Employing Compositional Semantics and Discourse Consistency in Chinese Event Extraction

Peifeng Li, Guodong Zhou, Qiaoming Zhu, Libin Hou
School of Computer Science & Technology
Soochow University, Suzhou, 215006, China
{pfli, gdzhou, qmzhu, 20094227021}@suda.edu.cn

Abstract

Current Chinese event extraction systems suffer much from two problems in trigger identification: unknown triggers and word segmentation errors to known triggers. To resolve these problems, this paper proposes two novel inference mechanisms to explore special characteristics in Chinese via compositional semantics inside Chinese triggers and discourse consistency between Chinese trigger mentions. Evaluation on the ACE 2005 Chinese corpus justifies the effectiveness of our approach over a strong baseline.

1 Introduction

Event extraction, a classic information extraction task, is to identify instances of a predefined event type and can be typically divided into four subtasks: trigger identification, trigger type determination, argument identification and argument role determination. In the literature, most studies focus on English event extraction and have achieved certain success (e.g. Grishman et al., 2005; Ahn, 2006; Hardy et al., 2006; Maslennikov and Chua, 2007; Finkel et al., 2005; Ji and Grishman, 2008; Patwardhan and Riloff, 2009, 2011; Liao and Grishman 2010; Hong et al., 2011).

In comparison, there are few successful stories regarding Chinese event extraction due to special characteristics in Chinese trigger identification. In particular, there are two major reasons for the low performance: unknown triggers [1] and word segmentation errors to known triggers. Table 1 gives the statistics of unknown triggers and word segmentation errors to known triggers in both the ACE 2005 Chinese and English corpora[2] using 10-fold cross-validation. In each validation, we leave 10% trigger mentions as the test set and the remaining ones as the training set. If a mention in the test set doesn't occurred in the training set, we regard it as an unknown trigger. It shows that these two cases cover almost 30% of Chinese trigger mentions while this figure reduces to only about 9% in English. It also shows that given the same number of event mentions, there are 30% more different triggers in Chinese than that in English. This justifies the low performance (specifically, the recall) of a Chinese event extraction system, which normally extracts those known triggers occurring in the training data as candidate instances and uses a classifier to distinguish correct triggers from wrong ones.

Language	Chinese	English
%unknown triggers	33.7%	18.5%
%unknown trigger mentions	20.9%	8.9%
%word segmentation errors to known trigger mentions	8.7%	0%
#triggers	763	586

Table 1. Statistics: a comparison between Chinese and English event extraction with regard to unknown triggers and word segmentation errors to known triggers. Note that word segmentation only applies to Chinese.

In this paper, we propose two novel inference mechanisms to Chinese trigger identification by employing compositional semantics inside Chinese triggers and discourse consistency between Chinese trigger mentions.

The first mechanism is motivated by the compositional nature of Chinese words, whose semantics can be often determined by the component characters. Hence, it is natural to infer

[1] In this paper, a trigger word/phrase occurring in the training data is called a known trigger and otherwise, an unknown trigger.

[2] The whole Chinese ACE corpus has about 3300 event mentions. For the sake of fair comparison, we choose the same number of event mentions from the English corpus as the cross-validation data.

unknown triggers by employing compositional semantics inside Chinese triggers.

The second mechanism is enlightened by the wide use of discourse consistency in natural languages, particularly for Chinese, due to its discourse-driven nature (Zhu, 1980). Very often, distinguishing true trigger mentions from pseudo ones is only possible with contextual information.

The rest of this paper is organized as follows. Section 2 overviews the related work. Section 3 introduces a state-of-the-art baseline system for Chinese event extraction. Sections 4 and 5 describe two novel inference mechanisms to Chinese trigger identification by employing compositional semantics inside Chinese triggers and discourse consistency between Chinese trigger mentions. Section 6 presents the experimental results. Section 7 concludes the paper and points out future work.

2 Related Work

Almost all the existing studies on event extraction concern English. While earlier studies focus on sentence-level extraction (Grishman et al., 2005; Ahn, 2006; Hardy et al., 2006), later ones turn to employ high-level information, such as document (Maslennikov and Chua, 2007; Finkel et al., 2005; Patwardhan and Riloff, 2009), cross-document (Ji and Grishman, 2008), cross-event (Liao and Grishman, 2010; Gupta and Ji, 2009) and cross-entity (Hong et al., 2011) information.

2.1 Chinese Event Extraction

Compared with tremendous efforts in English event extraction, there are only a few studies on Chinese event extraction.

