Situational Awareness for Low Resource Languages: the LORELEI Situation Frame Annotation Task

Stephanie M. Strassel, Ann Bies, and Jennifer Tracey

Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA
{strassel,bies,garjen}@ldc.upenn.edu

Abstract. The objective of the LORELEI Situation Frame task is to aggregate information from multiple data streams – including social media – into a comprehensive, actionable understanding of the basic facts needed to mount a response to an emerging situation. Rather than evaluating these capabilities in English, LORELEI is particularly concerned with advancing human language technology performance for low resource languages. The combination of domain, genre and language requirements make creation of linguistic resources for LORELEI in general, and the Situation Frame task in particular, especially challenging. Data is by definition relatively scarce for these languages, and real operational data may be impossible to come by, necessitating the use of "proxy" data sources. The annotation task itself, while superficially straightforward, requires navigating many difficult decisions involving the use of inference and the presence of widespread ambiguity and under-specification in the source data. We introduce the Situation Frame annotation task in the context of the goals of the larger LORELEI program, explore some of the most prevalent annotation challenges, and discuss the impact of various data types on annotation consistency. The data described in this paper will be made available to the wider research community after its use in LORELEI program evaluations.

Keywords: Low Resource Languages, Situational Awareness, Situation Frame Annotation, Linguistic Resources

1   Introduction

Over the past decade, human language technology (HLT) programs in the US have often included a focus on performance improvements in informal genres, with a particular emphasis on social media data including SMS and chat, blogs, newsgroups, and microblogs like Twitter [1, 2]. Another recent thrust has been HLT for low resource languages (LRLs) [3, 4, 5]. The DARPA LORELEI Program is particularly concerned with building HLT for low resource languages in the context of emergent situations like natural disasters or disease outbreaks. LORELEI systems are required to process information about topics, entities, events and sentiment in foreign language data, with the goal of providing situational awareness within days or even hours of the outbreak of an incident.

Linguistic resources are a core component of LORELEI, and the program is developing language packs comprising data, annotations, NLP tools, lexicons and grammatical resources for 23 Representative Languages (RLs) and 8-12 Incident Languages (ILs) [6]. Rather than providing training data tailored to a specific set of evaluation tasks and languages, the RLs are designed to enable system developers to leverage language-universal resources and project from related-language resources; to that end the RLs have been selected to provide broad typological coverage. The IL language packs are designed for development and testing of LORELEI system capabilities, and the choice of evaluation ILs remains unknown until the start of the evaluation.

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

Table 1. LORELEI Languages, with dev ILs in *italics* and eval ILs in **bold**

Language packs include several million words of monolingual text covering news, blogs, discussion forums, microblogs and similar genres, with a goal of at least half of the data coming from informal genres; the actual distribution reflects data availability for each language. A portion of the collected data is translated (using a combination of professional, crowdsourced and found parallel text), and a subset of the translated data is further annotated for entities, syntactic structure and predicates and argument roles. Language packs also include basic natural language processing (NLP) tools like tokenizers, sentence segmenters, taggers and encoding converters, along with a 10,000-lemma lexicon and an assortment of grammatical resources. The RL language packs include the full suite of components, while development IL language packs contain only a subset. Evaluation IL language packs add test data used to score system performance against a human reference. The first LORELEI evaluation in 2016 included three component tasks: Machine translation (MT), Named Entity Recognition (NER) and Situation Frame (SF). In 2017 the NER task will be replaced by an evaluation of Entity Discovery and Linking (EDL). While the LORELEI evaluation is restricted to Program performers, the National Institute of Standards and Technology (NIST) has introduced a parallel evaluation campaign, Low Resource Human Language Technologies (LoReHLT), that is open to all [7]. The 2017 evaluation includes two surprise Incident Languages, constrained and unconstrained training conditions, and three evaluation checkpoints to measure performance based on time and training resources provided.

