Incremental Dialogue Act Recognition: token- vs chunk-based classification

Eustace Ebhotemhen, Volha Petukhova, Dietrich Klakow

Spoken Language Systems Group, Saarland Informatics Campus, Saarland University, Germany

{eustace.ebhotemhen;v.petukhova;dietrich.klakow}@lsv.uni-saarland.de

Abstract

This paper presents a machine learning based approach to incremental dialogue act classification with a focus on the recognition of communicative functions associated with dialogue segments in a multidimensional space, as defined in the ISO 24617-2 dialogue act annotation standard. The main goal is to establish the nature of an *increment* whose processing will result in a reliable overall system performance. We explore scenarios where increments are tokens or syntactically, semantically or prosodically motivated chunks. Combing local classification with meta-classifiers at a late fusion decision level we obtained state-of-the-art classification performance. Experiments were carried out on manually corrected transcriptions and on potentially erroneous ASR output. Chunk-based classification yields better results on the manual transcriptions, whereas token-based classification shows a more robust performance on the ASR output. It is also demonstrated that layered hierarchical and cascade training procedures result in better classification performance than the single-layered approach based on a joint classification predicting complex class labels.

Index Terms: incremental processing, dialogue act classification

1. Introduction

Interpretation of human dialogue behaviour in terms of speaker intentions is of crucial importance for adequate computational dialogue modelling. While the general problem of theoretically and empirically sound dialogue modelling is far from solved, several aspects of dialogue modelling have been tackled rather successfully. For instance, in automatic dialogue act recognition significant progress has been achieved, see [1] for a brief overview. A dialogue act is a key concept in the semantic description of human dialogue behaviour, defined as communicative activity of a participant in dialogue, interpreted as having a certain *communicative function* and *semantic content*, and possibly also having certain semantic and rhetorical relations. Interpretation of dialogue behaviour is primarily based on the recognition of the speaker's intentions encoded in the communicative function. Thus, the dialogue act manual annotation and automatic classification tasks are typically narrowed down to the recognition of communicative functions that a dialogue unit may have in certain context [2]. Additionally, 'dimensions' have been introduced to classify dialogue acts in multidimensional space, see [3, 4].

Natural language processing including those of dialogue is, by its very nature, incremental [5, 6]. Recently, systems are developed where any minimal input triggers the system's processing which continues increment-by-increment till the complete input is recognized [7]. This creates possibilities for the system to show more interactive and pro-active behaviour (e.g. backchanneling, interrupting and completing the partner) and to minimize system response time, see [8, 9, 10]. While many researchers agree that natural language processing is largely incremental and should be modeled as such, there is no agreement on the nature of its minimal units, i.e increments. There is also no evidence that for all processing steps/types, increments should be of the same nature and size. For example, it is known that semantics is compositional, therefore it is reasonable to assume that semantic processing can be performed word-by-word where most content words would correspond to a semantic concept (i.e. event, participant in an event or their attributes). Pragmatic meaning is, by contrast, not compositional (see [11] for discussion). Therefore, for dialogue act classification, that have higher level of abstraction, token/word-based approach might be not the most adequate one. Bigger units may form the basis for incremental dialogue act processing. Such units, chunks, can be prosodically, syntactically or semantically motivated. In this study, we investigate the effects of different increment types on the dialogue act classification performance. We also considered two settings when trained classifiers operate on features extracted and computed (1) using near-perfect manually corrected speech transcriptions, and (2) from the word lists hypotheses generated by the ASR module.

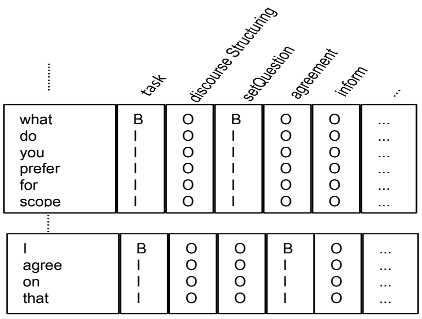
The paper is structured as follows. Section 2 discusses related work on incremental dialogue act classification. In Section 3 we outline our classification experimental setup describing the used data and different training procedures. In Section 4 the obtained results are presented. We wrap up the paper by summarizing our findings and outlining future research.