Tan et al. (2008) modeled event extraction as a pipeline of classification tasks. Specially, they used a local feature selection approach to ensure the performance of trigger classification (trigger identification + trigger type determination) and applied multiple levels of patterns to improve the coverage of patterns in argument classification (argument identification + argument role determination). Chen and Ji (2009a) proposed a bootstrapping framework, which exploited extra information captured by an English event extraction system. Chen and Ji (2009b) applied various kinds of lexical, syntactic and semantic features to address the specific issues in Chinese. They also constructed a global errata table to record the inconsistency in the training set and used it to correct the inconsistency in the test set. Ji (2009) extracted cross-lingual predicate clusters using bilingual parallel corpora and a cross-lingual information extraction system, and then used the derived clusters to improve the performance of Chinese event extraction.

2.2 Compositional Semantics

Almost all the related studies on compositional semantics focus on how to combine words together to convey complex meanings, such as semantic parser (Zettlemoyer and Collins, 2007; Wong and Mooney, 2007; Liang et al., 2011). However, the compositional semantics mentioned in this paper is more fined-grained and focuses on how to construct Chinese characters into a word and mine the semantics of words from the word structures, especially of verbs as event triggers.

To our knowledge, there is only one paper associated with compositional semantics inside Chinese words. Li (2011) discussed the internal structures inside Chinese nouns and used it in word segmentation.

2.3 Discourse Consistency

Discourse consistency is an important hypothesis in natural languages and has been applied to many natural language processing applications, such as named entity recognition and coreference resolution. Specially, several studies have successfully incorporated trigger or entity consistency constraint into event extraction.

Yarowsky (1995) and Yangarber et al. (Yangarber and Jokipii, 2005; Yangarber et al., 2007) applied cross-document inference to refine local extraction results for disease name, location and start/end time. Mann (2007) proposed some specific inference rules to improve extraction of personal information. Ji and Grishman (2008) employed a rule-based approach to propagate consistent triggers and arguments across topic-related documents. Gupta and Ji (2009) used a similar approach to recover implicit time information for events. Liao and Grishman (2011) also used a similar approach and a self-training strategy to extract events. Liao and Grishman (2010) employed cross-event consistency information to improve sentence-level event extraction. Hong et al. (2011) regarded entity type consistency as a key feature to predict event mentions and adopted this inference method to improve the traditional event extraction system.

3 Baseline

As a baseline, we re-implement a state-of-the-art system, which consists of four typical components (trigger identification, trigger type determination, argument identification and argument role determination), in a pipeline way and employ the same set of features as described in Chen and Ji (2009b).

Besides, the Maximum-Entropy (ME) model is employed to train individual component classifiers for the above four components. During testing, each word in the test set is first scanned for instances of known triggers from the training set. When an instance is found, the trigger identifier is applied to distinguish true trigger mentions from pseudo ones. If true, the trigger type determiner is then applied to recognize its event type. For any entity mentions in the sentence, the argument identifier is employed to assign possible arguments to them afterwards. Finally, the argument role determiner is introduced to assign a role to each argument.

One problem with Chen and Ji's system is its ignoring effective long-distance features. In order to resolve this problem and provide a stronger baseline, we introduce more refined and dependency features in four components:

- **Trigger Identification and Trigger Type Determination:** 1) syntactic features: path to the root of the governing clause, 2) nearest entity information: entity type of left syntactically/physically nearest entity to the trigger + entity, entity type of right syntactically/physically nearest entity to the trigger mention in the sentence + entity; 3) dependency features: the subject and the object of the trigger when they are entities.
- **Argument Identification and Argument Role Determination**: 1) basic features: POS of trigger; 2) neighboring words: left neighboring word of the entity + its POS, right neighbor word of the entity + its POS, left neighbor word of the trigger + its POS, right neighbor word of the trigger + its POS; 3) dependency feature: dependency path from the entity to the trigger; 4) semantic role features: Arg0 and Arg1 which tagged by semantic role labeling tool (Li, et al., 2010).

3.1 Experimental Setting

The ACE 2005 Chinese corpus (only the training data is available) is used in all our experiments. The corpus contains 633 Chinese documents annotated with 8 predefined event types and 33 predefined subtypes. Similar to previous studies, we treat these subtypes simply as 33 separate event types and do not consider the hierarchical structure among them.

Following Chen and Ji (2009b), we randomly select 567 documents as the training set and the remaining 66 documents as the test set. Besides, we reserve 33 documents in the training set as the development set, and follow the setting of ACE diagnostic tasks and use the ground truth entities, times and values for our training and testing.

For evaluation, we follow the standards as defined in Ji (2009):

- A trigger is *correctly* identified if its position in the document matches a reference trigger;
- A trigger type is *correctly* determined if its event type and position in the document match a reference trigger;
- An argument is *correctly* identified if its involved event type and position in the document match any of the reference argument mentions;
- An argument role is *correctly* determined if its involved event type, position in the document, and role match any of the reference argument mentions.

