Mach Translat (2018) 32:127–142
https://doi.org/10.1007/s10590-017-9204-4

The ELISA Situation Frame extraction for low resource languages pipeline for LoReHLT'2016

Nikolaos Malandrakis[1] · Anil Ramakrishna[1] · Victor Martinez[1] · Tanner Sorensen[1] · Dogan Can[1] · Shrikanth Narayanan[1]

Received: 31 May 2017 / Accepted: 2 October 2017 / Published online: 22 October 2017

Abstract This paper describes the Situation Frame extraction pipeline developed by team ELISA as a part of the DARPA Low Resource Languages for Emergent Incidents program. Situation Frames are structures describing humanitarian needs, including the type of need and the location affected by it. Situation Frames need to be extracted from text or speech audio in a low resource scenario where little data, including no annotated data, are available for the target language. Our Situation Frame pipeline is the final step of the overall ELISA processing pipeline and accepts as inputs the outputs of the ELISA machine translation and named entity recognition components. The inputs are processed by a combination of neural networks to detect the types of needs mentioned in each document and a second post-processing step connects needs to locations. The resulting Situation Frame system was used during the first yearly evaluation on extracting Situation Frames from text, producing encouraging results and was later successfully adapted to the speech audio version of the same task.

Keywords Situation Frames · Text classification · Topic classification

1 Introduction

The efficient and timely response to needs arising during mass emergencies, such as natural disasters, is of critical importance to the affected parties. Collecting reliable information on the unfolding events, that can be used to guide the Humanitarian Assistance-Disaster Relief (HA-DR) efforts, is a very significant and difficult part of

✉ Nikolaos Malandrakis
malandra@usc.edu

[1] Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles, CA 90089, USA

the process. Information collection can be hampered by the challenging prevailing conditions and becomes especially difficult when faced with language barriers, as in the case when emergency responders have to tend to a situation in a foreign country. Given the significance of the task, it is no surprise that the scientific community has been investigating methods of information extraction from digital media, including but not limited to the text from blogs and social media posts (Vieweg et al. 2010), to assist with situational awareness in emerging situations.

DARPA's LORELEI (2015) (Low Resource Languages for Emergent Incidents) Program focuses on the creation and adaptation of language technologies for low-resource languages, with a primary use case of information extraction for situational awareness and resource deployment in emergency situations. The development scenario is one of a sudden mass emergency in a geographical region with an unfamiliar language for which there are limited to no resources or tools, therefore resources need to be collected and tools developed under strict time constraints and used to extract information that can aid the humanitarian assistance response. This information takes the form of Situation Frames (SF), structures with fields identifying and characterizing needs. The program requirements mandate the rapid development of systems that can process text or speech audio from a variety of sources, including newscasts, news articles, blogs and social media posts, all in the local language, and populate these Situation Frames. While the task is very similar in nature to others in literature, it is defined by the very limited availability of data which is really the primary challenge: systems have to function with very little data overall and no annotations in the target language. This lack of data necessitates the use of simpler but more robust models and the utilization of comparable resources to augment the data available.

This paper describes the Situation Frame part of the processing pipeline developed by team ELISA, a large collaboration of laboratories participating in the DARPA LORELEI program. The overall ELISA pipeline contains components that perform machine translation (MT), named entity recognition (NER) and automatic speech recognition (ASR), in addition to the Situation Frame components described in this paper. The Situation Frame component uses all other components in the pipeline as inputs and produces Frames containing need types and locations as outputs. The following sections describe the problem, detail our approach and discuss the results of the first official evaluation (Tong et al. 2016) and some post-evaluation analysis.

2 Problem definition

This section describes the problem we are addressing, the task formed around it and the evaluation conditions and metric.

2.1 Situation Frames

Situational awareness information for DARPA LORELEI is organized in Situation Frames (Strassel and Tracey 2016). Situation Frames (SF) are structures with multiple fields, similar to the frames commonly extracted by natural language understanding (NLU) systems, with each frame corresponding to a single need affecting up to one

Table 1 Situation Frame types

Needs
Evacuation
Food supply
Search/rescue
Utilities, energy, or sanitation
Infrastructure
Medical assistance
Shelter
Water supply
Issues
Civil unrest or wide-spread crime
Elections, politics and regime change
Terrorism or other extreme violence

location. The definition of Situation Frames and of the fields comprising them has been evolving over time as the program matures. The current definition of a frame fields includes: a situation *Type* selected from the fixed inventory shown in Table 1, the *Place* affected by the situation (if a location is mentioned by name) and *Status* variables describing additional parameters (time, resolution and urgency). Types mostly correspond to *needs* that require action to be taken by the emergency responders, e.g. sending food or water, with some additional types describing *issues* that may adversely affect the assistance efforts, e.g. civil unrest may hinder the delivery of food or water. No types of needs are considered unless corresponding to one of the values in the need inventory. The Place field is populated by a named entity corresponding to the affected location or geopolitical entity, but may be vacant if no location is explicitly mentioned in the source data. Each frame can only contain a single Type and up to one place, so multiple SFs need to be produced if a need affects multiple places or multiple needs affect the same place.

