RPI_BLENDER TAC-KBP2015 System Description

Yu Hong[1,2], **Di Lu**[1], **Dian Yu**[1], **Xiaoman Pan**[1], **Xiaobin Wang**[2], **Yadong Chen**[2], **Lifu Huang**[1], **Heng Ji**[1]

[1] Computer Science Department, Rensselaer Polytechnic Institute
[2] Computer Science Department, Soochow University
tianxianer@gmail.com, jih@rpi.edu

1 Introduction

This year the RPI_BLENDER team participated and achieved top 1 in four tasks at KBP2015: Event Nugget Detection (section **??**), Event Nugget Coreference Resolution (section 4), Cold-start Slot Filling Validation Filtering (section 5) and Tri-lingual Entity Linking and top 2 in Tri-lingual Entity Discovery and Linking (section 2).

2 Tri-lingual Entity Discovery and Linking

2.1 Entity Mention Identification

To extract English name mentions, we apply a linear-chain CRFs model trained from ACE 2003-2005 corpora (Li et al., 2012a). For Chinese and Spanish, we use Stanford name tagger (Finkel et al., 2005). We also encode several regular expression based rules to extract poster name mentions in discussion forum posts. In this year's task, person nominal mentions extraction is added. There are two major challenges: (1) Only person nominal mentions referring to specific, individual real-world entities need to be extracted. Therefore, a system should be able to distinguish specific and generic person nominal mentions; (2) within-document coreference resolution should be applied to clustering person nominial and name mentions. We apply heuristic rules to try to solve these two challenges: (1) We consider person nominal mentions that appear after indefinite articles (e.g., *a/an*) or conditional conjunctions (e.g., *if*) as generic. The person nominal mention extraction F1 score of this approach is around 46% for English training data. (2) For coreference resolution, if the closest mention of a person nominal mention is a name, then we consider they are coreferential. The accuracy of this approach is 67% using perfect mentions in English training data.

2.2 Unsupervised Entity Linking

Our entity linking system is a domain and language independent system (Wang et al., 2015). This system is based on an unsupervised collective inference approach. Given a set of English entity mentions $M = \{m_1, m_2, ..., m_n\}$, our system first constructs a graph for all entity mentions based on their co-occurrence within a paragraph. Then, for each entity mention m, our system uses the surface form dictionary $< f, e_1, e_2, ..., e_k >$, where $e_1, e_2, ..., e_k$ is the set of entities with surface form f according to their properties (e.g., labels, names, aliases), to locate a list of candidate entities $e \in E$ and compute the importance score by an entropy based approach (Zheng et al., 2014). Finally, it computes similarity scores for each entity mention and candidate entity pair $< m, e >$ and selects the candidate with the highest score as the appropriate entity for linking. For Chinese and Spanish, we first translate mentions into English using name translation dictionaries mined from various approaches described in (Ji et al., 2009). If mentions cannot be found in the dictionaries, we use Pinyin for Chinese mentions and normalize special characters for Spanish mentions.

2.3 Linking Feedback for Typing

This year's pre-defined entity types are expanded to Person, Geo-political Entity, Organization,

Location and Facility. Therefore, we implement a fine-grained entity typing system based on linking feedback and map them back to these five types. We utilize Abstract Meaning Representation (AMR) corpus (Banarescu et al., 2013) which contains over 100 fine-grained entity types and human annotated KB titles. DBPedia[1] also provides rich types for each page. Therefore, we generate a mapping table between AMR type and DBPedia `rdf:type` (e.g., *university* - `TechnicalUniversitiesAndColleges`). Finally, we can get a list of typing candidates for each linking candidate. For example, give a mention ***RPI***, we can obtain a list of linking candidates [`Rensselaer Polytechnic Institute`, `Rensselaer at Hartford`, `Lally School of Management & Technology`, ...] using our entity linking system. Each linking candidate has a list of mapped AMR types, `Rensselaer Polytechnic Institute`: [*university* - `EducationalInstitution`, *university* - `UniversitiesAndCollegesInNewYork`, *organization* - `Organisation`, ...]. If the confidence value of the top 1 linking candidate is reliable, we only select its typing result. Otherwise, we merge the typing results of all candidates. The typing F1 score is 93.2% for perfect mentions in English training data.

2.4 Build EDL for a New Language Over Night

We also propose a novel unsupervised entity typing framework by combining symbolic and distributional semantics. We start from learning general embeddings for each entity mention, compose the embeddings of specific contexts using linguistic structures, link the mention to knowledge bases and learn its related knowledge representations. Then we develop a novel joint hierarchical clustering and linking algorithm to type all mentions using these representations. This framework doesn't rely on any annotated data, predefined typing schema, or handcrafted features, therefore it is highly extensible and can be quickly adapted to new languages. Different languages may have different linguistic resources available. For example, English has rich linguistic resources (e.g., Abstract Meaning Representation) that can be utilized to model local contexts while some languages don't. For these low-resource languages, we can utilize the embeddings of context words which occur within a limited size of window instead of rich linguistic resource based compositional specific-context embedding. In addition, for low-resource languages, there are not enough unlabeled documents to train word embeddings and KBs may not be available for these languages. In this case, we can utilize other feature representations such as bag-of-words tf-idf instead of embedding based representations. To prove this, we apply our framework to two low-resource languages: Hausa and Yoruba. The mention-level typing accuracy with perfect boundary is very promising: 85.42% for Hausa and 72.26% for Yoruba. Experiments on various languages show comparable performance with state-of-the-art supervised typing systems trained from a large amount of labeled data. Then we can apply the above unsupervised language-independent linking component to link each mention to the English KB and use the linking feedback to refine typing results.

3 Nugget Detection

3.1 Baseline Maximum Entropy Model

We utilize a Maximum Entropy model (MaxEnt) to predict the event type of realis type of each candidate event nugget, based on linguistic features as summarized in Table 1. They can be roughly divided into the following categories:

Lexical features include *nbr*, *sense* and *bro*. *nbr*s refer to the unigrams or bigrams within the text window of size 2. *bro* is a Brown cluster, which was learned from ACE English corpus (Brown et al., 1992). We used the clusters with prefixes of length 13, 16 and 20 for each token. The synonyms are the most possible synset in WordNet (George, 1995).