2   The Situation Frame Task

Of all the LORELEI component evaluations, the Situation Frame task is most directly aligned with the notion of situational awareness. The objective of SF is to enable information from many different data streams to be aggregated into a comprehensive, actionable understanding of the basic facts needed to mount a response to an emerging situation. In the LORELEI use case, being "actionable" requires that information distilled from these disparate data streams must be useful to downstream automated analytics and/or to mission planners responsible for designing assistance efforts. The SF evaluation is centered on the notion of a Situation Frame consisting of three information elements:

1. Characterization of the situation including its type and current status
2. Localization of the situation
3. Indication of any Sentiment, Emotion, or Cognitive State (SEC) for the Situation

Prior to the evaluation, human annotators compile a set of documents in the Incident Language that address one or more specific incidents (e.g. a particular earthquake). This evaluation set is approximately 200,000 words and covers a range of genres, including both formal news and informal social media data. For each incident document, annotators create zero or more Situation Frames that serve as the human reference against which LORELEI systems are scored. Two types of Frames can be created: Need Frames capture information about incident-related needs that (do, did or will) exist along with any response to those needs, while Issue Frames capture information about other societal events that could be relevant to planning a disaster response.

For each frame, annotators characterize the situation by indicating the type of need or issue that exists, selecting from the inventory of allowable types shown in Table 2.

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

Table 2. LORELEI Situation Frame Need and Issue Types

The need and issue types were defined with input from LORELEI stakeholders and by consulting existing annotation schemes including those used by the MicroMappers effort [8]. Every frame is further labeled for the status of the need or issue as well as the status of its resolution, as shown in Table 3. Finally, if known, 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[1].

Next, annotators localize the situation by specifying the place where it occurs. In the current task definition, this specification consists of selecting a single Location, Geopolitical or Facility entity for each situation, or an indication that no named place entity is mentioned in the document[2]. If the same situation in the same place with the same status is discussed more than once in the document, a single frame is created. Otherwise, multiple mentions of a situation (e.g. same place but different status; same situation in different places) require a new frame. To address some potential annotation challenges, annotators add a "Proxy Name" attribute when the exact place where the situation exists is not directly named in the document, but an entity containing that place is named and can stand in as a reasonable approximation of the exact place name.

Need or Issue Status	Need Resolution
Current	Sufficient
Future only	Insufficient/ Unknown sufficiency
Past only	No known resolution

Table 3. LORELEI Situation Frame Status Types

Slot	Value	Required?
Type	Fixed List	Needs + Issues
Description	Free Text	Optional
Place	0-1 named entities	Needs + Issues
Proxy Name	Yes/No	Needs + Issues
Status - Need	Fixed List	Needs + Issues
Urgent	Yes/No	Needs + Issues
Status - Resolution	Fixed List	Need Frames Only
Reported By	0+ named entities	Need Frames Only
Resolved By	0+ named entities	Need Frames Only

Table 4. LORELEI Situation Frame Template

Finally, annotators denote SEC for a situation by indicating whether it is urgent. In future iterations of the SF task annotators will also indicate whether any sentiment is associated with the situation, further labeling the sentiment holder, the target of the sentiment, the sentiment value (positive or negative) and other features.

In addition to characterization, localization and SEC, annotators also provide a brief description for each frame. The description is primarily intended to help annotators keep track of their work, and is not an official part of the frame, so there are no

---

[1] For the current SF task, only entities named in the document can be selected as "reporting" or "resolving" the issue. The rationale behind this is that non-named entities do not contribute actionable information about the situation.

[2] Here too, the current SF tasks limits these entities to those that are named in the document.

restrictions on the description content. The complete template for a Situation Frame is shown in Table 4.

3   Situation Frame Annotation Challenges

Although the SF annotation task is apparently straightforward, there are a number of issues that make it difficult to perform with a high degree of annotator consistency. For instance, it can be difficult to decide between potentially related need types (Infrastructure vs. Shelter, Infrastructure vs. Utilities, Energy, or Sanitation), but detailed annotation guidelines and formal annotator training are used to improve consistency in such cases. A more difficult challenge is ambiguity inherent in the data itself, rather than in the categories used to label the data.

Deciding when to annotate a need or issue as present in a document requires the use of some amount of inference in order to satisfy the goal of providing situational awareness. If too little (or no) inference is used, clear cases of need will be missed, while if too much inference is used, the mere mention of the word "earthquake" could result in a cascade of unwarranted frames for search/rescue, shelter, water supply, utilities/energy/sanitation, medical assistance and infrastructure even when the document does not support such conclusions.