A sample Situation Frame is shown in Fig. 1. It includes the identification of the document it was extracted from, the Type of need detected, a place described by character offsets in the document and an entity type (GPE for geopolitical entity or LOC for a location) and finally a confidence score in [0, 1].

2.2 Task parameters

Given text documents, from a variety of sources, in an incident language (IL), a Situation Frame system should be able to process them and return a collection of SFs containing the appropriate fields, plus the IDs of the documents containing relevant information and confidence scores. The development of such a system should be possible even with no annotated data in the IL, though annotated data in other languages are allowed. Text in the IL, including parallel text is available, but not SF annotations.

Fig. 1 A sample Situation Frame

```
{
   "DocumentID": "IL3_NW_0312",
   "Type": "water",
   "TypeConfidence": 0.76
   "PlaceMention": {
      "Start": 0,
      "End": 4,
      "EntityType": "GPE",
   },
}
```

The task is one of many supported by the DARPA LORELEI program, including NER and MT. The program regularly releases data packs in various languages and conducts competitive evaluations. The target IL(s) is announced on the start of the evaluation period and monolingual and parallel text are released to all participants, to be used on all tasks, but no annotations of any kind are provided. Using IL data acquired before the start of the evaluation period is allowed, but no further data may be collected after the IL has been announced. All participants have limited access to a native informant (NI), a native speaker of the IL who can be asked to perform virtually any task apart from annotating the evaluation set. The NI can be useful, but because he or she is not an expert in any relevant field there are limits to can be achieved in the allotted time. Performance is evaluated at predefined checkpoints from the beginning of the evaluation period on a fixed evaluation set and more development data are released to the participants after each checkpoint. Any system will need to be developed or adapted within a few days in order be ready for the first checkpoint, but further development is allowed for subsequent checkpoints.

2.3 Evaluation

Systems produce collections of SFs which are evaluated against a ground truth. Situation frames are evaluated in *layers* with each layer taking into account more information and requiring it to be correct for a sample to count as a hit. The first layer *Type* only checks the document ID and SF Type of each frame; all other elements are ignored. The second layer *Type+Place* includes the place mentions, so for a frame to be correct it needs to have the correct document ID, SF Type, entity character offsets and entity type. Further layers follow the same reasoning but are beyond the scope of this paper.

The metrics used by the program are: precision, recall, f-score and a custom metric called SF Error. A set of evaluation metrics is produced per system, per layer. Hopefully precision, recall and f-score are familiar to the reader, but SF Error probably is not. It is defined as:

$$SFError = \frac{spurious + deleted}{cardinality\ \ of\ \ ground\ \ truth}. \tag{1}$$

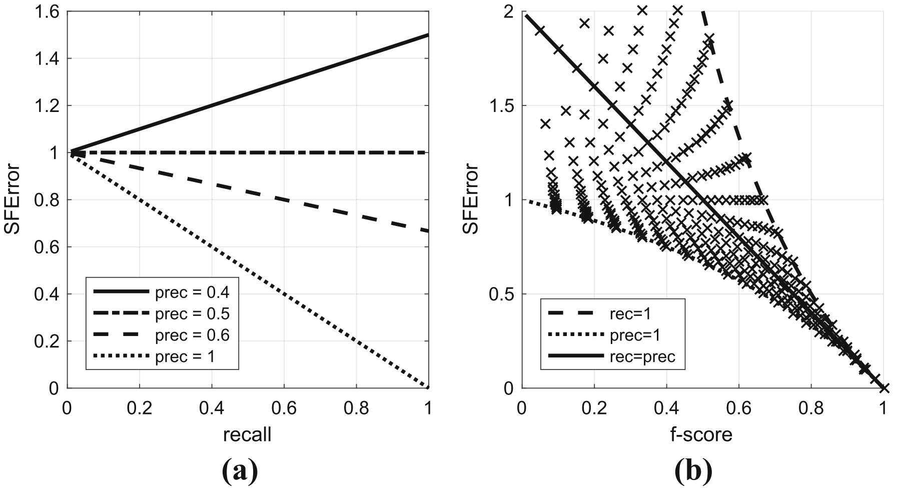

Fig. 2 SFError as **a** a function of precision and recall and **b** as a function of f-score

