Domain-Independent Quality Measures for Crowd Truth

Oana Inel, Lora Aroyo, Chris Welty and Robert-Jan Sips
Annotator disagreement is signal, not noise. It is indicative of the variation in human semantic interpretation of signs, and can indicate ambiguity, vagueness, over-generality, etc.

This paper: cross domain adaptation of metrics & spam detection
Background:
Crowd-Watson Framework for Medical Relation Extraction

Crowd-Watson Adaptation to Newspapers Event Extraction

- newspapers corpus
- identify events & role fillers (e.g. type, location, time, participants)
- understand the range of disagreements by creating a space of possibilities with frequencies & similarities
- Crowd-Watson framework
Greenpeace Protests Apple’s Energy Practices By Releasing Balloons

By C.J. HUGHES

The skies over the glass cube that serves as the entrance to the Apple store on Fifth Avenue turned a bit dark on Tuesday afternoon, at least for some shoppers inside. And a passing storm had nothing to do with it.

The sudden shadows were caused by bunches of balloons released by Greenpeace, the environmental advocacy group, in order to call attention to Apple’s use of coal power at its facilities.

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators entered the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

As they neared the bottom of the cube’s stairs, they released the balloons, which hit the ceiling of the cube and lodged themselves in place. Other demonstrators came later with paper shopping bags filled with balloons, which they let trickle out as they entered the store, before being escorted out.

The stunt, which did not lead to any arrests, according to Greenpeace, was timed to coincide with last week’s release of a report that ranked the eco-friendliness of top technology companies. Apple’s “clean energy index” of 15 percent was below that of Google, with 39 percent, and Dell, with 56 percent.

That relatively low score, according to Greenpeace, is partly because much of the electricity Apple uses at a new data center in Maiden, N.C., where the company’s iCloud storage is powered, comes from burning coal, a pollutant.
Greenpeace Protests Apple’s Energy Practices By Releasing Balloons

By C.J. HUGHES

The skies over the glass cube that serves as the entrance to the Apple store on Fifth Avenue turned a bit dark on Tuesday afternoon, at least for some shoppers inside. And a passing storm had nothing to do with it.

The sudden shadows were caused by bunches of balloons released by Greenpeace, the environmental advocacy group, in order to call attention to Apple’s use of coal power at its facilities.

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators entered the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

As they neared the bottom of the cube’s stairs, they released the balloons, which hit the ceiling of the cube and lodged themselves in place. Other demonstrators came later with paper shopping bags filled with balloons, which they let trickle out as they entered the store, before being escorted out.

The stunt, which did not lead to any arrests, according to Greenpeace, was timed to coincide with last week’s release of a report that ranked the eco-friendliness of top technology companies. Apple’s “clean energy index” of 15 percent was below that of Google, with 39 percent, and Dell, with 56 percent.

That relatively low score, according to Greenpeace, is partly because much of the electricity Apple uses at a new data center in Maiden, N.C., where the company’s iCloud storage is powered, comes from burning coal, a pollutant.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.
The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.

The police came to Apple’s glass cube on Fifth Avenue on Tuesday to enforce order after activists released black balloons inside the cube to protest the company’s environmental policies.
Role-Filler Taxonomies

- **event type:** Semafor
- **event location type:** GeoNames
- **event time type:** Allen’s time theory, KSL time ontology
- **event participant type:** based on proper nouns classes

<table>
<thead>
<tr>
<th>Role Filler</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Type</td>
<td>Purpose, Arriving or Departing, Motion, Communication, Usage, Judgment, Leadership, Success or Failure, Sending or Receiving, Action, Attack, Political, Other.</td>
</tr>
<tr>
<td>Location Type</td>
<td>Geographical - Continent, Country, Region, City, State, Area on Land - Valley, Island, Mountain, Beach, Forest, Park, Area on Water - Ocean, River, Lake, Sea, Road/Railroad - Road, Street, Railroad, Tunnel, Building - Educational, Government, Residence, Commercial, Industrial, Military, Religious, Other</td>
</tr>
<tr>
<td>Time Type</td>
<td>Before, During, After, Repetitive, Timestamp, Date, Century, Year, Week, Day, Part of Day, Other</td>
</tr>
<tr>
<td>Participants Type</td>
<td>Person, Organization, Geographical Region, Nation, Object, Other</td>
</tr>
</tbody>
</table>
How do we represent & measure disagreement in a way that it can be harnessed?
Events semantics are hard
Events have multiple dimensions
Events have multiple dimensions

In the sentence

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

does [ENTERED] refer to an EVENT or an ACTION?

