Information Modelling and Knowledge Bases XXXIII
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA210478

A Proposal for a Method of Determining Contextual Semantic Frames by Understanding the Mutual Objectives and Situations Between Speech Recognition and Interlocutors

Ryosuke KONISHI [a] Fumito NAKAMURA [a] Yasushi KIYOKI [b]
[a] *Generic Solution Corporation, Nampeidai-chou, Shibuya-ku, Tokyo, Japan*
[b] *Faculty of Environmental Information, KEIO University, Fujisawa, Kanagawa 252, Japan*

Abstract. In this study, we will examine how the speech sounds generated by one-to-one or one-to-n human communication are not only simple exchanges of intentions and opinions, but are also clearly divided into linguistic expression and linguistic understanding. In this course, we will discuss the problems of interaction between interlocutors and the complex interplay of phenomena that occur in individual interlocutors. We propose a method to determine the semantic frame of a dialogue contextually by integrating and calculating the features of speech and the features of the meaning of words.

Keywords. Mathematical Model of Meaning, Integration of heterogeneous information, Natural Language Process, Speech Recognition, Decision Inference,

1. Introduction

The acoustic model in speech recognition technology expresses the relationship between a series of acoustic features, which is a continuous quantity for each generation time, and a series of discrete symbols (words, phonemes, etc.).

The dramatic improvements achieved by deep learning in large-vocabulary continuous speech recognition tasks have attracted much attention. Almost all of the research that had been done on GMM, HMMs in the past has been revised at a rapid pace, and deep learning is now commonly used as the standard for acoustic modeling. Nowadays, deep learning is often used as a standard for acoustic modeling [1,2,3,4]. In the framework of deep learning, the relationship between discrete symbols is modeled on a continuous-valued vector space, and the relationship between variable-length input and output sequences is modeled directly. First, discrete symbols such as words and letters are treated as fixed-length continuous-valued vectors called Distributed Representation or Embedding. Instead of the conventional modeling based on sparse information (e.g., frequency), it is possible to model discrete symbols on a continuous-valued vector space.

It is now possible to capture the relationship between discrete symbols in a vector space of continuous values, rather than modeling based on sparse information (such as frequency).

Second, in addition to the extension techniques, it is now possible to flexibly handle variable-length In addition to the extension techniques, it is now possible to flexibly handle discrete symbolic sequences of variable length. In addition, it is now possible to flexibly handle discrete symbolic sequences of variable length, from fixed-length inputs (such as Bag-of-Words, which does not consider sequences, or local fixed-length context information) to fixed-length outputs. context information) as well as fixed-length output. It is now possible to treat a discrete symbolic series of variable length as a continuous vector of fixed length [5,6,7,8,9,10].

On the other hand, language models are also being extended in various ways with the advancement of techniques based on deep learning. Methods using deep learning have been reported to be overwhelmingly accurate in various tasks such as language model building [11], proper noun extraction [12], meaning construction based on constructive semantics [13], and reputation classification [14,15], and various other tasks, deep learning methods have been reported to have overwhelming accuracy [16,17].

Traditional word modeling involves word segmentation, parsing, and probabilistic language models. In the field adaptation, it is important to prepare texts in the adaptation field and to perform field adaptation of word segmentation and reading estimation. One of the applications of this method is the "Shabette Concierge" and language processing developed by NTT Docomo, but it is limited to task determination and keyword extraction. However, it is limited to task determination and keyword extraction, and semantic interpretation in simple limited situations [18].

Similar work has been done on slot filling and speech intention understanding. Typical methods include Support Vector Machine modeling using sparse features such as n-grams. Machine modeling using sparse features such as n-grams [19,20].

However, although modeling on discrete space and modeling on continuous space have their own challenges regarding their merits, there is some research on modeling that considers more complex structures as contexts, such as documents and discourse. There are few studies that interpret and determine semantic frames for acoustic and language models [21,22,23].

We propose a method for determining the semantic frame for the process as a result of the speaker's semantic understanding and dialogue in a complex specific situation.

Structure of this paper is organized as follows. First, we outline problem settings in this paper. Next, we briefly describe overall picture of our proposed method. Third, we present the basic theories to embody the components of the method. Fourth, we explain the details of the proposed method, and finally we conclude this paper.

