Simple Semantic Annotation and Situation Frames: Two Approaches to Basic Text Understanding in LORELEI

Kira Griffitt, Jennifer Tracey, Ann Bies, Stephanie Strassel
Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{kiragrif, garjen, bies, strassel}@ldc.upenn.edu

Abstract

We present two types of semantic annotation developed for the DARPA Low Resource Languages for Emerging Incidents (LORELEI) program: Simple Semantic Annotation (SSA) and Situation Frames (SF). Both of these annotation approaches are concerned with labeling basic semantic information relevant to humanitarian aid and disaster relief (HADR) scenarios, with SSA serving as a more general resource and SF more directly supporting the evaluation of LORELEI technology. Mapping between information in different annotation tasks is an area of ongoing research for both system developers and data providers. We discuss the similarities and differences between the two types of LORELEI semantic annotation, along with ways in which the general semantic information captured in SSA can be leveraged in order to recognize HADR-oriented information captured by SF. To date we have produced annotations for nineteen LORELEI languages; by the program's end both SF and SSA will be available for over two dozen typologically diverse languages. Initially data is provided to LORELEI performers and to participants in NIST's Low Resource Human Language Technologies (LoReHLT) evaluation series. After their use in LORELEI and LoReHLT evaluations the data sets will be published in the LDC catalog.

Keywords: semantic annotation, information extraction, linguistic resources

1. Introduction

Most of the world's languages are under-resourced for Human Language Technology (METANET, 2010; Rehm and Uszkoreit, 2012), but lack of resources does not correlate with lack of need for such technologies. The DARPA Low Resource Languages for Emerging Incidents (LORELEI) Program aims to advance the capabilities of NLP in low-resource languages, with a particular focus on using NLP to obtain situational awareness for an incident like a natural disaster involving a low-resource language within a short time of the emergence of that incident (DARPA, 2014).

Linguistic Data Consortium (LDC) is building a variety of linguistic resources for nearly three dozen low resource languages for the LORELEI program (see Table 1). Representative Language Packs – consisting of large volumes of formal and informal monolingual and parallel (with English) text with a variety of manual annotations to support situational awareness, plus a lexicon, grammatical sketch and basic processing tools – are designed to enable research into language universals and cross-language projection. Incident Language Packs contain manually labeled evaluation data designed to test system performance on tasks related to situational awareness for one or more surprise languages per year that remain unknown until the start of the annual evaluation (Strassel and Tracey, 2016).

Akan (Twi)	Hungarian	Tagalog	Vietnamese
Amharic	Indonesian	Tamil	Wolof
Arabic	Mandarin	Thai	Yoruba
Bengali	***Oromo***	***Tigrinya***	Zulu
English	Russian	Turkish	
Farsi	Somali	***Ukrainian***	
Hausa	Spanish	***Uyghur***	
Hindi	Swahili	Uzbek	

Table 1: LORELEI Representative and ***Incident*** Languages

This paper focuses on two semantic annotation tasks developed by LDC to support LORELEI research and evaluation: Simple Semantic Annotation (SSA) and Situation Frame (SF). Both SSA and SF label basic information relevant to humanitarian aid and disaster relief (HADR) scenarios. Situation Frame annotation directly corresponds to the LORELEI SF evaluation task, with a focus on actionable information contained in HADR-related documents, where "actionable" refers to the kind of content that mission planners might require in order to mount a response to an incident. SSA represents a more general approach to semantic annotation, albeit in the HADR domain.

By design, no training data of any kind is provided for LORELEI incident languages, since data of this type is unlikely to be available at the start of an incident involving a low resource language. Instead systems must make use of more general linguistic resources, through transfer learning, annotation projection and language universals, in order to rapidly respond to the need for situational awareness in a new language. SSA serves as a general semantic resource that can be used by system developers to build language-independent algorithms capable of labeling actionable HADR information in documents from a surprise incident language at test time. In the original LORELEI data plan SSA annotation is provided for all Representative Languages while no Situation Frame annotation is provided (apart from answer key annotations on the incident language test set). After the Year 2 evaluation the data plan was augmented to provide a small amount of SF annotation for all Representative Languages, though it remains the case that no training data -- whether SSA or SF -- is provided for any incident language.