2. Related work

Traditional approaches to dialogue act (DA) classification have mostly been based on attempts to classify complete utterances (or even speaker turns). Nakano et al. (1999) proposed an incremental approach to understanding user utterances called Incremental Significant Utterance Sequence Search (ISSS). This approach facilitates a word-by-word processing of utterances by finding plausible utterance units - significant utterances (SUs) - that play a role in changing the system's belief state. An SU can be an entire utterance or a sub-sentential phrase. ISSS holds multiple possible belief states which are updated when a word hypothesis arrives. Rather than trying to determine whether the whole input forms an SU, it determines where SUs are. The ISSS approach does however not deal with the possible multifunctionality of segments, and does not allow segments to overlap, to be discontinuous or spread over multiple turns. Lendvai and Geertzen (2007) proposed a token-based dialogue act segmentation and classification which takes dialogue data that is yet to be segmented into syntactic and semantic units. They assign a dialogue act label to each token in the transcribed speech stream of a dialogue participant, additionally classifying if the token is at the beginning of, inside, or outside the segment of that specific dialogue act. Classifiers are built for the recognition of multiple dialogue acts for each input token. They con-

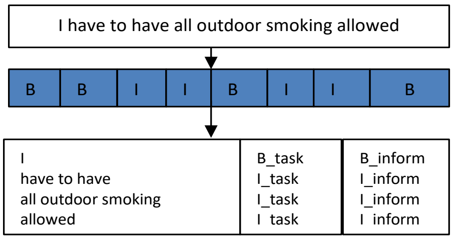
Table 1: *Functional segments distribution across dimensions (relative frequency, in %).*

Dimension **Functional tag**	Task (54.5)	Discourse Structuring (16.5)	Task Management (5.8)	Allo-feedback(2.1)	Auto-Feedback (21.1)
inform	26.8	-	2.2	-	-
offer	3.3	-	-	-	-
suggest	6.7	1.6	1.7	-	-
interactionStructuring	-	10.0	-	-	-
closing	-	1.2	-	-	-
checkQuestion	-	-	-	-	2.7
setQuestion	4.8	-	-	-	-
accept	9.8	1.4	1.9	2.1	-
decline	3.1	-	-	-	-
autoPositive	-	-	-	-	16.8
autoNegative	-	-	-	-	1.6
topicShift	-	2.3	-	-	-



	task	discourse Structuring	setQuestion	agreement	inform	...
what	B	O	B	O	O	...
do	I	O	I	O	O	...
you	I	O	I	O	O	...
prefer	I	O	I	O	O	...
for	I	O	I	O	O	...
scope	I	O	I	O	O	...
I	B	O	O	B	O	...
agree	I	O	O	I	O	...
on	I	O	O	I	O	...
that	I	O	O	I	O	...

Figure 1: *Token-based DA annotation and BIO encoding.*



I have to have all outdoor smoking allowed

B	B	I	I	B	I	I	B

I	B_task	B_inform
have to have	I_task	I_inform
all outdoor smoking	I_task	I_inform
allowed	I task	I inform

Figure 2: *Syntactic chunk-based DA annotation and BIO encoding.*

ducted experiments on the Monroe[1] and the MRDA[2] corpora using Conditional Random Fields and Memory Based Taggers with success in F-score between 47.7 to 81.7. Petukhova and Bunt (2011) presented an incremental approach for dialogue utterance understanding with a focus on the recognition of communicative functions in multidimensional space [14]. They combined local classifiers, which exploit local utterance features, and global classifiers which use the local classifier predictions in an adaptive training procedure. The F-scores obtained range between 71.8 to 98.6 for the different dimensions and with slight differences for HCRC Maptask[3] and AMI[4] data.

