Creating Large-Scale Argumentation Structures for Dialogue Systems

Kazuki Sakai[1], **Akari Inago**[2], **Ryuichiro Higashinaka**[3],
Yuichiro Yoshikawa[1], **Hiroshi Ishiguro**[1], **Junji Tomita**[3]
[1]Osaka University, [2]Ochanomizu University, and [3]NTT Media Intelligence Laboratories

Abstract

We are planning to develop argumentative dialogue systems that can discuss various topics with people by using large-scale argumentation structures. In this paper, we describe the creation process of these argumentation structures. We created ten structures each having more than 2000 nodes of five topics in English and five topics in Japanese. We analyzed the created structures for their characteristics and investigated the differences between the two languages. We conducted an evaluation experiment to ascertain that the structures can be applied to dialogue systems. We conducted another experiment to use the created argumentation structures as training data for augmenting the current argumentation structures.

Keywords: Argumentation, Large-scale argumentation structures, Dialogue systems

1. Introduction

Argumentation is a process of reaching consensus through premises and rebuttals and is important for making decisions and exchanging views. Argumentation has long been studied in the fields of rhetoric, informal logic, and more recently artificial intelligence. For example, there have been studies on automatically extracting conclusions and premises from documents (Rosenthal and McKeown, 2012; Yanai et al., 2016; Lippi and Torroni, 2016). Other studies have devised argumentation models (Toulmin, 1958; Reed and Rowe, 2004; Walton, 2013) and visualizations of the models (Gordon et al., 2007; Reed and Rowe, 2004; Snaith et al., 2010). To develop dialogue systems that can support humans in argumentation, we are planning to develop argumentative dialogue systems that can discuss various topics with people by using large-scale argumentation structures; we argue that large-scale argumentation structures are necessary for systems to respond appropriately to various arguments raised by users.

Recently, corpora containing argumentation structures of discussions/meetings have been made available (Janin et al., 2003; Renals et al., 2007). Other studies have extracted argumentation structures from corpora (Fernández et al., 2008; Bui et al., 2009). However, these structures are small; thus, insufficient as knowledge for dialogue systems. One of the largest argumentation databases currently available is AIFdb (Lawrence et al., 2012), which is an open database containing argumentation structures in argument interchange format (AIF). Although AIFdb contains many argumentation structures, each structure is small; for example, the largest contains about 250 nodes, and the average number of nodes per structure is 8.14 (See Table 1).

In this paper, we describe the creation process to construct large-scale argumentation structures for dialogue systems. We manually created several large-scale argumentation structures based on a conventional argumentation model (Walton, 2013), and each structure has more than 2000 nodes. We created the structures in two languages; English (major language) and Japanese (author language). To verify the effectiveness of the created argumentation structures, we conducted an evaluation experiment to ascertain that the structures can be applied to dialogue systems. We conducted another experiment to use the created argumentation structures as training data for augmenting the current argumentation structures.

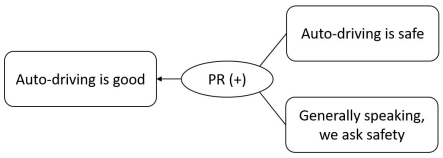

Figure 1: Argumentation model

2. Argumentation Structures

We describe the argumentation model on which our argumentation structures are based then discuss our process of creating them.

2.1. Argumentation Model

As shown in Figure 1, we use a simplified version of the model described in (Walton, 2013). The model has a graph structure, and nodes represent premises and edges represent relationship between nodes. Each node has a natural language statement representing the content of its premise. A node is connected to other nodes by directed arcs that represent a supportive (+) or non-supportive (−) relationship. If the logical connection is based on an argumentation scheme (Walton, 1996) (e.g. practical reasoning (PR), the logic is "if G is something good and action X leads to G, X should be done"), the scheme name is represented on the arcs.