To address this pervasive challenge, annotators are instructed to create a Frame whenever they believe that a need (or issue) is strongly implied or inevitable, even when neither the need nor its resolution is mentioned directly. For instance:

```
A typhoon has demolished the city of Tacloban.
```

Any reasonable reader of this passage would conclude that housing and other infrastructure in Tacloban must have been damaged or destroyed. A reasonable reader of the document might also believe other needs (such as food or water supply or search/rescue) are *possible*, but would probably not conclude that they are *strongly implied or inevitable* given the content of the document, and thus these additional need types should not be annotated. However, even with this additional guidance, reasonable readers of the same document may disagree.

In many cases, even if annotators agree on the question of whether a particular need type exists, the document text may be unclear or ambiguous with regard to designating the place, status, resolution, and urgency. A specific challenge that arises from the (necessary) use of inference combined with ambiguity in the data itself is deciding on the place localization for a given frame. An incident may be discussed in connection with several different places over the course of a document, 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. Because the Situation Frame task requires that each frame be localized to a single named place, if the same need or issue exists in multiple places, annotators must create a separate frame for each distinct place. This requirement means that annotators must find a way to resolve the ambiguity that exists in the text. Consider the following example:

```
Landslides hit Guinsaugon in south of Leyte. Village totally
flattened. Virtually all housing destroyed in Guinsaugon &
surrounding region.
```

Because the document explicitly states that housing was destroyed in Guinsaugon, it seems obvious that the annotator should create a Shelter Need Frame localized to Guinsaugon. However, this passage also discusses housing issues in "the surrounding region" without naming that region. In order to capture this additional information that further contributes to situational awareness, annotators should create a second Shelter Need Frame. The challenge is in deciding on the place name for this second frame: Is the place unnamed, or should the name of the island (Leyte) stand in as a "proxy" for the actual name of the surrounding region. Although annotation guidelines instruct the annotator to choose the most specific name that is applicable to the frame, the decision also rests in part on an annotator's world knowledge – someone who knows that Leyte is an island (as opposed to being, for instance, a district) may take a different approach than someone without that knowledge. The example above contains only one need type and a handful of places, but lengthy documents may contain dozens of inter-connected needs and places, with inexact names and imprecise descriptions of which needs exist where. This challenge is particularly prevalent in news texts, blog or forum posts, bulletins and other lengthy documents that may mention several related places affected by an incident, and it is especially acute in emerging situations where early reporting is often characterized by incomplete or inaccurate information.

In an effort to gain greater insight into the challenges of SF annotation and consider the most appropriate way to use human reference annotations in evaluating machine performance on the SF task, a massively multi-way human annotation exercise is currently in progress. LORELEI performers, evaluators and data providers have jointly annotated a set of 269 documents consisting of news articles, discussion forums, blogs, SMS, Twitter, and "situation reports" (SitReps). The data comprises three subsets: 1) English translations of the Year 1 LORELEI Evaluation Incident Language test set (originally in Uyghur); 2) additional English texts in the disaster domain; and 3) social media data from actual Humanitarian Assistance and Disaster Relief (HADR) operations. All 269 documents have been independently annotated by multiple people, with a core set of 42 documents all labeled by 36 individuals and the remaining documents labeled by at least 6 and up to 16 annotators. The goal of the exercise is both to gain insight into which parts of the annotation produced the largest amount of disagreement and to explore possible scoring methods that would make use of multiple independent reference annotations (as is frequently done in the evaluation of machine translation, for example). This exercise is ongoing and results are not yet available, but overall annotator consistency rates are comparable with the rates achieved by Uyghur annotators in the LORELEI Year 1 SF evaluation[3].

---

[3] We expect results to be available in time for inclusion in the final paper and presentation at the workshop.

4 Situation Frame Annotation in HADR Versus Non-HADR Data

The LORELEI Program, and the Situation Frame task in particular, are concerned with improved situational awareness for incidents involving low resource languages. One of the most central challenges in creating the LORELEI language packs, including the preparation of Incident Language data to support SF and other evaluation tasks, is the paucity of open source operational data for the languages of interest. While the use case for LORELEI is intended to address social media data of the type that existed for Haitian Creole via the Ushahidi platform during the Haitian earthquake in 2010 [9], the reality is that there is little to no pre-existing open source data of this type for most of the LORELEI languages. This means that LORELEI language packs must utilize other types of data, namely news articles (instead of SitReps) and Twitter (instead of SMS), plus blogs, discussion forums and similar types of informal web text – and even these proxy sources may be scarce for some languages. Because these (relatively) more prevalent data sources are standing in for actual operational data in the LORELEI language packs, it is important to understand their differences and similarities to operational data. The Situation Frame annotation exercise provides an opportunity to explore this question in some detail, given its inclusion of both HADR data (SitRep and SMS) and non-HADR data (news articles, Twitter, blogs and newsgroups).

