Harnessing Disagreement for Event Semantics

Lora Aroyo[1,2], Chris Welty[2]

[1] VU University Amsterdam
lora.aroyo@vu.nl
[2] IBM Watson Research Center
cawelty@gmail.com

Abstract. The focus of this paper is on how events can be detected & extracted from natural language text, and how those are represented for use on the semantic web. We draw an inspiration from the similarity between crowdsourcing approaches for tagging and text annotation task for ground truth of events. Thus, we propose a novel approach that harnesses the disagreement between the human annotators by defining a framework to capture and analyze the nature of the disagreement. We expect two novel results from this approach. On the one hand, achieving a new way of measuring ground truth (performance), and on the other hand identifying a new set of semantic features for learning in event extraction.

1 Introduction

Events play an important role in human communication. Our understanding of the world is transferred to others through stories, in which objects and abstract notions are grounded in space and time through their participation in events. In conventional narrative, these events unfold sequentially in a timeline. Upon inspection, however, our understanding of events is quite difficult to pin down. This can be seen in metaphysics, where theories range from events as the most basic kind of entity in the universe to events as an unreal fiction [1], and in Natural Language Processing (NLP), where the few annotation tasks for events that have been performed have shown very low inter-annotator agreement.

One of the simplest and most prevalent ontological views of the universe is that there are two basic kinds of entities, objects and events. They are distinguished in that events *perdure* (their parts exist at different time points) and objects *endure* (they have all their parts at all points in time) [2]. The distinction is sometimes phrased ”objects are wholly present at any point in time, events unfold over time.” This definition and distinction is not universally held, but it serves us here as a convenient reference point; we believe the conclusion holds regardless of the ontological status of events.

The importance of events and their interpretation is widely recognized in NLP, but solutions remain elusive, whereas NLP technology for detecting objects (such as people, places, organizations, etc.) in text has reached ”off the shelf” levels of maturity. In addition, there is comparatively little annotated data for training and evaluation of event detection systems, and the bulk of what is available is difficult to reproduce. Annotator disagreement is quite high in most cases, and since many believe this is a sign of a poorly defined problem, guidelines for these event annotation tasks are very precise in order to address and resolve specific kinds of disagreement. This leads to brittleness or over generality, making it difficult to transfer annotated data across domains or to use the results for anything practical.

One of the reasons for annotator disagreement is that events are highly compositional in the way they are described in language. Objects are compositional, too, but only in reality – in language we rarely refer to the parts of the object, only to the object itself. For events, we often describe where and when they take place, who or what the participants were, what the causes or results of the event were, and what type of event it was. More importantly events are usually referred to through their parts, e.g. we might talk about a terrorist event by using the word ”explosion”, which literally refers to only a small part of the overall event, making it sometimes difficult to determine whether two parts of one event refer to the same thing.

This highly compositional nature means that there are more potential ways in which two human annotators can disagree about a single event. Since agreement is never perfect for any annotation task, the agreement for a composite annotation task will necessarily degrade as the product of the agreement for the sub-tasks. In other words, if events are taken to be a time, place, actor, patient, and type, the agreement for the event task will be the product of the agreement on the five sub-tasks, which would be low since agreement for any task is between 0 and 1.

In our efforts to study the annotator disagreement problem for events, we began to realize that the disagreement didn't really change people's understanding of a news story or historical description. People seem to live with the vagueness of events perfectly well; the lack of precision and identity in event detection began to seem like artificial problems. This led us to the hypothesis of this paper, that the kind of annotator disagreement we see is a natural state, and that event semantics, both individual and social, is by its very nature imprecise and varied. We propose to harness this by incorporating disagreement as parameter of the annotated meaning of events using a crowdsourcing approach, which allows for capturing the wide range of interpretations of events with a minimal requirement for agreement (only for e.g. spam detection). We can then use a form of semantic clustering by defining a similarity space not of lexical features of language, but of dimensions that come from a classification of human disagreement on event interpretation.