2. Simple Semantic Annotation

Simple Semantic Annotation supports LORELEI's goal of situational awareness by labeling basic information about physical events and disaster-relevant situations, their participants, and their locations in text data. Given

LORELEI's low resource language setting combined with the need to simultaneously create resources for dozens of languages, the SSA task was designed with a naïve annotator in mind (i.e. without formal linguistic training or prior annotation experience). In this way it contrasts with more complex predicate-argument focused semantic representation schemes like Abstract Meaning Representation (AMR) (Banarescu, et al., 2013) Automatic Content Extraction (ACE) (Doddington, et al., 2004), PropBank (Palmer, et al., 2005), FrameNet (Baker, et al. 1998), Richer Event Description (RED) (Ikuta, et al., 2014), and Universal Decompositional Semantics (UDS) (White, et al., 2016), which require a background in linguistics and/or a long training period.

In order to make SSA feasible for non-experts to master quickly, we annotate a small number of broad, underspecified predicate and argument categories that do not require fine-grained semantic distinctions. We generally select names, pronouns, or heads of nominal phrases as annotation extents, but annotators are allowed to select "intuitive extents" if needed (e.g. for multiword expressions), meaning that strict rules about selecting head words are not enforced. Each sentence is annotated independently, with reference to the full document as needed for additional context, and there is no coreference of arguments or predicates. The following sections describe the predicate and argument categories annotated in SSA.

2.1 Predicate Categories

SSA has two coarse-grained predicate categories: Acts and States. Acts are event-like predications describing change, while States are situation-like, describing non-changing or ongoing circumstances. In SSA, Acts and States are semi-open classes. There is no fixed set of predicates made available to annotators and no typing of predicates beyond the broad categories of Act and State. Since exhaustive annotation of all Acts and States is impractical, and since SSA is focused on tagging information relevant to situational awareness, SSA annotators are instructed to restrict their annotations to capturing the following types of Acts and States:

- Physical Acts, which are events, actions, or activities that take place in the observable, material world (e.g. bombing) as opposed to events that do not (e.g. thinking)
- Disaster-Relevant States, which are situations that constitute, are caused by, or provide information relevant to a disaster or disaster-relief effort (e.g. scared, without water, etc.), but not those that bear no relationship to a disaster scenario (e.g. "married", "excited [about a movie]," etc.).

2.2 Argument Categories

We define three coarse-grained argument categories for SSA: Agents, Patients, and Places. SSA Agents are similar to the traditional linguistic notion of Agent, though slightly broader, encompassing the person or thing that does or performs an Act, or the person or thing that causes or enables an Act to occur or a State to arise. Agents are often typical entities like persons, organizations, and geopolitical entities (e.g. if the United Nations delivered supplies after a disaster, we would annotate "United Nations" as the Agent of the predicate "delivered"). SSA Patients include the traditional linguistic notion of Patient, but also include recipients, beneficiaries, experiencers, and purposes/goals. Arguments may be typical entities, or they may be other Acts or States. Place includes the physical location where an Act or State occurred, as well as directional locations.

3. Situation Frame Annotation

The Situation Frame annotation task was defined to support LORELEI technology evaluations, and is directly aligned with the goal of situational awareness in disaster response scenarios (Strassel, et al., 2016). The objective of SF is to aggregate information into a comprehensive, actionable understanding of the basic facts needed to mount a response to an emerging situation, including the following:

- Characterization of the situation type
- Status of need/issue and resolution of need
- Localization of the situation to a place
- Sentiment, Emotion, or Cognitive State (SEC).