Recently, deep neural networks gained a lot of attention. Hierarchical Recurrent Neural Networks (RNN) for learning sequences of dialogue acts are proposed [15]. In this study, two RNNs are trained one to capture dependencies at the conversational level (dialogue act sequences) and at the utterance level (token sequences within an utterance). The authors further incorporated attention mechanism to focus on salient tokens in utterances. Models were trained and tested on Maptask and Switchboard[5] data reporting accuracy of 74.5% for Switchboard and 63.3% for Maptask data. The approach is comparable to the one proposed in [14], and can be applied to incremental dialogue act classification modelling with local classification at the utterance and global at the conversational levels.

3. Classification experimental setup

We define the incremental dialogue act recognition task as a sequence learning task, for which we built Conditional Random Fields (CRFs) based classifiers [16]. CRFs, as discriminative undirected probabilistic graphical models, capture dependencies between certain output and input variables. Receiving mostly partial utterances as input either during training or testing and occasionally observing complete utterances, CRFs predict the most likely label sequence from any number of available input samples. This makes CRFs particularly suitable to simulate incremental processing procedures.

3.1. Data: collection, annotation, features

The data used in our experiments originates from the Metalogue Multi-Issue Bargaining (MIB) corpus [17]. Speech of 16 dialogue participants has been automatically transcribed and manually corrected. The MIB corpus has been manually segmented into functional segments [18] and annotated with ISO 24617-2 dialogue act information where each label consists of a dimension tag (d) and a communicative function tag (cf). The ISO 24617-2 taxonomy [4] distinguishes 9 dimensions, addressing information about a certain *Task*; the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*). 57 defined ISO 24627-2 communicative functions can be of two types: a general-purpose one addressing any of the nine dimension, like Answer, Agreement, or Correction, or a dimension-specific one addressing one particular dimension, such as ReturnGreeting, Accept Apology, Self-Correction and Completion. Two expert and one trained annotators performed annotations independently. The inter-annotator agreement obtained in terms of Cohen's kappa [19] of .90 on average on both segmentation and annotation tasks.

For learning plausible token sequences that form a functional segment, boundaries were marked by adding to the DA class label (<d;cf>) a prefix indicating whether an increment (either token or chunk) starts a segment (B), is inside a segment (I) or is outside a segment (O). Features computed from

