A Hybrid Model for Spontaneous Speech Understanding

Tong Zhang, Mark Hasegawa-Johnson, Stephen E. Levinson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
405 N. Mathews Ave., Urbana, IL 61801
{tzhang1, hasegawa, sel}@ifp.uiuc.edu

Abstract

This paper presents a hybrid model to understand spontaneous speech by combination of speech *connotation* and *denotation* analyses. The testbed of the approach is an intelligent tutoring system (ITS) collected by *Wizard-of-Oz* (WoZ) simulations. The children users are inexperienced and their utterances are often dysfluent and have loose grammar structure. To robustly understand spontaneous speech in the tutorial environment, we categorize the user utterances into 30 *tutoring events*, which can reflect the content meaning of utterances in a broad and shallow way. The objective of this study is to classify utterances into the target events. The tutoring event classification integrates speech connotation analysis and speech denotation analysis. The speech connotation analysis intends to model the *cognitive states* of students by three classes: *confidence*, *puzzlement*, and *hesitation*. The speech denotation analysis intends to compute the event-utterance similarity based on the *TF•IDF* vector of some pragmatically and semantically salient words embedded in the utterances. We define salient words by those words that contain novel information neither presupposed by the interlocutor nor denoted in the precedent part of the utterance. We used speech and transcribed text for experiments, and achieved 75.5% accuracy when the salient words were manually annotated. The accuracy reduced by 15.4% relative when the salient words were automatically extracted.

Introduction

The application platform for the work described in this paper is an intelligent tutoring system in basic math and physics, designed for children of elementary and early middle-school ages (Zhang, et al., 2004). The complete system does not exist; experiments described in this paper make use of data acquired through Wizard-of-Oz simulations, using a mock-up of the finished tutoring system.

The interpretation of children's speech in the ITS dialogue scenario requires robustness. Adult users of a typical dialogue system (e.g., for purchase of air travel or financial instruments) are usually able to learn, over a number of repeated interactions with the system, what actions are possible at each stage of the dialogue. By contrast, children users of our intelligent tutor are perpetually naïve with respect to the future content of the dialogue. Each child paticipates in at most three experiment sessions and each session has different tutorial content, because we do not ask children to relearn knowledge that they have mastered. Moreover, we encourage children to participate in our experiments and instigate their interestes in scientific learning by asking them open-ended questions rather than close-ended questions. For example, when the child is turning gears and the tutor wants to ask the child about the motion of the gears, he usually does not ask questions requring single correct/wrong answers, e.g., *In which direction are the gears turning?* Instead, he would ask *What are you noticing?*

Since children are not familiar with the experiment contents and the answers to open-ended questions are usually longer and more complicated than those to close-ended questions, their utterances are even more incoherent and dysfluent than is typical in interpersonal conversations. The utterances usually include loose grammar structure, fragmentary, restart, repair, meaningless speech (e.g., *That if the...*), and repetition. For example, *Ahmm when you...after it goes around once, the other one goes around the same, the same, I mean it goes around...you know you only have to spin it around once, and that makes sense basically because they are the same size.* Therefore, we need to understand children's heavily dysfluent and ungrammatical speech in a robust way. The existing SLU techniques usually concentrate on extracting semantic concepts embedded in spoken utterances. The extraction of semantic concepts requires appropriately segmentating text into syntactically legitimate units, and then selecting a semantically correct result from a potentially large number of syntactically legitimate candidates. Therefore, it is inefficient in dealing with the inputs that are outside of the system's grammatical coverage, if the grammar can be predefined.

In this tutoring scenario, we robustly understand spontaneous speech by categorizing user utterances into tutoring events. Two students worked together to categorize the content meaning of user utterances into 30 tutoring events. Their work was based on speech perception, transcription, and dialogue context. Some of the

Table 1: Some tutoring events and example utterances

Tutoring Event	*Description*	*Example*
IrreleventQuestion	The question is irrelevant to the experiment content	U: Do you guys have a TV here?
AskForPlayInstruction	A question requesting the instruction on how to play the Legos	U: Do I have to use all the beams?
IncompleteAnswer	Incomplete answer to a question on spinning speed or direction	T: You've tried all these 2 gear combinations. So why do you think it is? Why do you think it happens? U: The stronger the weaker, no, the bigger the weaker, the stronger...
SpinDirection	Talk about the spinning directions	T: In what direction does the medium gear push the small gear? U: Umm, it pushes it in the same, in the opposite direction that the left gear does the center gear.
SpinSpeed	Talk about the spinning speed and turning times	T: How did you know that? U: The big gear for every one turn, the small gear turns five times.
ColorLines	Each gear is painted with different pairs of colors to bring children the convenience to decide if gears are lined up.	T: What did you find? U: When you change them they change colors. When you match them up they stay the same colors when you match them up.
ArithmeticComputation	Perform arithmetic computation	T: Do you know the relationship between 8 and 24? U: 8 times 3 is 24.
Accept	Accept the suggestion, request, or opinions of the tutor	T: Can we start with the gears like this again? U: Yes, let's try again. Let's try.
RequestToCount	Request the tutor to count the spinning times	U: Let's count, let's count to two this time.
VoidMeaning	The utterance is not meaningful	T: What else do you notice? U: That if the…
ExplainAction	The user explains what is being done	U: I'm gonna replace the 40 gears with the 24 gears.