In this preliminary work we present the classification framework and annotation task, and describe how it will be used for event detection. This work is performed in the context of the DARPA's Machine Reading program (MRP)[3]

2 Classification Framework

Our classification of the multitude of event perspectives derives from, and forms the basis for understanding, the disagreement in the crowd-sourced event annotation task, and we use it further to define similarity between events identified by the annotators. Methodologically, the initial set of classifications in the framework were produced by observing disagreement in previous annotation tasks, and we expect to further extend and refine the set as we conduct new annotation tasks.

[3] http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx

We identify three high-level views to disagree on the annotation of events:

- *ontology*: disagreements on the basic status of events themselves as referents of linguistic utterances, for example are people events or do events exist at all.
- *granularity*: disagreements that result from issues of granularity, such as the location being a country, region, or city, the time being a day, week, month, etc.
- *interpretation*: disagreements that result from (non-granular) ambiguity, differences in perspective, or error in interpreting an expression, for example classifying a person as a terrorist or hero, the ”October Revolution” took place in September, etc.

2.1 Ontology Disagreements

We do not address ontological disagreements on events in this paper, and we assume annotation tasks to be defined by a particular ontology. The literature and history of event ontology is vast, see [1] for a good start. We assume for the purposes of this framework that events do exist (it is a particular ontological position that they don't), that they are located in space, occur over some time, have a prescribed type, have temporal parts, and have participants. This gives us five dimensions in which to classify possible annotator disagreements (space, time, classification, composition, and participation).

2.2 Granularity Disagreements

We consider disagreements on levels of granularity to be, for the most part, agreement about what the event refers to but disagreement about what level of detail is important to extract and identify the event.

- *Spatial* granularity disagreements occur when the location can be specified at sizes within some regional containment. If a sentence said, ”...a bombing in a downtown Beirut market...” the event might have taken place in ”downtown Beirut”, ”Beirut”, even ”Lebanon” or ”Middle East”. Each is correct, but typical gold standards define only one to be.
- *Temporal* granularity disagreements occur when the time can be specified at different durations of temporal containment. If a sentence said, ”...a bombing last Wednesday during the busy lunch hour...” might have taken place at ”lunch hour”, ”last Wednesday”, even ”last week”, ”2001”, etc.
- *Compositional* granularity disagreements occur when events are referred to by their parts at different levels of composition. Events are infinitely decomposable, and while this won't be reflected explicitly in a textual description, the compositionality does manifest as an abundance of ways of referring to what happened. If a sentence said, ”...a bombing took place last week, the explosion rocked the central marketplace...” we might say the event ”explosion” is part of the event ”bombing” and that the ”explosion” event is not the one of interest. There are many types of compositional disagreement (see section 2.3 below), here we refer only to disagreements in labeling the events in a way that affects *counting*, e.g. are there two events in the sentence or one? This category includes aggregate event mentions, such as ”5 bombings in Beirut”, for which annotators may disagree on whether the ”5 bombings” is one event with 5 parts, or 5 events.

- *Classificational* granularity disagreements occur when events are classified at different places in a given taxonomy, such that one class subsumes the other. If the annotators were provided with a taxonomy of events that specified bombing $\ll$ attack $\ll$ event, they may disagree on whether a particular event is a ”bombing” or ”attack”.
- *Participant* granularity disagreements occur when event participants are part of some group that can be identified at different levels. If a sentence said, ”... a shooting by Israeli soldiers ...” we might say the participants are ”soldiers”, ”Israeli soldiers”, ”Israeli Army”, or ”Israel”.

Thus, the identification of an event by human annotators can disagree in any of these granular dimensions with respect to the words used in the annotated text, while still representing a general agreement about the event itself. It is a peculiarity of NLP annotation tasks that this would be considered disagreement at all.

Often we observe disagreement in granularity when different levels of detail are needed to distinguish different events that share some property at some level. For example, if there were two bombings in Beirut on September 5th, some annotators would consider it more important to fix the time of day for each bombing or the participants mentioned by their role and name.