Syntactic features include *dg_words*, *dg_typ*, *be_pron* and *be_pron*. They represent the characteristics of dependency and coreference.

Entity features include *en_typ* and *en_typ(dg)*, which capture the participation of entities in events of specific types.

Statistical feature refers to *ev_typ*. The feature *ev_typ* is learned from the distributions of event

[1] http://dbpedia.org

types over nuggets in the training corpus.

Feature	Description
nbr	neighbor grams and POS
sense	lemma, synonyms and case of the target token
bro	Brown clusters (Brown et al., 1992)
dg_words	dependent and governor words
be_pron	whether the target token is a non-referential pronoun
be_mod	whether the target is a modifier of job title
en_typ	types of entities within the text window of size 3 if have
en_typ(dg)	types of dependent and governor entity of the target token
ev_typ	argmax event type of the target token in training data

Table 1: Features for Event Nugget Classification

Feature	Description
nbr	neighbor grams and POS
dg_words	dependent and governor words
dg_typ	dependency types associated of *dg_woods*
en_typ(dg)	types of dependent and governor entity of the target token
nug_attr	lemma, POS, event type and subtype of the nugget
fv	first verb within the clause containing the nugget
co_neg	whether occurred with a negative word (e.g., not)
co_unc	whether occurred with a modifier of uncertainty (maybe)
co_madv	whether occurred with a modal adverb (possibly)
dis	distance between the nugget and *co_neg* or *co_unc*

Table 2: Features for Realis Classification

Table 2 shows the features used in realis classification. Besides the basic features (*nbr*, *dg_words*, *dg_typ*, *en_typ(dg)* and *nug_attr*), we employed the words that express negation and uncertainty, such as *co_neg*, *co_unc*, and *co_madv*. The feature *dis* is used to identify the scope affected of negation and hypothesis. In addition, we considered the first verb of a clause (*fv*). An empirical finding shows that some *fvs* are capable of constraining the realis type in a clause.

3.2 Using Homogeneous Data to Improve Nugget Classification

Homogeneous data refers to a set of samples that have common attributes. The attributes can be predefined in terms of specific needs. In rhetorical analysis of discourse, for example, we employ the stylistic characteristics (narrative, argumentative, lyric, etc) as the attributes for homogeneity detection. The stylistically homogeneous discourses generally contain similar rhetorical structure. This helps an analyzer predict the rhetorical structure of a specific discourse sample by learning the homogeneous samples. In other fields, similarly, we may consider Part-of-Speech based homogeneity detection for machine learning on syntax, semantics for grammar, pragmatics for word sense, etc.

In this paper, we regard an occurrence of a nugget as a sample. We focus on the application of pragmatically homogeneous samples in event nugget classification.

Context is the most important aspect of pragmatics. In our case, it can be used to verify whether the samples of a nugget indicate the events of the same type. See the following sentences, where the samples of the nugget "*death*" in (1) and (2) occur in very similar contexts and both indicate an "*Die*" event. By contrast, the samples in (1) and (3) occur in different contexts and indicate different events. Accordingly, we define the homogeneous samples as the ones that occur in similar contexts. We name such samples as pragmatically homogeneous samples, which means that they convey similar senses in a specific context, triggering the events of the same type.

> 1) *UN puts the conflict's* death [Die/Actual] *toll at over 67,000.*
>
> 2) *I urge you to take in a variety of sources in researching the civilian* death [Die/Actual] *toll of the conflict.*
>
> 3) *Elliott was the third convict put to* death [Execute/Actual] *in the state since the start of the year and number 105 since 1976.* (here, *death* refers to *death penalty*)

It is worth noting that not just the pragmatically homogenous samples of a single nugget, but the ones among different nuggets can take advantage of the homogeneity in revealing the same event type. See sentences 4) and 5) where the samples of the nuggets "*executed*" and "*death*" occur in similar context, both of which evoke an "Execute" event.

> 4) *A convicted murderer was* $\underline{executed}$ [Execute/Actual] *in the electric chair Tuesday in Virginia.*
>
> 5) *Larry Elliott, 60, was the first person put to* $\underline{death}$ [Execute/Actual] *by electrocution since June of last year.*

In order to reduce the errors caused by uncertainty, a MaxEnt model always makes decision in terms of the most reliable priori knowledge. Accordingly, it always drives the classifier to assign a sample to the type class that represents its homogeneous samples. For example, suppose that the nugget "*sentence*" in either sentence (2) or (3) is a training sample, while (1) a test sample. As mentioned above, the nugget in (1) is pragmatically homogeneous with (2) but heterogeneous with (3). In a view of pragmatic features, therefore, the context in (2) provides reliable prior knowledge while (3) doesn't. To ensure the reliability of classification, the MaxEnt model will predict (1) as the same event type with (2) but not (3).

It is predictable that the nugget classification system can be optimized if trained on rich homogeneous samples. From here on, the remaining problems include 1) how to enrich homogeneous data and 2) enable learning among the homogeneous samples. To solve the problems, we respectively propose a feature oriented transfer learning method and a lexicon based enrichment method for homogeneous data.

3.2.1 Feature Oriented Transfer Learning

We detect pragmatically homogeneous samples of a nugget in terms of contextual similarity. The context appears as the co-occurred words with a sample, such as the ones in a chunk, text window and dependency tree. In some cases, a similar context means that two nuggets have very similar contextual words (see 1) and 2)). In other cases, the agreement in contexts can be reached only at the level of semantics. See 9) and 10) where the contexts have the same meaning but they are consisted of very different words.

> 9) *A good deal of this hatred is related to the fact that Congress has a tradition of preventing its own members convicted of crimes from ever going to* $\underline{jail}$ [Arrest-Jail/Other].
>
> 10) *Scholars of Brazil ' s judicial system say legislators in corruption scandals often avoid* $\underline{jail}$ [Arrest-Jail/Other].

We propose to employ semantic feature based transfer learning for training the MaxEnt model. Transfer learning is a process of learning homogeneous data. A feature based transfer learner develops uniform feature representation to characterize the common attributes of homogeneous samples and thus unifies them in the feature space. This usually facilitates the learning process of homogeneous samples.