One type of HADR data included in the Situation Frame exercise was Situation Reports, or SitReps. In many ways, SitReps are similar to news texts in the non-HADR data: they present a comprehensive view of an incident, though the primary focus is on the intervention effort itself. Unlike many breaking news reports, SitReps do not generally include any kind of "on the ground" reporting from the perspective of individuals experiencing the need or requiring assistance. Another limitation compared to non-HADR news texts is that SitReps by very their nature focus on information that is already processed and already receiving some sort of HADR response, which makes them less useful for identifying new or emerging needs or issues. (This means for instance that we would expect to see fewer "current/urgent/unresolved" needs expressed in SitRep data compared to other genres.)

Needs or issues are often mentioned in SitReps only in terms of the HADR response to that need, but inferring the frames is generally straightforward. For example, a SitRep might detail the construction of a temporary medical facility in a particular location:

```
A US Disaster Medical Assistance Team set up a field hospital
where the Israelis were based, at the edge of Cite Soleil,
and are now open and receiving patients.
```

Although there is no explicit mention of a Medical Assistance need, it is straightforward to infer the need from the response (and thus create a Need Frame).

To an even greater extent than non-HADR news texts, entities in SitReps are usually named, often in a complete or near complete form, and often with accompanying

acronym or abbreviated forms, which makes SF annotation somewhat more straight-forward than for other data types. This means that most frames created for SitRep data are complete (i.e. all possible "slots" in the Frame can be filled in) since the place and the entities involved with the frame are all clearly mentioned in each document. As a whole, SitReps are straightforward to annotate with reasonably high consistency.

Social media data differs from the SitRep and news text data in its prevalent inclusion of immediate, frequently local, often eyewitness information from individuals. This information is particularly valuable to situational awareness when those individuals are directly involved with the disaster situation. However, this type of data also presents many challenges for annotation. Both the HADR SMS data and the non-HADR Twitter data include prevalent use of abbreviations, typos, non-standard orthography and language use which may not be familiar to annotators. Many Tweets and SMS messages contain no named entities (people or places), and the names that do occur may not be easily identifiable due to unexpected spelling or abbreviations. This is also true of blogs and forum posts, though to a lesser degree. When they do occur, names often refer to hyper-local entities (e.g. people requiring assistance or directly involved with the situation), which can be difficult for annotators without deep world knowledge to contextualize. The result of these factors is that SMS, Twitter, blogs and discussion forum data on the whole yields many fewer named entities that can be associated with Situation Frames when compared with more formal news articles and SitReps.

Another way in which SMS and Twitter data differs from SitRep and news data is with respect to location information. Often very specific places (such as a particular local hospital where there is a medical need) are named in a Tweet or SMS message, e.g.

```
We received many wounded people, we have no staff and no sup-
plies, we need specialist at Immaculée Conception Hospital.
```

In this example it is straightforward to associate the name of the hospital as the place for a Medical Assistance need frame, but note that there is a lack of information about the city or region where it is located.

Because the Situation Frame annotation task was designed to support annotation of non-operational data of the type that is actually available for LORELEI languages, phenomena that are unique to the HADR data present special challenges. For instance, SMS messages may contain multiple mentions of hyper-local places, as in this example:

```
Hello, I am living in Lamentin on Route 54, zone Labelle, in
the Junette Hotel. Many people are sleeping here. No water,
please
```

Because the annotation guidelines require that each Frame have a single place, anno-tators struggled with several questions: what is the correct extent of the place name starting with "Lamentin"? Should there be one or multiple Water Supply Need frames? If only one frame, should the place be "Lamentin", "Route 54", "zone La-

belle", "Junette Hotel", "Lamentin on Route 54, zone Labelle, in the Junette Hotel", or something else? This type of example rarely if ever occurs in the non-operational data like Twitter, and the SF annotation guidelines do not address it.

