BBN ACCENT Event Coding Evaluation

Raytheon BBN Technologies

updated August 28, 2015 to include Recall & Machine Translation evaluations

Table of Contents

1 Overview

To support forecasting models for W-ICEWS (Worldwide Integrated Crisis Early Warning System), the BBN ACCENT event coder automatically extracts event data from news reports from around the world. BBN ACCENT is a state-of-the-art event coder based on BBN SERIF, a natural language analysis engine that extracts structured information (e.g. entities, relationships, and events) from text.

As part of the W-ICEWS effort, the performance of the BBN ACCENT event coder was evaluated in various ways. First, and most comprehensively, the precision of the event coder was evaluated: how many of the events produced by the system are actually correct? The overall results are as follows:

Event Code	BBN ACCENT Precision
01: Make Public Statement	71.1%
02: Appeal	71.4%
03: Express Intent To Cooperate	74.8%
04: Consult	80.6%
05: Diplomatic Cooperation	81.1%
06: Material Cooperation	65.9%
07: Provide Aid	73.9%
08: Yield	62.0%
09: Investigate	70.2%
10: Demand	58.7%
11: Disapprove	65.2%
12: Reject	74.6%
13: Threaten	66.0%
14: Protest	84.5%
15: Exhibit Force Posture	70.9%
16: Reduce Relations	69.9%
17: Coerce	88.1%
18/19: Assault/Fight	73.8%
20: Unconventional Mass Violence	83.6%
ALL (weighted by code frequency)	75.6%

Table 1: BBN ACCENT event coding precision

This document provides details on the methodology of this evaluation and its results.

We also performed a smaller, supplemental evaluation to assess the approximate recall of the event coder using a “gold standard” set of 1,000 documents coded by humans. This evaluation is designed to determine approximately how many events the automatic event coder is missing. Due to the sparsity of many types of events, this set of documents does not provide the same level of statistical significance as the precision evaluation. (For comparison, almost 20,000 individual events were judged during the precision evaluation, while the 1,000 documents in the gold standard contained only XX events, with fewer than a hundred for more than half of the top-level event codes.) Still, this evaluation provides important information about the estimated coverage of the BBN ACCENT event coder.

Finally, we performed a second supplemental evaluation to compare event coding performance over machine and human translations of the same French and Spanish documents. The goal of this evaluation was to give a sense of how much degradation in performance is experienced when one is forced to rely machine translations of source documents.

2 Task & Data

2.1 Ontology

The W-ICEWS program uses event data coded according to the Conflict and Mediation Event Observations (CAMEO) ontology, developed by Deborah J. Gerner, Philip A. Schrodt, Ömür Yilmaz, and Rajaa Abu-Jabr [1]. The version of the canonical CAMEO Codebook used as the foundation for this document can be viewed here[2] :

http://eventdata.parusanalytics.com/cameo.dir/CAMEO.CDB.09b5.pdf

The CAMEO ontology consists of twenty top-level categories of events:

- Make Public Statement (01)
- Appeal (02)
- Express Intent to Cooperate (03)
- Consult (04)
- Engage in Diplomatic Cooperation (05)
- Material Cooperation (06)
- Provide Aid (07)
- Yield (08)
- Investigate (09)
- Demand (10)
- Disapprove (11)
- Reject (12)
- Threaten (13)
- Protest (14)
- Exhibit Military Posture (15)

[1] D. Gerner, P. Schrodt, Ö. Yilmaz, R. Abu-Jabr . Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions. *Presented at the International Studies Association, New Orleans, and American Political Science Association, Boston* (2002).

[2] More recent versions of the CAMEO ontology can also be found online; the portion of the codebook that we drew upon appears largely unchanged, though we have not performed a line-by-line comparison. Note that the W-ICEWS program makes use only of the CAMEO event ontology, not the actor codebook also included in the original CAMEO codebook.

- Reduce Relations (16)
- Coerce (17)
- Assault (18)
- Fight (19)
- Engage in Unconventional Mass Violence (20)

Each top-level category is also broken down into sub-categories, e.g. Provide Economic Aid (071) and Hunger Strike (142).

Each CAMEO event has a source actor and a target actor, for example:

> *Demonstrators in Ukraine called for the resignation of Prime Minister Mykola Azarov.*
>
> Event code: 1411 (Demonstrate for leadership change)
> Source actor: Protester (Ukraine)
> Target actor: Mykola Azarov

All actors must be resolved to the canonical W-ICEWS actor database, which consists of about 55,000 named actors (e.g. *Mykola Azarov*) and about 700 types of described actor categories ("agents") which can then be linked to a country (e.g. *Protester,* which is linked above to Ukraine).

To support more consistent training data creation and event coding system evaluation, BBN expanded the original CAMEO codebook with additional guidelines and examples designed to clarify potential ambiguities and to resolve overlap between event codes and subcodes. This document (*ICEWS Expanded CAMEO Annotation Guidelines*) was iteratively reviewed and approved by participants in the W-ICEWS program and serves as the basis of the evaluation judgments described in this document. It is also available as part of the Dataverse release. Please note, however, that this document should not be taken as a canonical extension to the CAMEO framework, since feedback was limited to participants in the W-ICEWS program.

2.2 Corpus

The W-ICEWS program uses commercially-available news sources from roughly 300 different publishers, including a mix of internationally (e.g., Reuters, BBC) and nationally (e.g., O Globo, Fars News Agency) focused publishers. The W-ICEWS program filters the data stream to those news stories more likely to focus on socio-political topics and less likely to focus on sports or entertainment.

For the primary evaluation (precision), we used the portion of the W-ICEWS corpus published between 2011 and 2013. All BBN system development was performed using documents dated 2010 or earlier, to allow the evaluation to simulate the condition where the system is being evaluated on new, real-time data.

After removing duplicate and near duplicate[3] documents, there were approximately 5.5 million documents in the primary evaluation corpus. The size of the evaluation corpus was chosen to provide, where possible, at least 500 coded events in each top-level CAMEO event category for each evaluated system. This was successful in all but one case: the Jabari NLP system evaluated here produced only 336 events in the Unconventional Mass Violence category; these events are in general extremely rare.

According to standard W-ICEWS procedure, only the first six sentences of each document were included in the evaluation corpus.

3 Primary Evaluation: Precision

3.1 Systems Evaluated

Two systems were evaluated in this effort: BBN ACCENT and a baseline system, Jabari NLP. Jabari NLP was the previous event coder used by the W-ICEWS program; it was a Java port of the open-source event coder TABARI, with added NLP capabilities to improve its coding accuracy. The primary goal of this document is to present an analysis of the capabilities of BBN ACCENT. However, the inclusion of the baseline output from a second system (anonymized and shuffled together with BBN ACCENT's output) is helpful in ensuring a meaningful evaluation.

3.2 Evaluation Process

Both BBN ACCENT and Jabari NLP were run over the evaluation corpus in its entirety. For each top-level event code, 500 events were then randomly selected from each system's output[4]. These events were anonymized and shuffled together before being presented to trained evaluation staff for judgment. In the rare case that both systems produced exactly the same event and that event was randomly selected for evaluation for both systems, the event was only presented once to ensure consistency. (This is very rare simply because for most top-level event categories there are *far* more than 500 events produced by each system over the 5.5 million document corpus—sometimes hundreds of thousands—so the expected overlap between the 500 randomly selected for evaluation by each system is very small.)

Evaluators worked on one top-level event code at a time. So, for an example, an evaluator would evaluate all Investigate (09) events before moving on to all Demand (10) events. However, in the case of Assault (18) and Fight (19), the human confusion between the two categories was notable enough that we treated them as a single category ("Violence") for the purposes of the entire evaluation.[5]

The evaluation was performed by two members of BBN's contract annotation staff, who have no knowledge of BBN ACCENT, Jabari NLP, or any other NLP technology. Both were trained by a supervisor on the contract annotation staff using a set of 1900 doubly-annotated and adjudicated sample event

[3] A document was considered to be a near duplicate to another if more than 80% of its trigrams were covered by the other. If two documents each covered 80% of the other, the longer one was retained.

[4] The one exception, as noted above, was CAMEO Category 20 (Engage in Unconventional Mass Violence), where Jabari NLP only produced 336 events in the evaluation corpus.