[1] http://www.cs.rochester.edu/research/speech/monroe/

[2] http://www1.icsi.berkeley.edu/~ees/dadb/

[3] http://groups.inf.ed.ac.uk/maptask/

[4] http://groups.inf.ed.ac.uk/ami/corpus/

[5] http://groups.inf.ed.ac.uk/switchboard/

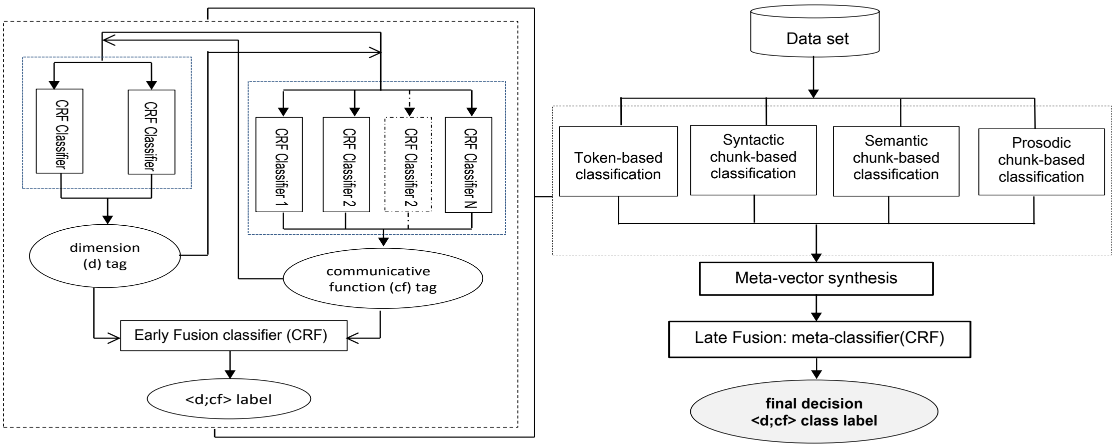
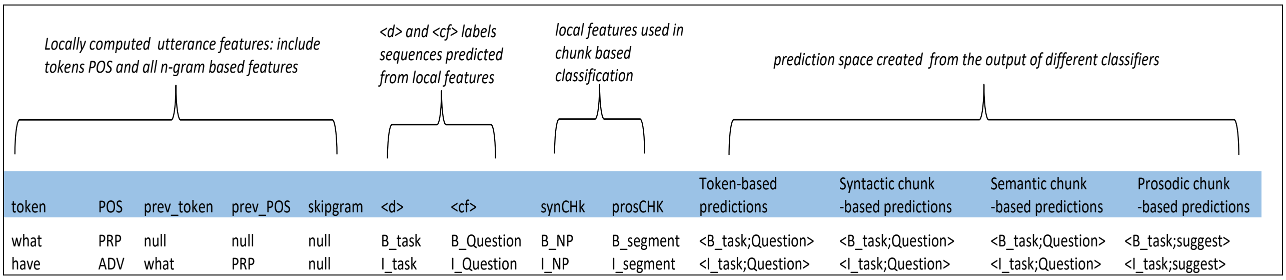

CRF Classifier
CRF Classifier
dimension (d) tag
CRF Classifier 1
CRF Classifier 2
CRF Classifier 2
CRF Classifier N
communicative function (cf) tag
Early Fusion classifier (CRF)
<d;cf> label
Data set
Token-based classification
Syntactic chunk-based classification
Semantic chunk-based classification
Prosodic chunk-based classification
Meta-vector synthesis
Late Fusion: meta-classifier(CRF)
final decision <d;cf> class label

Figure 3: *Hierarchical local classification procedures (left) and meta-classification (right) based on different increment types.*


Locally computed utterance features: include tokens POS and all n-gram based features
<d> and <cf> labels sequences predicted from local features
local features used in chunk based classification
prediction space created from the output of different classifiers

token	POS	prev_token	prev_POS	skipgram	<d>	<cf>	synCHk	prosCHK	Token-based predictions	Syntactic chunk -based predictions	Semantic chunk -based predictions	Prosodic chunk -based predictions
what	PRP	null	null	null	B_task	B_Question	B_NP	B_segment	<B_task;Question>	<B_task;Question>	<B_task;Question>	<B_task;suggest>
have	ADV	what	PRP	null	I_task	I_Question	I_NP	I_segment	<I_task;Question>	<I_task;Question>	<I_task;Question>	<I_task;suggest>

Figure 4: *Synthesised meta-vector used at late fusion decision making level.*

the data include bag-of-words, bag-of-lemmas, trigrams, skipgrams, part-of-speech (POS) tags, n-grams of POS tags using Penn Tree Bank parser [20], etc. The SENNA parser [21] was used for identifying syntactic constituents and performing semantic role labeling.

Our data consists of 5,781 functional segments (45,479 tokens). Table 1 presents functional tags distribution across dimensions.

3.2. Token- vs chunk-based processing

The transcribed speech files contain strings of words, disfluencies and non-speech events like silences and noise. Dialogue contributions are not grammatically well formed; interruptions, interjections, hesitations and repairs are frequent and contribute to discontinuity and overlap of many semantically meaningful segments. This motivates a dialogue processing based on token streams produced by multiple speakers in parallel, and this in multiple dimensions. The token-based dialogue act classification is defined as a sequence learning task where for each input token boundary prefix (BIO) and DA class label are learned. Figure 1 illustrates the token-based classification input.