Finally, all sentences in the corpus are divided into words using a word segmentation tool ICTCLAS[3] with all entities annotated in the corpus kept. Besides, we use Stanford Parser (Levy and Manning, 2003, Chang, et al., 2009) to create the constituent and dependency parse trees and employ the ME model to train individual component classifiers.

3.2 Experimental Results

Table 2 and 3 show the Precision (P), Recall (R) and F1-Measure (F) on the held-out test set. It shows that our baseline system outperforms Chen and Ji (2009b) by 1.8, 2.2, 3.9 and 2.3 in F1-measure on trigger identification, trigger type

[3] http://ictclas.org/

determination, argument identification and argument role determination, respectively, with both gains in precision and recall. This is simply due to contribution of the newly-added refined and dependency features.

Performance	Trigger Identification			Trigger Type Determination		
System	P(%)	R(%)	F	P(%)	R(%)	F
Chen and Ji (2009b)	71.5	51.2	59.7	66.5	47.7	55.6
Our Baseline	75.2	52.0	61.5	70.3	49.0	57.8

Table 2. Performance of trigger identification and trigger type determination

Performance	Argument Identification			Argument Role Determination		
System	P(%)	R(%)	F	P(%)	R(%)	F
Chen and Ji (2009b)	56.1	38.2	45.4	53.1	36.2	43.1
Our Baseline	58.4	42.7	49.3	55.2	38.6	45.4

Table 3. Performance of argument identification and argument role determination

For our baseline system, given the small performance gaps between trigger identification and trigger type determination (3.7 in F1-measure: 61.5 vs. 57.8) and between argument identification and argument role determination (3.9 in F1-measure: 49.3 vs. 45.4), the performance bottlenecks of our baseline system mainly exist in trigger identification and argument identification, particularly for the former one. While argument identification has the performance gap of 8.5 in F1-measure compared to trigger type determination (49.3 vs. 57.8), the former one, trigger identification, can only achieve the performance of 61.5 in F1-measure (in particular the recall with only 52.0). In this paper, we will focus on trigger identification to improve its performance, particularly for the recall, via compositional semantics inside Chinese triggers and discourse consistency between Chinese trigger mentions.

4 Employing Compositional Semantics inside Chinese Triggers

Language is perhaps the only communicative system in nature, which compositionally builds structured meanings from smaller pieces, and this compositionality is the cognitive mechanism that allows for what Humboldt called language’s “infinite use of finite means.” As usual, the lexical semantics is the smallest piece in most Chinese language processing applications. In this section, we introduce a more fine-grained semantics - the compositional semantics in Chinese verb structure - and unveil its effect and usage in Chinese language processing by employing it into Chinese event extraction.

4.1 Compositional Semantics inside Chinese Triggers

In English, a component character is just the basic unit to form a word instead of a semantics unit. In comparison, almost all Chinese characters have their own meanings and can be formed as SCWs (Single Character Words) themselves. If a Chinese word contains more than one character, its meaning can be often inferred from the meanings of its component characters (Yuan, 1998). Actually, it is the normal way of understanding a new Chinese word in everyday life of a Chinese native speaker. A general method to this problem is to systematically explore the morphological structures in Chinese words. In this paper, compositional semantics provides a simple but effective compromise to the general method and we leave the general method in the future work. Table 4 shows samples of such compositional semantics in Chinese words. For example, “会见” is composed of two characters: “会” and “见” which have their own semantics and the semantics of “会见” comes from that of its component characters “会” and “见”.

Words	Characters

会见 (interview[4])
击毙 (shoot and kill)
来到(come)
私信 (private letter)
	会 (meet) 见(meet)
	击(shoot) 毙 (kill)
	来 (come) 到 (to)
	私(private) 信(letter)

Table 4. Examples of compositional semantics in Chinese words

Therefore, it is natural to infer unknown triggers by employing compositional semantics inside Chinese triggers. Take following two sentences as examples:

(1) 4 名学生被玻璃**划伤**。(Known trigger)

[4] Most Chinese words have more than one sense. Here, we just give the one when it acts as a trigger.

(Four students were scratched by the glass.)

(2) 1名乘客被**刺伤**。(Unknown trigger)

(A passenger was stabbed.)

where “划伤” is a known trigger and “刺伤” is an unknown one.

In above examples, the semantics of “划伤” (injure by scratching) can be largely determined from those of its component characters “划” (scratch) and “伤” (injure) while the semantics of “刺伤” (injure by stabbing) from those of its component characters “刺” (stab) and “伤” (injure). Since these two triggers have similar internal structures, we can easily infer that “刺伤” is a trigger of *injure* event if “划伤” is known as a trigger of *injure* event. Similarly, we can infer more triggers for *injure* event, such as “灼伤” (injure by burning), “撞伤” (injure by hitting), “压伤” (injure by pressing), all with component character “伤” (injure) as the head and the other component character as the way of causing injury.