[5] Note that we evaluated subcode confusion on the Violence category the same way we did every other category, so a system would still be penalized for confusing Assault and Fight events in that context.

instances drawn from BBN ACCENT and Jabari NLP events found in documents published in 2009. Evaluators were trained on each top-level event code until either they achieved at least 80% agreement with the adjudicated sample set, or until the supervisor was satisfied that they understood and were correctly applying the coding guidelines and that the majority of their differences with the sample set were simply judgment calls rather than errors.

All events were seen by at least one evaluator, and 10% of events were seen by both evaluators. Specifically, for each top-level event code, a single annotator would perform the annotation for all such events, and the other annotator would then repeat the process for 10% of those events. (Annotator #1 was the primary annotator for twelve top-level event categories; Annotator #2 was the primary annotator for the other seven.)

When the evaluators disagreed on whether an event was correct, system was given half-credit. An analysis of inter-annotator agreement is presented in the Results section below.

3.3 Event Correctness

Detailed guidelines on event correctness are provided in the ICEWS Expanded CAMEO Annotation Guidelines. We will summarize the primary points of importance here.

In our primary precision metric, an event is judged "correct" if it has

- A correct top-level event code
- A correctly coded Source actor
- A correctly coded Target actor

3.3.1 Event Code Correctness

An event code is considered correct (in the primary metric) if it is in the "right part" of the ontology. Specifically, it is considered correct if:

- It is in the same top-level category as the correct event type, or
- It is in a closely related top-level category to the correct event type

Events with this kind of event code confusion still seem "more right" than wrong, e.g. when the system produces *Fight with artillery and tanks* (194) rather than *Use conventional military force* (190).

We defined three groups of "closely related" top-level categories:

- Diplomatic Cooperation (05) / Material Cooperation (06) / Provide Aid (07)
- Disapprove (11) / Reject (12) / Reduce Relations (16)
- Assault (18) / Fight (19)

We allow for confusion among these particular top-level codes in our primary metric because humans also appeared to have trouble distinguishing between them, and because they also still seem to be "more right" than wrong.

Exact subcode correctness is obviously still important, however, so we present separate secondary results on that dimension.

3.3.2 Actor Correctness

All actors must be resolved to the canonical W-ICEWS actor database. Named actor resolutions are either correct or not (does “*the president*” refer to actor 8839 [Vladimir Putin] or not?). A resolution for a described actor (an “agent”) is considered correct if it is either exactly correct or correct but not quite specific enough. An example of the latter situation would be if a system produced the actor “*Iraq*” when it should have produced “*Military (Iraq)*”. (However, the confusion of, e.g., “*Business (Iraq)*” for “*Military (Iraq)*” is not allowed.)

We separately evaluate systems’ ability to code actors specifically enough; performance for both systems was similar: BBN ACCENT actors were judged specific enough ~92% of the time, compared to ~89% for Jabari NLP.

3.3.3 Event Modality Correctness

Negative, failed, denied, and hypothetical events are all considered incorrect. For example, none of the following sentences contain a correct event:

- North Korea might bomb South Korea
- The police did not arrest the man’s accomplice
- Protesters denied plans to hold an anti-US rally next week

Future events are considered incorrect unless the event category guidelines explicitly allow them, e.g. in code 03 (*Express intent to cooperate*). This decision was based on guidance from the original CAMEO codebook, which indicates that for, e.g., code 07 (Provide Aid), aid must be actually delivered (not just promised) in order to be coded.

Generic events—references to a category of event rather than an explicit instance thereof—are also not considered codable, e.g. “In an average year, the Iraqi police arrest thousands of civilians.”

3.3.4 Garbled Documents

The goal of the primary evaluation is to quantify system performance over native English text. However, a small portion of the documents in the data stream are the product of machine translation (from French, Spanish, or Portuguese), which sometimes does not read as fluent English text. For the purposes of this evaluation, we instructed annotators to skip events that appeared in sentences that were markedly non-fluent and mark them as “garbled”. About 1%-2% of events were skipped because of garbling; the distribution was roughly equivalent across the two systems.

Below is an example of a sentence marked as garbled:

> *The Chamber of Deputies paraguaia endorsed this farm-fair a law to declare state of exception in the departments of Concepcion and San Pedro (north) to qualify the Armed Forces to combat a supposed foci guerrilla which attach several murders of police.*

We present separate results in Section 0 that compare performance on machine translated text (the source of almost all garbling) to native English text.

3.4 Results

The overall precision of both systems is presented below:

Event Code	BBN ACCENT Precision	Jabari NLP Precision
01: Make Public Statement	71.1%	29.3%
02: Appeal	71.4%	42.8%
03: Express Intent To Cooperate	74.8%	56.8%
04: Consult	80.6%	54.8%
05: Diplomatic Cooperation	81.1%	52.0%
06: Material Cooperation	65.9%	36.4%
07: Provide Aid	73.9%	48.2%
08: Yield	62.0%	20.5%
09: Investigate	70.2%	40.7%
10: Demand	58.7%	38.1%
11: Disapprove	65.2%	47.8%
12: Reject	74.6%	44.2%
13: Threaten	66.0%	33.0%
14: Protest	84.5%	44.2%
15: Exhibit Force Posture	70.9%	25.2%
16: Reduce Relations	69.9%	44.3%
17: Coerce	88.1%	45.7%
18/19: Assault/Fight	73.8%	49.6%
20: Unconventional Mass Violence	83.6%	40.3%
ALL	75.6%	45.7%

Table 2: BBN ACCENT and Jabari NLP event coding precision

BBN ACCENT improved performance over the baseline by a statistically significant margin ($p < .01$) for all top-level codes. The precision for ALL is generated by averaging across all categories, weighting each category by the frequency with which it was coded by the system in the evaluation corpus. (An average across categories that did not reflect category frequency would produce overall accuracies of 73.0% and 41.8%.)

Given the raw counts of events generated by the system and the estimated precision for each event code, we can estimate the number of correct events found for each category. For instance, since 62% of the randomly sampled BBN ACCENT Yield events were judged to be correct and BBN ACCENT produced 31059 total Yield events in the corpus, we can estimate that it produced 18698 correct Yield events. We round the numbers to the nearest ten in the table below to emphasize that this is an estimate rather than an exact count:

Event Code	BBN ACCENT Precision	Est. Number of Correct Events
01: Make Public Statement	71.1%	435510
02: Appeal	71.4%	137280
03: Express Intent To Cooperate	74.8%	190160
04: Consult	80.6%	634070
05: Diplomatic Cooperation	81.1%	185450
06: Material Cooperation	65.9%	11480
07: Provide Aid	73.9%	20550
08: Yield	62.0%	18700
09: Investigate	70.2%	21650
10: Demand	58.7%	29140
11: Disapprove	65.2%	118270
12: Reject	74.6%	34350
13: Threaten	66.0%	18010
14: Protest	84.5%	35870
15: Exhibit Force Posture	70.9%	7450
16: Reduce Relations	69.9%	16920
17: Coerce	88.1%	168450
18/19: Assault/Fight	73.80%	125500
20: Unconventional Mass Violence	83.60%	650
ALL	75.60%	2209470

Table 3: BBN ACCENT event coding precision and estimated numbers of correct events

Certain events are obviously much more common than others (e.g. Statement, Consult), though of course frequency does not necessarily correlate with importance.

Evaluating system output for precision cannot supply any measure of recall or coverage. That is, it is impossible to know from judging system output what events were missed by the system. However, since two systems were included in the evaluation, we can still measure relative recall by comparing the estimated number of correct events found by each system:

Event Code	BBN ACCENT	Jabari NLP
01: Make Public Statement	435510	174050
02: Appeal	137280	112710
03: Express Intent To Cooperate	190160	154410
04: Consult	634070	528910
05: Diplomatic Cooperation	185450	146950
06: Material Cooperation	11480	15340
07: Provide Aid	20550	18960
08: Yield	18700	18380
09: Investigate	21650	14740

10: Demand	29140	23290
11: Disapprove	118270	115350
12: Reject	34350	28060
13: Threaten	18010	9370
14: Protest	35870	12080
15: Exhibit Force Posture	7450	3160
16: Reduce Relations	16920	16670
17: Coerce	168450	82580
18/19: Assault/Fight	125500	73230
20: Unconventional Mass Violence	650	140
TOTAL	2209470	1548380

Table 4: BBN ACCENT and Jabari NLP estimated numbers of correct events

As seen in the table above, BBN ACCENT increases the number of estimated correct events by ~40% over the baseline.