It is the total number of errors in the evaluation set, divided by the total number of frames in the ground truth. It is unbounded due to the unconventional normalization. A system that outputs nothing achieves an SFError of 1 (number of deletions equal to the cardinality if the ground truth), while a system that outputs many false positives can have an error rate in the dozens. It can be represented in terms of confusion matrix elements (TP,FP,FN) and micro-averaged precision and recall as follows:

$$SFError = \frac{FP + FN}{TP + FN} = 1 - \left(2 - \frac{1}{prec}\right) rec, \qquad (2)$$

which means achieving an SF Error better than 1 requires producing more true positives than false positives or equivalently precision above 0.5. Visualizations of SFError versus precision, recall and f-score can be seen on Fig 2. If a system achieves precision under 0.5, then increased recall *increases* SFError, so if a system can not reliably reach 0.5 precision, the next best strategy is to target 0 recall - output nothing. If a system achieves precision higher than 0.5 then the optimal strategy is to aim for higher recall, which is both more beneficial to the overall score and easier to attain than extremely high precision.

SF Error was the primary evaluation metric for the first year of the program, so it was the metric we tuned for and had an effect on design decisions. It has since been supplanted by f-score, but keeping in mind the particular characteristics of SF Error helps understand some of the choices we made.

3 Our approach

In terms of requirements, the task of SF extraction seems very similar to NLU slot-filling, but with very strict limitations on data availability. In keeping with the program

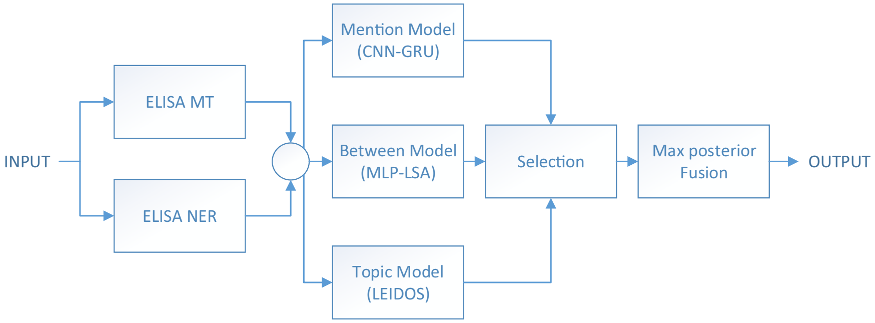

Fig. 3 Overview of the ELISA processing pipeline. The SF detection is performed by a combination of three models

goals, a system has to manage without task-specific data in the target language, however in the first iteration of the task we had access to very limited task data in *any* language. When designing the system and experiments described in this paper we chose to focus on the lack of task data rather than the projection across languages; we did not feel we had enough data or knowledge in any language to make the cross-lingual projection worthwhile.

Given the requirements of the program we started with some reasonable assumptions: we would have access to machine translation and named entity recognition developed by our team that could be used to pivot across languages and as a critical component of assigning locations to SFs, respectively. We could also potentially get very limited task data in the target language or English, using the native informant. Taking into account the data and time constraints we chose to develop a SF pipeline that works on translated English text, rather than one that processes the target language directly. An overview of the overall ELISA pipeline and the SF components is shown in Fig. 3. Describing the NER and MT components of the pipeline is beyond the scope of this paper, but more information about the MT components can be found in Papadopoulos et al. (2017) and about the NER components in Zhang et al. (2016) and Pan et al. (2017) and a more task oriented version of the pipeline can be found in Hermjakob et al. (2017). While this paper describes only the development and application of this pipeline to Situation Frame extraction from text, a simplified variant of the same approach was later adapted to perform the same task on speech audio. The speech version of the task is significantly different at this time and better described in Malandrakis et al. (2017), while the team ELISA speech pipeline for that evaluation is described in Papadopoulos et al. (2017).

The following sections describe the data used, including a corpus we annotated, and the models developed.

3.1 Data

This section describes all data used to develop and evaluate the models we used, as well as conduct all described experiments.

The ReliefWeb corpus (ReliefWeb 2016) of disaster-related documents was used to train models. The corpus contains disaster-related documents from various sources annotated for *theme* and *disaster type*, where theme labels are similar to topics discussed (food, water). Combining the theme and disaster type labels in a flat list created an inventory of 40 categories and the task of multi-label classification of documents into these categories was used as a proxy to the SF task, to train or pre-train models. Overall the corpus, in the version we used, contains 423,790 English, 30,728 Spanish and 46,115 French documents. In addition, to accommodate our use of machine translation as input we acquired translations, through Google Translate, of 1891 Spanish and 1292 French documents.

We used the publicly available GloVe word embeddings (Pennington et al. 2014) to initialize our neural networks. The HA/DR lexicon (Horwood and Bartrem 2016), containing terms separated into various disaster-related categories was used for data selection and to create the data we would ask the native informant to annotate. An internal dataset of about 3000 annotated English tweets was used to train models (detailed below). We also created a corpus of news articles related to disasters by querying Bing News search. The queries posed were manually authored to align with the SF Type inventory and the top results were collected and cleaned. The resulting corpus contains roughly 3000 documents each belonging to one or more topics. This corpus was used to train models, but was collected very late during the evaluation period so was not used by all models due to time constraints.