Since most triggers in Chinese event extraction are verbs[5], we focus on the compositional semantics in the verb structure. Statistics on the training set shows that 3.3% triggers (e.g. “公开信” (open letter), “事件” (event), “病情” (patient's condition), etc.) don’t contain a BV and all of them are nouns. Normally, almost all verbs contain one or more single-character verbs as the basic element to construct a verb (we call it basic verb, shorted as BV) and the semantics of such a verb thus can be inferred from its BV. There are some studies on the Chinese verb structure in linguistics. However, their structures are much more complex and there are no annotated corpora available. We define following six main structures from our empirical observations:

(1) BV (e.g. “看” (see), “杀” (kill))
(2) BV + verb (e.g. “会见” (meet))
(3) verb + BV (e.g. “解雇” (fire))
(4) BV + complementation (e.g. “杀了” (kill))
(5) BV + noun/adj. (e.g. “回家” (go to home))
(6) noun/adj. +BV (e.g. “枪击” (shoot using gun)).

From above structures, a BV plays an important role in the verb structure and most of semantics of a verb can be interred from its contained BV and two words normally have very similar semantics if they have the same BV (e.g. “会见” (meet) and “会晤” (meet)). Actually, sometime the verb can be shortened to its contained BV (e.g. “我**见**王教授” and “我**会见**王教授” have the same semantics.).

4.2 Inferring via Compositional Semantics inside Chinese Triggers

Here a simple rule is employed to infer triggers via compositional semantics inside Chinese triggers: **a verb is a trigger if it contains a BV which occurs as a known trigger or is contained in a known trigger**. Table 5 shows the distribution of the set of triggers (contains the same BV[6]) classified by number of triggers.

From Table 5, we can find out that 85.3% of BVs occur in more than one trigger and 56.2% of them in more than 4 triggers. As for trigger mentions, these percentages become 89.1% and 65.2% respectively. A extreme example is that 85.2% (75/88) of triggers of *Trial-Hearing* event mentions contain “审” (trial) and 85.4% (117/138) of triggers of *injure* event mentions contains “伤” (injure).

Number	Distribution over Triggers	Distribution over Trigger Mentions
1	14.7%	10.9%
2~4	29.1%	23.9%
5~9	28.1%	32.9%
>=10	28.1%	32.3%

Table 5. Distribution of BVs in the number of triggers/trigger mentions

In this paper, the inference is done as follows:

- Add all single-character triggers into the BV set if it’s a verb;
- Split all other triggers in the training set into a set of single characters and include all single characters into the BV set if it’s a verb;
- For each word in the test set, it is identified as a trigger if it contains a BV.

It is worthwhile to note that such inference works for unknown triggers and word

[5] Actually, in the ACE 2005 Chinese (training) corpus, more than 90% of triggers are either verbs al or verbal nouns (those verbs which act as nouns). For simplicity, we don’t differentiate these two types in this paper.

[6] We didn’t tag BVs in the training set and regards all single-character verbs contained in triggers as BVs.

segmentation errors to known triggers since in both cases, their BVs will always exist as either a SCW or a component of a word.

4.3 Noise Filtering

One problem with above inference is that while it is able to recover some true triggers and increase the recall, it may introduce many pseudo ones and harm the precision. To filter out those pseudo triggers, we propose following rules according to our intuition and statistics over the training set.

Non-trigger Filtering
A Chinese word will not be a trigger if it appears in the training set but never trigger an event. Statistics on the training set shows that this rule applies at 99.7% of cases.

POS filtering
A Chinese word will not be a trigger if it has a different POS from that of the same known trigger or similar known triggers[7] **in the training set.** In Chinese, a single-character verb has very high probability of composing words (e.g. “到” (come), “为” (act as), “并” (combine), etc) with different POS from the single-character verb itself, such as preposition (e.g. “为了” (for)), conjunction (e.g. “并且” (and)), etc. Statistics on the training set shows that this rule applies at 97.3% of cases.

Verb structure filtering
A Chinese word will not be a trigger if its verb structure is different from that of the same known trigger or similar known triggers in the training set. Figure 1 shows different distributions of three BVs over six verb structures as described in subsection 4.1. For example, we can find that all triggers including “解” (unbind) (e.g. “解聘” (fire), “解雇” (fire), “解散” (disband)) just have one verb structure (BV + verb) and those of “杀” (kill) have 4 structures. Obviously, we can use such distribution information to filter out pseudo triggers. For example, although both word “劝解” (console) and “分解” (decompose) are constructed form verb “解”, their verb structure (verb + BV) does not appear in the training set. Therefore, they will be filtered our via verb structure filtering. Statistics on the training set shows that this rule applies at 95.5% of cases.