We use frame semantics for characterizing the contexts of homogeneous samples. Given a sample along with its context, we transform the words in the context to their semantic frames. A semantic frame is the conceptual representation of a cluster of words, reflecting the general semantics of the words. See Table 4. Using such frames as features, we can generate a nearly uniform feature representation for the semantically similar contexts. See 11) which exhibits the common semantic frames of some key words in 9) and 10).

> Frame: *Leadership* (see Table 4)
> Lexical units in 9) and 10): *congress*, *legislator*
>
> Frame: *Preventing* (see Table 4)
> Lexical units in 9) and 10): *prevent*, *avoid*
>
> Frame: *Being_Incarcerate*
> Lexical units in 9) and 10): *jail*

Accordingly, the homogeneous samples that have similar context can be always transferred to the same region in the feature space, far from the heterogeneous samples. In terms of the space partition, the

*frm*_1 Preventing	*frm*_2 Leadership
prevent, stave off, avoid, avert, obviate, prohibit, obviate, upset, etc.	*congressman, legislator, administer, bishop, chairman, chief, etc.*

Table 3: Examples of semantic frames and the lexical units they contained

MaxEnt model is capable of assigning a nugget to the class type of the semantically-similar homogeneous samples, even if it has a very different context in content from the samples, such as that in 9) and 10).

In practice, we fulfill the semantic-level transfer learning for all kinds of homogeneous samples, i.e., the nuggets of 39 event types, including the 38 KBP event types and a N/A type (means "*Other*" type). In the feature space employed by the original classification system, we replace the contextual features by their semantic frames. The contextual features include the co-occurred words with a candidate nugget in the same chunk, text window and the first-level dependency subtree. We retrained the MaxEnt model using the revised feature space over the same training data.

3.2.2 Lexicon based Enrichment of Homogeneous Data

Because of the lack of training data, diverse nuggets in the test data and generally narrow scope of application of homogeneity, we suspect that there isn't any reliable homogeneous sample to use for some test samples during the process of machine learning. Our solution is to introduce more potential homogenous samples into the training data from external linguistic resources. We consider two classes of words for the enrichment of homogeneous samples: one can serve as a nugget to trigger an event even though it never occurred in the training data, while the other have occurred as known ground-truth nuggets but necessarily associated with some new context. We name the former cases as brand-new (BN) homogenous samples, by contrast, the latter half-new (HN).

In practice, for either BN or HN samples, it is necessary to add their contexts to the training data. That is because the MaxEnt model can learn the pragmatic features only from the contexts. From this perspective, our goal is actually to diversify the pragmatic context specific to a certain event type.

Acquisition of BN Samples

We propose to acquire BN samples from FrameNet, which organizes lexical units (words and phrases) into different clusters and assign a conceptual representation to each cluster[2]. The conceptual representation is also called semantic frame. Moreover, FrameNet provides an index structure. The index facilitates the search for the semantic frames given a lexical unit as query. By accessing frames, we can obtain all the lexical units semantically similar to the query.

Technically, we use a ground-truth nugget as query. By going through the index, we discover all the related semantic frames and further the semantically-similar lexical units. In terms of the priori correspondence between the ground-truth nugget and an event type, we associate the retrieved lexical units with the event type, generating event-type-specific candidate nuggets. By using all ground-truth nuggets as queries, we search all possible candidate nuggets in FrameNet. Based on this approach we can filter the duplicated candidates for each KBP event type to eventually produce unique new nuggets.

As mentioned above, the newly found nuggets are not yet eligible BN samples until we attach some contexts to each of them. Similarly, we acquire the contexts by using ad-hoc Information Retrieval (IR) technique. We show the detailed retrieval procedure as follows:

Input: a nugget *x*
Output: the set of contexts *C*

Step 1: Use *x* as query to search the related documents *D* in the KBP dataset.
Step 2: Pick up a document *d* from *D*.
Step 3: Extract a sentence S_{con} that contains *x* and the left and right neighbors S_{nei} from *d*.
Step 4: Use the co-occurred words with *x* in chunks and dependency trees in S_{con} respectively as the context c_{inc} and c_{dep}; Use the words in a radar-fixed

[2]https://framenet.icsi.berkeley.edu/fndrupal/

text window as the context c_{win} (the window may occupy contents in both S_{con} and S_{nei}).

Step 5: Verify whether *C* is empty. If yes, add the triple *T*(c_{inc}, c_{dep}, c_{win}) to *C*, else calculate the similarity of the current triple T_{cur} to every existing triple T_{exi} in *C*. We use VSM based Cosine metric in the similarity measurement. If the similarity to each T_{exi} is smaller than a threshold θ_{sim}, add T_{cur} to *C*, else skip to Step 6 directly.

Step 6: If the number of the context triples *T* in *C* is bigger than *n*, break out the loop.

Step 7: If there isn't any S_{con} in *d*, go to Step 2, else skip to 3.

In the procedure, we set *n* to 50 while θ_{sim} 0.05. Accordingly, we can obtain various contexts *T*(c_{inc}, c_{dep}, c_{win}) for each newly found nugget. We use every pair of nugget *x* and context *T* as a BN sample.

In the set of BN samples, however, there is a lot of noise. Most cases are caused by the ambiguous ground-truth nuggets. For example, the nugget "*strike*" can have many meanings, such as "*walk off the job and protest*", "*hit in a forceful way*" and "*be impressed*", corresponding to the semantic frames *"Political_actions"*, "*Attack/Cause_harm*" and "*Coming_to_believe*" respectively. The nugget indicates a "*Demonstrate*", "*Attack*" or "*Injury*" event only if it conveys the former two meanings. In this case, the eligible semantic frames are *"Political_actions"* and "*Attack/Cause_harm*". Undertaking the remaining meaning (i.e., "*Coming_to_believe*"), it cannot trigger any kind of KBP events.