The only top-level event category where BBN ACCENT does not increase the number of estimated correct events over the baseline is code 06 (Material Cooperation). On examination of the data, it appears that a number of events were coded by Jabari NLP as code 06 and by BBN ACCENT as code 05 (Diplomatic Cooperation). An example of this is the event found in the following sentence: *ISI officers collaborated with Lashkar-e-Taiba.* These two codes are very similar and often either could be considered correct.

3.4.1 Inter-Annotator Agreement

10% of events (i.e. 100 per top-level event category) were judged by two evaluators. Their overall level of agreement was 79.4% on the primary task ("Is this event correct or not?"). When broken down by top-level event category, agreement ranged from 72% to 86%, with no particularly strong outliers:

Event Code	IAA
01: Make Public Statement	85%
02: Appeal	76%
03: Express Intent To Cooperate	79%
04: Consult	80%
05: Diplomatic Cooperation	80%
06: Material Cooperation	78%
07: Provide Aid	76%
08: Yield	74%
09: Investigate	78%
10: Demand	75%
11: Disapprove	79%
12: Reject	82%
13: Threaten	72%

14: Protest	88%
15: Exhibit Force Posture	75%
16: Reduce Relations	83%
17: Coerce	83%
18/19: Assault/Fight	79%
20: Unconventional Mass Violence	86%
ALL	79.4%

Table 5: Inter-Annotator Agreement on the primary evaluation task

The same results, visually:

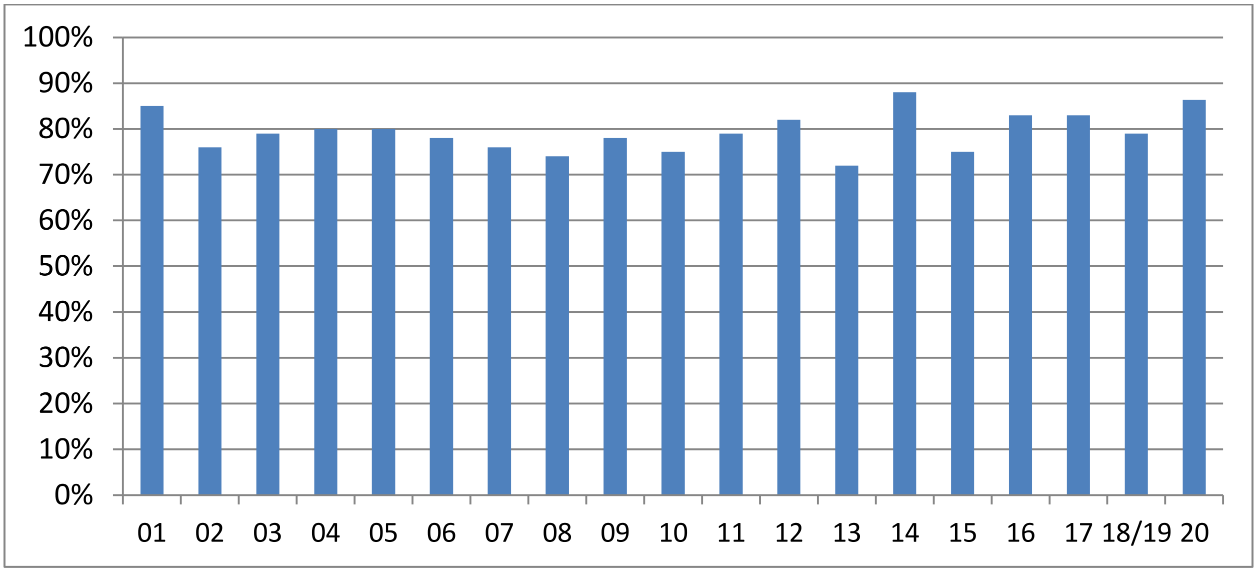

Figure 1: Inter-Annotator Agreement on the primary evaluation task

The lowest agreement was for event code 13 (Threaten), possibly because it can be hard to distinguish between a real “threat” and just a descriptive statement of a situation:

- “If you bomb us, we’ll bomb you back”… this is definitely a threat
- “If you bomb us, there will be political upheaval”… is this a threat? or just a statement of fact?

In an earlier event coding evaluation on the same ontology, we manually adjudicated a sample of evaluator disagreements and a third party judged that the event was correct approximately 43% of the time; this supports our decision to give half-credit in these cases.

3.4.2 Subcode Correctness

The primary metric requires only top-level event categories to be correct. However, event subcode correctness is also important. For instance, in the W-ICEWS context, the subcode determines the Goldstein score for each event, which represents the intensity of reported conflict or cooperation and is used by some of the predictive models. (Some examples of CAMEO events and their Goldstein scores, ranging from more negative/conflictual to more positive/cooperative: Assassinate (-10), Veto (-5), Make Statement (0), Make an Appeal (3), and Provide Aid (7)).

To measure this dimension of performance, whenever an evaluator marks an event as correct according to the primary metric, we also ask "Is the subcode also correct? If not, what is the correct subcode?".

This can be a somewhat difficult task, because some subcodes are easily confusable. The overall agreement (on subcode correctness given an otherwise correct event) is 74.3%, but there are significant outliers, specifically for codes 02 (Appeal) and 10 (Demand):

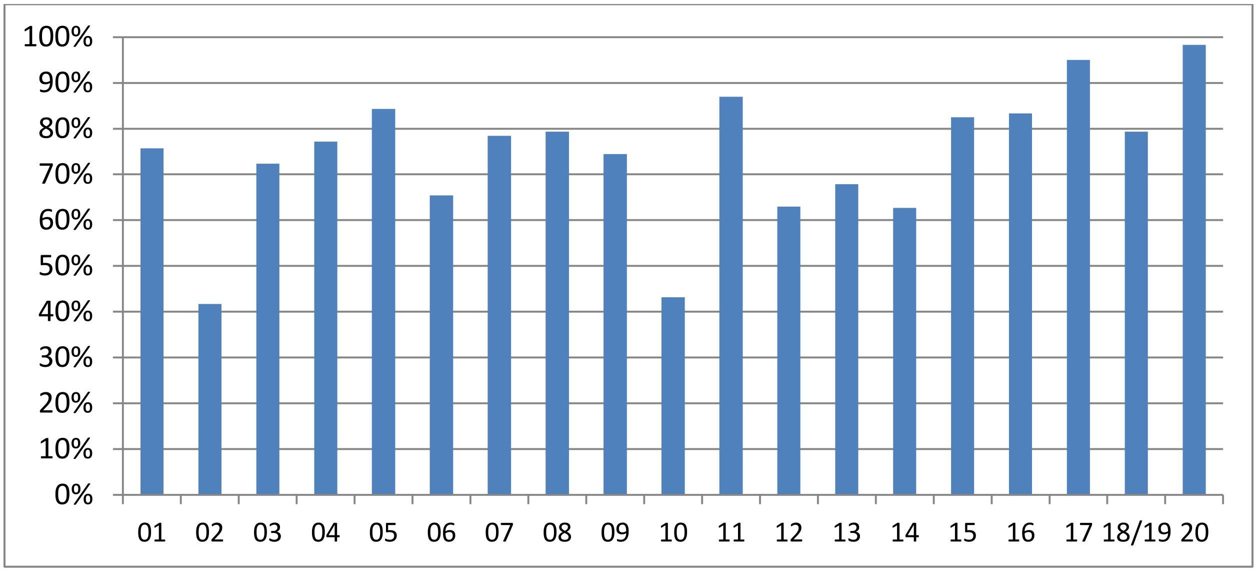

Figure 2: Inter-Annotator Agreement on the subcode correctness task

In the cases where agreement was very low, the most common scenario was that the system produced the most generic possible subcode (020 Appeal or 010 Demand) and only one of the two evaluators changed it to something more specific (e.g. 1042 Demand Policy Change).