Figure 2 shows the design of our argumentation structure with a specific purpose for dialogue systems. The structure has two parts represented by main issue nodes that enable the system to have opposing stances (we call the stances A and B). Below the main issue nodes, there are what we call viewpoints nodes that represent conversational topics. Under each viewpoint node, there are premise nodes that represent statements regarding each topic.

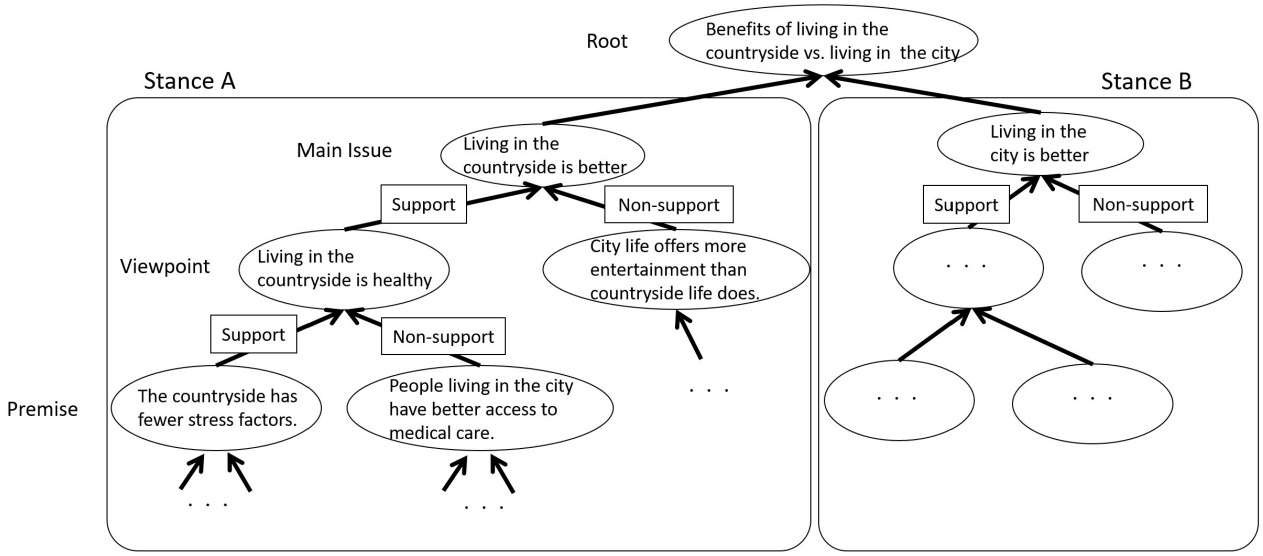

Figure 2: Design of our argumentation structures for dialogue systems

Let us suppose we have a structure about "Benefits of living in the countryside vs. living in the city", which is represented as the root node. The main issue nodes are "Living in the countryside is better" and "Living in the city is better". The viewpoint nodes would include "Living in the countryside is healthy". The premise nodes would include statements such as "The countryside has fewer stress factors" and "People living in the city have better access to medical care". With this design, we believe we can create dialogue systems that can discuss on a certain issue with a stance and from multiple viewpoints.

2.2. Creation Process

Now let us describe the process of creating large-scale argumentation structures. First, we determine the main proposition and two main issue nodes representing two stances. Next, we create viewpoint nodes that represent topics. Then, under the viewpoint nodes, we create premise nodes.

We recruited more than 30 annotators for creating the argumentation structures. First, we determined the propositions and their stances and created the viewpoint nodes to a certain extent. The annotators iteratively created supporting or non-supporting premises for existing premises under viewpoint nodes. When creating premises, they used argumentation schemes whenever possible. To maintain objectivity of the data, the logical relationships between two nodes were checked by other annotators. If the relationship was inappropriate, the annotators corrected or removed the corresponding nodes. This process was repeated until the relationship was appropriate. By repeating this process, we can create large-scale argumentation structures.