To filter the noise, we need to disable the ineligible semantic frames at the beginning of the acquisition process. We employed two rules to identify such frames:

> **r1:** FrameNet associates each semantic frame of a lexical unit with a POS tag. It means the POS is fixed when the unit indicates a specific meaning. For example, the word "*fine*" conveys the semantic frame "*Fining*" (which means "*penalty paid*") only if its POS is *noun* or *verb*. By contrast, the frame "*Fining*" (which means "*penalty paid*"). Similarly, a nugget generally has a fixed POS when triggering a type of KBP event. For a ground-truth nugget, accordingly, we detect its general POS in triggering specific types of events. Then we can determine the semantic frames unassociated with the POS to be ineligible.

> **r2:** Given all the ground-truth nuggets of a specific event type e_{typ} and a semantic frame s_{frm}, we measure the agreement of the nuggets in taking on the meaning of s_{frm}. If the agreement is high, we determine s_{frm} to be eligible. Of course, in this case, s_{frm} should be associated with the event type e_{typ}. It means that the lexical units in s_{frm} are very likely to trigger events of the type e_{typ}. Therefore, they should be used as new candidate nuggets.

We measure the agreement in rule 2 by calculating the average joint probability of the nuggets of E_{typ} occurring as a unit in the semantic frame S_{frm}:

$$P(e_{typ}|s_{frm}) = \sum_{i=1, g_i \Rightarrow e_{typ}}^{n_e} \frac{Bool(g_i, s_{frm})}{n_e n_s}$$

$$Bool(g_i, s_{frm}) = \begin{cases} 1, & g_i \in s_{frm} \\ 0, & otherwise \end{cases}$$

where, g_i is a ground-truth nugget indicting an event with type e_{typ}, n_e is the total number of the ground-truth nuggets known to trigger e_{typ}, while n_s is the number of lexical units in the semantic frame s_{frm}. The boolean logic indicates whether a g_i occurs as a unit in s_{frm} in FrameNet. The equation implies that a semantic frame is associated with a KBP event type only if many lexical units of such semantics have occurred as a nugget of the event type in the training data.

We only employ s_{frm} that has an agreement P($e_{typ}|s_{frm}$) bigger than θ_{agr} in acquiring BN samples for e_{typ}. We empirically set θ_{agr} equal to 0.01.

Acquisition of HN Samples

HN samples are the ground-truth nuggets attached with new context. To acquire HN samples, similarly, we employ an IR system to search for sentence-level

contexts. To ensure the diversity of the context, we only search and keep distinctive contexts. Different from the initial condition in searching BN, the set of contexts *C* for the acquisition of HN is nonempty. The scale of context in the sets is imbalanced: some can be loaded with rich contexts, others few. We focus on increasing the scale in the latter cases, adding new contexts to the sets until their scale reaches the level of the known largest set.

3.3 Topic based Nugget Disambiguation

There exist many ambiguous nuggets in reality. We define an ambiguous nugget as the word that may trigger multiple types of KBP events. Considering the feasibility of topic in word sense disambiguation, we suggest that topic is also an important clue to the real event type of a nugget.

Now the question is how we use topic modeling for nugget disambiguation (our goal). It is predictably useful to generate a relation network between topics and the KBP event types. The relationship strength, empirically, enables the determination of the exact event type. Therefore the utilization of the topic-event relationship as a discriminative feature is probably a feasible way to achieve the goal. However the size of the dataset is not large enough to provide robust estimates for the topic-event relationship, considering the condition that a topic in the test data may never occur in the training data.

We propose to implement a Case-Oriented Real-Time topic-event relation detection system (CORT for short). CORT fulfills a case study *S*(*g*,*d*) for each nugget *g* in a test document *d*. It collects a cluster *C* of related documents to *d* in real time. At the document level, CORT regards the documents in *C* as the same topic t_C, similar to the topic t_d of *d*. Under the precondition that *g* frequently occurs in C, CORT detects the most probable event type e_C of *g* and associates it with the topic t_C. According to the strong similarity between the single-document topic t_d and the multiple-document t_C, CORT propagates the association of e_C with t_C to t_d.

The main characteristics of CORT include:

> ***Reliable:*** CORT analyzes the topic-event relation in a dataset containing abundant related documents. Compared to the original single document, the related documents provide richer topic-related information for surveying the closest topic-event relation. The estimation, therefore, is more reliable.
>
> ***Case oriented:*** CORT regards a test document as an independent scenario. A nugget in the document only evokes events of the type specific to the sole topic of the scenario. This facilitates a case study for every nugget in each test document.
>
> ***Unsupervised:*** During the case study, CORT collects related documents in real time for the test document by using a content-based local text search engine. Over the documents, CORT analyzes topic-event pairs in terms of their probability distributions.

CORT is used as post-processing behind the supervised nugget classification system. Its input is a document in the test data. The available priori knowledge is a list of ambiguous nuggets found by the classifier. CORT looks through the test document, word by word, and enforces disambiguation for every occurrence of an ambiguous nugget. Specifically CORT re-estimates the event types in terms of the topic-event association strength, using the new estimate to replace the original if they are different from each other.

We employed a Lucene-based search engine to support the acquisition of the related documents. Toward a test document, we use the keywords to formulate the query. In terms of the content similarity, we retrieved the related documents from the KBP dataset. In the ranking list of the search results, we uniformly contained top *n* (*n=100*) most similar documents as the reliable related documents.

To determine the topic-related event type, we introduce a margin model in topic-event association strength determination. Given an ambiguous nugget *g* and the event types *E* it evoked in the cluster *C* of the retrieved related documents, the margin *M* is calculated as:

$$M(e_i, e_j) = \frac{O(e_i) - O(e_j)}{O(e_i)O(e_j)},\ \forall e_i, e_j \in E$$

$$R(e_*, T_C) = \begin{cases} 1, & M(e_*, e_i) > \theta_M \\ 0, & otherwise \end{cases}$$

where, e_* is an event type evoked by *g*, *O*(*) denotes the occurrence frequency of e_* in *C*, T_C is the common topic of the related documents, and θ_M is a threshold for the margin *M*. Actually, the margin is used to reflect whether there is an event type of *g* occurred in *C* much more frequently than any other type. Constrained by the threshold (θ_M=0.7), the most frequently occurred event type in *C* is determined related to the topic T_C. The rest will be regarded as some unreliable estimates. Accordingly, for each ambiguous nugget *g*, CORT only remains only one reliable topic-related event type as the determination result.