Finally, we used some SF corpora provided by the program (Strassel and Tracey 2016). The main development set was a corpus of 132 Mandarin documents, including reference translations and entity and SF annotations. We augmented the Mandarin set by getting translations for the documents from Google Translate, Bing translate and the ELISA MT system, creating four English versions of each document (including the reference translations). Evaluation was conducted on a corpus of Uyghur documents, of undisclosed size. After the evaluation period was over, the organizers released a subset of the evaluation set, including reference translations, entities and SF annotations. This Uyghur *unsequestered* set contains 199 documents and was used post-evaluation to conduct further analysis.

3.1.1 Internally annotated tweets

We performed annotations on English tweets according to LDC guidelines. A total of 4030 tweets were collected from a set of about 50 million tweets related to hurricane Sandy. The selection was performed using terms selected from the HA/DR lexicon: the intent was to collect tweets that simple keyword searches would identify as containing a Situation Frame, but which most probably did not contain any, and in the process create a collection of strong negative exemplars that would boost our systems' precision.

The task of annotating SFs was deemed too complicated to crowd-source and was instead performed by four members of our development team, with each tweet annotated by two people. Initially we wanted to create a ground truth by enforcing complete agreement, ignoring any tweet where the two annotations differ in any way. That was revised to complete agreement at the Type layer, but not including localization, due to the resulting very high disagreement. While a conscious attempt was made to limit annotations to situations explicitly mentioned, with little allowance for implied or inferred situations, in practice it proved difficult to enforce and annotation agreement suffered accordingly.

The derived ground truth contains 2934 tweets and, despite the problems encountered, fulfilled the intended function of providing us with negative samples and improving overall system performance.

3.2 The models

The following section describes the models and combinations thereof developed for the task. Overall we developed six models, though only three are part of the main pipeline, while the other three were developed as baselines.

The main three models were designed so as to target SFs with different scopes, requiring different amounts of information to trigger. At the top level we have a compositional topic model (LEIDOS), that will only trigger if an SF Type is one of the main topics of the document. On the other end of the spectrum we have a compositional SF model (CNN-GRU) that can trigger with as little as a single Type-related keyword. The middle ground in terms of sensitivity is covered by a bag-of-words SF model (MLP-LSA). These models were expected to produce SF sets with limited overlap that would facilitate model combinations. All model hyper-parameters were tuned by grid searching with SF Error rate as the tuning criterion. They were intended to be applied to machine translated text, and therefore were largely trained using translated text, effectively incorporating any translation noise into the training process.

The LEIDOS model is a compositional CNN-GRU, similar to the model described in Lai et al. (2015), that accepts input documents as sequences of 1-hot vectors and uses a CNN to compose word embeddings into sentences and a single forward GRU to compose sentences into documents, as shown in Fig 4a. The choice of using a CNN rather than a recursive component was task-driven: we designed this model to be applied to translated text and the stricter word order required by an RNN or LSTM would become a problem. The final layer is composed of 40 binary classifiers, each corresponding to one topic or disaster type from the ReliefWeb inventory. To apply to the LORELEI SF task we simply created a deterministic mapping from some ReliefWeb categories to LORELEI Types. The initial mapping was:

- ‘Food and Nutrition’ → ‘Food Supply’
- ‘HIV/Aids’, ‘Health’, ‘Epidemic’ → ‘Medical Assistance’
- ‘Drought’, ‘Water Sanitation Hygiene’ → ‘Water Supply’
- ‘Shelter and Non-Food Items’ → ‘Shelter’,

which meant it was only capable of producing a small subset of all possible Types. It was eventually augmented with more labels, tagged using manually authored combinations of regular expressions, e.g. “terroris.*”. After this expansion the model could also produce the “Evacuation”, “Terrorism or other Extreme Violence” and “Search/Rescue” Types. It was trained on the ReliefWeb corpus and the word embeddings were initialized using GloVe.