2. Problem Settings

Here, we briefly specify the problem. This paper considers involving communication between consumers and a clerk, such as during a business negotiation or sales promotion. In such cases, the outcome of the communication, such as whether the negotiation was successful, unsuccessful, or put on hold, is considered to depend on the tension and activity of the spoken words and voice. This paper considers the problem of inferring the speaker's semantic understanding using linguistic and acoustic information.

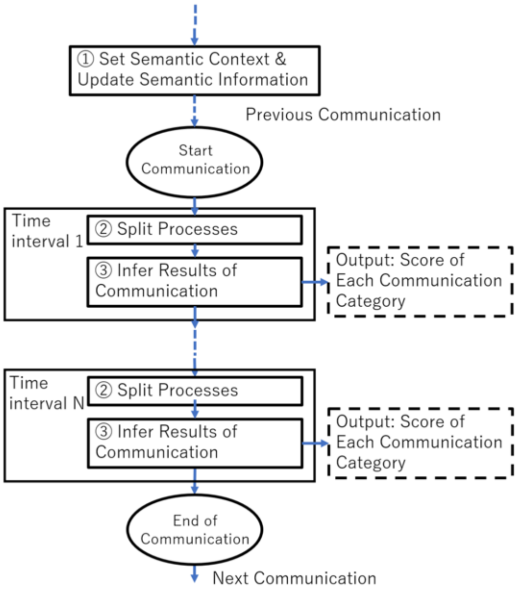
① Set Semantic Context & Update Semantic Information
Previous Communication
Start Communication
Time interval 1
② Split Processes
③ Infer Results of Communication
Output: Score of Each Communication Category
Time interval N
② Split Processes
③ Infer Results of Communication
Output: Score of Each Communication Category
End of Communication
Next Communication

Figure 1. Semantic frame of the proposed method

3. Overview of our method

This section introduces a whole architecture, which is referred to as semantic frame, to solve the problem in section 2. Figure 1 shows the frame. First, our system sets assumed semantic contexts that represent the situation of the communication and updates semantic information that projects linguistic and acoustic information on a semantic space to evaluate the speaker's semantic understanding. Note that the frequency of the calculations, referred to as the period, is days, weeks, etc., because the update requires a little bit of computational cost for the matrix product and the eigenvalues computation, as you can be seen later.

For each communication, our system splits the entire communication into processes by streaming, where the unit of the segmentation is called the time interval.

For each time interval, the system estimates the result of the communication using linguistic and acoustic information, and it outputs the quantized score for each successful, unsuccessful, or pending communication category. We achieve the following functions by applying the next section's methodologies to realize the frame:

1. Set the Semantic Contexts and Update Semantic Information
2. Split Conversations
3. Infer the Results of Communication

4. Basic Theories

Here, we explain basic theories utilized in the proposed method. In this paper, we utilize the mathematical model of meaning (MMM) to extract semantic information. Moreover, we explain how to process linguistic and acoustic information because both kinds have to be quantized to calculate the score.

4.1. Mathematical Model of Meaning

The MMM was first proposed to extract semantic information behind data deterministically [24]. Let $X \in \mathbb{R}^{N\times M}$ be a data matrix, where N is the number of data, and M is the number of features. We use 2-norm normalization in each column of the matrix, and denote the resulting normalized matrix by $\tilde{X}$, i.e., the (i, j)th element of the matrix is

$$\tilde{X}_{i,j} = \frac{X_{i,j}}{\sqrt{\sum_{k=1}^{N} X_{k,j}^2}}. \tag{1}$$

This is referred to as the fundamental data matrix. The product of the fundamental data matrix $\tilde{X}^{\mathrm{T}}\tilde{X}$ represents a similarity matrix among features, and a subspace spanned using a combination of the eigenvectors extracted from the product represents a semantic space. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0$ be the eigenvalues and $v_1, v_2, \ldots, v_M$ be the corresponding eigenvectors. Nevertheless, because the similarity matrix is a symmetric matrix, the eigenvalues are non-negative and real, and the number of the combination is 2^M. A context matrix and a threshold are used to determine the subspace from the combination. Let $Y = [Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, \ldots, Y_t^{\mathrm{T}}] \in \mathbb{R}^{t\times M}$ $t \times M$ be the context matrix, and $Q = [v_1, v_2, \ldots, v_M] \in \mathbb{R}^{M\times M}$. Then, for the threshold ϵ_{mmm}, MMM determines the index set of the chosen eigenvectors $\Lambda_{\epsilon_{mmm}}$ using

$$\Lambda_{\epsilon_{mmm}} = \Big\{ j \in \{1, 2, \ldots, M\} | 0 < \epsilon_{mmm} < 1, \frac{(Q_Y)_j}{\|Q_Y\|_\infty} > \epsilon_{mmm} \Big\}. \tag{2}$$