Another scenario explores syntactic chunking. For example, parsing *I have to have all outdoor smoking allowed* utterance syntactically results in:

(1) *('i', 'B-NP'), ('have', 'B-VP'), ('to', 'I-VP'), ('have', 'I-VP'), ('all', 'B-NP'), ('outdoor', 'I-NP'), ('smoking', 'I-NP'), ('allowed', 'B-VP')*

where B and I prefixes indicate chunk boundaries. Figure 2 demonstrates how the parsed utterance in (1) is annotated with dialogue act information. Similarly, semantic chunks are constructed when parsing with predicate-argument structures using SENNA parser. For prosodic chunk-based classification, speech signals were analysed computing Mel-Frequency Cepstral Coefficients (MFCCs) [22] for each elementary speech unit (e.g. allophone) including silence units. When a silence unit of >200ms is identified, the ASR cuts segments. The resulted chucks correspond roughly to inter-pausal units [23].

3.3. Experimental procedures

Various classifiers were built to operate on tokens and chunks as an input. For such complex task as multidimensional incremental dialogue act classification, splitting up the (in-) and output structures may make the task more manageable. Knowing that a particular dimension is addressed makes a decision on certain communicative functions much easier, and vice versa. For this, stratified 3-fold cross-validation experiments were performed using *cascade* and *hierarchical* classification procedures. The first set of local classifiers was trained to segment and classify based on dimension labels. Dimension class predictions were added as features to perform communicative functions classification (cascading). The second set of classifiers was trained to segment and classify communicative functions, and the classification continued on a higher level grouping related functions together per dimension (hierarchical) as defined in ISO 24617-2 DA hierarchical taxonomy. Additionally, input features and predictions from various cascade and hierarchical classifiers were fused to build local multi-class classifiers predicting complex class <BIO-prefix_d;cf> labels (we called it the *early fusion* (EF) decision level). The performance of early fusion prediction models obtained applying one of the specified layered training procedure has been compared with joint classification (JC) of complex label sequences. Local classification design is depicted

Table 2: *Classification results in terms of F-scores obtained in 'real' (ASR-based) and 'simulated' (based on manual transcriptions) experimental settings on different type of tested input applying hierarchical and cascade local classification procedures, early fusion (EF) and join classification (JC), and late fusion (LF) methods.*

Setting	Simulated							Real						
Classification task	cascade		EF	hierarchical		EF	JC	cascade		EF	hierarchical		EF	JC
	d	cf	$< d; cf >$	d	cf	$< d; cf >$	$< d; cf >$	d	cf	$< d; cf >$	d	cf	$< d; cf >$	$< d; cf >$
Token-based	0.98	0.81	0.80	0.97	0.80	0.80	0.79	0.99	0.79	0.77	0.96	0.77	0.71	0.7
chunk-based (syntactic)	0.98	0.85	0.84	0.96	0.83	0.82	0.8	0.98	0.78	0.70	0.96	0.74	0.64	0.69
chunk-based (semantic)	0.98	0.84	0.84	0.96	0.82	0.82	0.8	0.98	0.75	0.70	0.95	0.74	0.65	0.69
chunk-based (prosodic)	na							0.98	0.75	0.72	0.94	0.73	0.66	0.66
LF: Majority Voting	na		0.85	na		0.82	0.79	na		0.78	na		0.76	0.72
LF: Meta-classification	na		0.86	na		0.82	0.80	na		0.80	na		0.77	0.72

in the left part of Figure 3.

