INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Extracting Situation Frames from non-English Speech: Evaluation Framework and Pilot Results

Nikolaos Malandrakis[1], *Ondrej Glembek*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
[2]Brno University of Technology, Bozetechova 2, 61266 Brno, Czech Republic

malandra@usc.edu, glembek@fit.vutbr.cz, shri@sipi.usc.edu

Abstract

This paper describes the first evaluation framework for the extraction of Situation Frames - structures describing humanitarian assistance needs - from non-English speech audio, conducted for the DARPA LORELEI (Low Resource Languages for Emergent Incidents) program. Participants in LORELEI had to process audio from a variety of sources, in non-English languages, and extract the information required to populate Situation Frames describing whether any need is mentioned, the type of need present and where the need exists. The evaluation was conducted over a period of 10 days and attracted submissions from 6 teams, each team spanning multiple organizations. Performance was evaluated using precision-recall curves. The results are encouraging, with most teams showing some capability to detect the type of situation discussed, but more work will be required to connect needs to specific locations.

Index Terms: speech recognition, speech analysis, performance evaluation, natural language processing

1. Introduction

During times of mass emergency, such as natural disasters, a variety of critical needs arise that require appropriate response. The effective and efficient deployment of available resources is great importance to Humanitarian Assistance - Disaster Relief (HA-DR) and depends on the timely acquisition of reliable information. Collecting such information can be difficult given the prevailing conditions and language barriers. In recent years researchers have been investigating the extraction of information from digital media, mostly text in the form of blogs or social media posts [1], to assist with situational awareness.

DARPA's LORELEI (Low Resource Languages for Emergent Incidents) Program [2] focuses on the creation and adaptation of language technologies for low-resource languages, with a main use case of information extraction for situational awareness and resource deployment in emergency situations. The operating scenario, of a sudden disaster in a region of the world for the language of which there are limited or no resources, requires the rapid development or adaptation of tools that can extract information used to guide humanitarian assistance efforts. The difficulty of the task should not be under-estimated, as any efforts will have to combine information extraction and methods of knowledge transfer across languages [3][4] to achieve a satisfactory result.

This paper describes the first pilot evaluation, conducted for DARPA LORELEI, on the extraction of information relevant to humanitarian assistance from speech audio. The goal was to develop technologies for processing speech audio into actionable information.

2. Situation Frames

Situational awareness information for DARPA LORELEI is organized in the form of Situation Frames [5]. Situation Frames (SF) are structures, similar in nature to frames used in Natural Language Understanding (NLU) systems, each corresponding to a single incident at a single location. The definition of Situation Frames and the fields comprising a frame have been evolving. At the time of writing a frame includes a situation *Type* taken from the fixed inventory shown in Table 1, the *Location* where the situation exists (if a location is mentioned) and extra variables clarifying the *Status* of the situation (time, resolution & urgency). For the purposes of the pilot evaluation we only took into account the type and location fields. Most types correspond to specific needs that map to the type of assistance required, with some being indicators of prevailing conditions (Issues) that may need to be taken into account when tending to the needs, e.g., civil unrest may hamper medical assistance efforts. Locations are named entities referring to either a location or geopolitical entity, e.g., a city or country. Each frame can reference up to a single location, but may also reference no location if none is named. In the case of a situation affecting multiple named locations, multiple frames would be required.

Table 1: *Situation Frame Types*

Needs
Evacuation
Food Supply
Urgent Rescue
Utilities, Energy, or Sanitation
Infrastructure
Medical Assistance
Shelter
Water Supply

Issues
Civil Unrest or Wide-spread Crime
Elections and Politics
Terrorism or other Extreme Violence

3. Task definition

Given short speech audio segments in the native language of a geographical region where a major incident occurred, a Situation Frame system should automatically identify any SFs included in these segments. A segment may include multiple SFs, although most include zero or one. Each frame produced has to include the audio segment id, the situation Type and optionally a

http://dx.doi.org/10.21437/Interspeech.2017-226

”DocumentID”:	”EVAL_096_004”,
”Type”:	”Medical Assistance”,
”PlaceMention”:	”北京”,
”TypeConfidence”:	0.558

Figure 1: *A sample Situation Frame.*

location. All fields are represented as strings, with locations being the transcribed location names in the native language script. We also required each frame to include a confidence score in $[0, 1]$, corresponding to confidence in its existence, to allow for a curve-based evaluation. A sample Situation Frame is shown in Fig. 1.