2.3. Constructed Structures

We manually created five English and five Japanese structures each with five different topics. Figure 3 shows an example of visualizing the argumentation structures. Each constructed structure has more than 2000 nodes that are supporting or not supporting other nodes. The structures are also hierarchical, as described in Section 2.1.

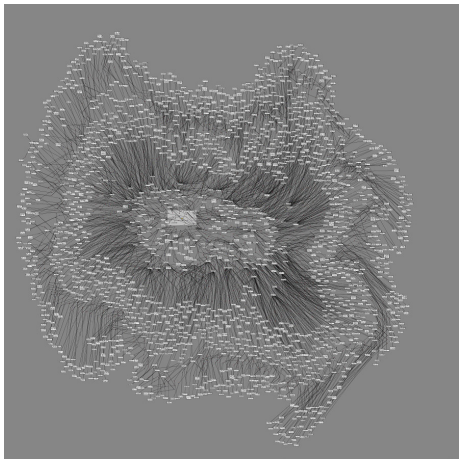

Figure 3: Visualization of constructed structure. Topic is countryside vs. city (English).

The five topics of the graphs written in English are as follows: the pros and cons of driving automobiles (Auto driving), benefits of living in the countryside vs. living in the city (Countryside), who is the greater pop icon, Lady Gaga or Taylor Swift? (Lady Gaga), which is the better Japanese meal, sushi or ramen? (Sushi), and which is the better living environment, east or west coast? (East coast).

The five topics of the structures written in Japanese are as follows: the pros and cons of driving automobiles (Auto driving), benefits of living in the countryside vs. living in the city (Countryside), which is the better place to travel to in Japan, Hokkaido or Okinawa? (Hokkaido), which is the better breakfast, bread or rice? (Bread), and which is the better theme park, Tokyo Disney Resort (TDR) or Universal Studio Japan? (TDR).

	No. of files	No. of nodes per structure		Depth per node		No. of words per node		No. of sentences per node		Branches per node	
		ave.	var.	ave.	var.	ave.	var.	ave.	var.	ave.	var.
AIFdb	7066	8.14	165.52	2.18	1.13	32.44	5389.81	1.95	11.39	1.50	0.78
Our argumentation structures (English)	5	2281.60	5214.64	4.46	0.39	16.90	50.59	1.01	0.01	3.74	19.20
Our argumentation structures (Japanese)	5	2253.80	15885.36	4.42	0.25	19.20	56.33	1.00	0.00	3.63	8.70

Table 1: Statistics of constructed Structures. AIFdb is representative of currently available data. Note that we used nltk (English) for tokenizing words and sentences, and MeCab (Japanese) for tokenizing words.

	English		Japanese	
	scheme	frequency	scheme	frequency
1	CE (+)	141	CE (+)	374
2	EH (+)	99	CO (+)	244
3	EO (+)	59	EX (−)	242
4	PO (+)	47	CO (−)	142
5	VC (+)	43	PO (+)	118
6	AO (−)	40	EH (+)	88
7	AD (+)	33	EO (+)	79
8	CO (+)	30	EC (+)	75
9	PR (+)	28	CE (−)	72
10	EC (+)	26	PO (−)	47

Table 2: Top 10 schemes used in each English and Japanese structures. Sign (+) or (−) with schemes means node supports or refutes another node by using logic represented by scheme, respectively.

A Life in the countryside is better.
B But there is a lot more entertainment facilities in the city than in the countryside.
A But you cannot enjoy the splendor of nature or bountiful natural resources in cities like you can in the country. The countryside offers more areas for exploring.
B But not as many places are available for adventure in the countryside because most of the land is owned by someone as part of their ranch or farm.
A But adventures first started by exploring the natural wonders of the world.
A For many cities, the concept of green space in urban development is relatively new, not to mention being a luxury.
B But cities make full use of the natural resources around them.

Figure 4: First part of dialogue text on topic of countryside vs. city. Stance A is countryside; stance B is city.