Where inter-annotator agreement (IAA) on subcodes is very low, system subcode scores should be taken with a grain of salt. For example, BBN ACCENT's subcode was marked correct for only 36% of Demand events—but even humans only agreed with each other 43% of the time. (The human task is different than the system's—grading system output rather than generating it from scratch—so the system and human scores should not be considered directly quantitatively comparable, but the comparison is still qualitatively informative.) Fortunately, human agreement on subcodes is lowest for codes that share Goldstein scores—i.e. those that are, at least on that dimension, relatively similar. For instance, all Demand/10 events share the same Goldstein score (-5).

The following table presents subcode correctness scores for both systems, as well as the inter-annotator agreement for each code:

Event Code	**BBN ACCENT**	**Jabari NLP**	**Human (IAA)**
01: Make Public Statement	83.5%	68.2%	75.7%
02: Appeal	32.7%	25.7%	41.7%
03: Express Intent To Cooperate	77.6%	78.3%	72.3%

04: Consult	72.5%	68.1%	77.2%
05: Diplomatic Cooperation	95.5%	94.2%	84.3%
06: Material Cooperation	78.9%	52.0%	65.4%
07: Provide Aid	91.9%	88.2%	78.4%
08: Yield	88.4%	70.2%	79.3%
09: Investigate	71.9%	77.1%	74.5%
10: Demand	35.8%	32.4%	43.2%
11: Disapprove	72.6%	73.0%	87.0%
12: Reject	81.0%	70.9%	63.0%
13: Threaten	63.7%	75.4%	67.9%
14: Protest	83.0%	70.2%	62.7%
15: Exhibit Force Posture	91.0%	75.2%	82.5%
16: Reduce Relations	88.1%	85.3%	83.3%
17: Coerce	97.0%	98.4%	95.0%
18/19: Assault/Fight	86.5%	88.6%	79.3%
20: Unconventional Mass Violence	98.4%	87.9%	98.3%
AVERAGE	**78.4%**	**72.6%**	**74.3%**

Table 6: BBN ACCENT and Jabari NLP subcode correctness +
Inter-Annotator Agreement on the subcode correctness task

Scores for both systems are very low for codes 02 (Appeal) and 10 (Demand), but as noted above, these are also codes where humans disagree the majority of the time.

BBN ACCENT usually outperforms the previous solution, with statistically significant differences marked in red (p < 0.05). However, performance between both systems is not truly comparable, because the metric only measures subcode correctness given the otherwise-correct events produced by the system in each category—which are obviously different for each system. For instance, the baseline system statistically significantly outperforms BBN ACCENT on this measure for code 13 (Threaten). Looking at the data, it appears that BBN ACCENT identifies a large number of generic Threaten events (130) which should have really been coded more specifically as Threaten Military Force (138). However, the baseline system simply missed these events entirely, so although its subcode correctness is technically higher, that correctness only applies to half as many events.

The table below shows the ten most commonly confused subcodes for the BBN ACCENT system, along with the Goldstein scores for the system and correct event subcodes. (The abbreviation NSB stands for "not specified below", meaning that this is the generic catch-all for a top-level event category.)

System Subcode	**Correct Subcode**	**System Goldstein Score**	**Correct Goldstein Score**
020: Make appeal (NSB)	022: Appeal for diplomatic cooperation	3	3.4
090: Investigate (NSB)	091: Investigate crime	-2	-2
040: Consult (NSB)	046: Engage in negotiation	1	7

130: Threaten (NSB)	138: Threaten with military force	-4.4	-7
100: Demand (NSB)	105: Demand that target yields	-5	-5
010: Make statement (NSB)	013: Make optimistic comment	0	0.4
020: Make appeal (NSB)	025: Appeal to yield	3	-0.3
100: Demand (NSB)	101: Demand material cooperation	-5	-5
112: Accuse (NSB)	1123: Accuse of aggression	-2	-2
100: Demand (NSB)	102: Demand diplomatic cooperation	-5	-5

Table 7: Most frequently confused subcodes (BBN ACCENT)

The top 10 most frequent confusions all occur when the system returns a generic code that should have been more specific, e.g. “Demand” rather than “Demand that target yields”. It is much more rare for a system to confuse two distinct subcodes, e.g. 161 (Reduce diplomatic relations) vs. 164 (Halt negotiations). The table above also shows that Goldstein scores are often (though not always) similar when subcodes are frequently confused.

4 Supplemental Evaluation: Recall

As stated above, evaluating the precision of an event coder does not give any sense of recall: what events is the coder missing? To evaluate recall, one typically needs to create a “gold standard” of events coded by humans, against which the system output can be compared. In 2014, BBN was tasked with the creation of a small gold standard for this purpose. There are many challenges in creating a gold standard for such a large ontology, so the goal of this effort was not to achieve the same level of statistical significance as the precision evaluation but rather to create a smaller, initial gold standard that could provide a starting point for understanding this dimension of automatic event coding.

4.1 Data Selection

The W-ICEWS CAMEO gold standard developed under this effort consists of the following:

- 1,000 randomly-selected, doubly-annotated documents:
- 1,100 singly-annotated documents selected to target the eleven “less common” top-level event codes (100 such documents per code)

All documents were randomly selected from the most recent six months of the W-ICEWS corpus (January 2014 – June 2014; the gold standard effort was begun in July 2014). In all cases, we manually coded the first six sentences of each document, per standard W-ICEWS practice.

The documents in the gold standard were coded by humans for all event types except for the very common 010 (*Make Statement*), 012 (*Make Pessimistic Comment*), and 013 (*Make Optimistic Comment*) events. Because these event types are so common, coding them would have increased the investment required to create the gold standard beyond the scope of this effort.

The supplemental set of documents targeting less common event codes were annotated *only* for the top-level event code for which they were selected. Results on these documents are not included in our primary analysis but are discussed in Section 4.6.

All of the gold standard data (documents and events) has been kept blind from system developers with the intent that it be re-usable to evaluate future progress. (System scores over the gold standard, as seen below, have been released to system developers, but not the data itself.)

We also doubly-annotated an additional 1,000 randomly-selected English documents drawn from the same corpus, which we have not kept blind: these are not part of the formal gold standard and are intended for use in system development. Examples in this evaluation document are drawn from the open document set.

4.2 Annotation Process

All annotation (human coding) was performed using ENote, a BBN tool designed for single-document relation or event annotation. In additional to the extended CAMEO guidelines discussed above, we provided annotators with an additional set of gold standard guidelines, provided here as "ICEWS Gold Standard CAMEO Annotation Guidelines". In addition to identifying the Source and Target of CAMEO events, the coders were also asked to mark Location and Time arguments where appropriate; these are not evaluated here. Where possible, the coder was also asked to provide the best antecedent (or "resolution") for each Source or Target actor. That is, if the coder marked "*he*" as the Source for an event, the coder would also indicate that "*he*" referred to "*Joe Smith*".

In addition to event annotation, the participants in each event needed be mapped to the actors and agents in the W-ICEWS database. This process is not part of the ENote tool, since it requires direct access to the W-ICEWS database. To make this process tractable within the (small) scope of this effort, we provided a single annotator with all the actor/agent codes produced by BBN ACCENT for each event participant; the annotator provided any necessary corrections. Further details and analysis of this process and its effects on scoring (including a discussion of possible desirable or undesirable biases introduced by this process) are presented in Appendix B.

We only include in the gold standard the human-coded events where both actors were assigned an ICEWS actor code by the human coder, since these are the events targeted by the W-ICEWS program and correspondingly the machine coder. This excludes events like "*Putin met with others yesterday*", since the identity of *others* is unknown and is therefore not codable.

4.3 Scoring

Our primary scoring metric considers two events equivalent if they match event code, Source actor, Target actor, and sentence number.

