtype of its Arg0.
  - The semantic type/subtype of it Arg1.

Linguistic Extensions

- 6 types of features
  - Character-Based Features
  - Semantic Role Labeling
  - Trigger Probability Feature
  - Zero Pronoun Features
  - Trigger Type Consistency Features
  - Argument Consistency Features

Trigger Probability Feature

- Trigger probability of a word ***w*** is the probability that ***w*** appears as a true trigger in the training set.
- Hence, a word ***w*** with a higher probability is more likely to be a true trigger.
- Create a new feature for the trigger related classifiers, whose value is the trigger probability of the word under consideration.

Linguistic Extensions

- 6 types of features
  - Character-Based Features
  - Semantic Role Labeling
  - Trigger Probability Feature
  - Zero Pronoun Features
  - Trigger Type Consistency Features
  - Argument Consistency Features

Zero Pronoun: An Overview

- Example:

国家主席江泽民今天晚上乘专机*离开深圳*前往文莱

(President Jiang Zemin took the plane tonight, *left Shenzhen and *went to Brunei)

No overt subject for the verbs离开(left) and 前往(went).

The gaps before离开 and前往 are called zero pronouns.

A zero pronoun has an antecedent, which is a mention that can fill the gap.

The mention 江泽民(Jiang Zemin) should be used to fill the gap because it is coreferent with the two zero pronouns.

Why are zero pronouns useful for event extraction?

- If an event trigger happens to have an zero pronoun preceding it, then the antecedent of the zero pronoun can be this event’s argument
  – Helpful for argument identification

Zero Pronoun Resolution Method

- We employ a simple rule-based method for
  - Detecting zero pronouns
  - Finding the antecedents of zero pronouns

Zero Pronoun Features

- Encode zero pronoun output as two features for the argument related classifiers
  - Whether there is a zero pronoun before this trigger
  - Whether the candidate argument under consideration is coreferent with the zero pronoun

Linguistic Extensions

- 6 types of features
  – Character-Based Features
  – Semantic Role Labeling
  – Trigger Probability Feature
  – Zero Pronoun Features
  – Trigger Type Consistency Features
  – Argument Consistency Features

Trigger Type Consistency

- Observation:
  - Documents in the ACE 2005 Chinese corpus are mostly news articles, each of which describes one theme and most of true triggers are compatible with this document theme
  - Example:

    In a document that describes a fire accident, most of the annotated triggers are of type ***DIE***.

Trigger Type Consistency

- If a candidate trigger's type is the same as that of the majority of the triggers in the document, we say that it is being ***type-consistent*** with the other triggers in the document

Why is trigger type consistency useful for event extraction?

- A candidate trigger that is type-consistent with other triggers is more likely to be a true trigger
  - Helpful for trigger identification

Trigger Type Consistency Features

- Encode type consistency as features
  - Create 33 features for trigger related classifiers, each feature corresponding to one of the 33 predefined trigger types.
  - If , for example, one trigger has type ***DIE***, then
    - The value of the feature corresponding to ***DIE*** is <u>the probability that a trigger in this document has type ***DIE***</u>
    - The values of the remaining 32 trigger type are all zero.

Trigger Type Consistency Features

- To calculate the probability that a trigger in this document has a certain type
  - Run the baseline trigger identifier and the trigger type classifier to identify triggers and predict their types on each document.

Linguistic Extensions

- 6 types of features
  - Character-Based Features
  - Semantic Role Labeling
  - Trigger Probability Feature
  - Zero Pronoun Features
  - Trigger Type Consistency Features
  - Argument Consistency Features

Argument Consistency

- Observation
  - True triggers typically correspond to events that are related to the main person or some major entities mentioned in the documents.
  - If a candidate trigger has arguments that are coreferent with the arguments of true triggers, the candidate trigger is likely to be a true trigger

Argument Consistency

- Example:

  [一家三口]在昨天深夜集体喝下**农药**[自杀]

  [A family of three] drank pesticide to [suicide] last night

  [三个人]**总算是稳住了**[病情]

  [Three people] finally stabilize the [patient's condition]

  It is fairly easy to detect自杀, a verb, as a trigger in the first sentence. Knowing 一家三口 and三个人 are coreferent ,病情, a noun, which is hard to be classified as true trigger, can be detected by the classifier.