2.4. Analysis of Constructed Structures

2.4.1. Comparison with Existing Data

Table 1 shows the statistics of the created structures. We chose AIFdb as the existing data; we crawled some larger structures from the AIFdb website (`http://www.aifdb.org/search`). As can be seen from the table, our structures have more nodes per structure than AIFdb. The number of sentences in the node is comparatively smaller than that of AIFdb, suggesting the possibility of appropriateness for dialogue system utterances. The average number of branches is larger than that of AIFdb and the variance is higher, which indicates the possibility of generating more various utterances for user utterances.

2.4.2. Comparison between Languages

To investigate the difference between languages, we compared the usage of schemes because schemes are likely to reflect the way of thinking in a language. Table 2 shows the top ten schemes used in English and Japanese. There were some differences depending on the language; for example, the scheme called argument from composition (CO), whose logic is "A is part of X, A has property Y, therefore, X has property Y", used in Japanese is more frequent than that used in English. An example argument using CO is that a premise "emergence of automobiles leads to law amendment" concludes "technical innovation leads to law amendment." On the other hand, the scheme called argument from verbal classification (VC), whose logic is "P has property F, and for all x, x has property G if x has F, then P has G", used in English is more frequent than that used in Japanese. An example argument using VC is that two premises "automobiles require the latest electronics" and "the latest electronics are expensive" conclude "auto driving makes vehicles more expensive." Although further examination is needed, if we can create a structure with a certain distribution of schemes, we may be able to develop dialogue systems appropriate to a certain language.

3. Towards Development of Argumentative Dialogue Systems

To apply the created argumentation structures to dialogue systems, we first verified whether the structures can be applied to dialogue generation then developed a dialogue system prototype that used the created structures.

3.1. Dialogue Generation

We generated dialogue text based on argumentation structures. We verified that appropriate argumentative dialogue can be created by just traversing along with the structures if the structures are constructed based on logical relationships. Note that we used only English argumentation structures for this study.

3.1.1. Generation Methods

We created two dialogue texts from each of the five English structures. This was done by traversing structures un-

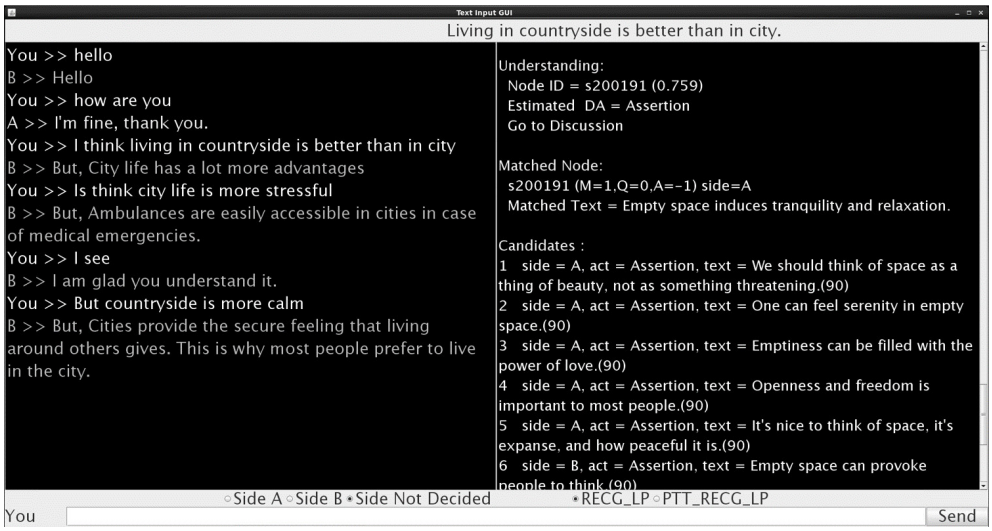

Figure 5: Text chat interface of our dialogue system prototype. The prototype works either in English or Japanese. Here, the topic is "the benefits of living in the countryside vs. living in the city".