An event code match is allowed when both events are part of the same top-level CAMEO event code. However, cross-code matches are allowed for the following sets of top-level codes which are often confused by humans as well as machines:

- 05/Diplomatic Cooperation + 06/Material Cooperation + 07/Provide Aid
- 11/Disapprove + 12/Reject + 16/Reduce Relations
- 18/Assault + 19/Fight

We allow Source and Target actors to match in three ways:

- By character offsets
    - Vladimir Putin criticized Bashar al-Assad.
        - Annotator #1: Source = “Vladimir Putin”
        - Annotator #2: Source = “Vladimir Putin”
- By character offsets for actor resolutions
    - Police announced today that they arrested Joe Smith.
        - Annotator #1: Source = “Police”
        - Annotator #2: Source = “they” (“Police”)
- By W-ICEWS actor dictionary code:
    - The delegation arrived at the Adan Adde airport in Mogadishu.
        - Annotator #1: Target = Adan Adde airport (CAMEO code = Somalia (40214))
        - Annotator #2: Target = Mogadishu (CAMEO code = Somalia (40214))
    - When matching actor/agent codes, we allow a match if one coder produces just a named actor and the other the same named actor plus an agent, e.g. *Iraq* vs. *Military (Iraq)*. However, we do not allow a match if both produce the same named actor but non-matching agents (e.g. *Rebel (Iraq)* vs. *Military (Iraq)*). Note that this does exclude some matches which are reasonably considered correct, e.g. *Vladimir Putin vs. Administration (Russia)* or *Military (Russia)* vs. *Military Personnel (Russia).*

The goal of allowing the second two types of matching is that there are sometimes multiple correct event participants in the same sentence, and we want to capture that (e.g. “Police” vs. “they” or “Adan Adde airport” vs. “Mogadishu”).

In our analysis we also present results with relaxed (or tightened) matching criteria, e.g. results for the task of finding an event of the correct type in a sentence, ignoring the arguments. Before computing mappings between coder A and coder B for any set of matching criteria, we first collapse all annotated events for each coder that meet the matching criteria. For instance, if coder A found two 173/Arrest events involving *Law Enforcement (Syria)* and *Protestor (Syria)* in the same sentence, this would be considered a single event for the purposes of matching with the primary set of criteria.

4.4 Inter-Annotator Agreement

Coding documents from scratch is much more difficult to do consistently than evaluating system output. (This has historically been shown to be true for a wide variety of natural language tasks.) There are two primary ways that inconsistency is greater when annotating from scratch rather than evaluating system output.

First, the mere act of explicitly presenting a single decision point to a judge will result in a more consistent decision on that decision point. In contrast, coding a whole document from scratch could be thought of as hundreds or thousands of different decision points. (For each sentence and for each of the ~300 types in the ontology, is there an event here? If an event exists, what are its arguments?) The human cannot realistically explicitly consider each decision point, and therefore oversights are far more common.

Second, events coded by automatic systems tend fall more squarely into the target category than those coded by humans, and they are thus easier to evaluate consistently. This tendency arises because the automatic coder is designed to extrapolate from the types of events it has seen in training; so, when it tags an event, it is probably similar to the kind of event that has been seen multiple times before, i.e. a construction that is at least somewhat common for this event type. In contrast, a human coder can identify events even when they are expressed in far less conventional, even "one-off", ways—but these are also more likely to be cases where even humans disagree. For example, annotator #1 marked the following Accuse event, which is certainly defensible in context but is less "conventional"; indeed, annotator #2 did not mark this event.

- *Welle: Who do you consider to be possible suspects? Schneider: I assume there's a connection with the self-styled "Emir of the Caucasus Emirate," Doku Umarov.*
  - Annotator #1: Accuse(I, Doku Umarov)

In the process of creating this gold standard, we trained both annotators by having them annotate the same set of documents and discuss their disagreements. However, there was clearly still some distance between the two of them, as evidenced by the fact that annotator #1 produced 2651 events and annotator #2 produced only 1871.[6] Of these, only 1026 events were found by both annotators. Scoring each annotator against the other results in a recall of 55% and 39%, respectively, for an average of 47%. (One could also compute the F-Measure of one annotator's response scored against the other; this would be 45%.)

Although low, this figure is not inconsistent with what we have seen in other event coding tasks. For instance, in the government-sponsored ACE evaluations in 2005, the Linguistic Data Consortium (which employs highly experienced professional human coders) annotated events for a much smaller ontology and reported an ACE score of 21.3 when comparing first-pass annotation to first-pass annotation and an ACE score of 31.5 when comparing one annotator to a dually-annotated corpus. ACE scores are *not* directly comparable to straight percentages (a perfect score is 100, but a negative score is possible), but it is still an informative point of reference. More recently, the TAC event argument extraction evaluation included a time-limited human coder as a participant in the evaluation. In this context, the human recall was still only ~25% compared to the total pool of correct event arguments extracted by the union of all participants. (The human's precision was ~80%.) Clearly the time limit is an important confounding factor here—humans had 30 minutes to annotate each newswire document (capped at 800 words), resulting in an average of 15 event arguments marked per document—but this also provides another point of reference.

[6] This difference is consistent across virtually all top-level codes, meaning there is likely simply a fundamental difference in the way one annotator is looking at the task, perhaps with regard to how explicit evidence for an event must be before it is tagged, or how far from the core definition of any event a coder can move.

For our task, the average inter-annotator agreement goes up from 47% to 59% if one ignores argument selection and just requires the two annotators to agree on the existence of an event of the same top-level type in a sentence. For instance, in the following sentence, both annotators marked a Demand event, but they marked different Source/Target participants.

- *Leaders of the Right Sector nationalistic radical group have demanded from the country's authorities to open arsenals for their self-defence forces.*
    - Annotator #1: Demand(Leaders, authorities)
    - Annotator #2: Demand(group, country)

Average inter-annotator agreement is 65% if one considers only the question of whether the two annotators agree that there is an event of a particular top-level type in the document at all.

Annotator disagreement is not spread evenly across the top-level codes. The table below shows the average of the two annotators' scores for each top-level code.

Event Code	Event Type + Args	Event Type only (sentence)	Event Type only (document)
01: Make Public Statement	9%	20%	29%
02: Appeal	48%	57%	61%
03: Express Intent To Cooperate	44%	50%	62%
04: Consult	49%	63%	70%
05: Diplomatic Cooperation	48%	54%	59%
06: Material Cooperation	40%	60%	69%
07: Provide Aid	46%	58%	65%
08: Yield	20%	15%	17%
09: Investigate	42%	55%	58%
10: Demand	40%	42%	50%
11: Disapprove	46%	59%	65%
12: Reject	45%	54%	67%
13: Threaten	35%	34%	41%
14: Protest	50%	68%	72%
15: Exhibit Force Posture	12%	35%	43%
16: Reduce Relations	50%	63%	69%
17: Coerce	58%	71%	78%
18: Assault	51%	79%	82%
19: Fight	48%	67%	77%
20: Unconventional Mass Violence	0%	0%	0%
AVERAGE	47%	59%	65%

Table 8: Average number of events found by one annotator, compared to the other

Future work in this area should certainly examine those categories where agreement is particularly low, e.g. 01 and 08.

4.5 Results

Given the level of disagreement among annotators, we report recall both in absolute terms and also as a percentage of human performance. Specifically, when using annotator #1 as the gold standard, we compare the performance of BBN ACCENT to that of annotator #2, and vice versa.

Using these metrics, BBN ACCENT's average relative recall is 34%. (Table Table 9 below gives the breakdown by subtype and shows both relative and absolute scores.)

Event Code	# Events Found by BBN ACCENT	Annotator #1		Annotator #2	
		Absolute	Relative	Absolute	Relative
01: Make Public Statement	21	13%	400%	43%	300%
02: Appeal	54	24%	55%	33%	65%
03: Express Intent To Cooperate	90	17%	53%	9%	16%
04: Consult	164	13%	34%	16%	28%
05: Diplomatic Cooperation	49	12%	30%	15%	26%
06: Material Cooperation	2	6%	27%	15%	26%
07: Provide Aid	8	7%	14%	6%	13%
08: Yield	3	0%	0%	8%	33%
09: Investigate	11	12%	29%	8%	19%
10: Demand	13	3%	10%	0%	0%
11: Disapprove	56	16%	54%	28%	45%
12: Reject	8	17%	45%	31%	57%
13: Threaten	4	0%	0%	0%	0%
14: Protest	8	11%	21%	12%	26%
15: Exhibit Force Posture	5	4%	50%	23%	150%
16: Reduce Relations	13	14%	30%	12%	23%
17: Coerce	46	22%	38%	19%	33%
18: Assault	35	20%	46%	18%	32%
19: Fight	39	18%	42%	17%	32%
20: Unconventional Mass Violence	0	0%	n/a	0%	n/a
AVERAGE	629	15%	38%	17%	30%

Table 9: BBN ACCENT recall with respect to human annotations

Note that this metric only requires BBN ACCENT to find the event and its arguments correctly in the text, not to code its actors correctly with respect to the actor dictionary. If we require correct actor/agent codes, BBN ACCENT's average relative recall is 24%.