Systems were evaluated in layers, with each layer taking into account more information included in the frame. At the first layer, *Relevance*, we evaluated systems on the separation of relevant segments (including at least 1 SF) from non-relevant segments (including no frames). We only took into account the segment id, so a frame was correct if it referred to a segment that included any frame. At the second layer, *Type*, we evaluated systems on the detection of SF types. We took into account the segment id and frame type and both needed to be correct (included in the ground truth). At the third layer, *Type+Place* we added the requirement for localization: for a frame to be correct its segment id, type and location should match with a frame in the ground truth. Note that frames that did not include a location were ignored at this layer.

4. The data

The data are segmented audio recordings from a variety of sources, including newscasts. The data selection process for the evaluation is not public at this time. Data for multiple languages were collected and annotated by Appen [6], resulting in data packs containing roughly 14 hours of audio for each language and the corresponding SF annotations. Participants were provided development data in Amharic (687 segments), Hausa (915 segments), Russian (787 segments), Turkish (2096 segments) & Uzbek (1416 segments). The evaluation was conducted on Mandarin Chinese (724 segments) and Uyghur (883 segments). The datasets contain only segmented audio and the corresponding Situation Frame annotations. No transcripts were provided.

4.1. Annotation

Annotations were performed by native speakers of each language, who were responsible for segmenting and annotating the original audio clips. Segmentation was performed using primarily a semantic criterion, attempting to maintain thematic coherence as represented by SF Types, so ideally each segment would include a maximum of one SF (through that was not always possible). Segments both start and end on a suitable pause, to avoid the truncation of words. If no significant shift in SFs is found, then a maximum segment length of 120 seconds is imposed.

5. Evaluation Measures

The desired operating point (precision-recall trade-off) of an SF system is not fixed and may be affected by many variables, such as the availability of assistance resources, which would lead to varying costs for different types of errors. To evaluate systems at various operating points, we performed a curve based evaluation using precision-recall (PR) curves, with area under the curve (AUC) used as the summary statistic and for ranking overall system performance. Other curves (ROC, DET) were considered, however their use at the Type+Place was problematic, due to the requirement for a True Negative estimate. For each system submission & for each layer of the evaluation a PR curve was generated, with each point of the curve corresponding to a combination of micro-averaged recall and precision. The curves were produced by iteratively sweeping across the confidence values in the system outputs, using 500 quantiles at 0.2% intervals.

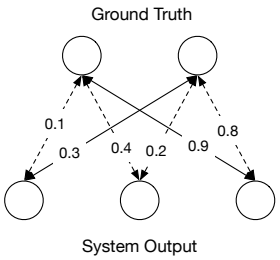

Figure 2: *Alignment of frames based on maximum sum of similarity. Pair-wise similarities shown on arrows. Resulting alignment represented by solid arrows.*

The process which generates a single precision-recall point, for a given confidence threshold, is the following:

1. Remove all frames with confidence scores below the current threshold
2. Transform the remaining frames to the current evaluation layer, by removing extraneous attributes and merging duplicates.
3. Align the ground truth and output frames via maximum similarity
4. Calculate True Positives (TP), False Positives (FP) and False Negatives (FN), then Precision and Recall

5.1. Frame similarity

We expected that location detection would be performed through Automatic Speech Recognition (ASR) which would lead to a high incidence rate of partially correct strings. We wanted to allow that and give partial credit.