The information is arranged into "frames," which represent *needs* that may require a response (e.g., food, shelter, etc.), or *issues* that may affect the ability to deliver aid (e.g., widespread crime in the area). The frame contains all of the type, status, place, entity and SEC information elements for a given need or issue. Note that the term "frame" is used here in the general sense of a linguistic frame, rather than with any more specific reference to PropBank frame files (Palmer, et al. 2005) or the FrameNet lexical database (Baker, et al. 1998). Situation Frame annotation captures information about needs and issues at the document level, rather than capturing semantics at the event or word level.

3.1 Situation Type

For each frame, annotators characterize the situation by indicating the type of need or issue that exists, selecting from the types shown in Table 2. Multiple needs or issues in a document result in multiple frames, one for each unique combination of type, status, and place. Need and issue types were defined with input from LORELEI stakeholders and from existing annotation schemes such as MicroMappers (Imran, et al., 2014).

Need Types	**Issue Types**
• Evacuation • Food Supply • Search/Rescue • Utilities/Energy/Sanitation • Infrastructure • Medical Assistance • Shelter • Water Supply	• Civil Unrest/ Widespread Crime • Regime Change • Terrorism/Extreme Violence

Table 2: Situation Frame Types

Situation Frame annotation requires the use of inference, since annotators must be able to recognize that an implied need exists even when it is not explicitly stated (e.g. when a document about a hurricane says that "housing across the island was destroyed", annotators should label a shelter need). However, inference is a slippery slope, and too much use of inference can lead annotators to create frames for all possible needs typically associated with an incident – even when they are not implied by the document (e.g. creating search/rescue, shelter, water supply, utilities/energy/sanitation, medical assistance and infrastructure needs anytime an earthquake is mentioned, even when the document does not imply such needs currently exist.) A major challenge for SF annotation is creating guidelines and training annotators to use the right amount of inference such that annotators create frames for needs (or issues) that are strongly implied or inevitable, but not when it is merely possible or even likely. Because of the inherent challenge in achieving highly consistent annotations involving inference, SF evaluation data is labeled by a panel of annotators, which is reflected in system scoring (NIST, 2017).

3.2 Situation Status

Every situation frame is labeled for the status of the need or issue (current or not) as well as the status of the resolution (sufficient or not) for needs. Annotators also label the source(s) of information about the need and/or its resolution, as well as the entity/entities involved in resolving the need. For example, if the Red Cross and the government of Mexico are both mentioned in the document as contributing to the relief of a food need, "Red Cross" and "Mexico" would be added to the frame's "resolved by" element. Only entities named in the document can be selected as "reporting" or "resolving" the need or issue.

3.3 Situation Location

Annotators localize the situation by specifying the place where it occurs, selecting a single Location or Geopolitical Entity for each frame, or an indication that no named place entity relevant to the frame is mentioned in the document. Only named entities can be selected as locations for a situation frame. If there are multiple mentions of a situation (e.g. same place but different status; same need/issue in different places) multiple frames are created. Labeling SF place can be challenging due to the fact that news reports, tweets, etc. may be ambiguous or vague about the exact location of a need, even when the type of need is clear. An incident may be discussed in connection with several different (related or adjacent) places, but it is not always clear whether they are all affected by the same set of needs or whether the status or urgency differs among the various locations.

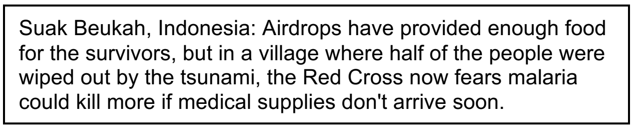
3.4 Situation SEC

Finally, annotators denote SEC for a situation by indicating whether it is urgent. Urgency can be both a property of the emotional/cognitive state of those affected as well as a property of the situation itself, regardless of any emotional component. For example, both "We're in desperate need of water. It's awful!" and "Officials say it is imperative that the drinking water be brought to the area immediately" are tagged as urgent. In pilot experiments the SF task has been augmented with additional SEC information, including positive or negative sentiment and two specific emotions: anger, and fear. These are labeled if they are present in the document and related to the frames, and annotation includes the sentiment holder, target, and sentiment/emotion value (positive, negative, anger, fear).