1. Choose one option:
   - [ENTERED] refers to an EVENT in this sentence
   - [ENTERED] refers to an ACTION in this sentence
   - None of the above

2. Motivate your answer by explaining why do you think it refers to an event or an action.

Do not copy & paste words from the text above as an explanation. Answers such as, 'It is an event', or 'It is an action', etc. are not acceptable for
Events have multiple dimensions

3. Select the type of the event

- [PURPOSE]
- [ARRIVING_OR_DEPARTING]
- [MOTION]
- [COMMUNICATION]
- [USAGE]
- [JUDGMENT]
- [LEADERSHIP]
- [SUCCESS_OR_FAILURE]
- [SENDING_OR_RECEIVING]
- [POLITICAL]
- [ACTION]
- [ATTACK]
- [OTHER]

Selected type(s)

[PURPOSE] [ARRIVING_OR_DEPARTING]

Here are the types that you selected.
Events have multiple dimensions

In the sentence

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

does [ENTERED] refer to an EVENT or an ACTION?

1. Choose one option:
   - [ENTERED] refers to an EVENT (or ACTION) in this sentence
   - [ENTERED] does not refer to an EVENT (or ACTION) in this sentence

2. Is there a LOCATION mention in the sentence for this EVENT (or ACTION)?
   - Yes, there is a location reference in the sentence for this event/action
   - No, there is no location reference in the sentence for this event/action

3. Select the words referring to LOCATION in the sentence
   - the cube

To HIGHLIGHT a LOCATION expression double-click each separate word in the text. To de-select word(s) single-click on already highlighted word.
Each dimension has different granularity
Each dimension has different granularity

In the sentence

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

does [ENTERED] refer to an EVENT or an ACTION?

1. Choose one option:
   - [ENTERED] refers to an EVENT (or ACTION) in this sentence
   - [ENTERED] does not refer to an EVENT (or ACTION) in this sentence

2. Is there TIME mention in the sentence for this EVENT (or ACTION)?
   - Yes, there is a time reference in the sentence for this event/action
   - No, there is no time reference in the sentence for this event/action

3. Select the words referring to TIME in the sentence

2:30 p.m.

To HIGHLIGHT a TIME expression double-click each separate word in the text. To de-select word(s) single-click on already highlighted word.
Each dimension has different granularity

In the sentence

*Around 2:30 p.m.*, as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

does [ENTERED] refer to an EVENT or an ACTION?

1. Choose one option:
   - [ENTERED] refers to an EVENT (or ACTION) in this sentence
   - [ENTERED] does not refer to an EVENT (or ACTION) in this sentence

2. Is there TIME mention in the sentence for this EVENT (or ACTION)?
   - Yes, there is a time reference in the sentence for this event/action
   - No, there is no time reference in the sentence for this event/action

3. Select the words referring to TIME in the sentence

   Around 2:30 p.m.

To HIGHLIGHT a TIME expression double-click each separate word in the text. To de-select word(s) single-click on already highlighted word.
People have different points of views
People have different points of views

In the sentence:

Around 2:30 p.m., as if delivering birthday greetings, several **Greenpeace demonstrators** [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

1. **Choose one option:**
   - [ENTERED] refers to an EVENT (or an ACTION) in this sentence
   - [ENTERED] does not refer to an EVENT (or an ACTION) in this sentence

2. **Is there PARTICIPANT mentioned in the sentence for this EVENT/ACTION?**
   - Yes, there is a participant mentioned in the sentence for this event/action
   - No, there is no participant mentioned in the sentence for this event/action

3. **Select (highlight) the words in the sentence that refer to the PARTICIPANT.**
   - **Greenpeace demonstrators**
   - To SELECT a PARTICIPANT double-click each separate word in the text. To de-select word(s) single-click on already highlighted word.

4. **Select the TYPE of the PARTICIPANT**
   - [NATION]
   - [GEOGRAPHICAL_REGION]
   - [PERSON (or PEOPLE)]
   - [ORGANIZATION]
   - [OBJECT]
   - [OTHER]
People have different points of views

In the sentence

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

does [ENTERED] refer to an EVENT or an ACTION in this sentence?

1. Choose one option:
   - [ENTERED] refers to an EVENT (or an ACTION) in this sentence
   - [ENTERED] does not refer to an EVENT (or an ACTION) in this sentence

2. Is there PARTICIPANT mentioned in the sentence for this EVENT/ACTION?
   - Yes, there is a participant mentioned in the sentence for this event/action
   - No, there is no participant mentioned in the sentence for this event/action

3. Select (highlight) the words in the sentence that refer to the PARTICIPANT.
   Greenpeace demonstrators

To SELECT a PARTICIPANT double-click each separate word in the text. To de-select word(s) single-click on already highlighted word.