To drive CORT to make the decision on topic-event relationship, it is important to note the event types of the ambiguous nuggets beforehand. However, there are no ground-truth event type annotations in the real-time retrieved documents. To solve this problem, we can use a weaker IE system to detect the event types beforehand. Obviously, the detection results definitely involve some errors. The event types used in the margin calculation, therefore, are some pseudo-relevant data in reality, easily causing a wrong decision on the topic-related event type. The way to reduce the risk of making mistake is to enhance CORT using adaptive boosting method. In detail, the IE system can be used to generate the event types in the retrieved related documents, while CORT can serve as the post-processing to improve the IE system. The iterative cycle can proceed iteratively

3.4 Multi-word Nugget Identification

A nugget can also be a phrase or chunk, such as *shoot up* (phrase) and *thrown in jail* (chunk). A basic step to identify such nuggets, accordingly, is to parse sentence constituents and extract phrases and chunks. However, it is difficult to solve the problem of jump-over connection, such as *throw in jail* in *throw him in jail*. Instead of syntactic parsing, we employ dependency parsing to extract candidate multi-gram nuggets. Specifically, we use the dependency structures of the ground-truth multi-gram nuggets as the templates to extract the nuggets from the test data. By using these templates, we extract nearly all eligible event nuggets (0.99 recall score). Nevertheless, the precision is very low due to two problems:

> *Some general dependency structures introduce large-scale noise, such as to be, to throw, in the, over and, by step, etc. In contrast, the eligible grams should be as to be, throw out, in the jail, over and over, step by step, etc.*
>
> *Most candidate multi-grams are not KBP event related nuggets.*

To filter the noise, we propose a bilingual word alignment based multi-gram qualification verification method. It determines the eligibility of an English multi-gram by checking whether the gram has a consistently aligned Chinese word or phrase in bilingual data.

For type-specific multi-gram nugget detection, we evaluated three methods, including rule based detection, supervised classification using linguistic features and semantics based clustering. Instead of words, we use the multi-grams as the objects, no matter in the stage of training or test. In the clustering, we generate the semantic vector for both candidate multi-gram nuggets and the ground-truth word-level ones. On the basis, we partition the nuggets in different clusters in terms of semantic similarity. In each cluster, we use the word-level nuggets and their event types as references to determine the event type of those semantically similar multi-gram nuggets. Similarly, we employ the propagation algorithm in the determination procedure.

3.5 Experiment and Analysis

3.5.1 Data and Evaluation Metrics

We trained and tested our system based on the TAC KBP 2015 Event Nugget Training Data set, which contains 158 documents. We chose 126 documents as training data and 32 documents as testing data. We evaluated our system on four metrics: CEAFe, B^3, Muc and Blanc.

3.5.2 Analysis of New Methods for Event Nugget Detection

We separately evaluated the proposed methods on the KBP 2015 nugget detection training data by 4-

System	*plain*	*typ*	*rls*	*typ+rls*
Baseline(Bas)	63.19	53.28	44.91	37.35
Bas+FTL	64.84	60.05	45.95	42.21
Bas+BN	67.68	62.01	47.25	42.99
Bas+TND	67.93	63.09	47.57	43.73
all	68.16	62.64	47.52	43.34

typ: mention type; rls: realis status

Table 4: Improvement Achieved over the Baseline

fold cross validation. See the performance in Table 4. BN denotes the collected type-specific Bran-New homogeneous samples, including BN nuggets acquired from FrameNet and their contexts retrieved from local KBP dataset. We added BN homogeneous samples to the original training data and retrain the MaxEnt classifier. The goal is to increase priori knowledge about potential nuggets as well as the contexts. The remaining results show that the supervised MaxEnt classifier achieves better discriminative ability by learning from rich knowledge.

By contrast, FTL doesn't rely on rich homogeneous samples. It is illustrated that FTL achieves significant improvements over the baseline using the same training data. It proves that the MaxEnt model obtains a real-time reasoning ability by transfer learning from the homogeneous data at the semantic level.

TND yields the most significant improvement than the baseline. It performs particularly well in determining the mention types. It proves that topic-event relationship specific to a nugget benefits the disambiguation of the nugget in different scenarios.

4 Event Nugget Coreference

4.1 Approach

We propose a method that views the event nugget coreference space as an undirected weighted graph in which the nodes represent all the event nuggets and the edge weight indicates coreference confidence between two event nuggets. There're two modules in our system: Maximum Entropy Model and clustering module.

4.1.1 Pairwise Model

We train a Maximum Entropy model to generate the cofidence matrix W. Each confidence value indicates the probability that there exists a coreference link C between event nuggets en_i and en_j.

$$P(C|en_i, en_j) = \frac{e^{(\Sigma_k \lambda_k g_k(en_i, en_j, C))}}{Z(en_i, en_j)}$$

where $g_k(en_i, en_j, C)$ is a feature and λ_k is its weight; $Z(en_i, en_j)$ is the normalizing factor.
The feature sets used during train are listed in Table 5.

4.1.2 Clustering

Let $EN = \{en_n : 1 \leq n \leq N\}$ be N event nuggets for one event type in a document and $EH = \{eh_k : 1 \leq k \leq K\}$ be K event hoppers. Let $f : EN \rightarrow EH$ be the function mapping from an event nugget $en_n \in EN$ to an event hopper $eh_k \in EH$. Let $coref : EN \times EN \rightarrow [0, 1]$ be the function that computes the coreference confidence value between two event nuggets $en_i, en_j \in EN$.
For each event type in the document, we construct a graph $G(V, E)$, where $V = \{en_n | f(en_n).en_n \in EN\}$ and $E = \{(en_i, en_j, coref(en_i, en_j)) | en_i, en_j \in EN\}$.
We then apply a hierarchical clustering algorithm to the graph. Because the number of hoppers K, which has to be set in advance, is unknown, we define a parameter ϵ to quantify the performance of the clustering results by varying the number of hoppers K from 1 to N. N is the number of event nuggets. For each K, ϵ is the ratio of the number of conflicting edges(N_c) to the number of all edges(N_e). The smaller ϵ is, the better the result of clustering is.

$$\epsilon_K = \frac{N_e}{N_a}$$

An edge between two event nuggets is defined as conflicting in either of the following two cases, where δ is the confidence threshold:

1. $f(en_i) = f(en_j)$ but $coref(en_i, en_j) < \delta$
2. $f(en_i) \neq f(en_j)$ but $coref(en_i, en_j) > \delta$

Here's an example to demonstrate how to compute ϵ. In Figure 1, five event nuggets are classified into three event hoppers, and there are three conflicting edges($coref(en_1, en_5), coref(en_2, en_4), coref(en_3, en_4)$)

Features	Remarks(EN1: the first event nugget, EN2: the second event nugget)
type_subtype_match	1 if the types and subtypes of the event nuggets match
trigger_pair_exact_match	1 if the spellings of triggers in EN1 and EN2 exactly match
stem_of_the_trigger_match	1 if the stems of triggers in EN1 and EN2 match
similarity_of_the_triggers(wordnet)	quantized semantic similarity score (0-5) using WordNet resource
similarity_of_the_triggers(word2vec)	quantized semantic similarity score (0-5) using word2vec embedding
token_dist	how many tokens between triggers of EN1 and EN2
realis_conflict	1 if the realis values of EN1 and EN2 exactly match
Sentence_match	1 if the sentences of EN1 and EN2 exactly match
extent_match	1 if the extents of EN1 and EN2 exactly match
POS_match	1 if two sentences have the same NNP, CD

Table 5: Features for the Pairwise Model

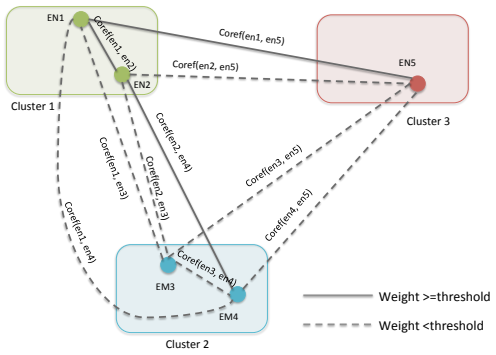

Figure 1: Clustering Example

according to our definition. Thus

$$\epsilon_3 = \frac{3}{10} = 0.3$$

4.1.3 Adding Cross-Media Features

In our recent work (Zhang et al., 2015), we proved that additional visual similarity features can be used to improve cross-document event coreference resolution for news videos. In this event nugget coreference resolutio we extend it to retrieve related images automatically, though we didn't use it for the evaluation because we are not allowed to use web access. The rich visual concepts in images can help us to identify event hoppers, even when they are difficult to be identified just based on the text resources. For example, when there are many different arguments in two event nuggets, it ' s very challenging for text features to tell if they are coreferential or not. However, we can use arguments as keywords to search for images online. If we can get similar or even the same images, it ' s quite likely that two event nuggets refer to the same event hopper.

For example, the following two event nuggets refer to the same event hopper. However it's very difficult to make the correct judgement just based on text information.

- EN1: Nigeria's graft-facing ex-governor {arrested} in Dubai.
- EN2: London's Metropolitan Police confirmed Ibori's {arrest}.

Let's make the use of powerful multimedia information. Figure 2 shows the top retrieved images using ["Nigeria", "ex-governor", "arrested", "Dubai"] as the query via Google image search, and Figure 3 shows the top retrieved images using ["London", "Metropolitan", "Police", "Ibori", "arrest"] as the query. We can see that the retrieved images of these two event nuggets share a lot of visual features, which indicate they might talk about the arrest of the same person. Thus it's highly possible that they are coreferential.

Figure 4 and Figure 5 shows another example including the following two event nuggets. The queries for image search are ["attacked", "oil", "facility", "Niger", "Delta"] and ["attacked", "Agip", "facility"] respectively.

- EN3: A militant group says it {attacked} an oil facility in the Niger Delta.
- EN4: A statement sent to reporters from the Joint Revolutionary Council claims militants

Figure 2: Retrieved Images for EN1

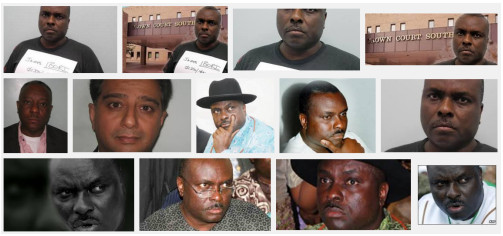

Figure 3: Retrieved Images for EN2

{attacked} the Agip facility early Wednesday morning.

Figure 4: Retrieved Images for EN3

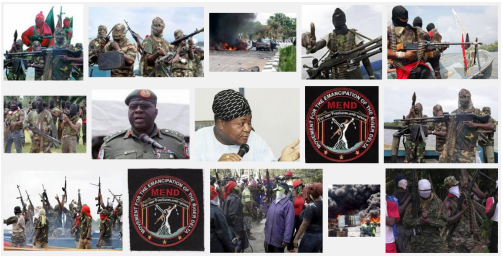

Figure 5: Retrieved Images for EN4

4.1.4 Feedback for Realis Improvement

The realis classification accuracy in the event nugget detection system is quite low (around 0.5). We hypothesize that event nuggets which refer to the same event hopper should have the same realis type. So we trained another classifier without the feature of realis_conflict, and used the event coreference results from this classifier to improve realis classification.

For example, the following three event nuggets from the evaluation data are automatically clustered into the same event hopper:

- EN1: E3 prison Justice_Arrest-Jail Generic 0.54
- EN2: E20 prison Justice_Arrest-Jail Actual 1.00
- EN3: E21 prison Justice_Arrest-Jail Actual 1.00

And according to the output of our event nugget detection system, the realis calues of these three event nuggets are ("*Generic*", 0.54), ("*Actual*", 1.00) and ("*Actual*", 1.00) respectively. While the realis confidence of EN1 is low, that of EN2 and EN3 are pretty high. Thus we modify the realis of EN1 to "*Actual*" by a majority voting. This feedback approach yields substantial improvement on realis classification.