To give partial credit we defined Frame similarity, indicated by a number in $[0, 1]$ with 1 indicating a perfect match. The frame similarity between two frames is the product of all field-wise similarity scores. For the Type and segment id fields the field similarity is binary (zero or one) corresponding to a perfect match or no match. For the Location field we used the *Levenshtein ratio* [7] between the 2 location fields, so the similarity $S(w_1, w_2)$ between locations w_1 and w_2 is defined as:

$$S(w_1, w_2) = \frac{l(w_1) + l(w_2) - e(w_1, w_2)}{l(w_1) + l(w_2)}, \qquad (1)$$

where $l(w)$ the length (in characters) of string w and $e(w_1, w_2)$ the character level minimum edit distance between strings w_1 and w_2. The Levenshtein ratio takes values in $[0, 1]$ and the minimum edit distance is calculated using costs of 1 for insertions and deletions and 2 for substitutions.

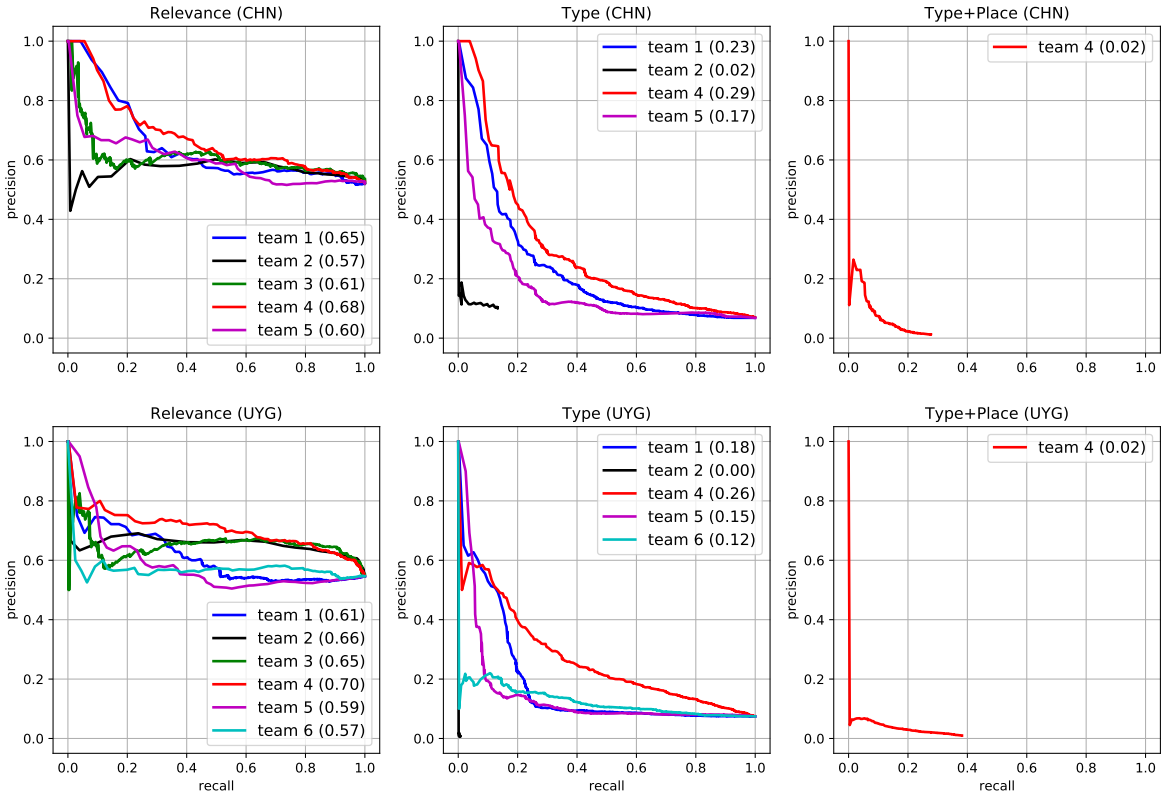

Figure 3: *Results on the Mandarin Chinese (CHN) and Uyghur (UYG) datasets, for the Relevance, Type and Type+Place layers. AUC values in parentheses.*

5.2. Frame alignment

Given that a system output may have different cardinality than the ground truth and the soft similarity used to compare all frames in the ground truth data with all frames in a system output, we needed a way to align the system output and ground truth. To accomplish that we used a maximum similarity criterion: we want the 1-to-1 mapping that maximizes the sum of similarities across the entire dataset. To align we used the Kuhn-Munkres linear sum assignment algorithm [8] applied to a matrix of all pair-wise frame similarities to assign each frame in a system output to a frame in the ground truth. An example of the alignment is shown in Fig. 2.