A significant gap for BBN ACCENT indeed relates to the question of actor coding. BBN ACCENT is relatively conservative in its actor coding; if it cannot identify an actor or agent's correct code with a high-level of confidence, it will not code the actor, and therefore will not code any events involving that actor. If we score BBN ACCENT on only those events where it correctly selected actor/agent codes for both participants in an event, its average relative recall goes up to 49%, confirming that this is indeed a significant factor.

There are also, of course, many events found by BBN ACCENT that are not found by one or both annotators. If we attempt to make our scoring condition as analogous as possible to the conditions of the precision evaluation (so, we match actors solely based on actor/agent codes and require BBN ACCENT to find the correct such codes to get credit), annotator #1 finds only 43% of BBN ACCENT's events and annotator #2 finds 34%. If we assume that the overall accuracy of BBN ACCENT is ~75% as seen in the primary precision evaluation, one must conclude that each human annotator is missing a significant number of correct events found by BBN ACCENT. Looking at the union of the two annotators' responses, still only 51% of BBN ACCENT's events are found by a human; meaning that there are likely a number of correct events found only by BBN ACCENT and *neither* human annotator. Here are two of examples of what that can look like:

Example #1: *Israeli forces raided the house of Ahmad Younis Abu Ayyash*

- Annotator #1: Seize or damage property (171), Source: *forces*, Target: *house*
- Annotator #2: n/a
- BBN ACCENT: Use conventional military force (190), Source: *forces*, Target: *Ahmad Younis Abu Ayyash*

Here, both a human and BBN ACCENT essentially found the same event, but they coded it differently enough that there was no match in the scoring process. (Yet both are probably reasonable interpretations of the sentence.) As a result, it appears to the scorer that BBN ACCENT missed a human-coded event and Annotator #1 missed an automatically coded event—when in reality this is not the case. (It is the case, however, that Annotator #2 clearly made an error of oversight in not coding something in this sentence.) These kind of examples—which are relatively frequent—should be borne in mind when assessing the recall scores, since they deflate performance for both humans and machines.

Example #2: *Authorities say Rizza assaulted the inmate this past May*

- Annotator #1: n/a
- Annotator #2: n/a
- BBN ACCENT: Use unconventional violence (180), Source: *Rizza*, Target: *the inmate*

In this example, we see a clear case where BBN ACCENT found a valid event missed for some reason by both humans.

Still, it is clear from these results that the automatic coder is only finding a fraction of the events that are actually reported in the news. It is still extracting more than the previously deployed solution against which it was originally compared, but there is obviously significant room for possible improvement here.

For reference to the broader state of the art in event extraction, the most recent open TAC KBP event argument evaluation assessed performance of one time-limited human and dozens of automatic event coders targeting an ontology of 30 event types (grouped into seven high-level classes: business, conflict, life, personnel, movement, transaction, and justice events). Recall was assessed by comparing against a pool of correct results from all submissions. In this context, the highest recall achieved by an automatic coder on newswire data was just under 30% (in this evaluation, each event argument is scored independently). The highest observed precision was around 60%, though the submission with the ~30% recall had a precision lower than 50%. (These are all absolute numbers, since there is no non-time-limited human against which to compare.)

4.6 Targeted Annotation

Certain event types occurred very infrequently even in 1000 documents. To gather more information about system performance on these event types, we also coded a supplementary set of documents specifically selected to attempt to target certain low-frequency event types. To select documents for a particular target event type, we first generated keyword sets made up of words that disproportionately appeared in documents coded with the target event type by either Jabari NLP, BBN SERIF, or by human annotators. We then randomly selected 100 documents (per target event code) using those keywords. Please see Appendix A: Targeted Document Selection for the details on how these keywords and documents were selected.

Due to the role played by BBN ACCENT and Jabari NLP in the creation of the keyword lists, there is some amount of bias in the selection process. (The other obvious alternative would have been to manually generate keyword lists, which has its own sort of unconscious bias towards particular types of events.) Regardless, these document sets should be considered supplementary to the primary (completely random) gold standard. The goal here was simply to provide additional information about system coverage for these less common event codes; without some sort of pre-selection, these events are simply otherwise too sparse.

The following table shows the distribution of documents containing events for each top-level event type (marked by annotator #1) on the randomly-selected 1,000 documents:

CAMEO code	Event Count
01: Make Public Statement	25
02: Appeal	105
03: Express Intent To Cooperate	96
04: Consult	268
05: Diplomatic Cooperation	103
06: Material Cooperation	31
07: Provide Aid	33

08: Yield	16
09: Investigate	40
10: Demand	27
11: Disapprove	128
12: Reject	28
13: Threaten	15
14: Protest	27
15: Exhibit Force Posture	13
16: Reduce Relations	34
17: Coerce	117
18: Assault	44
19: Fight	75
20: Unconventional Mass Violence	2
TOTAL	25

Table 10: Number of events of each event type found in the 1,000-document randomly-selected gold standard

We considered any event code that occurred in fewer than 40 documents to be “less common”, namely:

- 01: Make Public Statement (010/012/013 excluded)
- 06: Material Cooperation
- 07: Provide Aid
- 08: Yield
- 10: Demand
- 12: Reject
- 13: Threaten
- 14: Protest
- 15: Exhibit Military Posture
- 16: Reduce Relations
- 20: Unconventional Mass Violence

In most cases the targeting did result in significantly more events of each type annotated, as seen below comparing counts of documents containing the target event types in the randomly-selected 1000-document set and in the targeted 100-document set.

Event Code	**Random selection**	**Targeted selection**
01: Make Public Statement	2.5%	39%
06: Material Cooperation	3.1%	16%
07: Provide Aid	3.3%	18%
08: Yield	1.6%	24%
10: Demand	2.7%	8%
12: Reject	2.8%	13%
13: Threaten	1.5%	20%

14: Protest	2.7%	29%
15: Exhibit Force Posture	1.3%	5%
16: Reduce Relations	3.4%	23%
20: Unconventional Mass Violence	0.2%	2%

Table 11: Percentage of annotated documents containing an event of this target type

The targeted documents were each only annotated for a single event code and were only annotated by a single annotator. We can therefore only report absolute recall for BBN ACCENT. We compare against recall with respect to annotator #1 on the random set, since annotator #1 performed the targeted annotation.

Event Code	**Random**		**Targeted**	
	# Annotated Events	**Recall**	**# Annotated Events**	**Recall**
01: Make Public Statement	32	13%	76	11%
06: Material Cooperation	102	6%	59	0%
07: Provide Aid	46	7%	25	8%
08: Yield	19	0%	26	8%
10: Demand	31	3%	8	13%
12: Reject	30	17%	23	4%
13: Threaten	21	0%	27	19%
14: Protest	35	11%	40	13%
15: Exhibit Force Posture	24	4%	6	0%
16: Reduce Relations	57	14%	22	27%
20: Unconventional Mass Violence	3	0%	4	0%

Table 12: Comparison of BBN ACCENT absolute recall on targeted and random sets

With such a small set, it is not clear that we can draw any statistically significant conclusions from this data; this data might be best used as an open set for system development in the future.

5 Supplemental Evaluation: Machine Translation

5.1 Overview

To compare BBN ACCENT performance on French and Spanish machine translation (part of the W-ICEWS data stream) to its performance on native English data, we relied on a corpus where human translations of French and Spanish news articles were available, specifically the 2008-2013 development sets from the ACL Workshop on Statistical Machine Translation. These sets contain a total of 606 news documents with parallel versions in English, French, and Spanish. (The original text may have been written in English, French, Spanish, German, Czech, Hungarian, or Italian; each document was then manually translated into the other languages involved in the workshop.) We confirmed that these documents were not contained in the training sets for the machine translation models used for W-ICEWS.