4. Mapping from SSA to SF

While SSA and SF annotation have different scopes (SSA is sentence-level while SF is document-level), because they both target disaster-relevant semantic information it is conceivable that the general-purpose semantic information captured by SSA could be utilized for improved performance on the incident language Situation Frames task through transfer learning. In this section we consider the possibility of direct mappings or at least inferential correspondences between the more general-purpose, sentence-level information captured in SSA and the use-case oriented, document-level information labeled in SF. Note that there are no automated techniques used to generate the SSA-SF mappings in the following sections. All mappings discussed in this paper have been generated by manual comparison of the annotations.

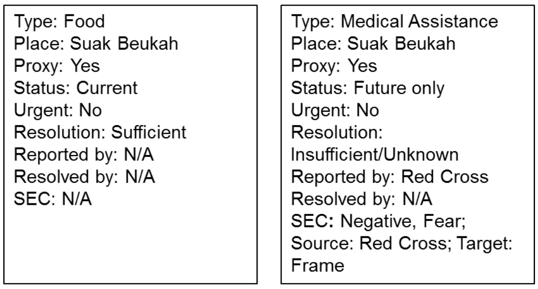
Let us take the following document excerpt as an example and point of comparison for SSA and SF annotation:



Suak Beukah, Indonesia: Airdrops have provided enough food for the survivors, but in a village where half of the people were wiped out by the tsunami, the Red Cross now fears malaria could kill more if medical supplies don't arrive soon.

Figure 1: Sample document excerpt

For this document excerpt, SF would annotate two frames, one a Food Need, the other a Medical Assistance Need:

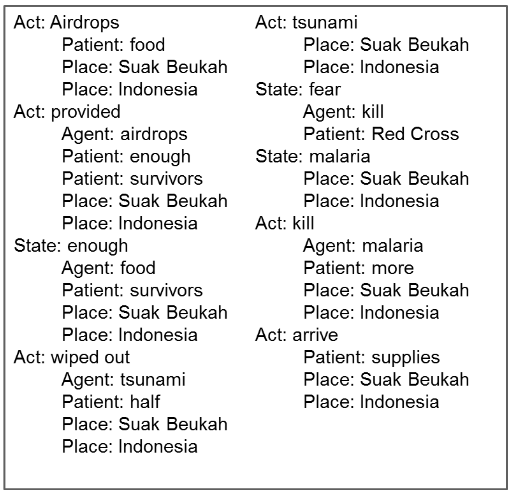


Type: Food
Place: Suak Beukah
Proxy: Yes
Status: Current
Urgent: No
Resolution: Sufficient
Reported by: N/A
Resolved by: N/A
SEC: N/A

Type: Medical Assistance
Place: Suak Beukah
Proxy: Yes
Status: Future only
Urgent: No
Resolution: Insufficient/Unknown
Reported by: Red Cross
Resolved by: N/A
SEC: Negative, Fear; Source: Red Cross; Target: Frame

Figure 2: Sample document with SF annotation

This information provides an aggregated, actionable understanding that there are food and medical assistance needs that exist in Suak Beukah as a result of the tsunami described in the document.



```
Act: Airdrops
        Patient: food
        Place: Suak Beukah
        Place: Indonesia
Act: provided
        Agent: airdrops
        Patient: enough
        Patient: survivors
        Place: Suak Beukah
        Place: Indonesia
State: enough
        Agent: food
        Patient: survivors
        Place: Suak Beukah
        Place: Indonesia
Act: wiped out
        Agent: tsunami
        Patient: half
        Place: Suak Beukah
        Place: Indonesia
Act: tsunami
        Place: Suak Beukah
        Place: Indonesia
State: fear
        Agent: kill
        Patient: Red Cross
State: malaria
        Place: Suak Beukah
        Place: Indonesia
Act: kill
        Agent: malaria
        Patient: more
        Place: Suak Beukah
        Place: Indonesia
Act: arrive
        Patient: supplies
        Place: Suak Beukah
        Place: Indonesia
```