Here, $Q_Y = \sum_{i=1}^{t} Y_i Q$, $(Q_Y)_j$ is the j-th element of Q_Y, and $\|Q_Y\|_\infty = \max\limits_{1\leq j\leq M} |(Q_Y)_j|$. We obtain the semantic projection by arranging the chosen eigenvectors

$$P(Y) = [v_j^{\mathrm{T}} | j \in \Lambda_{\epsilon_{mmm}}] \in \mathbb{R}^{|\Lambda_{\epsilon_{mmm}}|\times M}, \tag{3}$$

where $|\Lambda_{\epsilon_{mmm}}|$ represents the cardinality of $\Lambda_{\epsilon_{mmm}}$. When we apply the projection to each datum $g \in \mathbb{R}^M$, the datum is regarded as an element of a semantic space, denoted by $P(Y)g$. In this semantic space, we can measure the distance between the datum and the semantic centroid $\bar{D}$:

$$\bar{D} = \frac{1}{\|P_Y\|_\infty} P_Y, \tag{4}$$

where $P_Y = \sum_{i=1}^{t} P(Y) Y_i$. Let $\bar{D}_j$ and $(P(Y)g)_j$ be the j-th elements of the vectors; the distance is calculated as the weighted Euclidean distance $dist(\bar{D}, g)$,

$$dist(\bar{D}, g) = \sqrt{\sum_{j=1}^{|\Lambda_{\epsilon_{mmm}}|} c_j \big(\bar{D}_j - (P(Y)g)_j\big)^2}, \tag{5}$$

where $c_j = \frac{(P_Y)_j}{\|P_Y\|_\infty}$.

4.2. Configuring Linguistic Information

Assuming that we only utilize speech information, we need a system that generates text from speech to apply linguistic techniques. Such a speech recognition technique can be used in the case of both a cloud[25,26] and a premise[27,28]. Next, we need a linguistic technique to deal with the generated text quantitatively. In this paper, we utilize the following well-known features as the linguistic features.

- Bag-of-Words(BoW)
- Word embeddings

Let W be a set of words and H the dimension of the embeddings; then, we create a linguistic feature having the dimension of $|W|+H$.

4.3. Configuring Acoustic Information

Famous acoustic features obtained from a speech signal are the following ones [29,30].

- power of speech signal
- fundamental frequency
- Mel-Frequency Cepstrum Coefficient

Note that these heterogeneous features need to be aggregated because the features are calculated differently. For example, the power is obtained for each digitalized timing, while the fundamental frequency is obtained for each window size. In this paper, we summarize each feature by the maximum, minimum, mean, and standard deviation with each common period.

5. Proposed Method

In this section, we explain how to achieve the functions we explained in section 3 one by one.

5.1. Set the Semantic Contexts and Update Semantic Information

In this process, our system sets semantic contexts and updates semantic linguistic and acoustic projections. Algorithm 1 shows an overview of the process.

Let S be a set of a semantic context such as in general cases like a business negotiation and a sales promotion and in specific cases such as a business negotiation in which a customer decides to purchase something, and let C be a set of communication categories such as successes, failures, or reservations. Then, our system first establishes the set S of the semantic context for each period. Next, our system obtains the past linguistic and acoustic fundamental data matrices X_s^l and X_s^a for each element $s \in S$ with the time interval in the row direction and the features in the column direction, where the element of each feature is the one described in section 4.3 and 4.3. Then, our system calculates the eigenvectors $(v_{1s}^l, \ldots, v_{Ms}^l)$ and $(v_{1s}^a, \ldots, v_{Ms}^a)$ of the product $\tilde{X_s^l}^{\mathrm{T}} \tilde{X_s^l}$ and $\tilde{X_s^a}^{\mathrm{T}} \tilde{X_s^a}$ of the normalized matrices X_s^l and X_s^a. Using the combination of the eigenvectors, our system prepares the semantic projections for each communication category $c \in C$. The linguistic