In previous attempts to define event annotation tasks, researchers have typically “perfumed” annotator disagreement on granularity by forcing one choice in particular contexts. Examples include fixing the granularity for all events to a day, if a day is unavailable, the week, then month, then year, then decade. This is regardless of whether that choice is believed by the annotator to be the most relevant level of detail, or even correct. These choices may reduce disagreement according to some measure, but we argue that they do not fix the problem, they simply cover it up: they are brittle in that they cannot be reused for applications requiring a different granularity, they make the task harder to learn (for machines) as they force an interpretation that people may not consistently have, and they occasionally force annotators to make the wrong choices in certain situations, even when they know its wrong.

2.3 Interpretation Disagreements

Disagreements on interpretation reflect genuine disagreement about what the event refers to. As with granularity, the disagreement can come from an event’s relation to other entities, and we break interpretation disagreements into the same five dimensions. Interpretation disagreements also include errors and misunderstandings by the annotators.

- *Spatial* interpretation disagreements occur when the location is vague, controversial, has some context that may change the coordinates, or perspectives that change some element of the spatial containment across annotators. For example, the location of a bombing could be ”the front lines”, which may be shifting and difficult to pin down latitude and longitude, or ”Prussia” which is still the name of a region but once also the name of a much larger country. A location, such as Taiwan, may be considered by one annotator to be part of the People’s Republic of China, and by another not to be.

- *Temporal* interpretation disagreements, similar to spatial, may occur when the time is vague or has some context that changes the actual time points. For example, the time of a bombing may be reported in a country whose time zone makes the time or even the day of the event different, or expressions like ”the past couple days” in which one annotator may take it to be a duration of two days, and another may take as a different duration. Relative dates like ”the end of world war II” or ”the October Revolution” (which took place in September) can also cause genuine disagreement among annotators if required to normalize the date to a specific year, month and day.
- *Compositional* interpretation disagreements occur when events are referred to by their parts and the annotators disagree on what the parts are. This includes the direction of the composition, e.g. ”bombing” is part of the ”explosion”, or ”explosion” is part of the ”bombing” in the previous example. This also includes the placement by annotators of implied events that contain, or are contained by, the mentioned ones.
- *Classificational* interpretation disagreements occur when events are classified under different classes, and one class does not imply the other (as opposed to granularity). This includes cases where the two classes are logically disjoint, and cases where they are not disjoint but in different branches of the taxonomy.
- *Participant* interpretation disagreements occur when the participants are vague (e.g. ”Western Authorities”), or controversial (e.g. ”Pakistan denied responsibility for the bombing”), or has some context that causes an annotator to differ from others. For example, in ”Saddam Hussein’s top advisor called the bombing an outrage” an annotator might assume that the advisor would not have spoken unless it was what he was told to say, and attribute ”Saddam Hussein” as the participant in the ”called” event, whereas a stricter reading would have the advisor as the participant.

The most common form of interpretation disagreements are ones that stem from misreadings of the text. It is important to note that most of the time, human readers are very tolerant of these kinds of errors in forming their understanding of what happened. It is more reasonable to try and ”correct” these errors to reduce disagreement, but we claim that if annotation is to scale, we need to be tolerant of them.

Interpretation disagreements are more difficult to account for than the granularity disagreements. Thus, we start with the first version of this crowdsourced annotation experiment by focussing on granularity disagreements only.

3 Annotation Task

NLP systems typically use the ground truth of an annotated corpus in order to learn and evaluate their output. Traditionally, the ground truth is determined by humans annotating a sample of the text corpus with the target events and entities, with the aim to optimize the inter-annotator agreement by restricting the definition of events and providing annotators with very precise guidelines. In this paper, we propose an alternative approach for the event annotation, which introduces a novel setting and different perspective on the overall goal.

Table 1. Annotation Matrix for Putative Event$_i$

Event$_i$	Temporal						Spatial						Participants						Compositional						Classificational					
	1	2	3	4	5	ø	1	2	3	4	5	ø	1	2	3	4	5	ø	1	2	3	4	5	ø	1	2	3	4	5	ø
ann$_1$	1	0	1	1	0	1	0	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
ann$_N$	1	1	1	1	0	1	0	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	1	0	1	1	1	0

By analogy to image and video tagging crowdsourcing games, e.g. Your Paintings Tagger [4] and Yahoo! Video Tag Game [3], we envision that a crowdsourcing setting could be a good candidate to the problem of insufficient annotation data. However, we do not exploit the typical crowdsourcing agreement between two or more independent taggers, but on the contrary, we harness their disagreement. Our goal is to allow for a maximum disagreement between the annotators in order to capture a maximum diversity in the event expressions.