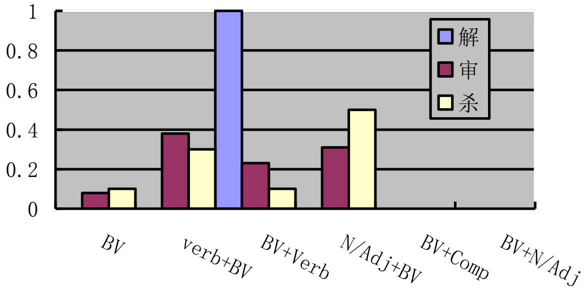

Figure 1. Distribution of three BVs (“解” (unbind), “审” (trial) and “杀” (kill)) over six verb structures in constructing triggers

5 Employing Discourse Consistency between Chinese Trigger Mentions

Chinese event extraction may suffer much from the errors propagated from upstream processing such as part-of-speech tagging and parsing, especially word segmentation. To alleviate word segmentation errors to known triggers, Chen and Ji (2009b) constructed a global errata table to record the inconsistency in the training set and proved its effectiveness. In this paper, a merge and split method is applied to recover those known triggers. In this way, word segmentation errors can be alleviated to certain extent.

For unknown triggers, we can merge two or more neighboring short words or single characters as a trigger candidate. In this paper, for each single-character verb in a document after word segmentation, this single-character verb can be merged with either previous SCW or next SCW to form a trigger candidate if this single-character verb has occurred in the training set with the same verb structure.

Given above recovered triggers for both known and unknown triggers, the key issue here is how to distinguish true triggers from pseudo ones. In this paper, we employ discourse consistency between Chinese trigger mentions for Chinese event extraction. Previous studies on English event extraction have proved the effectiveness of both cross-entity and cross-document consistency.

5.1 Discourse Consistency between Chinese Trigger Mentions

As a discourse-driven language, the syntax of

[7] Similar triggers are those ones which have the same BV and verb structure.

Chinese is not as strict as English and sometime we must infer from the discourse-level information to understand the meaning of a sentence. Kim (2000) compared the use of overt subjects in English and Chinese and he found that overt subjects occupy over 96% in English, while this percentage drops to only 64% in Chinese. Similarly, argument missing is another issue in Chinese event extraction and almost 55% of arguments are missing in the ACE 2005 Chinese corpus. Normally, using a feature-based approach to distinguish true triggers from pseudo ones is very difficult from the sentence level if some of related arguments are missing from the trigger-occurring sentence. Take following two contingent sentences as examples:

(3) 美国与北韩 3 号在吉隆坡结束飞弹**会谈。**
(The United States and the Democratic People's Republic of Korea finished missile **talks** in Kuala Lumpur.)

(4) **会谈**的气氛严肃。
(The **talks** are serious.)

While it is relatively easy to determine that mention "会谈" in sentence (3) indicates a *meet* event from the contained information in itself (there are many entities, such as agents, time and place in the sentence) and difficult to determine that mention "会谈" in sentence (4) is a *meet* event from the contained information in itself, we can easily infer from sentence (3) that sentence (4) also indicates a *meet* event, using discourse consistency: if one instance of a word is a trigger mention, other instances in the same discourse will be a trigger mention with high probability.

Language	Discourse-based	Instance-based
English	70.2%	87.5%
Chinese	90.5%	95.4%

Table 6. Comparison of discourse consistency between Chinese and English trigger mentions

Table 6 compares the probabilities of discourse consistency between Chinese and English trigger mentions in the ACE 2005 Chinese and English corpora. A trigger may appear many times in a discourse. It's considered discourse-consistent when all the appearances of a trigger have the same event type while instance-based consistency refers to pair-wired cases. It shows that within the discourse, there is a strong consistency in both Chinese and English between trigger mentions: if one instance of a word is a trigger, other instances in the same discourse will be a trigger of the same event type with very high probability.

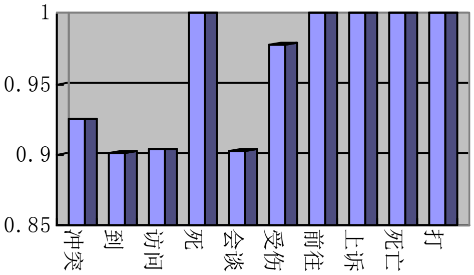

Figure 2. Probabilities of discourse-level consistency of top 10 frequent triggers

It also shows that discourse consistency in Chinese triggers holds much more likely than the English counterpart. Figure 2 give the probabilities of discourse-level consistency of top 10 frequent triggers, which occupy 18% of event mentions in the ACE 2005 Chinese corpus.