Figure 3. Sample document with SSA annotation

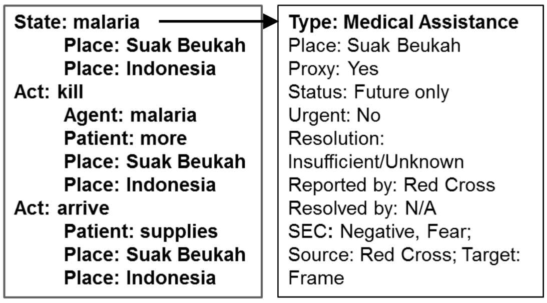
In contrast, SSA captures several specific, individual physical Acts or disaster-relevant States, including: airdrops of food, there being enough food, a tsunami, being wiped out by a tsunami, the occurrence of malaria, and deaths from malaria.

The Place arguments for these SSA annotations also indicate that the events and situations described in the document are located in Suak Beukah. These SSA annotations provide a detailed, but uncategorized picture of the individual events and situations that are occurring in the wake of the tsunami as described in the document.

Given each task's annotations for this excerpt, we will now identify and examine informational correspondences between the SSA annotations and two SF frames.

4.1 Mapping SF Frame Type from SSA

Looking at the second column of SSA annotations, we identify several correspondences between the SSA elements and SF frame type (indicated by shaded boxes): The SSA annotations do not contain the lexical items "medical" or "assistance", and so direct mapping based on the lexical items identified in SSA and the names of the SF need type is not possible. However, using world knowledge, we can identify "malaria" as a disease or medical condition that causes deaths, and thus infer that the document is describing a medical situation. Further, we can use world knowledge to identify the Red Cross as an organization providing assistance, and observe that supplies are being provided in the area where a medical situation exists. These two inferences let us infer a Medical Assistance frame type.



State: malaria
Place: Suak Beukah
Place: Indonesia
Act: kill
Agent: malaria
Patient: more
Place: Suak Beukah
Place: Indonesia
Act: arrive
Patient: supplies
Place: Suak Beukah
Place: Indonesia

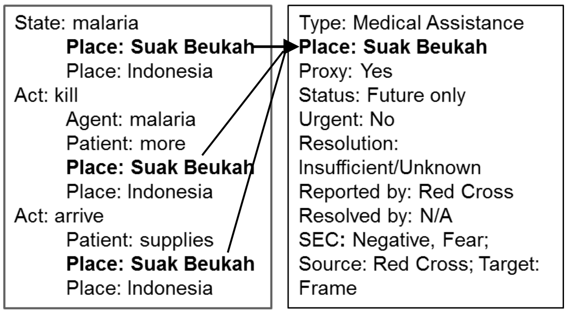
Type: Medical Assistance
Place: Suak Beukah
Proxy: Yes
Status: Future only
Urgent: No
Resolution: Insufficient/Unknown
Reported by: Red Cross
Resolved by: N/A
SEC: Negative, Fear; Source: Red Cross; Target: Frame

Figure 4: Mapping frame type

4.2 Mapping SF Place from SSA

Turning to the Place information type, we can see that it is possible to directly ascertain Place information for the Medical Assistance Frame from the SSA annotations:

Here, as shown in Figure 5, all the medically-relevant SSA Predicates have "Suak Beukah" as their Place, which lets us map Suak Beukah as the Place for the Medical Assistance frame.