4.2 Experiments

4.2.1 Analysis on Metrics

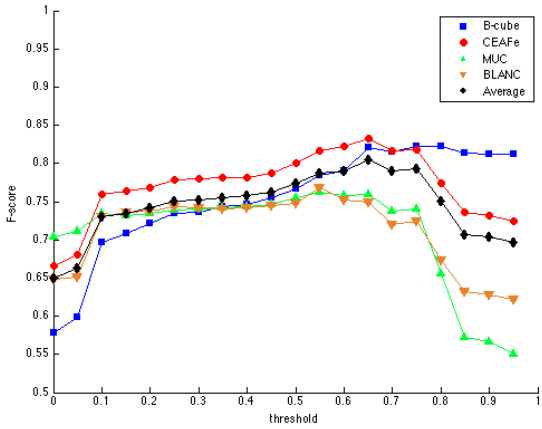

Figure 6: F-socres based on four metrics versus confidence threshold

Fig 6 shows the F-scores versus threshold δ based on four evaluation metrics. The maximal CEAFe score 0.83 was obtained when the threshold was 0.68. Thus we chose 0.68 as the cofidence threshold. However, it's obvious that the F-scores based on Muc and Blanc metrics are highly sensitive to δ.

This is similar to the observation from previous work on ACE event coreference resolution (Chen et al., 2015). Event nugget coreference is a new task, so we think it's not necessary to use them just for comparison with previous work. We recommend to use B^3 and CEAFe only to evaluation event nugget coreference resolution.

4.2.2 Remaining Challenges

Even though we achieved top 1 in end-to-end event nugget detection and coreference resolution, some challenges remain. We observed that some coreferential event nuggets share very few features. For example, the following two event nuggets refer to the same hopper:

- EN1: After months of speculation, the Simon Property Group on Tuesday finally made an unsolicited $10 billion offer for General Growth Properties, its {bankrupt} rival.
- EN2: it is not sufficient to pre-empt the process we are undertaking to explore all avenues to emerge from {Chapter 11} and maximize value for all the Company's stakeholders

However, it's hard for the system to tell that they refer to the same event. First of all, it's challenging to determine that two trigger phrases "bankrupt" and "Chapter 11" are semantically similar. Second, these two sentences don't share any obvious patterns or arguments.

5 Slot Filling Validation - Filtering

Our basic assumption is that a response is more likely to be true if it is supported by multiple strong teams. In order to validate this assumption, we first propose an unsupervised method to roughly estimate the performance of teams with little/no prior knowledge. After that, we focus on categorizing runs into multiple tiers based on their estimated performance and apply a tier-specific voting strategy incorporating linguistic constraints.

5.1 Tier Classification

Our objective is to classify a team as strong, relatively strong or relatively weak. The performance of a team is usually consistent regardless of individual queries. Thus, we can take advantage of the team ranking based on the preliminary assessments to estimate the overall performance/rank of a run.

Unfortunately, preliminary assessment results are often not available. But we can still obtain reliable initial credibility scores of runs by analyzing the common characteristics among various runs. Given the set of runs $R = \{r_1, \ldots, r_m\}$, we initialize their credibility scores $c(r)$ based on their interactions on ***claims*** (*i.e.*, a combination of query, slot type and filler). Suppose each run r_i generates a set of claims M_{r_i}. The similarity between two runs r_i and r_j is defined as follows (Mihalcea, 2004).

$$simlairty(r_i, r_j) = \frac{|M_{r_i} \cap M_{r_j}|}{\log(|M_{r_i}|) + \log(|M_{r_j}|)} \quad (1)$$

Then we construct a weighted undirected graph $G = \langle R, E \rangle$, where $R(G) = \{r_1, \ldots, r_m\}$ and $E(G) = \{\langle r_i, r_j \rangle\}$, $\langle r_i, r_j \rangle = similarity(r_i, r_j)$, and apply TextRank algorithm (Mihalcea, 2004) on G to obtain $c(r)$.

In our task, the classification problem can be regarded as finding two intervals within a set of credibility scores $C = \{c(r_1), \ldots, c(r_m)\}$ with optimal interval borders. We implemented the Jenks optimization method to determine the best arrangement of runs into three tiers. This is done by minimizing each tier's average deviation from the tier mean, while maximizing each tier's deviation from the means of the other groups (McMaster and McMaster, 2002).

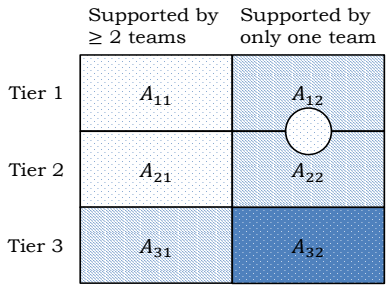

Figure 7: Tier-specific constraints.

5.2 Linguistic Constraints

We analyze the evidence sentence extracted for each claim by checking if it satisfies trigger constraints or the candidate filler satisfies type constraints.

Trigger Constraints

A trigger is defined as the smallest extent of a text which most clearly indicates a relation type. For trigger-driven slot types such per:city_of_birth, a response without sufficient lexical support will be directly judged as wrong by annotators.

We generally follow our previous work (Yu et al., 2015) on trigger mining. We mined fact-specific trigger lists based on patterns (Chen et al., 2010; Min et al., 2012; Li et al., 2012b) and correct evidence sentences from KBP 2012-2014 training corpus. In our experiment, we use 8,237 triggers and 392 triggers on average for each trigger-driven slot types[3].

Type Constraints

In Slot Filling (SF)/Cold Start Slot Filling (CSSF) task, slots are labeled as Name, Value, or String based on the content of their fillers.

Name slots (*e.g.*, per:spouse, org:parents) are required to be filled by a person, organization or geopolitical entity (GPE). We rely on name tagging results to validate type constraints for name slots. We also use city/state/coutry dictionaries to further validate if an entity belongs to a GPE subtype.

Value slots should be filled by a numerical value (*e.g.*, per:age, per:date_of_death and org:website). We use regular expressions and/or the name tagging result to verify the correctness of a value format.

For a string slot (*e.g.*, per:religion and per:origin), we collected category dictionaries from SF source corpus such as religion, origin, disease, title and crime which can help us make a rough judgement whether a candidate filler belongs to a specific category or not. We also mined dictionaries from NELL (Carlson et al., 2010) annotated KBP corpus which contains rich semantic category labels for millions of noun phrases. We mapped these semantic categories to slot types and keep high-confidence noun phrases in order to generate clean dictionaries.