5.2 Inference via Discourse Consistency between Chinese Trigger Mentions

Given a discourse and different mentions of a trigger returned by the trigger identifier, we can simply accept those mentions with high probability as true mentions of the trigger and discard those with low probability[8]. However, for those mentions in-between, an additional discourse-level trigger identifier is further employed to determine whether a trigger mention is true or not from the discourse level by augmenting the normal trigger identifier with several features to explore the consistency information between trigger mentions in the discourse (first three features) and the related information returned from the trigger type identifier (last two features).

- Probability of the discourse consistency of the candidate trigger mention in the training set. If it doesn't exist in the training set, we infer its probability from that of all of its similar triggers
- Number of candidate trigger mentions being a trigger in the same discourse via trigger identification
- Number of candidate trigger mentions being a non-trigger in the same discourse via trigger identification

[8] The high and low probability thresholds are fine-tuned to 95% and 5% respectively, using the development set.

- ➢ Event type of candidate trigger mention via trigger type determination
- ➢ Confidence of trigger type determination

6 Experiments

In this section, we evaluate our two inference mechanisms in Chinese trigger identification and its application to overall Chinese event extraction, using the same experimental settings as described in Subsection 3.1.

6.1 Chinese Trigger Identification

Table 7 shows the impact of compositional semantics in trigger identification. Here, the baseline just extracts those triggers occurring in the training data. It justifies the effectiveness of our compositional semantics-based inference mechanism in recovering true triggers and its three filtering rules in removing pseudo triggers.

Numbers Approaches	Triggers	Non-triggers
Baseline	266	629
+Compositional semantics without filtering	334	1885
+ Non-trigger filtering	328	1062
+ POS filtering	325	974
+ Verb structure filtering	302	444
Gold	367	-

Table 7. Impact of compositional semantics in trigger identification

To reduce those pseudo triggers after above inference process, three rules are introduced.

The first rule, the non-trigger filtering rule, filters out those pseudo ones in the test set which do not frequently occur as trigger mentions in the training set. In particular, to keep true triggers in our candidate set as many as possible, we just filter out those candidates which occur as non-triggers more than 5 times in the training set according to our validation on the development set. Table 7 shows that 43.7% (823) of pseudo triggers are filtered out while only 1.8% (6) of true ones is wrongly filtered out.

The second rule, the POS filtering rule, just filters out 8.3% (88) of pseudo triggers, due to POS errors in word segmentation and constituent parsing (e.g. 9.4% of candidate triggers have wrong POS tags in the development set.). Manual inspection shows that if we correct those wrong POS tags, that percentage will be increased to 14.5%.

The third rule, the verb structure filtering rule, is deployed in following steps: 1) keeping all candidates if they act as a trigger in the training set; 2) if the candidate is a SCW, removing it when it does not occur as a BV in any triggers in the training set; 3) if the candidate is not a SCW, calculating the condition probability of its similar trigger words as triggers in the training set[9] and then deleting all candidates whose conditional probabilities are less than a threshold θ , which is fine-tuned to 0.5. Figure 3 shows the effect on precision, recall and F1-measure of varying the thresholdθ on the development set.

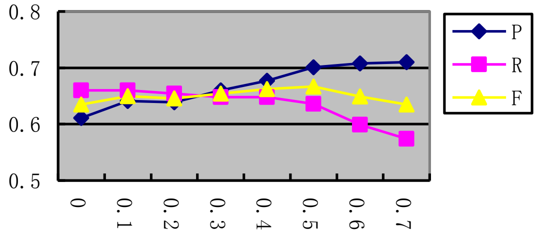

Figure 3. Effect of threshold θ on the development set

Performance System	Trigger Identification		
	P(%)	R(%)	F
Baseline	75.2	52.0	61.5
+Compositional semantics without filtering	34.8	66.8	45.8
+ Non-trigger filtering	49.4	66.5	56.7
+ POS filtering	50.2	65.9	57.0
+ Verb structure filtering	73.5	62.1	67.4
+Discourse consistency	**79.3**	**63.5**	**70.5**

Table 8. Contribution to Chinese triggers identification (incremental)

Table 8 shows the contribution of employing compositional semantics and discourse consistency to trigger identification on the held-out test set. We can find out that our approach dramatically enhances F1-measure by 9.0 units, largely due to a dramatic increase of 11.5% in recall, benefiting from both compositional semantics and discourse consistency mechanisms. We expect that the precision will also increase since our filtering approach successfully filters out almost 30% more

[9] If there are more than one BV in a candidate, we calculate the average one.

non-triggers and the number of non-trigger mentions is less than that of the baseline. Unfortunately, the resulting set of 444 non-trigger mentions (after all filtering) is not a subset of original 629 non-trigger ones. Our observation shows that our compositional semantics inference adds almost 10% new non-triggers into candidates which are very hard to distinguish.