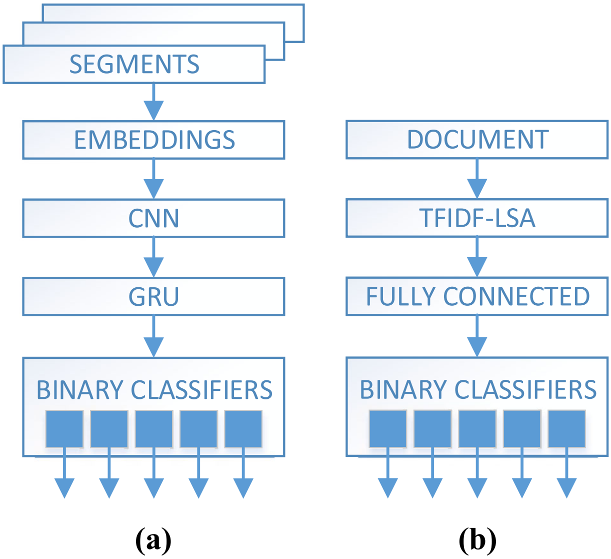

Fig. 4 Model architectures overview, for **a** the LEIDOS and CNN-GRU models and **b** the MLP-LSA model

The CNN-GRU model is the LEIDOS model, re-trained specifically for the SF task. It shares the same topology as the LEIDOS model, however to accommodate usage with very limited amounts of data the network components have much lower dimensionalities. It was trained using the ReliefWeb corpus and GloVe embeddings, then re-trained using the Mandarin development set (including all translations) and the internal set of annotated tweets. Re-training was performed by replacing the final layer of binary classifiers and re-training the entire network.

The MLP-LSA model is a multi-layered perceptron applied to Latent Semantic Analysis (LSA) (Deerwester et al. 1990) document vectors, as shown in Fig 4b. The TF-IDF and LSA transformations were learned using the ReliefWeb corpus, which was also used to train the network. It was then adapted to SF, by replacing the final layer of binary classifiers and re-training using the Mandarin development set (including all translations) and the internal set of annotated tweets.

The entity-based model is a gradient boosting ensemble of L2-regularized decision trees designed for learning document level Situation Frames from the context in which named entities are used within a document. This was our only model using a bottom-up approach, that is, trying to label a document starting from the localization. It was trained on the entity mentions found in the Mandarin SF corpus and uses sentence-bound context windows around found location entity mentions as inputs. The features used include word embeddings and LSA vectors extracted over the context windows.

The lexicon based models (svm and logreg) are simple linear classifiers trained on bag of word features extracted at the document level from two lexical resources: the HA/DR lexicon and a ReliefWeb-derived term lexicon. These models were meant to act as baselines; they should give us an idea of what kind of gains (or losses) we are achieving by using more complicated machine learning algorithms. The final submissions use an SVM model for the ReliefWeb features and logistic regression (logreg) for the HA/DR features. They were trained using the Mandarin development set, the internal set of annotated tweets and the corpus of Bing News articles.

3.3 Model combinations

The main models were developed to be highly biased towards precision and combinations were used to boost recall and the overall SF Error score. We used two combinations over the course of the evaluation. The first, $SYSCOMB$, was composed of whichever of the main models were deemed to have high precision at the current experimental conditions, which for the purposes of LoReHLT'16 meant it was a combination of LEIDOS and MLP-LSA; the CNN-GRU appeared not robust to the MT produced by our team. The second, $Optimist$, was merely composed of all three main models, unconditionally. In both cases the combination was performed by maximum posterior probability fusion: the posterior probability of each SF Type was set to the maximum value assigned by any of the constituent models. The result was equivalent to the union of the SF sets produced by each model at the Type layer, but not so at the Type+Place layer.

3.4 Localization

Most of the models described above are top-down: they consume the entire document and produce document-level Type labels. To localize, we use a simple solution of creating location-specific sub-documents and attempting to classify them using the same models. For each detected location mention, provided by the ELISA NER system, we collected all sentences/segments that contain said entity mention and formed a dummy "document" per entity. These dummy documents were passed through the SF model again, creating a set of Type labels per location. The entity-level Types were filtered by the document-level Types: Types not detected during the document-level pass were not allowed at the entity level. Finally, a third pass was performed to restrict the localization of each Type to a maximum of two locations each, selected by a maximum posterior criterion if there were more than two candidate entity mentions per Type. If no entity mention was connected to a Type that was detected at the document level, then a non-localized frame was created instead.

3.5 Native informant use

We considered using the native informant to annotate a few documents and use their input as part of reinforcement learning. That idea was abandoned due to time constraints: training an annotator to perform the SF task requires significant time. Instead we devoted our allotted NI time to improving the machine translation as it pertains to the detection of Situation Frames. We used the ReliefWeb corpus, HA/DR lexicon and

Mandarin SF corpus to select English terms relevant to the task, mostly bigrams with a few unigrams and trigrams, and had the native informant translate them to Uyghur. The term list was selected by using a combination of class-relevance and frequency and post-edited by hand. It was then adjusted to the development data provided by the organizers with each checkpoint: terms existing in the released parallel data were removed and terms associated to Types appearing frequently in the same data (types determined by our models and visual inspection) were moved up the list. In total we had three NI sessions, spanning 1, 2 and 2 h respectively. In all three sessions the task was the same: translate salient n-grams from English to Uyghur to help with MT. We collected translations for about 125, 150 and 200 terms during the three sessions, respectively.