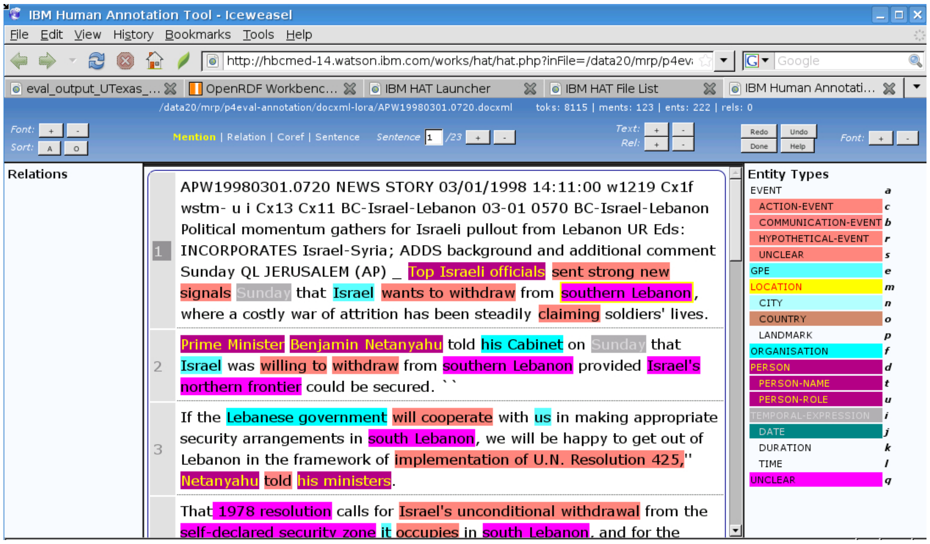
Annotation Matrix: In section 2 we introduced a classification framework to understand the disagreement between annotators. In our annotation task we only consider the granularity-based disagreement – with five axes and five levels of granularity for each axis. Following this, for each putative event (a marked verb or nominalized verb), we build an *Annotation Matrix* (Table 1) from the input of all annotators. We can then subsequently use these annotation matrices for an analysis over the whole collection of events, e.g. for determining similarity between different events and thus recognizing missed coreferences. We can also use the matrices for an analysis of the annotation space of each individual event. For example, the highest agreement in each axis level could indicate the most likely granularity for this event, while still giving a sense of the range of acceptable granularities in each dimension. Such in-depth analysis of the annotations can allow us to identify a new set of features that can help to improve the event extraction. For example, we could thus expect to find dependencies between the type of events and the level of granularity for its spatial or temporal entities.

Annotation Setting: For the proposed annotation task we plan to use a sample of the $10,000$ documents taken from the Gigaword corpus (used in the context of the DARPA's Machine Reading program (MRP)[5]) together with several sources for background knowledge. The background knowledge includes, for example, the IC++ Domain Ontology for Violent Events (identifying event types and binary relations), geographical and temporal resources as well as general lexical resources such as WordNet and DBpedia.

A pre-annotation is performed by automatically marking all the verbs and nominalized verbs as putative events (Fig. 2): this would include both events from the IC++ ontology, as well as reporting and other communication events. The IBM Human Annotation Tool (HAT) was used as an initial annotation interface. Our background knowl-

[4] http://tagger.thepcf.org.uk/

[5] http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx


IBM Human Annotation Tool - Iceweasel
File Edit View History Bookmarks Tools Help
http://hbcmed-14.watson.ibm.com/works/hat/hat.php?inFile=/data20/mrp/p4ev
Google
eval_output_UTexas_...
OpenRDF Workbenc...
IBM HAT Launcher
IBM HAT File List
IBM Human Annotati...
Font: + -
Sort: A O
Mention | Relation | Coref | Sentence
Sentence 1 /23 + -
Text: + -
Rel: + -
Redo Undo
Done Help
Font: + -
Relations