4. Select the TYPE of the PARTICIPANT
   - [NATION]
   - [GEOGRAPHICAL_REGION]
   - [PERSON (or PEOPLE)]
   - [ORGANIZATION]
   - [OBJECT]
   - [OTHER]
Why do people disagree?

Diagram:
- Sign
- Referent
- Observer
Why do people disagree?

Sentence

Ontology

Worker
Disagreement Analytics

- **sentence metrics**: sentence clarity
- **ontology metrics**: (future work)
- **worker content-based metrics**:
  - worker-sentence disagreement
  - worker-worker disagreement
  - avg number of annotations per sentence
- **worker explanation-based metrics**:
  - valid words in explanation text
  - same explanation across contributions
  - “[OTHER]” + different type

AAAI2013 Fall Symposium on Semantics for Big Data (in print) (2013)
Disagreement Analytics

- **sentence metrics**: sentence clarity
- **ontology metrics**: (future work)
- **worker content-based metrics**:
  - worker-sentence disagreement
  - worker-worker disagreement
  - avg number of annotations per sentence
- **worker explanation-based metrics**:
  - valid words in explanation text
  - same explanation across contributions
  - “[OTHER]” + different type

Adapted from Medical Relation Extraction are those applicable also for event extraction.

AAAI2013 Fall Symposium on Semantics for Big Data (in print) (2013)
Experimental Setting

• 2 batches of 35 putative events: total of 70 putative events

• 8 experiments: 2 for each event role filler

• annotations: 15 per each putative event

• maximum annotations per worker: 10

• workers: native English speakers on CF
Annotation Example

Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.

Overall annotation & granularity distribution:
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators entered the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
Around 2:30 p.m., as if delivering birthday greetings, several Greenpeace demonstrators [ENTERED] the cube clutching helium-filled balloons, which were the shape and color of charcoal briquettes.
The high disagreement for event type **across all sentences** likely indicates problems with the ontology. These event types are difficult to distinguish between. The event classes may overlap, be confusable, too vague, etc.
The high disagreement for event type across all sentences likely indicates problems with the ontology. These event types are difficult to distinguish between. The event classes may overlap, be confusable, too vague, etc.
Comparative Annotation Distribution

Location Type Distribution

Participant Type Distribution
Comparative Annotation Distribution

Location Type Distribution

Participant Type Distribution
# Sentence Clarity

Identifies sentences that are unclear or ambiguous based on the distribution of types.