4 Results

The official evaluation was conducted over a period of one month in July-August 2016. The target IL (Uyghur) was announced on July 6. System submission were required for three checkpoints on July 13, July 20 and finally August 3. After the IL announcement and every checkpoint, apart from the final one, further development data were released to the participants.

At the beginning of the evaluation we had trained versions of all SF models, to be modified as deemed necessary. Over the course of the evaluation, the greater ELISA team was working in parallel on the available tasks of Machine Translation, Named Entity Recognition and Situation Frames and, as the final consumers of everything produced, we had to adapt to these constant modifications. The main concern was Machine Translation and how the SF models would cope, particularly at the early stages when MT performance was expected to be low. Our approach was to continuously evaluate the SF models against each other on every MT revision, using the Uyghur development parallel data. While we had no SF annotations for these documents, we could inspect the model outputs manually and we could compare them to each other: they were expected to detect different frames, but the relative cardinality of the produced SF sets was a good indicator of whether each model remained precise given the current MT. If a model was trusted to produce in a high precision output, it was included in the primary SYSCOMB system.

We attempted to improve the SF models during the evaluation period, with little success. The main incident of the evaluation scenario was an earthquake (this information provided by the organizers), so we created earthquake-specific models by filtering the training data for earthquake related documents, but that failed to produce an improvement on the Mandarin SF dataset (also about an earthquake): the datasets contain documents that are not about the primary incident, limiting the utility of incident-specific models. Multiple strategies of supervised model fusion were evaluated, but invariably lead to over-fitting, so we returned to the unsupervised maximum posterior fusion.

We submitted results for all checkpoints for all models apart from logreg. The main three models were virtually unchanged during the evaluation, but there were some changes in the combinations. The SYSCOMB combination was our primary submis-

Table 2 Official evaluation results at the Type layer

System	SFError			f-Score		
	CP1	CP2	CP3	CP1	CP2	CP3
CNN-GRU	1.763	2.000	2.405	0.150	0.205	0.205
MLP-LSA	**0.994**	**0.983**	**0.966**	0.013	0.245	0.215
LEIDOS	1.195	1.294	1.507	0.217	0.222	0.216
SYSCOMB	1.189	1.294	1.505	**0.224**	0.222	**0.296**
Optimist	1.975	2.346	2.918	0.215	**0.271**	0.248
Logreg		1.065	1.086		0.153	0.213
svm	1.147	1.289	1.371	0.078	0.142	0.164
Entity-based	14.853	3.331	2.864	0.056	0.075	0.072

System	Precision			Recall		
	CP1	CP2	CP3	CP1	CP2	CP3
CNN-GRU	0.145	0.170	0.153	0.155	0.258	0.310
MLP-LSA	**1.000**	**0.528**	**0.573**	0.006	0.159	0.132
LEIDOS	0.315	0.278	0.225	0.166	0.184	0.208
SYSCOMB	0.323	0.278	0.278	0.172	0.184	0.317
Optimist	0.178	0.197	0.167	0.270	**0.436**	**0.482**
Logreg		0.374	0.387		0.096	0.147
svm	0.198	0.212	0.210	0.048	0.107	0.134
Entity-based	0.030	0.052	0.053	**0.440**	0.134	0.111

Bold indicates the best achieved performance

sion and it was composed of the MLP-LSA and LEIDOS models for checkpoint 1 and 3, but only LEIDOS for checkpoint 2. The results for all models and all checkpoints are shown in Table 2 for the Type layer and Table 3 for the Type+Place layer.

Starting from the Type layer results in Table 2, our systems were highly competitive overall. It is important to note the very different scaling of SF Error compared to the other metrics: it consistently gets worse across the checkpoints, because most models did not achieve the required 0.5 precision. We expected the three primary systems to be much closer to 0.5 precision, but apparently the evaluation set contained a higher proportion of documents containing no frames than anticipated, leading to a higher than estimated ratio of false positives. Most of these models were unchanged during the evaluation, with improvements in performance mostly attributed to the continuously improving machine translation input, the main effect of which was improved recall. Our intuition was that the CNN-GRU model was not precise enough to be included in the SYSCOMB and that was indeed the case, though there are potential benefits with respect to recall and f-score, as seen in the Optimist results. Overall the MLP-LSA proved most precise by a wide margin, the LEIDOS model is designed to be more conservative still, but the imperfect correspondence between the ReliefWeb categories and LORELEI Types clearly has an effect. Finally we should note the relation between the recall scores of the model combinations and their constituent models. SYSCOMB