Given various local training conditions, we created possible output prediction (hypotheses) space. Often local predictions once made are never revisited. Humans, by contrast, may revise their previous decisions while interpreting utterances. Technically, the revision of previously made decisions may be a rather costly procedure, i.e. backtracking can get easily very complex. On the other hand, it is even more undesirable and costly to update the dialogue context model with locally computed hypotheses that may contain many errors. This may lead to wrong or unexpected dialogue system behaviour which is much harder to correct requiring initiation of a clarification sub-dialogue and/or error recovery strategy which is incorporated into the dialogue management procedures. One option is to search the local partial output space for the best predictions using, for example, *majority voting* [24] methods. This may not be always the best strategy, however, since the highest-ranked predictions are not always correct in a bigger (or global) context. A rather straightforward and efficient solution which requires minimum computations is to provide a CRF classifier with more contextual information, but still keep it local as possible (i.e. avoid to look too much into the future). To optimize the overall classification decision taking process, feature vectors were automatically resynthesised (fused) combining local features and the predicted output history from multiple local classifiers of various types, see Figure 4. Subsequently, *meta-classifier* is trained to make the final decision, see the right part of Figure 3. Making use of the partial output predicted so far, i.e. of the history of previous predictions, and taking this as features into the next classification step helps discovering and learning from mistakes, correcting errors, and making more accurate predictions.

4. Classification results

We conducted two series of experiments. Token-based classification was compared with syntactic and semantic chunk-based classification using manual speech transcriptions. The token-based classifier yields an F-score of .80 over a baseline of .63[6] In syntactic and semantic chunk-based classification F-score of .84 was achieved. This shows that using syntactic and semantic chunks as increments for DA recognition may improve recognition results significantly ($p < .05$, McNemar's test [25]).

We assessed classifiers performance using multi-and single-layered training approaches. We experimented using both cascading and hierarchical procedures as discussed above. Results show that the cascading approach outperforms the hierarchical approach although the differences are not statistically significant in the simulated setting. It suggests that cascading approach is a better classification strategy; information about what dimension is addressed is beneficial for communicative function identification. Two-layered approach is more powerful than the joint classification of complex labels sequences.

For the second set of experiments, we included prosodic chunk-based classification in our work flow based on ASR output with a word error rate (WER) of 31.59%.[7] ASR errors present a challenge especially to the chunk-based classification. ASR output often differs for one and the same utterance produced by different speakers or even by the same speaker in slightly different acoustic conditions making chunking rather inconsistent.

While token-based classification can be modelled rather robustly using skip n-gram leaving out one or more potentially misrecognised tokens, syntactic and semantic analyses are troublesome when almost every third token is misrecognised. Prosodic chunking is often inconsistent as the same utterance produced by a different speaker and/or under slightly changed acoustic conditions may result in different prosodic structures. Cascading approach in this case showed significant improvements over the hierarchical approach and the single-layered approach. Table 2 summarises our best experimental results.

It can be also observed that meta-classification strategy is superior to individual local classifiers and outperforms the majority voting strategy in a real ASR-based setting. The incremental classifiers performance is comparable to the one of non-incremental SVM-based classifiers which reached the F-score of .85 on average [1] for transcribed data.

5. Conclusions and future research

Incremental dialogue act recognition is acknowledged to have the advantage that parts of an utterance are interpreted by the system before the last utterance token is processed. We have presented a machine learning based approach to incremental dialogue act classification with a focus on recognising dimensions and communicative function. We explored different local classification procedures assessing classifier performance, and proposed a meta-classification approach with meta-features synthesized from local classifiers. Compared to token-based incremental classification our syntactic and semantic chunk-based classification produce better results on manual transcriptions. In reality, where ASR output contains many errors, token-based incremental recognition is proven to be more reliable and robust. The proposed methodology accounts for empirically motivated and technically sound classification procedures that may reduce training costs significantly.

In future, a full-scale implementation and testing will be performed where the dialogue system will be able to manage partial update processes.

[6] In all experiments, the classifiers performance on a single easy computable feature, namely bag-of-tokens, has been used as a baseline.

[7] It should be noticed that the corpus contains a significant proportion of non-native English speakers. The used ASR system is the state-of-the-art open source Kaldi-based ASR system [26] and is set to send the most likely recognized sequence of words (1-best hypothesis) for the further processing.

6. References

[1] D. Amanova, V. Petukhova, and D. Klakow, "Creating annotated dialogue resources: Cross-domain dialogue act classification," in *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Paris, 2016.