Argument Consistency Feature

- Encode argument consistency Feature
  - The feature is the ***role*** of the argument that is coreferent with a predicted true trigger’s argument.
  - To obtain predicted true trigger’s argument, we run the baseline classifiers to identify triggers, predict their types, arguments and also the argument roles on each document

Evaluation

- **Goal**: determine whether the Baseline system
  - Li et al.'s pipeline system architecture
  - Li et al.'s feature set (the Baseline feature set)

  can be improved by using
  - our joint learning architecture
  - our rich linguistic features to augment the Baseline feature set

- Performance will be measured on the 4 event extraction subtasks

Evaluation Measures

- Report recall, precision, F-score for each subtask

How do we determine correctness for the 4 subtasks?

- **Trigger identification:** A trigger is correctly identified if its *offset* exactly match a reference trigger
- **Trigger type determination:** Trigger type is correctly determined if its trigger *type* and *offsets* exactly match a reference trigger
- **Argument identification:** An argument is correctly identified if its *offsets*, related trigger *type* and trigger's *offsets* exactly match a reference argument
- **Argument role determination:** An argument role is correctly determined if its *offsets*, *role*, related trigger *type* and trigger's *offset* exactly match a reference argument

Evaluation Dataset

- All 633 Chinese document in Automatic Content Extraction (ACE) Evaluation 2005 training corpus

- We performed 10-fold cross-validation to obtain a more accurate estimate of system performance
  - Previous work typically evaluated on 10% of documents

Feature Selection

- To obtain better performance, we use feature selection to select different feature groups for different classifiers based on development data

Feature Selection

- Totally 7 feature groups to be selected
  - Discourse consistency feature (Li et al., 2012) (G1)
  - Semantic role labeling feature (G2)
  - Trigger probability features (G3)
  - Character-based features (G4)
  - Argument consistency feature (G5)
  - Trigger type consistency feature (G6)
  - Zero pronoun features (G7)

Feature Selection Procedure

- Backward elimination
  - Start with full 7 feature groups, together with baseline features
  - Remove in each iteration the feature group, whose removal yields the best performance
  - Run iterations till all 7 feature groups are removed and identify the feature subset that yields the best performance

Feature Selection Result on Development Set

Approach	Classifier	Selected Features
Pipeline	Trigger Identification	G1, G2, G3, G4, G5, G6
	Trigger Type Determination	G4
	Argument Identification	G2, G7
	Argument Role Determination	---
Joint	Trigger Component	G2, G3, G4, G6
	Argument Component	G2, G7

- Discourse consistency feature (Li et al., 2012) (G1)
- Semantic role labeling feature (G2)
- Trigger probability features (G3)
- Character-based features (G4)
- Argument consistency feature (G5)
- Trigger type consistency feature (G6)
- Zero pronoun features (G7)

Pipeline modeling results on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.6	75.5	60.6	47.5	70.8	56.8	35.1	52.3	42.0	31.2	46.5	37.4
Baseline features with DC	55.6	72.7	63.0	52.0	67.9	58.9	38.9	50.2	43.8	34.8	45.0	39.2
Our selected features	60.5	70.1	**64.9**	56.6	65.6	**60.8**	43.8	50.2	**46.8**	39.3	45.1	**42.0**

- Pipeline modeling system can be improved by our extension features

Joint modeling results on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
Baseline features with DC	55.3	75.6.	63.9	52.6	71.8	60.7	38.2	57.4	45.9	34.3	51.5	41.1
Our selected features	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

- Joint modeling system can be improved by our extension features

Comparison of Pipeline and Joint Model

- Pipeline modeling

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.6	75.5	60.6	47.5	70.8	56.8	35.1	52.3	42.0	31.2	46.5	37.4
Baseline features with DC	55.6	72.7	63.0	52.0	67.9	58.9	38.9	50.2	43.8	34.8	45.0	39.2
Our selected features	60.5	70.1	**64.9**	56.6	65.6	**60.8**	43.8	50.2	**46.8**	39.3	45.1	**42.0**

- Joint modeling

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
Baseline features with DC	55.3	75.6.	63.9	52.6	71.8	60.7	38.2	57.4	45.9	34.3	51.5	41.1
Our selected features	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Comparison of Li et, al. and Our system

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Li et. al, 2012's system	55.6	72.7	63.0	52.0	67.9	58.9	38.9	50.2	43.8	34.8	45.0	39.2
Our joint system with rich linguistic features	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Our system outperforms Li et.al,2012's system by 3.7%, 4.3%, 5.7% and 5.4% on 4 subtasks