questionnaire	item	ave.
Q1	Understandability of content	3.81
Q2	Naturalness of dialogue	3.01
Q3	Understandability of stances	3.52

Table 3: Averaged questionnaire scores

der viewpoint nodes[1] in a depth-first search fashion. When visiting each node, the statement associated with it was appended to the dialogue text. The dialogue-generation process finished when 20 utterances were generated. Figure 4 shows a sample dialogue (on the benefits of living in the countryside versus the city). Note that when transiting to a node from stance A to stance B, or vise-versa, "But" is inserted to make the dialogue look natural.

We recruited 19 participants (12 males and 7 females; average age: 28.9). They read ten dialogue texts (two samples $\times$ five structures) that were randomly ordered and answered a questionnaire. The questionnaire included three questions: (Q1) "Was the meaning of individual utterances easy to understand?", (Q2) "Did the dialogue look natural?", and (Q3) "Was it easy to understand how the two systems agreed or disagreed?" They answered these three questions on a five-point Likert scale, where 5 meant the highest degree of agreement.

3.1.2. Results

Table 3 shows the questionnaire results. The score for Q1 was much higher than 3, so the dialogue texts were understandable to all participants. The score for Q2 and Q3 was higher than 3, so the texts successfully exhibited the characteristics of argumentative dialogue. The evaluation results indicate that the dialogue texts generated by argumentation structures were reasonable, suggesting the effectiveness for use in dialogue systems.

[1]For each structure, we selected the top two viewpoint nodes in the number of descendants.

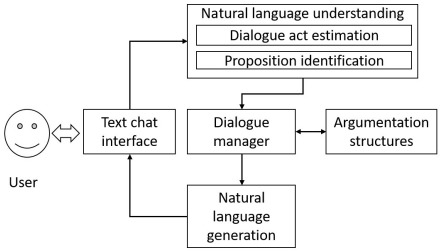

Figure 6: Architecture of our dialogue system prototype. Prototype works either in English or Japanese.

3.2. Dialogue System Prototype

We constructed a dialogue system prototype that argues on the basis of the constructed argumentation structures used in English and Japanese (Higashinaka et al., 2017). Figure 5 shows a text chat interface of our dialogue system prototype, and Figure 6 shows its architecture. When the user types, the sentence is input to two main modules: dialogue act estimation and proposition identification. The dialogue-act-estimation module estimates four types of dialogue acts; assertion, question, concession, and retraction. The proposition-identification module finds, on the basis of word2vec (Mikolov et al., 2013), the corresponding node that has the statement whose meaning is the closest to the input sentence. Finally, the system uses the found node to generate the response text by using the statement of the node. We confirmed that the prototype system can return reasonable responses when the user utterance is within the topic in question.

4. Towards Automatic Creation of Structures

Although the constructed structures are useful for dialogue systems, it is expensive to expand the structures and create new structures of other topics since our current cre-

Table 4: Classification results for English argumentation structures. Bold font indicates the highest score in each column.

	Auto driving	Countryside	Gaga	Sushi	East coast	Average
baseline	0.502	0.524	0.537	0.513	0.511	0.517
SVM (uni + bi + tri)	0.539	0.564	0.575	0.554	0.566	0.560
SVM (uni + uni pair + bi pair)	0.548	0.578	0.597	0.575	**0.580**	0.576
ERT (uni + bi + tri)	0.543	**0.596**	**0.601**	0.583	0.574	**0.579**
MNB (uni + bi + tri)	0.522	0.539	0.533	0.551	0.531	0.535
MNB (uni + uni pair + bi pair)	0.509	0.559	0.468	0.545	0.526	0.522
CBOW (no w2v)	0.532	0.580	0.575	0.587	0.563	0.567
CBOW (use w2v)	0.546	0.576	0.553	0.579	0.567	0.564
LSTM (no w2v)	0.522	0.561	0.527	0.531	0.511	0.530
LSTM (use w2v)	**0.554**	0.568	0.585	0.587	0.558	0.571
BLSTM (no w2v)	0.541	0.585	0.583	0.561	0.536	0.561
BLSTM (use w2v)	0.551	0.586	0.586	**0.592**	0.562	0.576

Table 5: Classification results for Japanese argumentation structures. Bold font indicates the highest score in each column.