/data20/mrp/p4eval-annotation/docxml-lora/APW19980301.0720.docxml toks: 8115 | ments: 123 | ents: 222 | rels: 0

1 APW19980301.0720 NEWS STORY 03/01/1998 14:11:00 w1219 Cx1f wstm- u i Cx13 Cx11 BC-Israel-Lebanon 03-01 0570 BC-Israel-Lebanon Political momentum gathers for Israeli pullout from Lebanon UR Eds: INCORPORATES Israel-Syria; ADDS background and additional comment Sunday QL JERUSALEM (AP) _ Top Israeli officials sent strong new signals Sunday that Israel wants to withdraw from southern Lebanon, where a costly war of attrition has been steadily claiming soldiers' lives.

2 Prime Minister Benjamin Netanyahu told his Cabinet on Sunday that Israel was willing to withdraw from southern Lebanon provided Israel's northern frontier could be secured. ``

3 If the Lebanese government will cooperate with us in making appropriate security arrangements in south Lebanon, we will be happy to get out of Lebanon in the framework of implementation of U.N. Resolution 425,'' Netanyahu told his ministers.

That 1978 resolution calls for Israel's unconditional withdrawal from the self-declared security zone it occupies in south Lebanon, and for the

Entity Types

EVENT	*a*
ACTION-EVENT	*c*
COMMUNICATION-EVENT	*b*
HYPOTHETICAL-EVENT	*r*
UNCLEAR	*s*
GPE	*e*
LOCATION	*m*
CITY	*n*
COUNTRY	*o*
LANDMARK	*p*
ORGANISATION	*f*
PERSON	*d*
PERSON-NAME	*t*
PERSON-ROLE	*u*
TEMPORAL-EXPRESSION	*i*
DATE	*j*
DURATION	*k*
TIME	*l*
UNCLEAR	*q*

Fig. 1. Annotation interface

edge base allows us to pre-label temporal, spatial, and participant entities with granularities (e.g. city, region, country), and we provide an *a-priori* mapping from these to the numbers in the annotation matrix. The annotators do not need to know the granularity level, they are presented with all the possible choices and they select one (or more), and their choices are automatically mapped into the matrix. For example, for the sentence, ”A bomb exploded in Beiruit, Lebanon last Friday,” the annotator would be presented with ”exploded” as the putative event, and could select between Beirut and Lebanon (or both) as the location. Since our background knowledge includes that Beirut is a city and Lebanon a country, if selected as a location for the event these are mapped to granularity levels 2 and 3, resp.

We ran explorative annotation experiments with the IBM Human Annotation Tool (Fig. 1), and proceeded further with using larger annotator pool at Amazon Mechanical Turk and CrowdFlower. Annotation data was collected according to the stages sketched in Fig. 2. As presented in the figure, the process comprises of four Phases (I-IV). Each Phase is split in two main steps: (A) collecting initial set of annotations (in each Phase different types of annotations) and (B) performing spam filtering step. In each phase we select from the A results items that can be used as Gold Standard items in step B.

4 Related Work

This work derives directly from our efforts in the Machine Reading Program (MRP) to define an annotation task for event coreference. The process of developing guidelines is very iterative - starting with an initial set of requirements from simple examples, the guidelines are then applied by a small group and the disagreements, in particular, are studied and the guidelines modified to address them. The process is repeated until the