Incremental addition of features to joint model on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
+Semantic Role Labeling	52.1	77.7	62.4	49.8	74.4	59.7	36.9	61.7	46.2	33.2	55.4	41.5
+Trigger Probability	56.0	75.3	**64.3**	53.3	71.5	**61.1**	39.2	59.7	**47.3**	35.2	53.7	**42.5**
+Character Features	59.8	73.8	**66.1**	56.6	69.6	**62.6**	41.2	57.9	**48.2**	37.2	52.3	**43.5**
+Trigger Type Consistency	62.2	71.9	**66.7**	58.9	68.1	**63.2**	42.7	56.5	**48.6**	38.5	50.9	**43.8**
+Zero Pronouns	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Semantic Role Labeling increases 4 subtasks by 1.7%, 1.9%, 3.0% and 3.0%

Incremental addition of features to joint model on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
+Semantic Role Labeling	52.1	77.7	62.4	49.8	74.4	59.7	36.9	61.7	46.2	33.2	55.4	41.5
+Trigger Probability	56.0	75.3	**64.3**	53.3	71.5	**61.1**	39.2	59.7	**47.3**	35.2	53.7	**42.5**
+Character Features	59.8	73.8	**66.1**	56.6	69.6	**62.6**	41.2	57.9	**48.2**	37.2	52.3	**43.5**
+Trigger Type Consistency	62.2	71.9	**66.7**	58.9	68.1	**63.2**	42.7	56.5	**48.6**	38.5	50.9	**43.8**
+Zero Pronouns	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Trigger Probability increases 4 subtasks by 1.9%, 1.4%, 1.1% and 1.0%

Incremental addition of features to joint model on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
+Semantic Role Labeling	52.1	77.7	62.4	49.8	74.4	59.7	36.9	61.7	46.2	33.2	55.4	41.5
+Trigger Probability	56.0	75.3	**64.3**	53.3	71.5	**61.1**	39.2	59.7	**47.3**	35.2	53.7	**42.5**
+Character Features	59.8	73.8	**66.1**	56.6	69.6	**62.6**	41.2	57.9	**48.2**	37.2	52.3	**43.5**
+Trigger Type Consistency	62.2	71.9	**66.7**	58.9	68.1	**63.2**	42.7	56.5	**48.6**	38.5	50.9	**43.8**
+Zero Pronouns	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Character-based features increases 4 subtasks by 1.8%, 1.5%, 0.9% and 1.0%

Incremental addition of features to joint model on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
+Semantic Role Labeling	52.1	77.7	62.4	49.8	74.4	59.7	36.9	61.7	46.2	33.2	55.4	41.5
+Trigger Probability	56.0	75.3	**64.3**	53.3	71.5	**61.1**	39.2	59.7	**47.3**	35.2	53.7	**42.5**
+Character Features	59.8	73.8	**66.1**	56.6	69.6	**62.6**	41.2	57.9	**48.2**	37.2	52.3	**43.5**
+Trigger Type Consistency	62.2	71.9	**66.7**	58.9	68.1	**63.2**	42.7	56.5	**48.6**	38.5	50.9	**43.8**
+Zero Pronouns	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Trigger type consistency increases 4 subtasks by 0.6%, 0.6%, 0.4% and 0.3%

Incremental addition of features to joint model on test set

	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline Features w/o DC	50.0	77.0	60.7	47.5	73.1	57.6	34.1	58.7	43.2	30.4	52.3	38.5
+Semantic Role Labeling	52.1	77.7	62.4	49.8	74.4	59.7	36.9	61.7	46.2	33.2	55.4	41.5
+Trigger Probability	56.0	75.3	**64.3**	53.3	71.5	**61.1**	39.2	59.7	**47.3**	35.2	53.7	**42.5**
+Character Features	59.8	73.8	**66.1**	56.6	69.6	**62.6**	41.2	57.9	**48.2**	37.2	52.3	**43.5**
+Trigger Type Consistency	62.2	71.9	**66.7**	58.9	68.1	**63.2**	42.7	56.5	**48.6**	38.5	50.9	**43.8**
+Zero Pronouns	62.2	71.9	**66.7**	58.9	68.1	**63.2**	43.6	57.3	**49.5**	39.2	51.6	**44.6**

Zero pronouns increases 4 subtasks by 0.0%, 0.0%, 0.9% and 0.8%

Summary

- Joint-learning, knowledge rich approach that extends Li et al.'s (2012) state-of-the-art Chinese event extraction system
- Outperformed Li et al.'s system by 3.7-5.7% on the four event extraction subtasks