Table 8 also justifies the impact of the discourse consistency between trigger mentions in trigger identification and the effect of the additional discourse-level trigger identifier, with a big gain of 5.8% in precision and a small gain of 1.4% in recall.

6.2 Chinese Event Extraction

Table 9 shows the contribution of trigger identification with compositional semantics and discourse consistency to overall event extraction on the held-out test set. In addition, we also report the performance of two human annotators (The human annotator 1 is a first year postgraduate student with no background to Chinese event extraction while the human annotator 2 is a third year postgraduate student working on Chinese event extraction) on 33 texts (a subset of the held-out test set). From the results presented in Table 9, we can find that our approach can improve the F1-measure for trigger identification by 9.0 units, trigger type determination by 9.1 units, argument identification by 6.0 units and argument role determination (i.e. overall event extraction) by 5.4 units, largely due to the dramatic increase in recall of 11.5%, 11.2%, 7.5% and 7.2%.

Performance System/Human	Trigger Identification			Trigger Type Determination			Argument Identification			Argument Role Determination		
	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F
Our Baseline	75.2	52.0	61.5	70.3	49.0	57.8	58.4	42.7	49.3	55.2	38.6	45.4
+Compositional semantics	73.5	62.1	67.4	70.2	59.1	64.2	58.0	48.9	53.0	54.7	44.5	49.1
+Discourse consistency	**79.3**	**63.5**	**70.5**	**75.2**	**60.2**	**66.9**	**61.6**	**50.2**	**55.3**	**56.9**	**45.8**	**50.8**
Human annotator1(blind)	63.3	62.9	63.1	61.7	59.5	60.6	64.6	54.1	58.9	60.9	48.2	53.8
Human annotator2(familiar)	72.6	74.3	73.4	69.1	70.2	69.6	71.5	65.9	68.6	66.4	54.6	59.9
Inter-Annotator Agreement	45.8	42.9	44.3	45.3	42.5	43.8	60.4	49.7	54.5	55.1	45.9	50.1

Table 9: Overall contribution to Chinese event extraction

In addition, the results of two annotators show that Chinese event extraction is really challenging even for a well-educated human being. As shown in Table 9, the inter-annotator agreement on trigger identification and trigger type determination is even less than 45%. Although this figure is very low, it is not surprising: the results on the English ACE 2005 corpus show that the inter-annotator agreement on trigger identification is only about 40% (Ji and Grishman, 2008). Detailed analysis shows that a human annotator tends to make more mistakes in trigger identification for two reasons. The first reason is that a human annotator always misses some event mentions when a sentence contains more than one event mention. The second reason is that it is hard to identify an event mention due to the failure of following specified annotation guidelines, as mentioned in Ji and Grishman (2008). Table 9 also shows the performance gaps of human annotators between trigger identification and trigger type determination is very small (2.5% and 3.8% in F1-measure). It ensures that trigger identification is the most important step in Chinese event extraction for a human being. For human annotators, it’s much easier to determine the event type of a trigger, identify its arguments and determine the role of each argument, all with more than 90% in accuracy, once a trigger is identified correctly.

6.3 Discussion

Compared with English, the word structures in Chinese are much more complex and diverse, causing a lot of troubles in Chinese language processing. We ensure that compositional semantics in Chinese words is very useful for many Chinese language processing applications, such as machine translation, semantic parser, etc. For example, many actions (e.g. “砍” (hack), “咬” (bite), “踢” (kick), etc) can combine with “伤” (injure) to form words and most of those words have similar semantics. The results in table 8 show its contribution in Chinese event extraction. Although our approach is simple, the result is promising enough for further efforts in this direction.

This paper shows that the compositional semantics in the verb structure provides an ideal way to expand the coverage of triggers. As a discourse-driven language, ellipsis is very common in Chinese, causing inference from the discourse-level information is a fundamental requirement to understand the meaning of a clause, sentence or discourse.

7 Conclusion

In this paper we propose two novel inference mechanisms to Chinese trigger identification. In particular, compositional semantics inside Chinese triggers and discourse consistency between Chinese trigger mentions are used to resolve two critical issues in Chinese trigger identification: unknown triggers and word segmentation errors to known triggers. We give good reasons why this should be done, and present effective methods how this could be done. It shows that such novel inference mechanisms for Chinese event extraction are linguistically justified and pragmatically beneficial to real world applications.

In future work, we will focus on how to introduce the discourse information into the individual classifiers to capture those long-distance features and joint learning of subtasks in Chinese event extraction.

Acknowledgments

The authors would like to thank three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant No. 61070123 and No. 90920004, the National 863 Project of China under Grant No. 2012AA011102.

References

David Ahn. 2006. The Stages of Event Extraction. In Proc. COLING/ACL 2006 Workshop on Annotating and Reasoning about Time and Events. Pages 1-8, Sydney, Australia.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In Proc. Third Workshop on Syntax and Structure in Statistical Translation, pages 51-59.