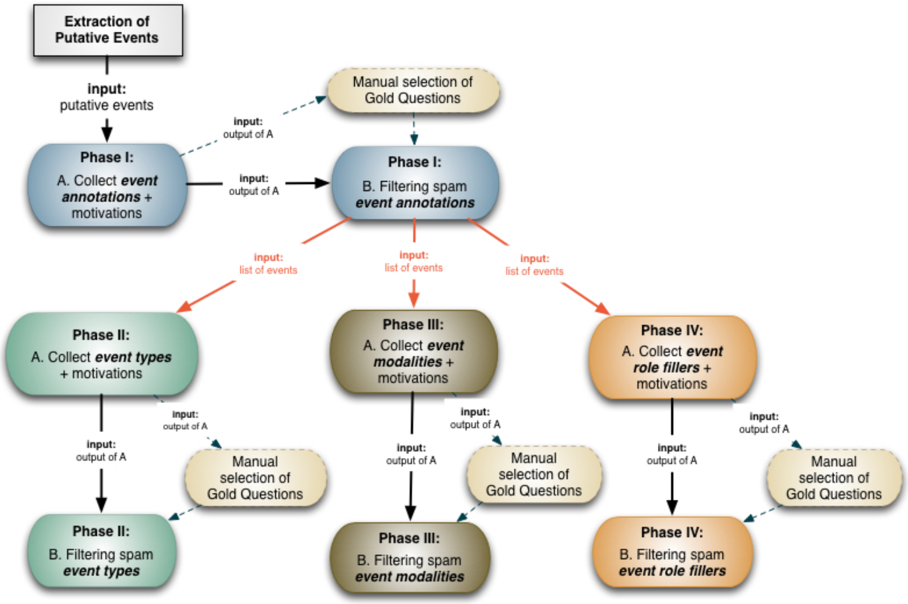

Extraction of Putative Events
input: putative events
Phase I: A. Collect event annotations + motivations
input: output of A
Manual selection of Gold Questions
input: output of A
Phase I: B. Filtering spam event annotations
input: list of events
input: list of events
input: list of events
Phase II: A. Collect event types + motivations
Phase III: A. Collect event modalities + motivations
Phase IV: A. Collect event role fillers + motivations
input: output of A
input: output of A
Manual selection of Gold Questions
Phase II: B. Filtering spam event types
input: output of A
input: output of A
Manual selection of Gold Questions
Phase III: B. Filtering spam event modalities
input: output of A
input: output of A
Manual selection of Gold Questions
Phase IV: B. Filtering spam event role fillers

Fig. 2. Crowdsourcing Annotation Process

agreement (typically a κ score) reaches an acceptable threshold, and then is distributed to the actual annotators. Developing the guidelines usually takes several months and requires language experts.

The idea of analyzing and classifying annotator disagreement on a task is therefore not new, but part of the standard practice in developing guidelines, which are widely viewed as necessary for human annotation tasks. However, the goals of classifying disagreement, in most previous efforts, has been designed to eliminate it, not to exploit it. This can be seen in most annotation guidelines for NLP tasks. For example, in [4], the instructions include that all modality annotations should, "ignore temporal components of meaning. For example, a belief stated in the future tense (Mary will meet the president tomorrow) should be annotated with the modality 'firmly believes' not 'intends' or 'is trying'." [4]. Here the guideline authors repeat that these choices should be made, "even though other interpretations can be argued."

Similarly, in the annotator guidelines for the MRP Event Extraction Experiment (aiming to determine a baseline measure for how well machine reading systems extract attacking, injuring, killing, and bombing events) [5] show examples of restricting humans to follow one interpretation, for example for location, in order to ensure higher chance for the inter-annotator agreement. In this case, the spatial information is restricted only to "country", even though other more specific location indicators might be present in the text, e.g. the Pentagon.

There are many annotation guidelines available on the web and they all have examples of "perfuming" the annotation process by forcing constraints to reduce disagreement (with a few exceptions). In [6] and subsequent work in emotion [7], disagreement is used as a trigger for *consensus-based annotation* in which all disagreeing annotators are forced to discuss and arrive at a consensus. This approach acheives very high κ scores (above .9), but it is not clear if the forced consensus really achieves anything meaningful. It is also not clear if this is practical in a crowdsourcing environment.

A good survey and set of experiments using disagreement based semi-supervised learning can be found in [8]. However, they use disagreement to describe a set of techniques based on bootsrapping, not collecting and exploiting the disagreement between human annotators. The bootstrapping idea is that small amounts of labelled data can be exploited with unlabelled data in an iterative process [9], with some user-relevance feedback (aka active learning).

Disagreement harnessing and crowdsourcing has previously been used by [10] for the purpose of word sense disambiguation, and we will explore similar strategies in our experiments for event modeling. As in our approach, they form a confusion matrix from the disagreement between annotators, and then use this to form a similarity cluster. In addition to applying this technique to events, our work adds a novel classification scheme for annotator disagreement that provides a more meaningful feature space for the confusion matrix; it remains to be demonstrated whether this will have impact.