5.3. Scoring by soft cardinality

The calculation of TP, FP and FN takes into account the soft matching and utilizes soft cardinality. True Positives are calculated as the sum of all similarity scores, False Positives as the cardinality of the system output minus TP and False Negatives as the cardinality of the ground truth minus TP. For the example of Fig. 2 that yields: $TP = 0.9 + 0.3 = 1.2$, $FN = 2 - 1.2 = 0.8$, $FP = 3 - 1.2 = 1.8$. The micro-averaged precision and recall are then calculated using these TP, FP & FN values.

6. Evaluation Process

The pilot evaluation was conducted on two languages, Chinese Mandarin and Uyghur. The participants knew the languages they would be evaluated on before the evaluation started. The two languages represent two different scenarios in terms of available resources. For the Mandarin evaluation participants were allowed to use the multitudes of tools and resources available for the language, such as the large vocabulary ASR that some teams already possessed. For the Uyghur evaluation participants were only allowed to use any Uyghur resource that they had collected before the languages were announced, plus a pre-existing corpus of Uyghur text created for the program by the LDC [5]. Restrictions only applied to the evaluation languages and any resources or tools from different languages could be used freely.

The evaluation was conducted over a 10-day period. The evaluation datasets were released on day one and there was a single submission checkpoint on day 10. There were also development audio datasets released for both languages (951 segments for Uyghur, 901 segments for Mandarin), that included only segmented audio and no annotations of any kind. Each team was allowed to submit 1 primary and 2 contrastive systems for each evaluation language, with the primary systems used for inter-team comparisons.

Over the course of the evaluation period all teams had access to a *native informant (NI)*. The native informant is a native speaker of the incident language, a non-expert, naive to the task. Teams were free to ask the NI to perform any task they found useful, such as annotations of any kind, including transcriptions and translations, but were not allowed to have the NI annotate or otherwise access the evaluation data. Access to the NI was provided over voice call and each team was assigned two hours of NI time.

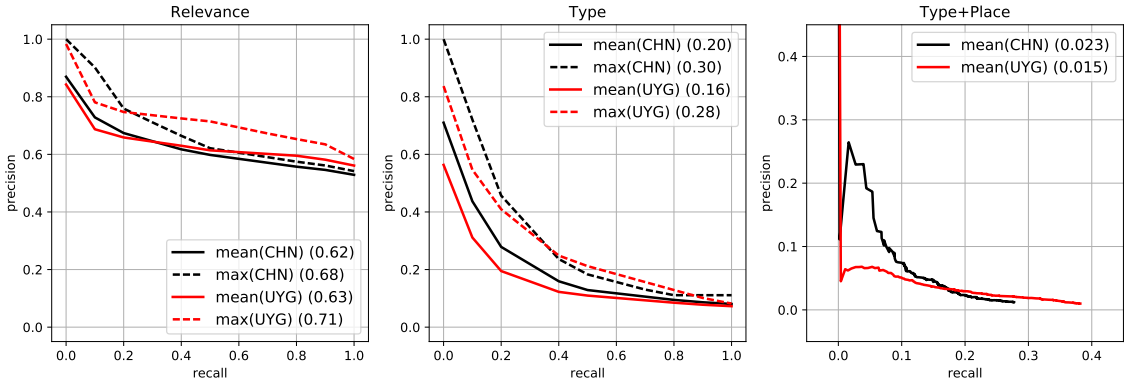

Figure 4: *Comparison of mean and maximum performance per language. AUC values in parentheses.*

7. Evaluation results

In total we received submissions from 6 teams, de-identified for the purposes of this paper. All teams submitted relevance results and 5 of 6 submitted SF Type outputs, but only one team attempted generating localized SFs for the Type+Place layer.