Table 3 Official evaluation results at the Type+Place layer

System	SFError			f-Score		
	CP1	CP2	CP3	CP1	CP2	CP3
CNN-GRU	1.443	1.521	1.614	0.021	0.044	0.030
MLP-LSA	**0.999**	**1.086**	**1.072**	0.003	0.095	0.033
LEIDOS	1.185	1.191	1.32	0.107	0.116	0.099
SYSCOMB	1.184	1.191	1.384	**0.109**	0.116	**0.108**
Optimist	1.614	1.788	2.022	0.090	**0.144**	0.091
Logreg		1.450	1.568		0.106	0.082
svm	1.684	1.660	1.663	0.041	0.087	0.066

System	Precision			Recall		
	CP1	CP2	CP3	CP1	CP2	CP3
CNN-GRU	0.032	0.059	0.038	0.015	0.035	0.025
MLP-LSA	**1.000**	**0.285**	**0.167**	0.001	0.057	0.018
LEIDOS	0.217	0.225	0.156	0.071	0.078	0.072
SYSCOMB	0.220	0.225	0.152	0.072	0.078	0.084
Optimist	0.103	0.138	0.083	**0.079**	**0.15**	**0.102**
Logreg		0.139	0.098		0.086	0.070
svm	0.048	0.097	0.075	0.036	0.079	0.058

Bold indicates the best achieved performance

is effectively the union of MLP-LSA and LEIDOS at checkpoints 1 and 3 and the achieved recall is almost equal to the sum of the partial recalls, so the MLP-LSA and LEIDOS produced SF sets with very little overlap. Optimist achieved a much higher recall still, by including the CNN-GRU, but the improvement over SYSCOMB indicates that the CNN-GRU has significant overlap with the other two models.

Looking at the Type+Place layer results in Table 3, the first observation is that the results are significantly worse than at the Type layer. That is expected given the increase in task difficulty and the structure of our models: they are top-down so the performance at the Type layer becomes the upper bound of performance at the Type+Place layer. Despite the drop in terms of absolute performance, these model were competitive in terms of relative performance. Most of the observations made at the Type layer hold at the Type+Place layer, however in this case the performance improvement across checkpoints is much smaller. This can be attributed to the simplistic nature of the localization process.

After the evaluation period was over we had the opportunity to explore the effect of machine translation quality on SF creation. We were given access, by the program organizers, to 97 machine translation outputs produced by participants of the LORELEI evaluation which we could process with our SF models to find out whether SF performance improved with MT performance given fixed SF models. To that end, we used our three main models plus the SYSCOMB combination in their final versions and applied to the 97 MT outputs and the 4 reference translations provided by the program. The SF and MT outputs were evaluated using the corresponding metrics, 4 reference BLEU in the case of MT, and we generated graphs and estimated the Pearson

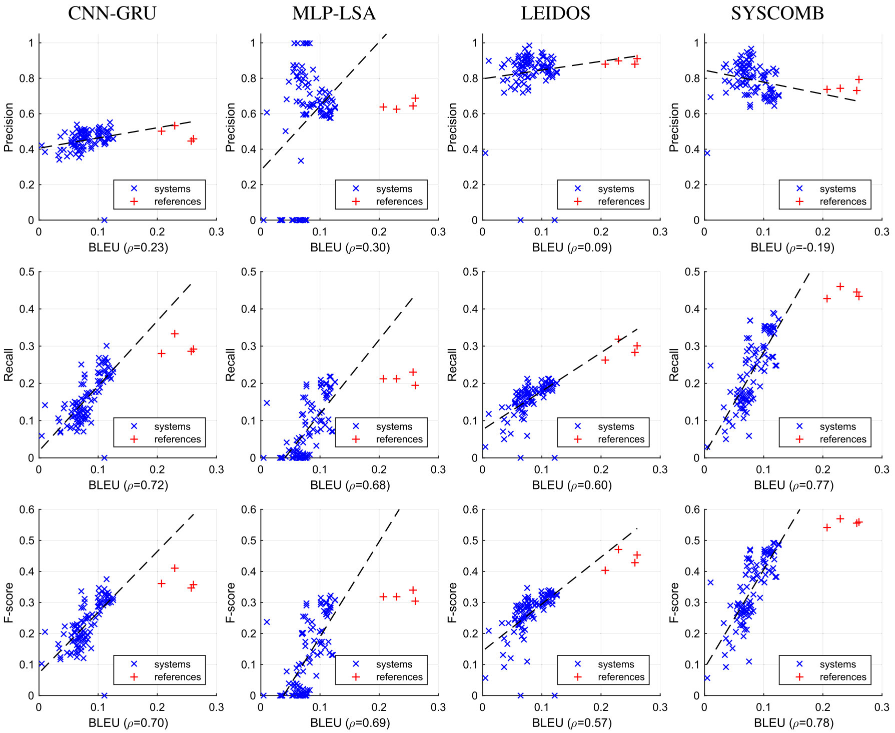

Fig. 5 Situation Frame precision, recall, and f-score as functions of MT BLEU, for the CNN-GRU, MLP-LSA, LEIDOS and SYSCOMB SF models, at the Type layer. Results shown for 97 MT systems and 4 MT references. The Pearson correlation ρ between BLEU and each SF performance metric in parentheses