<table>
<thead>
<tr>
<th>unit_id</th>
<th>P</th>
<th>GR</th>
<th>N</th>
<th>OBJ</th>
<th>ORG</th>
<th>OTH</th>
<th>NA</th>
<th>sentClarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>291103360</td>
<td>0.57</td>
<td>0.00</td>
<td>0.08</td>
<td>0.08</td>
<td>0.81</td>
<td>0</td>
<td>0.08</td>
<td>0.81</td>
</tr>
<tr>
<td>291103361</td>
<td>0.87</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.47</td>
<td>0</td>
<td>0.16</td>
<td>0.87</td>
</tr>
<tr>
<td>291103362</td>
<td>0.49</td>
<td>0.12</td>
<td>0.08</td>
<td>0.37</td>
<td>0</td>
<td>0.6</td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>291103363</td>
<td>0.19</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>291103364</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0</td>
<td>0.16</td>
<td>0.98</td>
</tr>
<tr>
<td>291103365</td>
<td>0</td>
<td>0.11</td>
<td>0.64</td>
<td>0.01</td>
<td>0.75</td>
<td>0</td>
<td>0.75</td>
<td>0.98</td>
</tr>
<tr>
<td>291103366</td>
<td>0.55</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.82</td>
<td>0</td>
<td>0.18</td>
<td>0.82</td>
</tr>
<tr>
<td>291103367</td>
<td>0.68</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.68</td>
<td>0</td>
<td>0.26</td>
<td>0.68</td>
</tr>
<tr>
<td>291103368</td>
<td>0.83</td>
<td>0.31</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.41</td>
<td>0.41</td>
<td>0.83</td>
</tr>
<tr>
<td>291103369</td>
<td>0.87</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.1</td>
<td>0.48</td>
<td>0.48</td>
<td>0.87</td>
</tr>
<tr>
<td>291103370</td>
<td>0.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.3</td>
<td>0</td>
<td>0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>291103371</td>
<td>0.91</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.18</td>
<td>0</td>
<td>0.37</td>
<td>0.91</td>
</tr>
<tr>
<td>291103372</td>
<td>0</td>
<td>0.11</td>
<td>0.68</td>
<td>0.2</td>
<td>0.68</td>
<td>0</td>
<td>0.68</td>
<td>0.91</td>
</tr>
<tr>
<td>291103373</td>
<td>0.12</td>
<td>0.61</td>
<td>0.61</td>
<td>0.97</td>
<td>0</td>
<td>0.49</td>
<td>0.49</td>
<td>0.97</td>
</tr>
<tr>
<td>291103374</td>
<td>0</td>
<td>0.24</td>
<td>0.87</td>
<td>0.08</td>
<td>0.38</td>
<td>0</td>
<td>0.38</td>
<td>0.87</td>
</tr>
<tr>
<td>291103375</td>
<td>0</td>
<td>0.08</td>
<td>0.97</td>
<td>0.08</td>
<td>0.24</td>
<td>0</td>
<td>0.24</td>
<td>0.97</td>
</tr>
<tr>
<td>291103376</td>
<td>0</td>
<td>0.08</td>
<td>0.97</td>
<td>0.08</td>
<td>0.24</td>
<td>0</td>
<td>0.24</td>
<td>0.97</td>
</tr>
<tr>
<td>291103377</td>
<td>0</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.26</td>
<td>0</td>
<td>0.26</td>
<td>0.95</td>
</tr>
<tr>
<td>291103378</td>
<td>0.17</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.24</td>
<td>0</td>
<td>0.24</td>
<td>0.96</td>
</tr>
<tr>
<td>291103379</td>
<td>0</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
<td>0.45</td>
<td>0</td>
<td>0.45</td>
<td>0.89</td>
</tr>
<tr>
<td>291103380</td>
<td>0</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0</td>
<td>0.36</td>
<td>0.89</td>
</tr>
<tr>
<td>291103381</td>
<td>0.89</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0</td>
<td>0.27</td>
<td>0.89</td>
</tr>
<tr>
<td>291103382</td>
<td>0.95</td>
<td>0.11</td>
<td>0.76</td>
<td>0.11</td>
<td>0.55</td>
<td>0</td>
<td>0.55</td>
<td>0.95</td>
</tr>
<tr>
<td>291103383</td>
<td>0</td>
<td>0.11</td>
<td>0.44</td>
<td>0.11</td>
<td>0.44</td>
<td>0</td>
<td>0.44</td>
<td>0.77</td>
</tr>
<tr>
<td>291103384</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
<td>0.89</td>
</tr>
<tr>
<td>291103385</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
<td>0.92</td>
</tr>
<tr>
<td>291103386</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
<td>0.92</td>
</tr>
<tr>
<td>291103387</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Spam Detection

- *filter sentences based on their clarity score* in order to avoid penalizing workers for contributing on difficult or ambiguous sentences
- *apply the worker metrics to analyze worker agreement* on a specific putative event or across all the putative events that s(he) solved
- *apply explanation-based filters* in order to assess the overall quality of each worker rational
the explanation-based filters are able to increase the accuracy of detecting spam and low-quality workers with at least 5%, which leads to a better interpretation of the crowdsourced data and representation of the events.
Spam Detection

Worker Metrics Evaluation

Worker Metrics and Evaluation-based Filters Evaluation

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>event type</td>
<td>88</td>
<td>91</td>
<td>89</td>
<td>90</td>
</tr>
<tr>
<td>event location</td>
<td>86</td>
<td>82</td>
<td>97</td>
<td>88</td>
</tr>
<tr>
<td>event time</td>
<td>88</td>
<td>86</td>
<td>97</td>
<td>91</td>
</tr>
<tr>
<td>event participants</td>
<td>92</td>
<td>94</td>
<td>94</td>
<td>94</td>
</tr>
</tbody>
</table>

the explanation-based filters are able to increase the accuracy of detecting spam and low-quality workers with at least 5%, which leads to a better interpretation of the crowdsourced data and representation of the events.
Spam Detection

Worker Metrics Evaluation

Worker Metrics and Evaluation-based Filters Evaluation

the explanation-based filters are able to increase the accuracy of detecting spam and low-quality workers with at least 5%, which leads to a better interpretation of the crowdsourced data and representation of the events.
What more …

Understand human disagreement on event extraction with focus on *ambiguity*:

○ Would *different classification (ontology)* of putative events perform better?
○ Does the *overlapping of the types (ontology)* influence the results?
○ Identify *the right role fillers (per event) for multiple putative events*.
○ Would *event clustering* help with determining the most appropriate structure of the event and its role fillers?
Conclusions

- **disagreement metrics adaptable across domains** - helped us to understand a bit more the vagueness and the clarity of a sentence/putative event

- clarity or the vagueness of sentences help **select the good cases for automated training**

- **micro-task template design was most difficult process** as it aimed at harnessing diversity and disagreement, while making the task understandable and affordable for the crowdsourcers

- **understanding disagreement can help us understand event semantics**