	Auto driving	Countryside	Hokkaido	Bread	TDR	Average
baseline	0.515	0.508	0.504	0.498	0.498	0.505
SVM (uni + bi + tri)	0.586	0.592	0.606	0.581	0.595	0.592
SVM (uni + uni pair + bi pair)	**0.622**	0.611	0.642	0.626	0.630	0.626
ERT (uni + bi + tri)	0.617	**0.627**	0.637	**0.626**	**0.638**	**0.629**
MNB (uni + bi + tri)	0.589	0.591	0.596	0.606	0.592	0.595
MNB (uni + uni pair + bi pair)	0.592	0.602	0.598	0.601	0.603	0.599
CBOW (no w2v)	0.569	0.584	0.613	0.575	0.582	0.585
CBOW (use w2v)	0.579	0.594	0.595	0.569	0.592	0.586
LSTM (no w2v)	0.575	0.589	0.604	0.578	0.582	0.586
LSTM (use w2v)	0.603	0.616	0.624	0.586	0.609	0.608
BLSTM (no w2v)	0.576	0.593	0.614	0.586	0.607	0.595
BLSTM (use w2v)	0.608	0.625	**0.646**	0.596	0.617	0.618

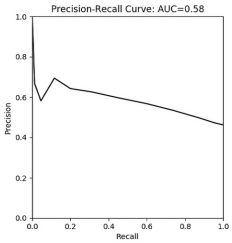

Figure 7: Precision-Recall curve of method that had highest average score in English

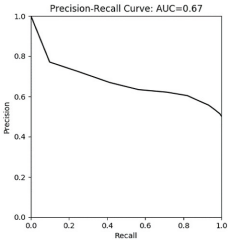

Figure 8: Precision-Recall curve of method that had highest average score in Japanese

ation process is carried out manually. As preliminary work, we are conducting research on estimating the support/non-support relationship between two statements by using machine learning methods and our large-scale argumentation structures as training data.

4.1. Classification Methods

We classified the relationship between two statements as support or non-support. The data were pairs of statements: a node that has a statement directly connected to another node that has the other statement. We examined the following machine learning methods:

Support Vector Machine (SVM) Two statements were represented as a feature vector. Then, the vector was input to a linear SVM. See below for how we created the feature vectors.

Extremely Randomized Tree (ERT) Two statements were represented as a feature vector. Then, the vector was input to an ERT.

Multinomial Naive Bayes (MNB) Two statements were represented as a feature vector. Then, the vector was input to an MNB.

Continuous Bag-of-Words (CBOW) Each statement was embedded, and each vector was averaged. Then two vectors were concatenated, and their relationship was estimated using the softmax function.

Long Short-Term Memory (LSTM) Each statement was embedded, and each vector was input to LSTM. Then two vectors were concatenated, and their relationship was estimated.

Bidirectional LSTM (BLSTM) Each statement was embedded, and each vector was input to BLSTM. Then two vectors were concatenated, and their relationship was estimated.

Two sets of features were used for SVM and MNB; (1) the word uni-, bi-, and tri-grams of statements (uni + bi + tri) and (2) the word uni-grams of statements and word uni-, and bi-grams pairs between two statements (uni + uni

pair + bi pair). The word uni-, bi-, and tri-grams of statements were used for ERT. On the other hand, a sequence of one-hot word vectors were used for CBOW, LSTM, and BLSTM. These methods used two ways to embed each statement; the pre-trained word2vec model (use w2v) and not using this model (no w2v). The Japanese word2vec was learned from the data of Wikipedia, while the English word2vec was learned from the data of Google News. The classifiers were trained using pairs of statements in four argumentation structures written in the same language then were tested using pairs of statements in the other argumentation structure. Specifically, for training in CBOW, LSTM, and BLSTM, training data were divided into two parts; for training (90%) and development to tune the parameters (10%).