5.3 Tier-specific Voting based on Constraints

We divided all the responses into 6 fields as shown in Figure 7. A_{i1} represents the responses submitted by at least two teams in Tier i and A_{i2} denotes the responses submitted by only one team in Tier i.

[3]The trigger lists are publicly available for research purposes at: http://nlp.cs.rpi.edu/data/triggers.zip

Method	τ	Z_τ
PageRank	0.69	8.35
Preliminary Assessment	0.79	9.59
Golden Standard	1.00	N/A

Table 6: Performance of estimating the ranks.

Note that a team can submit at most five runs during the evaluation. In this step, we keep the responses which are submitted by multiple strong or relatively strong teams. In other words, the responses from A_{11}, A_{21} and common responses of A_{12} and A_{22} are annotated as correct.

We discard all the responses in A_{32} since these responses are extremely noisy and we only lost less than 3% correct claims. A response in the remaining fields will be annotated as correct if it satisfies the above slot-specific trigger and type constraints.

5.4 Evaluation

Based on the released CSSF assessment result (69 runs in total), we can use the Kendall rank correlation coefficient τ (Kendall, 1948) to evaluate the degree of similarity between our estimated ranking and the standard ranking given the same set of N runs. The symmetric difference distance between two sets of ordered pairs $\mathcal{P}_1$ and $\mathcal{P}_2$ is denoted $d_\Delta(\mathcal{P}_1, \mathcal{P}_2)$. For example, the ordered set of $N = 3$ runs [r_1, r_2, r_3] gives the ranks [1,3,2] and can be decomposed into $\frac{1}{2}N(N-1)$ ordered pairs $P = \{[r_1, r_3], [r_1, r_2], [r_3, r_2]\}$. For $N \geq 10$, a null hypothesis test can be performed by transforming τ into a Z value as shown in Formula 3 (Abdi, 2007).

$$\tau = 1 - \frac{2 \times [d_\Delta(\mathcal{P}_1, \mathcal{P}_2)]}{N(N-1)} \quad (2)$$

$$Z_\tau = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \quad (3)$$

From Table 6, we can see that the value of Z of both methods is large enough to reject the null hypothesis and therefore we can conclude that our predicted ranking and the final ranking displayed a significant agreement. In addition, the tier classification method can successfully annotate the top 31 runs as tier 1 and the bottom 15 runs as tier 3.

CSSF run	F-score (%)	
	Original	Filtered
SFV2015_KB_12_4	30.1	**36.7**
SFV2015_KB_12_1	30.4	**34.9**
SFV2015_KB_12_5	30.9	**34.3**
SFV2015_KB_12_3	29.6	**33.7**
SFV2015_SF_03_3	31.4	**31.4**
SFV2015_KB_12_2	**33.8**	33.1
SFV2015_KB_03_2	30.5	**34.1**

Table 7: Performance upon top runs (CSLDC level).

CSSF run	F-score (%)	
	Original	Filtered
SFV2015_KB_12_4	27.6	**30.3**
SFV2015_KB_12_1	27.5	**28.4**
SFV2015_KB_12_5	27.1	**27.3**
SFV2015_KB_12_3	26.7	**27.3**
SFV2015_SF_03_3	27.5	**29.1**
SFV2015_KB_12_2	**28.8**	26.9
SFV2015_KB_03_2	22.9	**26.8**

Table 8: Performance upon top runs (CSSF level).

In Table 7 and 8, we show that our method can improve upon the top CSSF runs. Here we used the final scores which are computed considering both hops. Compared with the SF task, CSSF runs have relatively lower recall and therefore more sensitive to the filtering process. In addition, the majority voting method has a relatively worse performance since there are no pre-assigned queries. In this case, a correct claim is more likely to be submitted by only one team. Our strategy of discarding all the responses in A_{32} lead to the failure of our method in annotating responses of weak runs.

Acknowledgments

This work was supported by the U.S. DARPA LORELEI Program, DARPA DEFT Program No. FA8750-13-2-0041, DARPA Multimedia Seedling award, ARL NS-CTA No. W911NF-09-2-0053, NSF CAREER Award IIS-1523198, AFRL DREAM project, gift awards from IBM, Google and Bosch. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

H. Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proc. ACL 2013 Workshop on Linguistic Annotation and Interoperability with Discourse*.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R Hruschka Jr, and T. M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI 2010)*.

Z. Chen, S. Tamang, A. Lee, X. Li, W. Lin, M. Snover, J. Artiles, M. Passantino, and H. Ji. 2010. Cuny-blender tac-kbp2010 entity linking and slot filling system description. In *Proc. Text Analysis Conference (TAC 2012)*.

Zheng Chen, Heng Ji, and Robert Harallick. 2015. A pairwise coreference model, feature impact and evaluation for event coreference resolution. In *Proc. RANLP 2009 workshop on Events in Emerging Text Types*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

H. Ji, R. Grishman, D. Freitag, M. Blume, J. Wang, S. Khadivi, R. Zens, and H. Ney. 2009. Name extraction and translation for distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

Maurice George Kendall. 1948. Rank correlation methods.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012a. Joint bilingual name tagging for parallel corpora. In *21st ACM International Conference on Information and Knowledge Management*.

Y. Li, S. Chen, Z. Zhou, J. Yin, H. Luo, L. Hong, W. Xu, G. Chen, and J. Guo. 2012b. Pris at tac2012 kbp track. In *Proc. Text Analysis Conference (TAC 2012)*.

R. McMaster and S. McMaster. 2002. A history of twentieth-century american academic cartography. *Cartography and Geographic Information Science*, 29(3):305–321.

R. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proc. Association for Computational Linguistics (ACL 2004)*.

B. Min, X. Li, R. Grishman, and A. Sun. 2012. New york university 2012 system for kbp slot filling. *Proc. Text Analysis Conference (TAC 2012)*.

Han Wang, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*.

D. Yu, H. Ji, S. Li, and C. Lin. 2015. Why read if you can scan? trigger scoping strategy for biographical fact extraction. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*.

Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. 2015. Cross-document event coreference resolution based on cross-media features. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*.