Algorithm 1 Algorithm of Setting the Semantic Contexts and Updating the Semantic Information

```
Require: $\epsilon_{mmm}$: Threshold to determine the semantic projection, $C$: Set of communication category
procedure SET THE SEMANTIC CONTEXT AND UPDATE THE SEMANTIC INFORMATION
    Set the semantic context $S$
    for each $s \in \mathcal{S}$ do
        Construct linguistic and acoustic fundamental data matrices $X^l_s$ and $X^a_s$ in $s$, respectively.
        Calculate $\tilde{X}^l_s$ and $\tilde{X}^a_s$ using (1).
        Calculate eigenvector $(v^l_{1s}, \dots, v^l_{Ms})$ and $(v^a_{1s}, \dots, v^a_{Ms})$ of $\tilde{X}^{l\,\mathrm{T}}_s \tilde{X}^l_s$ and $\tilde{X}^{a\,\mathrm{T}}_s \tilde{X}^a_s$, respectively.
        for each $c \in C$ do
            Update linguistic and acoustic context matrices $Y^l_{s,c}$ and $Y^a_{s,c}$ of $c$ in $s$, respectively.
            Determine semantic projections $P(Y^l_{s,c})$ and $P(Y^a_{s,c})$ by (3).
        end for
    end for
end procedure
```

and acoustic context matrices $Y^l_{s,c}$ and $Y^a_{s,c}$ of a semantic context s and a communication category c are updated when those features utilized for the previous period are labeled by c. Our system updates the semantic projections $P(Y)^l_s$ and $P(Y)^a_s$ using the context matrices, the eigenvectors, and the threshold ϵ_{mmm}.

Bear in mind that the computational complexity of this process is $O(M^3 + N^2)$, where M is the number of features, and N is the number of data of the fundamental data matrix. Therefore, we assume the period to be days, weeks, and so on.

5.2. Split Processes

In this process, our system determines whether or not the conversation is broken based on a streamed speech signal $S(t)$. Algorithm 2 shows an overview of the process.

Algorithm 2 Algorithm of Split Processes

```
Require: $S(t)$: Signal of streamed acoustic information, $\epsilon_a$: Threshold of silence length, $\epsilon_l$: Threshold of text distance
procedure SPLIT PROCESSES
    Aggregate silence interval in $S(t)$.
    Evaluate silence interval using $\epsilon_a$.
    Enable speech recognition and generate text $text(t)$ from $S(t)$.
    Vectorize $text(t)$, denoted by $V(t)$.
    Evaluate distance between $V(t)$ and previous vector $V(t-1)$ by $\epsilon_l$.
    Integrate two evaluated results and determine whether or not the conversation is broken.
end procedure
```

Both linguistic and acoustic determinations can be made on whether or not the conversation is broken. In the acoustic component, our system evaluates the length of the silence interval. Therefore, our system aggregates the length from $S(t)$ and evaluates that the interval exceeds a threshold ϵ_a. In the linguistic component, our system first generates a text $text(t)$ from $S(t)$ using speech recognition. Then, our system transforms $text(t)$ into a vector $V(t)$ using word embeddings. Finally, our system evaluates the distance between $V(t)$ and previous vector $V(t-1)$ and determines whether or not the distance exceeds a threshold ϵ_l. By combining the results, our system determines whether or not the conversation is broken.

5.3. Infer the Results of Communication

In this process, our system estimates the results of the communication in a split process. Algorithm 3 shows an overview of the process.