4.2. Results

Table 4 lists the accuracy of the support/non-support classification for English structures, and Table 5 lists the accuracy of the support/non-support classification for Japanese structures. In Japanese and English, an extremely randomized tree method with uni-, bi-, tri-grams had the highest score. The score for Japanese was higher than that for English. This is because in Japanese, there seems to be more linguistic constructs that denote/infer discourse or logical relationships.

In Japanese and English, we confirmed that both highest scoring methods could classify the relationships more accurately than each baseline (McNemar's test, Japanese: $p < .001$, English: $p < .001$).

However, as shown in Figures. 7 and 8, the value of recall is very small in the high precision region (more than 90%). Therefore, we consider it is currently difficult to automatically augment argumentation structures. The progress in the field of argumentation mining would help in the future (Lippi and Torroni, 2016).

5. Conclusion

We created large-scale argumentation structures for dialogue systems. We compared the structures for their characteristics to conventional argumentation structures and usage of schemes between English and Japanese. We conducted a subjective evaluation of dialogue generation by using the argumentation structures and described the development of a dialogue system prototype. For automatic augmentation of such structures, we conducted an experiment of automatic support/non-support classification. For future work, we will develop a method for automatically constructing such structures. We will also evaluate a dialogue system prototype and improve the dialogue system; for example, we will develop dialogue strategies for natural argumentation.

6. Bibliographical References

Bui, T. H., Frampton, M., Dowding, J., and Peters, S. (2009). Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proc. SIGDIAL*, pages 235–243.

Fernández, R., Frampton, M., Ehlen, P., Purver, M., and Peters, S. (2008). Modelling and detecting decisions in multi-party dialogue. In *Proc. SIGDIAL*, pages 156–163.

Gordon, T. F., Prakken, H., and Walton, D. (2007). The carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10):875–896.

Higashinaka, R., Sakai, K., Sugiyama, H., Narimatsu, H., Arimoto, T., Fukutomi, T., Matsui, K., Ijima, Y., Ito, H., Araki, S., Yoshikawa, Y., Ishiguro, H., and Matsuo1, Y. (2017). Argumentative dialogue system based on argumentation structures. In *Proc. SemDial*, pages 154–155.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., and Stolcke, A. (2003). The ICSI meeting corpus. In *Proc. ICASSP*, volume 1, pages 364–367.

Lawrence, J., Bex, F., Reed, C., and Snaith, M. (2012). AIFdb: Infrastructure for the argument web. In *Computational Models of Argument (COMMA)*, pages 515–516.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119.

Reed, C. and Rowe, G. (2004). Araucaria: software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13:961–980.

Renals, S., Hain, T., and Bourlard, H. (2007). Recognition and understanding of meetings the AMI and AMIDA projects. In *Proc. ASRU*, pages 238–247.

Rosenthal, S. and McKeown, K. (2012). Detecting opinionated claims in online discussions. In *Proc. 2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.

Snaith, M., Lawrence, J., and Reed, C. (2010). Mixed initiative argument in public deliberation. In *Proc. 4th International Conference on Online Deliberation*, pages 2–13.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge university press.

Walton, D. (1996). *Argumentation schemes for presumptive reasoning*. Routledge.

Walton, D. (2013). *Methods of argumentation*. Cambridge University Press.

Yanai, K., Kobayashi, Y., Sato, M., Yanase, T., Miyashi, T., Niwa, Y., and Ikeda, H. (2016). Debating artificial intelligence. *Hitachi Review*, 65(6):151.