Zheng Chen and Heng Ji. 2009a. Can One Language Bootstrap the Other: A Case Study on Event Extraction. In Proc. NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, pages 66-74, Boulder, Colorado.

Zheng Chen and Heng Ji. 2009b. Language Specific Issue and Feature Exploration in Chinese Event Extraction. In Proc. NAACL HLT 2009, pages 209-212, Boulder, CO.

Jenny Rose Finkel, Trond Grenager and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proc. ACL 2005, pages 363-370, Ann Arbor, MI.

Prashant Gupta and Heng Ji. 2009. Predicting Unknown Time Arguments based on Cross-Event Propagation. In Proc. ACL-IJCNLP 2009, pages 369-272, Suntec, Singapore.

Ralph Grishman, David Westbrook and Adam Meyers. 2005. NYU's English ACE 2005 System Description. In Proc. ACE 2005 Evaluation Workshop, Gaithersburg, MD.

Hilda Hardy, Vika Kanchakouskaya and Tomek Strzalkowski. 2006. Automatic Event Classification Using Surface Text Features. In Proc. AAAI 2006 Workshop on Event Extraction and Synthesis, pages 36-41, Boston, MA.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou and Qiaoming Zhu. 2011. Using Cross-Entity Inference to Improve Event Extraction. In Proc. ACL 2011, pages 1127-1136, Portland, OR.

Heng Ji. 2009. Cross-lingual Predicate Cluster Acquisition to Improve Bilingual Event Extraction by Inductive Learning. In Proc. NAACL HLT Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, pages 27-35, Boulder, CO.

Heng Ji and Ralph Grishman. 2008. Refining Event Extraction through Cross-Document Inference. In Proc. ACL-08: HLT, pages 254-262, Columbus, OH.

Young-Joo Kim. 2000. Subject/object drop in the acquisition of Korean: A Cross-linguistic Comparison. Journal of East Asian Linguistics, 9(4): 325-351.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In Proc. ACL 2003, pages 439-446, Sapporo, Japan.

Shasha Liao and Ralph Grishman. 2010. Using Document Level Cross-Event Inference to Improve Event Extraction. In Proc. ACL 2010, pages 789-797, Uppsala, Sweden.

Zhongguo Li. 2011. Parsing the Internal Structure of Words: A New Paradigm for Chinese Word Segmentation. In Proc. ACL 2011, pages 1405-1414, Portland, OR.

Percy Liang, Michael I. Joedan and Dan Klein. 2011. Learning Dependency-Based Compositional

Semantics. In Proc. ACL 2011, pages 590-599, Portland, OR.

Gideon Mann. 2007. Multi-document Relationship Fusion via Constraints on Probabilistic Databases. In Proc. HLT/NAACL 2007, pages 332-229, Rochester, NY.

Mstislav Maslennikov and Tat-Seng Chua. 2007. A Multi Resolution Framework for Information Extraction from Free Text. In Proc. ACL 2007, pages 592-599, Prague, Czech Republic.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In Proc. EMNLP/CoNLL 2007, pages 717-727, Prague, Czech Republic.

Siddharth Patwardhan and Ellen Riloff. 2009. A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In Proc. EMNLP 2009, pages 151-160, Singapore.

Hongye Tan, Tiejun Zhao, Jiaheng Zheng. 2008. Identification of Chinese Event and Their Argument Roles. Proc. of the 2008 IEEE 8th International Conference on Computer and Information Technology Workshops, pages 14-19, Sydney, Australia.

Yuk Wah Wong and Raymond J. Mooney. 2007. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus. In Proc. ACL 2007, pages 960-967, Prague, Czech Republic.

Roman Yangarber, Clive Best, Peter von Etter, Flavio Fuart, David Horby and Ralf Steinberger. 2007. Combining Information about Epidemic Threats from Multiple Sources. In Proc. RANLP 2007 workshop on Multi-source, Multilingual Information Extraction and Summarization. Borovets, pages 41-48, Borovets, Bulgaria.

Roman Yangarber and Lauri Jokipii. 2005. Redundancy-based Correction of Automatically Extracted Facts. In Proc. EMNLP 2005, pages 57-64, Vancouver, Canada.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proc. ACL 1995, pages 189-196, Cambridge, MA.

Minglin Yuan. 1998. Studies on Valency in Modern Chinese. Chinese Commerce and Trade Press, Beijing, China.

Luke S. Zettlemoyer and Michael Collins. 2007. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. In EMNLP/CoNLL 2007, pages 678-687, Prague, Czech Republic.

Dexi Zhu. 1980. Research on Chinese Modern Grammars. Chinese Commerce and Trade Press, Beijing, China.