7.1. Results on Mandarin

The PR curves for the Mandarin language dataset are shown in Fig. 3. At the Relevance layer the systems are very competitive. The Team 1 system performs best at low recall, with the Team 4 system taking over at higher recall values. All systems converge to the majority class baseline as recall goes to 1, so they are virtually equal at very high recall values. The Team 4 system has the best performance overall with an AUC of 0.676. At the Type layer comparisons are more straightforward: the curves are clearly ranked, with the Team 4 system performing best for any recall value. Only Team 4 submitted localized SFs, so they were the only ones evaluated at the Type+Place layer.

7.2. Results on Uyghur

The PR curves for the Uyghur language dataset are shown in Fig. 3. At the Relevance layer things are less competitive than in Mandarin. The Team 5 system performs best at low recall values, with the Team 4 system taking the lead after that. While all systems converge to the majority baseline at high recall, some do it better than others, so true equalization only comes when recall is almost 1. Best performance overall is achieved by the Team 4 system. At the Type layer we see much of the same, with the Team 5 system being particularly good at low recall values, but the Team 4 system performing best everywhere else. Finally at the Type+Place layer there is again only one system achieving notably worse performance than on Mandarin.

7.3. Overall

The two languages correspond to different scenarios in terms of resource availability and we expected that would be reflected in the results. To compare we combined the results of all teams and used recall quantization and precision interpolation to estimate the maximum and mean performance PR curves. The results are shown in Fig. 4.

We expected performance to be higher for the Mandarin dataset and that holds at the Type and even more so at the Type+Place layer, but not for the Relevance layer. That may be an artifact of different class distributions, since the Mandarin set is slightly more balanced. At the Type layer there is a significant difference at low recall values, corresponding to roughly 0.1 higher precision for the Mandarin set, that predictably disappears at high recall. The largest difference is observed at the Type+Place layer, presumably due to better ASR performance, but there are no safe conclusions to be made due to only one team submitting localization results.

8. Conclusions

Overall the first pilot was successful. We received results from 6 teams, though most teams did not tackle every aspect of the problem. Results at the Type level are encouraging given the difficulty of the task, with teams achieving up to 0.3 AUC. The use of large amounts of data for Mandarin resulted in improved results, particularly in the case of localization, probably because of improved speech recognition. The evaluation process and metric worked well and the PR curves provide more insight than a single statistic, though the requirement for meaningful confidence scores contributes to the difficulty of the task.

As far as the evaluation process is concerned, the main area of improvement should be in localization. Only one team submitted results and the performance achieved is not high. Perhaps the requirement that systems produce an exact string for location is too hard to fulfill at this point and alternatives should be investigated. We could possibly use existing knowledge bases to allow for alternative location names as long as they refer to the same location or perhaps add an evaluation layer allowing phonetic transcriptions.

Given that this was the first iteration of the task, we expect great strides to be made in the next iterations, moving us closer to the program goals.

9. Acknowledgements

The authors would like to thank Marjorie Freedman and Florian Metze for their valuable input during the evaluation design phase, as well as Jonathan Fiscus and the team at NIST for their work on the evaluation of the similar Situation Frames from text task.

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.

10. References

[1] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "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*, 2010, pp. 1079–1088.

[2] "Darpa lorelei website, retrieved october 25, 2015," http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents.

[3] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of ICASSP*, 2013, pp. 7304–7308.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] S. Strassel and J. Tracey, "Lorelei language packs: Data, tools, and resources for technology development in low resource languages," in *Proceedings of LREC*, 2016, pp. 3273–3280.

[6] "Appen website, retrieved march 5, 2016," http://appen.com/.

[7] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[8] H. W. Kuhn, "Variants of the hungarian method for assignment problems," *Naval Research Logistics Quarterly*, vol. 3, no. 4, pp. 253–258, 1956.