correlation ρ between SF performance metrics and MT BLEU. In addition to the MT outputs we applied the SF models to the reference translations. MT BLEU for the 4 references was estimated as follows: first we estimated 3 and 4 reference BLEU scores for all MT outputs and took the average delta between 3 and 4 reference BLEU, then we estimated 3 reference BLEU for each of the 4 references and added the previously calculated delta to get 4 reference estimate. The results are shown in Fig 5. Despite the very different nature of the models, all react remarkably different to improved MT performance: precision is relatively unaffected with minor improvements or drops, but recall improves significantly leading to higher f-scores. The trends observed with improved MT outputs extend to the reference translations, that predictably yield the best results. The SF performance delta between MT outputs and references is not as large as expected, but that is probably an artifact of using SF models designed from the ground up for imperfect inputs. Unfortunately we have no access to any information about the MT systems that produced these outputs, which limits the conclusions that can be drawn from these results, but we can confidently state that improved MT performance leads to improved SF creation performance.

Beyond the effect of MT on SF, it is also worth noting that the performance our SF models achieved on this subset of the complete evaluation set is much higher than that achieved on the complete set: we assume that is because the unsequestered subset contains only documents with SFs, whereas the complete set probably contains some documents that don't include any SFs. Improving the handling of these non-relevant documents should lead to significant performance benefits.

5 Conclusions and future work

We developed a variety of models for the task of Situation Frame creation, utilizing machine translation and named entity recognition as black box inputs. The results are encouraging, with the models achieving competitive performance at the Type and Type+Place evaluation layers, though there is great potential for improvement. We expect all our models will see significant gains as more training data becomes available, since the current task data availability is very poor, but there are other obvious improvements being worked on. The models presented were not well equipped to handle non-disaster documents due to their training process, which could lead to many false positives depending on evaluation data composition. Performance improves significantly with machine translation quality, which may pose problems when machine translation is poor, a problem perhaps addressed by adding a translation-independent path to the process. Finally, our approach to localization, while relatively effective, is very simple and could use some expansion.

Regardless, the progress so far is heartening and we are looking forward to the challenges arising as the program nears its goals.

Acknowledgements This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0115 with the University of Southern California. 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.

References

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Hermjakob U, Li Q, Marcu D, May J, Mielke S, Pourdamghani N, Pust M, Shi X, Knight K, Levinboim T, Murray K, Chiang D, Zhang B, Pan X, Lu D, Lin Y, Ji H (2017) Incident-driven machine translation and name tagging for low-resource languages. Machine translation special issue : NLP in low-resource languages

Horwood G, Bartrem K (2016) Lorelei humanitarian assistance/disaster relief lexicon v1 [data file]. Leidos, Inc. Distributed through https://xnet2.nextcentury.com/confluence/pages/viewpage.action?pageId=10650520

Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, pp 2267–2273

LORELEI (2015) DARPA LORELEI website. http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents. Retrieved 25 Oct 2015

Malandrakis N, Glembek O, Narayanan S (2017) Extracting situation frames from non-english speech: evaluation framework and pilot results. In: Proceedings of Interspeech

Pan X, Zhang B, May J, Nothman J, Knight K, Ji H (2017) Cross-lingual name tagging and linking for 282 languages. In: Proceedings of ACL

Papadopoulos P, Travadi R, Vaz C, Malandrakis N, Hermjakob U, Pourdamghani N, Pust M, Zhang B, Pan X, Lu D, Lin Y, Glembek O, Baskar MK, Karafiát M, Burget L, May J, Ji H, Knight K, Narayanan

S (2017) Team ELISA system for DARPA LORELEI speech evaluation 2016. In: Proceedings of Interspeech

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, pp 1532–1543

ReliefWeb (2016) ReliefWeb website. http://reliefweb.int/. Retrieved 31 Mar 2016

Strassel S, Tracey J (2016) Lorelei language packs: data, tools, and resources for technology development in low resource languages. In: Proceedings of LREC, pp 3273–3280

Tong A, Diduch L, Fiscus J, Haghpanah Y, Huang S, Joy D, Peterson K, Soboroff I (2017) Overview of the nist 2016 lorehlt evaluation. Machine translation special issue: NLP in low-resource languages

Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1079–1088

Zhang B, Pan X, Wang T, Vaswani A, Ji H, Knight K, Marcu D (2016) Name tagging for low-resource incident languages based on expectation-driven learning. In: Proceedings of the NAACL-HLT, pp 249–259