Algorithm 3 Algorithm of Inferring the Results of Communication

```
Require: $U(t)$: Signal of acoustic information at time interval $t$, $s$: Semantic context, $\alpha$: Mixing ratio of distance
procedure Infer Results of Communication
    Convert $U(t)$ into linguistic feature $X^l(t)$ and acoustic feature $X^a(t)$.
    for each $c \in C$ do
        Project the linguistic feature into semantic space in $c$, denoted by $P(Y^l_{s,c})X^l(t)$.
        Calculate the distance between $P(Y^l_{s,c})X^l(t)$ and the semantic centroid.
        Project the acoustic feature into the semantic space in $c$, denoted by $P(Y^a_{s,c})X^a(t)$.
        Calculate the distance between $P(Y^a_{s,c})X^a(t)$ and the semantic centroid.
        Integrate the two distance by mixing the ratio $\alpha$.
    end for
end procedure
```

To estimate the results of communication at time interval t, our system transforms the speech signal into linguistic and acoustic features $X^l(t)$ and $X^a(t)$ using methodologies in 4.2 and 4.3. Moreover, in a semantic context s and communication category c, each type of feature is projected using semantic projection$P(Y^l_{s,c})$ and $P(Y^a_{s,c})$. Our system calculates the distance between the projected feature and semantic centroid $dist(D^{\bar{l}}_{s,c}, X^l(t))$ and $dist(D^{\bar{a}}_{s,c}, X^a(t))$ in (5) for each s and c. Our system integrates the distance using

$$d_{s,c}(X^l(t), X^a(t)) = \alpha \times dist(D^{\bar{l}}_{s,c}, X^l(t)) + (1-\alpha) \times dist(D^{\bar{a}}_{s,c}, X^a(t)), \tag{6}$$

where $0 \leq \alpha \leq 1$. Finally, our system estimates the results of communication based on (6).

6. Conclusion

We proposed a method for determining the semantic frame for the process as a result of the speaker's semantic understanding and dialogue in a complex specific situation.

Our future work involves discussing improvements such as feature selection for both linguistic and acoustic features and alternative ways to split processes. In addition, we will conduct numerical experiments and compare the performance between candidates to validate the method.

References

[1] N. Kanda, "Deep learning based acoustic modeling for speech recognition," *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 73, pp. 31–38, 2017.

[2] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[3] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," *Twelfth annual conference of the international speech communication association*, 2011.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, pp. 30–42, 2011.

[5] R. Masumura, "Language modeling and spoken language understanding based on deep learning," *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 73, pp. 39–46, 2017.

[6] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE*, vol. 88, pp. 1270–1278, 2000.

[7] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Combinations of various language model technologies including data expansion and adaptation in spontaneous speech recognition," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] P. Haffner, G. Tur, and J. H. Wright, "Optimizing SVMs for complex call classification," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1, pp. 632–635, 2003.

[9] S. Yaman, L. Deng, D. Yu, Y. Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1207–1214, 2008.

[10] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 234–239, 2012.

[11] C. D. Manning and H. Schutze, *Foundations of statistical natural language processing*. 1999.

[12] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The journal of machine learning research*, pp. 1137–1155, 2003.

[13] M. Wang and C. D. Manning, "Effect of non-linear deep architecture in sequence labeling," *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1285–1291, 2013.

[14] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection," *NIPS*, vol. 24, pp. 801–809, 2011.

[15] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks," *ICML*, 2011.

[16] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 151–161, 2011.

[17] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 1201–1211, 2012.

[18] T. Yoshimura, "Shabette-Concier Service realized by Natural Language Processing," *SLP*, pp. 1–6, 2012.

[19] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," *arXiv preprint arXiv:1506.06726*, 2015.

[20] J. Li, M. T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," *arXiv preprint arXiv:1506.01057*, 2015.

[21] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," *arXiv preprint arXiv:1511.01432*, 2015.

[22] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," *arXiv preprint arXiv:1507.04808*, vol. 7, pp. 434–441, 2015.

[23] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[24] T. Kitagawa and Y. Kiyoki, "A mathematical model of meaning and its application to multidatabase systems," in *Proceedings RIDE-IMS93: Third International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems*, pp. 130–135, 1993.

[25] https://cloud.google.com/speech-to text, "Google Cloud Speech."

[26] https://www.ibm.com/jp-ja/cloud/watson-speech-to text, "Watson Speech to Text."

[27] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," *EUROSPEECH*, pp. 1691–1694, 2001.

[28] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," *APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pp. 131–137, 2009.

[29] A. Cheveigné and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," *Seventh European Conference on Speech Communication and Technology*, 2001.

[30] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr{\'e}, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and T. Khiet, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, pp. 190–202, 2015.