5.2 Approach

Our approach to this evaluation was as follows:

1. We ran the French and Spanish versions of each document through the machine translation engine used by W-ICEWS.
2. We ran BBN ACCENT over three English versions of each document: native English, machine translation of the Spanish, and machine translation of the French.
3. We manually assessed the correctness of all BBN ACCENT events produced over this data, excluding code 010 (generic public statements, which are very common and rarely used by the W-ICEWS program).
    a. Events were presented for judgment in the context of the native English document. This has two advantages. First, the human evaluator could not tell whether the event had been identified from the native English document or from a machine translation of the Spanish or French, thus avoiding any possible unconscious bias in evaluation. Second, this approach ensures that the event was actually judged correct or incorrect given the real meaning of the document, not the perhaps-incorrect meaning conveyed by the machine translation.
    b. Where the system produced the same event in more than one version of a document, this event was only assessed once, to guarantee consistency across versions. Within a single top-level event code, events were ordered by document and sentence when presented to the human evaluator. This ensured that the evaluator saw events extracted from the same sentence at the same time, ideally leading to greater consistency. For instance, BBN ACCENT might have found a Cooperate event between *Iran* and *Russia* in the native English text and between *Iran* and the *Russian government* in the machine translation of the French text. These two events are not exactly the same, so they are both presented to the evaluator, but they should probably be given the same judgment. Presenting them consecutively makes this much more likely than if the events were presented in random order.

5.3 Results

The following table shows the correct and total events for each top-level CAMEO category[7] for each condition:

	English			French MT			Spanish MT		
	Correct	Total	Accuracy	Correct	Total	Accuracy	Correct	Total	Accuracy
01	14	21	66.7%	1	4	25.0%	3	8	37.5%
02	19	24	79.2%	10	17	58.8%	8	8	100.0%
03	32	37	86.5%	18	32	56.3%	13	18	72.2%
04	56	72	77.8%	20	30	66.7%	14	28	50.0%
05	13	26	50.0%	10	18	55.6%	9	21	42.9%
06	2	3	66.7%	0	1	0.0%	2	4	50.0%

[7] There were no extracted instances of Unconventional Mass Violence (20) in this small test set.

07	2	3	66.7%	3	5	60.0%	1	3	33.3%
08	3	5	60.0%	0	2	0.0%	0	0	0.0%
09	2	5	40.0%	1	2	50.0%	1	1	100.0%
10	4	8	50.0%	2	4	50.0%	3	3	100.0%
11	22	29	75.9%	13	22	59.1%	10	21	47.6%
12	11	15	73.3%	3	5	60.0%	4	4	100.0%
13	7	7	100.0%	2	3	66.7%	3	3	100.0%
14	11	13	84.6%	11	12	91.7%	12	13	92.3%
15	1	1	100.0%	1	1	100.0%	1	1	100.0%
16	0	1	0.0%	0	1	0.0%	1	1	100.0%
17	16	21	76.2%	9	10	90.0%	14	17	82.4%
18/19	16	20	80.0%	12	19	63.2%	5	10	50.0%
ALL	**231**	**311**	**74.3%**	**116**	**188**	**61.7%**	**104**	**164**	**63.4%**

BBN ACCENT performance on the native English documents in this test set is consistent with observed performance on the W-ICEWS data stream. The precision on this data set is 74.3%; the analogous precision observed on W-ICEWS data stream in the primary evaluation was 75.6%. In the table above, the precision on French machine translation is slightly lower than on Spanish machine translation, but this difference is not statistically significant.

As expected, precision is somewhat lower on machine-translated text (though still above 60% in both cases). However, the degradation is more noticeable in the number of correct events found, which drops by a factor of two. It appears that the noise introduced by machine translation results in relatively few additional false alarms, but many more misses. This is consistent with our experience using machine translation for other information extraction tasks: it is common for the system to fail to find something due to the machine translation's dropping of an important verb or name, or its mangling of English syntax, but more rare for translation errors to cause it to imagine something that isn't really there.

6 Contact Information

If you have any questions about the evaluation process or about BBN ACCENT, please feel free to contact Elizabeth Boschee (eboschee@bbn.com) at Raytheon BBN Technologies.

Appendix A: Targeted Document Selection

To generate the "targeted" document sets for low-frequency event codes, we used keyword sets made up of words that disproportionately appeared in documents coded with the target event type by either Jabari NLP, BBN SERIF, or by human annotators.

Specifically, keywords were harvested from two document sets:

- 50,000 documents randomly selected from the 2008 W-ICEWS corpus and coded with events by BBN SERIF and Jabari NLP.
- The 2,000 documents randomly selected from 2014 and coded with events by a human annotator. (This includes both the 1,000 documents in the gold standard and the 1,000 documents coded for eventual use in system development.)

We stripped punctuation and stopwords (using the Natural Language Toolkit's default list) and then stemmed using the Natural Language Toolkit's SnowballStemmer[8]. As a simple method for minimizing the inclusion of proper names, we only included lowercase words.

For each coder C, each event code E, and each word W, we calculated the probability that coder C will find an event of type E in a document that contains word W:

$$P_C(E|W) = \frac{\#\ of\ documents\ with\ word\ W\ where\ coder\ C\ finds\ event\ type\ E}{\#\ of\ documents\ with\ word\ W}$$

(For E, we consider both full event codes, e.g. 1721, as well as top-level event codes, e.g. 17.)

To be able to normalize across event codes, we also calculated $P_C(E)$ as the probability that coder C will find an event of type E in any given document:

$$P_C(E) = \frac{\#\ of\ documents\ where\ coder\ C\ finds\ event\ type\ E}{\#\ of\ documents}$$

We then considered the "score" of word W in the context of coder C and event code E to be:

$$Score_{CEW} = \frac{P_C(E|W)}{P_C(E)}$$

For a given coder C and event code E, we discarded all words whose scores were less than 2.0, i.e. those where the existence of word W does not make it at least twice as likely that this document contains event code E. We also discarded all words that did not occur at least N times in documents where coder C found event code E. (For SERIF and Jabari, N=10; for the human coder, N=5, since the document set is so much smaller.)

We harvested 25 keywords for each top-level event type, by doing the following:

[8] The Natural Language Toolkit can be found at www.nltk.org.

- Select the highest-scoring word for each coder for each event subcode and for the top-level code. (For each top-level code we considered, there were less than 25 such unique words.)
- Select the next highest-scoring words from the top-level code until we have reached 25 words.

As an example, here are the words selected for code 14 (Protest):

- Best for each subcode and/or for the top-level code:
  - protest (Jabari/141, Human/1411, Human/1412, Human/144, Human/145)
  - chant (SERIF/141, Jabari/14)
  - block (SERIF/144, Human/144)
  - dispers (SERIF/145)
  - hunger (SERIF/142)
  - boycot (SERIF/143)
  - loot (Jabari/145)
  - antigovern (Human/14)
  - slogan (Human/141)
- Remaining words:
  - protestor (SERIF)
  - demonstr (Human, SERIF, Jabari)
  - march (SERIF, Jabari)
  - hurl (SERIF)
  - monk (SERIF)
  - ralli (Human, SERIF, Jabari)
  - riot (SERIF, Jabari)
  - tear (SERIF)
  - banner (SERIF)
  - shout (SERIF)
  - crackdown (Jabari)
  - flag (Jabari)
  - activist (Jabari)
  - stone (SERIF)
  - boycott (SERIF)
  - angri (SERIF)

In this list, we note that a word was from a particular coder if it was observed by the selection process with that coder before the 25-word cutoff was hit. For instance, all three coders ranked "*demonstr*" highly enough that it was seen while selecting the top 25 keywords. On the other side, only Jabari ranked "*activist*" within the top 25, so it is shown as only (Jabari); however, SERIF also ranked it within the top 50.

Somewhat fewer of the keywords come from the human annotation when compared to the Jabari and SERIF sets; this is because there are relatively few documents annotated by humans and therefore it is more difficult to observe any statistically significant patterns.

For the very rare event code 20 (Mass Violence), there were no words that met any of our criteria. We selected words manually from the Jabari NLP rules that fired on the entire W-ICEWS corpus[9] and the descriptions of the events in the CAMEO ontology. These words are: (*drive, empti, fled, massacr, expuls, expel, ethnic, cleans, mass, destruct, nuclear, chemic, biolog*).