Also, the extremely limited context of each individual Tweet or SMS message means that annotators may lack the larger situational context to draw on in making decisions about each frame. Compared to the SitRep and news text data, there are considerably more decisions involving inference with Twitter/SMS data, and annotators may be more inclined to infer frames when there is so little context to the document. In some cases, this involves lexical inference regarding both taggability and type. For example, words like "explosion" or "rebels" in short messages may lead to annotators to infer the existence of Terrorism or Regime Change issues, even without additional context in the tweet. Increased use of inference may also take the form of annotators giving greater weight to potentially external context. For example, a message such as the following:

```
We are in blancha. We have nothing.
```

could imply a lack of water, food, shelter or something else. Annotators who are more inclined to use inference will likely create one or more Need Frames for this Tweet, while more conservative annotators may not create any frames. An annotator's approach will also be affected by their world knowledge about the situation and/or the place. These factors can all contribute to lower annotation consistency, and while more extensive guidelines and/or additional annotator training could remediate this to some extent, this remains a very challenging problem.

5 Conclusions and Future Work

We have introduced the LORELEI Situation Frame task, whose goal is to aggregate information from multiple data streams, including social media, into a comprehensive, actionable understanding of the basic facts needed to mount a response to an emerging situation. LORELEI's particular focus on improved HLT for situational awareness in low resource languages creates unique challenges for data creation. Data scarcity is a real issue, particularly with respect to the types of streaming social media data available to HADR personnel in a real mission scenario. And while the annotation task design is straightforward, the necessary use of inference in decision-making, and the prevalence of ambiguity or lack of specificity in the data itself, mean that annotation consistency remains an elusive goal.

The Situation Frame task was defined in 2016 to support the Year 1 LORELEI component evaluations, and is currently undergoing review with a special focus on questions of annotator consistency and the possible use of multiple human references in system scoring. The basic task definition is expected to remain stable for the Year 2 evaluation, while future evaluations will enhance the task with richer SEC annotation. In addition to the Year 1 test set comprising 200,000 words of Uyghur Incident Data, we have also produced multi-way annotation for an additional 269 English documents that include both HADR and non-HADR data. For the Year 2 evaluation we will an-

notate up to 200,000 words for two separate Evaluation Incident Languages, whose identity will be disclosed at the start of the evaluation in July 2017.

Representative and Development language packs are delivered to LORELEI at the end of each program year. Representative language packs are also deposited in the LDC Catalog as they are completed, while Evaluation Incident language packs (including Situation Frame annotation) are published after they are no longer sequestered for use in LORELEI or LoReHLT evaluations. LORELEI Performers and LDC members receive language packs at no cost. Members of the general research community will pay a minimal fee to defray the costs of data curation, storage and distribution. All deliverables are provided to the government under LDC's existing government-wide license. The first LORELEI language packs will be published by LDC in 2017.

References

1. Olive J, Christianson C, McCary J (eds) (2011) Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation. Springer, New York. doi: 10.1007/978-1-4419-7713-7
2. DARPA BOLT Program. http://www.darpa.mil/program/broad-operational-language-translation. Retrieved January 27, 2017.
3. IARPA BABEL Program. https://www.iarpa.gov/index.php/research-programs/babel. Retrieved January 27, 2017.
4. DARPA LORELEI Program. https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents. Retrieved January 27, 2017.
5. IARPA MATERIAL Program. https://www.iarpa.gov/index.php/research-programs/material. Retrieved January 27, 2017.
6. Strassel S, Tracey J (2016) LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France.
7. NIST LoReHLT Evaluations Website. https://www.nist.gov/itl/iad/mig/lorehlt-evaluations. Retrieved January 27, 2017.
8. Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) AIDR: Artificial Intelligence for Disaster Response. In: WWW'14 Companion. International World Wide Web Conference Committee (IW3C2). doi: 10.1145/2567948.2577034
9. Norheim-Hagtun I, Meier P (2010) Crowdsourcing for Crisis Mapping in Haiti. Innovations: Technology, Governance, Globalization Fall 2010, Vol. 5, No. 4: 81–89. doi: 10.1162/INOV_a_00046