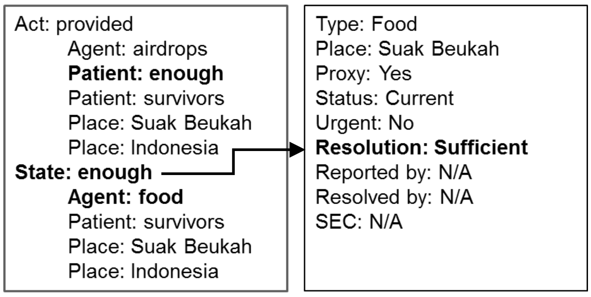
State: malaria
Place: Suak Beukah
Place: Indonesia
Act: kill
Agent: malaria
Patient: more
Place: Suak Beukah
Place: Indonesia
Act: arrive
Patient: supplies
Place: Suak Beukah
Place: Indonesia

Type: Medical Assistance
Place: Suak Beukah
Proxy: Yes
Status: Future only
Urgent: No
Resolution: Insufficient/Unknown
Reported by: Red Cross
Resolved by: N/A
SEC: Negative, Fear; Source: Red Cross; Target: Frame

Figure 5: Mapping place information

4.3 Understanding SF Resolution

Looking at the SF Food need frame, we see that it is also possible to ascertain Resolution information:



Act: provided
Agent: airdrops
Patient: enough
Patient: survivors
Place: Suak Beukah
Place: Indonesia
State: enough
Agent: food
Patient: survivors
Place: Suak Beukah
Place: Indonesia

Type: Food
Place: Suak Beukah
Proxy: Yes
Status: Current
Urgent: No
Resolution: Sufficient
Reported by: N/A
Resolved by: N/A
SEC: N/A

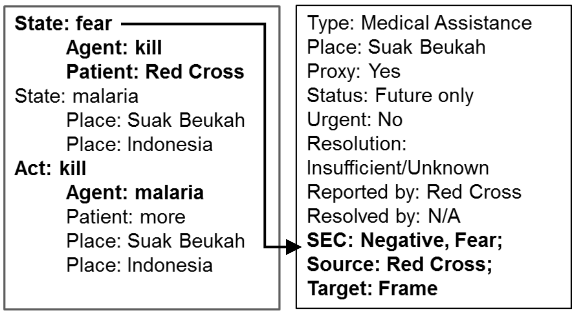
Figure 6: Mapping resolution information

Looking at the SSA annotations in the first column of Figure 6 that are relevant to food needs, we can see in the "Act: provided" Predicate that "enough" of something has been provided, and in the "State: enough" Predicate that the thing there is enough of is food.

Based on this information, we can use lexical knowledge to infer that there being "enough" of something corresponds to there being a sufficient amount of that thing, which allows us to recognize that the Resolution of the Food need frame is "Sufficient".

4.4 Mapping SF SEC from SSA

Finally, returning to the Medical Assistance frame, we can observe correspondences that allow us to identify SEC information for the Medical Assistance frame from SSA.



State: fear
Agent: kill
Patient: Red Cross
State: malaria
Place: Suak Beukah
Place: Indonesia
Act: kill
Agent: malaria
Patient: more
Place: Suak Beukah
Place: Indonesia

Type: Medical Assistance
Place: Suak Beukah
Proxy: Yes
Status: Future only
Urgent: No
Resolution: Insufficient/Unknown
Reported by: Red Cross
Resolved by: N/A
SEC: Negative, Fear; Source: Red Cross; Target: Frame

Figure 7: Mapping SEC information

Looking at the SSA Predicates with correspondence to the Medical Assistance frame in Figure 7, we can use the SSA Predicate "State: fear" to directly recognize an SEC value of Fear for the frame, and use world knowledge to identify "fear" as a negative emotion, which allows us to map a Negative sentiment for the frame. Further, the SSA Argument "Patient: Red Cross" to identifies the Red Cross as the experiencer "State: fear", which we can use to recognize the Red Cross as the Source of the Negative sentiment and Fear SEC values.

Finally we can use the SSA Argument "Agent: kill" to identify people being killed as the reason the Red Cross is experiencing fear. We can then trace through the other SSA annotations "Act: kill, Agent: malaria" and "State: malaria" to understand that malaria as the cause of people being killed. Since this set of Predicates and Arguments all correspond to the Medical Assistance frame, we can infer the frame itself as the SEC Target value for this frame.