To select documents using a keyword set, we did the following:

- Identify all documents from the corpus that were not already randomly selected for annotation. Exclude those which are not native English.
- For each code:
    - Randomly sort these documents.
    - Create a "target keyword list" by repeatedly iterating through its keywords and adding them to the target list until the target list contains 100 words. (For all codes but 20, this target list will contain exactly four copies of each keyword, since there are 25 keywords for each code.)
    - Iterate through the documents. Whenever one matches one or more keywords on the target list, select the document for annotation. Randomly choose one of the matching keywords and remove that word from the target list. Stop when all the words have been removed from the target list. (This produces 100 documents for annotation.)

The actual keyword lists used for selection were as follows:

- 01: Make Public Statement (010/012/013 excluded)
    - consid, plead, reject, courtesi, condol, agre, comment, reconsid, acknowledg, refut, wreath, deni, mark, sympathi, mourn, tribut, greet, funer, convey, sincer, 60th, gun, condemn, anniversari, commemor
- 06: Material Cooperation
    - export, invest, polic, troop, provid, assist, sell, militari, cooper, suspect, accus, sold, appoint, justic, intellig, relev, weapon, electr, resolv, final, sanction, partner, bilater, cabinet, mutual

[9] DRIVE - + * OUT_ OF_ ^ BY_ $
EMPT - $ * ^ OF_ +
FLED - THOUSANDS * OFFENSIVE AGAINST + IN_ $
MASSACRE - * CIVILIAN
MASSACRE - * IN_ RAID
MASSACRE - + * BY_ $
MASSACRE - + * IN_ $
MASSACRE - GUNMEN * + IN_ $
CLEAN - + ETHNIC * BY_ $
CLEAN - ETHNIC *

- 07: Provide Aid
  - airlift, peacekeep, countri, donat, dollar, caption, send, rescu, treat, million, cyclon, hostag, captiv, food, humanitarian, relief, medicin, refuge, aid, survivor, guerrilla, accid, ton, businessmen, assist
- 08: Yield
  - freed, surrend, peacekeep, acquit, resign, withdrawn, ceasefir, pardon, send, releas, forc, overturn, captiv, hostag, guerrilla, swap, withdrew, ransom, jail, abduct, kidnap, detaine, withdraw, gestur, fled
- 10: Demand
  - told, order, truce, resign, declar, investig, decre, unilater, demand, withdraw, ethnic, remov, highlight, enrich, instead, breakaway, warrant, genocid, urg, thorough, recognis, sovereignti, fall, pressur, ceasefir
- 12: Reject
  - reject, allow, defi, court, met, fail, veto, attend, rule, upheld, defect, dismiss, refus, oppos, spokesperson, unconstitut, request, deni, interfer, ground, unilater, treati, plea, reason, elabor
- 13: Threathen
  - threaten, given, plot, warn, fight, caution, vow, threat, sanction, stand, blow, pull, premier, prepar, attempt, intellig, allianc, action, milit, suicid, rebel, terrorist, clash, withdraw, attack
- 14: Protest
  - dispers, block, hunger, boycot, chant, loot, protest, antigovern, slogan, protestor, demonstr, hurl, march, monk, riot, tear, banner, ralli, shout, crackdown, flag, activist, stone, boycott, angri
- 15: Exhibit Military Posture
  - deploy, secur, troop, alert, battalion, beef, personnel, missil, peacekeep, patrol, tension, milit, border, combat, send, mission, clash, soldier, withdraw, amid, rebel, militari, unit, sent, forc
- 16: Reduce Relations
  - talk, expel, suppli, sanction, suspend, oust, embassi, resign, would, uranium, freez, boycott, coup, cancel, impos, appoint, diplomat, recognit, dismiss, recal, postpon, evacu, recogn, eas, disput
- 20: Unconventional Mass Violence
  - drive, empti, fled, massacr, expuls, expel, ethnic, cleans, mass, destruct, nuclear, chemic, biology

Appendix B: Actor Coding Analysis

As described in section 4.2, we generated actor and agent codes for the gold standard by having an annotator manually correct the codes provided by BBN ACCENT for each event participant (in events generated by either the system or by humans).

Over the set of 1,000 randomly selected documents, the actor annotator's corrective actions were as follows:

- For uncoded actors:
    - 158 were correct
    - 242 were changed to an actor
    - 911 were changed to an agent+actor
- For named actors:
    - 1810 were correct
    - 106 had an agent added, e.g. *Iraq* should have been *Rebel(Iraq)*
    - 46 were changed to a different named actor
    - 47 were changed to a different named actor *and* had an agent added
    - 20 were removed (shouldn't have been coded)
- For agent+actors:
    - 1162 were correct
    - 101 had their actor changed
    - 314 had their agent changed
    - 99 had both actor and agent changed
    - 12 had their agent removed (should have just been the named actor)
    - 24 were removed (shouldn't have been coded at all)

(The numbers above are derived from the set of 1,000 randomly selected documents, but the same general patterns held roughly across the board for the much smaller MT and targeted sets.)

As we can see, the primary source of error for BBN ACCENT comes in the actors it does not code. This is likely because BBN ACCENT takes a fairly conservative approach to actor coding, particularly when dealing with coreference. For instance, if it is not fairly sure which person "*he*" refers to, it will choose not to code that actor at all.

We had a second annotator perform this actor coding task for 100 randomly selected event participants, and the two annotators agreed 77% of the time. Almost all of the disagreement was over the agent categories rather than the named actors. There were four cases of annotator disagreement on named actors. Two were annotator oversights with a clear correct answer and two were judgment calls as to how a particular actor should be tagged (e.g. is it OK to tag X's brigade as X?). Of the remaining 19 cases of disagreement, all involved agent labels and all were cases where BBN ACCENT produced no agent label or produced a generic default label like Citizen or Man; in other words, an informative agent label had to be selected manually from the dictionary. With 700+ possible labels and a great deal of overlap, this was often a judgment call, e.g.:

- Domestic Affairs vs. Ministry
- Municipal Court vs. Public Courts
- Armed Gang vs. Armed Rebel
- Military vs. Army National Guard
- Military vs. Military Personnel
- Regional Governor vs. Governor

For completeness, the breakdown of the 77 cases where the two annotators agreed was as follows:

- No change to BBN ACCENT coding: 53
    - Named actor: 34
    - Actor+agent: 16
    - Not codable: 3
- BBN ACCENT produced spurious or incomplete result; annotators corrected it: 8
    - Change in named actor: 2
    - Change just in agent: 6
- BBN ACCENT produced no actor or agent; annotators added one or both: 16
    - Added named actor: 7
    - Added actor+agent: 9

We expect there is bias in using BBN ACCENT's actor coding as the starting place for this correction. Given that named actor coding is a relatively clear-cut process (the pronoun "*he*" either does or does not refer to Vladimir Putin; rarely is this terribly ambiguous in newswire documents), we expect the effect there is relatively minimal. We would only see bias in the cases where the system makes an error that the annotator overlooks, e.g. the system tags "*he*" as Vladimir Putin and the annotator does not read carefully enough to notice that in the sentence, "*he*" actually means Bashar al-Assad. These errors will occur but relatively infrequently. (Concretely, in the 100 cases annotated by both annotators, it appears that each annotator made such an error once.)

On the other hand, it appears that the task of agent coding is much more susceptible to bias. If the system codes an agent as "Governor", it is far less likely that the annotator will change that tag to "Regional Governor", whereas if the annotator had been generating codes from scratch, either might have been equally likely. Indeed, in the 100 cases annotated by both annotators, we see this happen in cases where BBN ACCENT failed to produce a code for an actor. For instance, in one such case, one annotator chose "Governor" while the other chose "Regional Governor". In this case, we would expect to see a noticeable bias in favor of the agent label chosen by the system.

Fortunately, for our *current* purposes, this bias is not entirely disadvantageous, because our scoring metric requires the system's agent label to exactly match the one proposed by the human but we have no way to say that two agent labels are close enough to be considered equivalent. It is actually therefore preferable that the gold standard include the label the system produced, if it is indeed one of several possible correct labels. It would be unfortunate if the system were penalized for producing "Regional Governor" rather than "Governor" if both are essentially equally valid.

Where this bias would be very problematic is if a second system were scored on the same gold standard. Indeed, in order for this gold standard to be applied to multiple systems, we would very strongly recommend that the actor and agent codes would need to be generated from scratch. We would also recommend that some more thought be put into the “agent correctness” metric to account for the fact that many labels are difficult to distinguish from one another and in some cases many can be plausibly correct for a particular event participant.