[2] V. Petukhova, *Multidimensional Dialogue Modelling. PhD Thesis*. The Netherlands: Tilburg University, 2011.

[3] H. Bunt, "Multifunctionality in dialogue," *Computer, Speech and Language*, vol. 25, pp. 222–245, 2011.

[4] ISO, *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. Geneva: ISO Central Secretariat, 2012.

[5] N. Haddock, "Computational models of incremental semantic interpretation," *Language and Cognitive Processes*, vol. 14 (3), pp. SI337–SI380, 1989.

[6] D. Milward and R. Cooper, "Incremental interpretation: applications, theory, and relationship to dynamic semantics," in *Proceedings COLING 2009, Kyoto, Japan*, 2009, pp. 748–754.

[7] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 710–718.

[8] G. Aist, J. Allen, E. Campana, C. Gomez Gallo, S. Stoness, M. Swift, and M. K. Tanenhaus, "Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods," in *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, R. Arstein and L. Vieu, Eds., Trento, Italy, 2007, pp. 149–154.

[9] D. DeVault and M. Stone, "Domain inference in incremental interpretation," in *Proceedings of the Workshop on Inference in Computational Semantics*, INRIA Lorraine, Nancy, France, 2003, pp. 73–87.

[10] M. Paetzel, R. Manuvinakurike, and D. DeVault, "So, which one is it? the effect of alternative incremental architectures in a high-performance game-playing agent," in *Proceedings of the SIGDIAL 2015 Conference*, Prague, Czech Republic, 2015, p. 7786.

[11] Z. Szabó, "Compositionality," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2017.

[12] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka, and T. Kawabata, "Understanding unsegmented user utterances in real-time spoken dialogue systems," in *Proceedings of the 37th Annual Conference of the Association of Computational Linguistics, ACL*, 1999, pp. 200–207.

[13] P. Lendvai and J. Geertzen, "Token-based chunking of turn-internal dialogue act sequences," in *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, 2007, pp. 174–181.

[14] V. Petukhova and H. Bunt, "Incremental dialogue act understanding," in *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, J. Bos and S. Pulman, Eds., Oxford, UK, 2011, pp. 235–245.

[15] Q.-H. Tran, Z. I., and G. Haffari, "A hierarchical neural model for learning sequence of dialogue acts," in *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2017.

[16] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML '01*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[17] V. Petukhova, C. Stevens, H. de Weerd, N. Taatgen, F. Cnossen, and A. Malchanau, "Modelling multi-issue bargaining dialogues:data collection, annotation design and corpus," in *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Paris, 2016.

[18] J. Geertzen, V. Petukhova, and H. Bunt, "A multidimensional approach to utterance segmentation and dialogue act classification," in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. Antwerp, Belgium: Association for Computational Linguistics, 2007, pp. 140–149.

[19] J. Cohen, "A coefficient of agreement for nominal scales," *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[20] M. Marcus, M. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[21] M. Bansal, K. Gimpel, and K. Livescu, "Tailoring continuous word representations for dependency parsing." in *ACL (2)*, 2014, pp. 809–815.

[22] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the mfcc in speech and speaker recognition," *Acta Scientiarum Naturalium-Universitatis Pekinensis*, vol. 37, no. 3, pp. 371–378, 2001.

[23] Y. Ishimoto, T. Tsuchiya, H. Koiso, and Y. Den, "Towards automatic transformation between different transcription conventions: Prediction of intonation markers from linguistic and acoustic features," in *Proc. 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

[24] E. Morvant, A. Habrard, and S. Ayache, "Majority vote of diverse classifiers for late fusion," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer Berlin Heidelberg, 2014, pp. 153–162.

[25] P. Lachenbruch and C. J. Lynch, "Assessing screening tests: extensions of mcnemar's test," *Statistics in medicine*, vol. 17, no. 19, pp. 2207–2217, 1998.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.