The key idea behind our work is that harnessing disagreement brings in multiple perspectives on data, beyond what experts may believe is salient or correct. This concept has been demonstrated previously in the Waisda? video tagging game [11], in which lay (non-expert) users provided tags for videos in a crowdsourcing game. The Wasida? study showed that only 14% of tags provided by lay users could be found in the professional video annotating vocabulary (GTAA), which indicates a huge gap between the professional and lay users' views on what is important in a video. The study showed the lay user tags were meaningful (as opposed to useless or erroneous), and the mere quantity of tags was a success factor in retrieval systems for these multimedia objects. Similarly, the steve.museum project [12] studied the link between a crowdsourced user tags folksonomy and the professionally created museum documentation. The results showed that users tag artworks from a different perspective than that of museum professionals: again in this seperate study only 14% of lay user tags were found in the expert-curated collection documentation.

5 Conclusions

When considering approaches for detecting and extracting events in natural language text and representing those extracted events for use in the Semantic Web, we see the implications of what differentiates events from objects. When it comes to annotation tasks, the compositional nature of events plays an important role in the way in which annotators perceive the events, annotate them and agree in their existence.

For the goal of improving event detection, we have chosen to leverage the annotator disagreement in order to obtain an event description that allows machine readers to better identify and detect events. In this way, we do not aim for annotator agreement (as in many tagging scenarios where similarity is an indicator for success), but on the contrary we hypothesized that annotator disagreement for even annotation actually could provides us with a better event description from the perspective of automatic event detection. By factoring in the different viewpoints that annotators can have, the likelihood of identifying events that have been represented with such viewpoints is higher.

In this paper we have contributed a classification framework of the variety of ways in which people can perceive events, with a matrix for the identification of patterns of agreement and disagreement (with the aim to be able later to exploit them in the MR of events), and with a description of the design of the experiment to verify the effect of using the matrix in the annotation task.

6 Acknowledgments

The authors gratefully acknowledge the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0172. Any opinions, findings, conclusion or recommendations ex-pressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government. We would like to thank Sid Patwardhan from IBM Research for his contribution to the implementation.

References

1. Higginbotham, J., Pianesi, F., Varzi, A.: Speaking of Events. Oxford University Press, USA (2000)
2. Lewis, D.K.: On the Plurality of Worlds. Blackwell Publishers (1986)
3. van Zwol, R., Garcia, L., Ramirez, G., Sigurbjornsson, B., Labad, M.: Video tag game. In: 17th International World Wide Web Conference (WWW developer track), ACM (April 2008)
4. Baker, K., Bloodgood, M., Diab, M., Dorr, B., Hovy, E., Levin, L., McShane, M., Mitamura, T., Nirenburg, S., Piatko, C., Rambow, O., Richardson, G.: Simt scale 2009 modality annotation guidelines. Technical Report 4, Human Language Technology Center of Excellence (2010)
5. Hovy, E., Mitamura, T., Verdejo, F.: Event coreference annotation manual. Technical report, Information Sciences Institute (ISI) (2012)
6. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: in Proc. ICSLP 2002. (2002) 2037–2040
7. Litman, D.J.: Annotating student emotional states in spoken tutoring dialogues. In: In Proc. 5th SIGdial Workshop on Discourse and Dialogue. (2004) 144–153
8. Zhou, Z.H., Li, M.: Semi-supervised learning by disagreement. Knowl. Inf. Syst. **24**(3) (2010) 415–439
9. Riloff, E., Jones, R.: Learning dictionaries for information extraction by multi-level bootstrapping. In: AAAI/IAAI. (1999) 474–479
10. Chklovski, T., Mihalcea, R.: Exploiting agreement and disagreement of human annotators for word sense disambiguation. In: UNT Scholarly Works. UNT Digital Library (2003)
11. Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., Aroyo, L.: On the role of user-generated metadata in audio visual collections. In: K-CAP. (2011) 145–152
12. Leason, T.: Steve: The art museum social tagging project: A report on the tag contributor experience. In: Museums and the Web 2009:Proceedings. (2009)