5. Conclusion

We have presented two types of semantic annotation that were developed at LDC to support the humanitarian aid and disaster relief use case for the LORELEI program: Simple Semantic Annotation and Situation Frames. We have further presented some ways in which the general-purpose semantic information captured in SSA can be leveraged in order to map to the use-case information captured by SF. This is an area of on-going research for both system developers and annotation creation.

The linguistic resources described here have been distributed to LORELEI performers and to participants in the NIST Open Low Resource Human Language Technologies (LoReHLT) evaluation (NIST, 2017). As the data sets are completed under LORELEI they will be published in the LDC catalog, making them generally available to the broader research community.

We have produced manually annotated SSA data for 25K words in each of 9 representative languages (Amharic, Arabic, Chinese, Hungarian, Farsi, Russian, Spanish, Vietnamese, Yoruba). This data has been used as training data as part of the representative language packs. We have also manually annotated SF for 3 incident languages thus far: 200Kw in Uyghur, and 50Kw in each of Tigrinya and Oromo. This data has been used as evaluation data as part of the LoReHLT 2016 (Uyghur) and 2017 (Tigrinya and Oromo) evaluations. These resources have been released in language packs to participants in the LORELEI program and LoReHLT evaluation participants, and all will be made available to the larger research community as part of the LDC catalog starting in February 2018.

In addition, we have produced a multi-way annotated SF dataset in English, as part of an experiment on the degree of inference that is possible with SF annotation. Annotation for both SSA and SF is on-going in additional languages, and we expect to complete 25Kw of SSA for an additional 12 languages, and 25Kw of SF for an additional 24 languages by the end of 2018.

9. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0123. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

6. Bibliographical References

Baker, C., Fillmore, C., and Lowe, J. (1998). The Berkeley FrameNet project. In Proceedings of COLING/ACL, pages 86–90, Montreal.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider N. (2013). Abstract Meaning Representation for Sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse.

DARPA. 2014. Broad Agency Announcement: I2O Low Resource Languages for Emergent Incidents (LORELEI). Defense Advanced Research Projects Agency, DARPA-BAA-15-04.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, May 24-30. European Language Resource Association (ELRA).

Ikuta, R., Styler, W., Hamang, M., O'Gorman, T., and Palmer, M. (2014). Challenges of adding causation to richer event descriptions. In Proceedings. of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 12–20.

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S. (2014). AIDR: Artificial Intelligence for Disaster Response. In Proceedings of the 23rd International World Wide Web Conference, Seoul, South Korea, April 7-11.

METANET. (2010). META-NET White Paper Series: Press Release, http://www.meta-net.eu/whitepapers/press-release-en, accessed March 16, 2016.

NIST. (2017). Low Resource Human Language Technologies (LoReHLT) evaluations. https://www.nist.gov/itl/iad/mig/lorehlt17-evaluations. August 2017. Accessed September, 2017.

Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, 31:1.

Rehm, Georg, Hans Uszkoreit, eds. 2012. META-NET White Paper Series: Europe's Languages in the Digital Age, URL: www.meta-net.eu/whitepapers.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From Light to Rich ERE: Annotation of Entities, Relations, and Events. In Proceedings of the Fourteenth Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Third Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Denver, Colorado, June 4.

Strassel, S., Bies, A. and Tracey, J. (2017). Situational Awareness for Low Resource Languages: the LORELEI Situation Frame Annotation Task. In Proceedings of the First Workshop on Exploitation of Social Media for Emergency Relief and Preparedness, Aberdeen, Scotland, April 9.

Strassel, S., and Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3273-3280, Portoroz, Slovenia, May 23-28. European Language Resource Association (ELRA).

White, A., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., and Van Durme, B. (2016). Universal decompositional semantics on universal dependencies. In Proceedings. of EMNLP. pages 1713– 1723.