tutoring events and their sample utterances are listed in Table 1. The tutoring events reflect the content meaning of speech in a shallow and broad way: (1) Sometimes an utterance is long, dysfluent, loosely grammatical, and even incoherent, then it is very hard to derive the exact meaning of the utterance. In this case, we require interpretation to be only *approximately correct*, which only require the gist (i.e., the basic content) of the utterance to be detected. In the following example (T is tutor and U is user),

T: Why do they turn in the same direction?
U: Because of the color, well, the way you put on them. Because if I took both of the big ones like this, and turn them, um...the yellow side, and I spin both of them, and they would still be on the same side.

it is unable to obtain the exact meaning of the user's utterance. We can only get a rough estimate on what the user is saying: he is talking about the line up of gears using their color pattern. (2) The overall meaning of the spoken messages sometimes provides sufficient information for response. For example, when the tutor is giving a suggestion on the user's action (e.g., *Can you make it so both the red parts of gears are closest to you, like this?*), the tutoring event *reject* (e.g., *I don't know I can't move I can't easily move this big one.*) or *accept* (e.g., *Yes, I can.*) are able to summarize the user's response to the tutor's suggestion so as to help the computer make appropriate response. The objective of this study is to robustly understand spontaneous utterances by categorizing each utterances into one of the 30 tutoring events.

Related Work

The strategy of using tutoring event to understand spoken language in the tutoring scenario is *topic identification* in spoken language understanding. The purpose of topic identification is to detect semantic concepts from an utterance, and then classify the utterance to one of the predefined categories. A semantic concept can be composed of a syntactic constituent or a single word. Topic identification has been successfully used in automatic call centers or call routers. For example, spontaneous speech is understood by automatic detection of a set of salient grammar fragments, each of which is a finite-state-machine cluster of semantically similar salient phrases with variable length. The salient grammar fragments are then used as textual features for call-type classification (Arai, et al., 1998; Gorin, et al., 2002). The vector space model is another method applied to the automatic call type classification (Chu-Carroll and Carpenter, 1999). The call router uses desired destinations and non-stop words extracted from caller requests to compose a feature vector space. In the routing matrix R, each element $R_{m,n}$ of R represents the degree of association between the m^{th} relevant term and the n^{th} destination computed by the $tf \bullet idf$ measure. In addition, researchers extracted semantic concepts by weighted finite-state transducer, and then identified the calls with a multi-layer perceptron neural network given a vector of binary values indicating the existence of the predefined concepts (Wutiwiwatchai and Furui, 2004). Unlike parsing methods, topic identification ensures utterances to be classified regardless of the utterance length and the complexity of linguistic structure.

Corpus Analysis

Our ITS corpus consists of 714 utterances, containing approximately 50 mins of relatively clean speech. On average each utterance has 4.2s speech and 8.1 words. The vast majority of the utterances contain between 1 and 20 words, while the longest utterance has 57 words. The 714 utterances of the ITS corpus can be partitioned into 58 single-word utterances, 26 single-phrase utterances (such as *how-many?*), 22 utterances that are merely repetitions of the human wizard, 19 utterances that are not semantically meaning, and 589 non-single word/phrase utterances containing novel information. We use the 630 (589 novel-meaning utterances, 22 repetitions of the tutor, 19 semantically void utterances) non-single word/phrase utterances for tutoring event classification. The task perplexity provides a measure of the difficulty in classifying samples drawn from their distribution. The probability distribution of the tutoring events has an entropy of 4.49. Therefore, our task perplexity is $2^{H(x)} = 22.52$.

Strategy

Speech has broad spectrum in terms of information revealing the intentions of the speaker: explicit literal meaning and implicit connotation. For a spoken message, the semantic content explicitly elicited is referred to as its denotation, while the internal conditions of the speaker, such as emotion, attention and attitude, are part of the message's connotation. Speech in our ITS corpus carries information closely related to the student's cognition, which reflects the student's mental activities during the process of knowledge acquisition. We categorize the cognitive activities of students into three states: *confidence*, *puzzlement*, and *hesitation*. Confidence means the users answer questions or explains their actions in confident mode, or make commands. Puzzlement means the users ask questions or states the lack of knowledge. Hesitation means the users answer questions in hesitant or uncertain mode. The cognitive states are classified based on the lexical, prosodic, spectral, and syntactic analyses (Zhang, et al., 2004), and then information fusion by a decision tree (Rulequest Research, 2004).

The tutoring events can be recognized more efficiently if one has knowledge of the user's cognitive state. This is because:

1. Many tutoring events are typically associated with particular cognitive states. For example, *hesitation* is more likely to accompany a tutoring event related to an incomplete answer (e.g. *I saw the …yellow part…*) or a wrong answer (e.g., *It…I think it goes around… one and a half times*) than a correct answer (e.g., *The large gear has five times as many teeth as the small ones*); *puzzlement* is a strong indicator of question-related tutoring events (e.g., *Do I have to use all the beams?*).
2. Compared with adult's speech, children's speech has very different acoustic characteristics, which cause degradation in speech recognition performance (Narayanan and Potamianos, 2002). The high word error rate of speech recognition brings difficulty and inaccuracy to subsequent understanding. The case is worse especially when the verbal content is hard to recognize or interpret. For example, when a user is not certain how to answer a question, his/her speech may therefore be dysfluent and incomplete, resulting in higher speech recognition word error rate. However, *hesitation* is easy to get recognized using prosodic and spectral clues that are directly derived from the speech signals. The *hesitation* recognition helps to classify tutoring events and understand the user's situation.

As shown in Figure 1, the tutoring event classification system uses a hybrid model that integrates speech connotation and speech denotation by a decision tree. The speech denotation analysis module consists of three components: salient word extraction, tutoring event

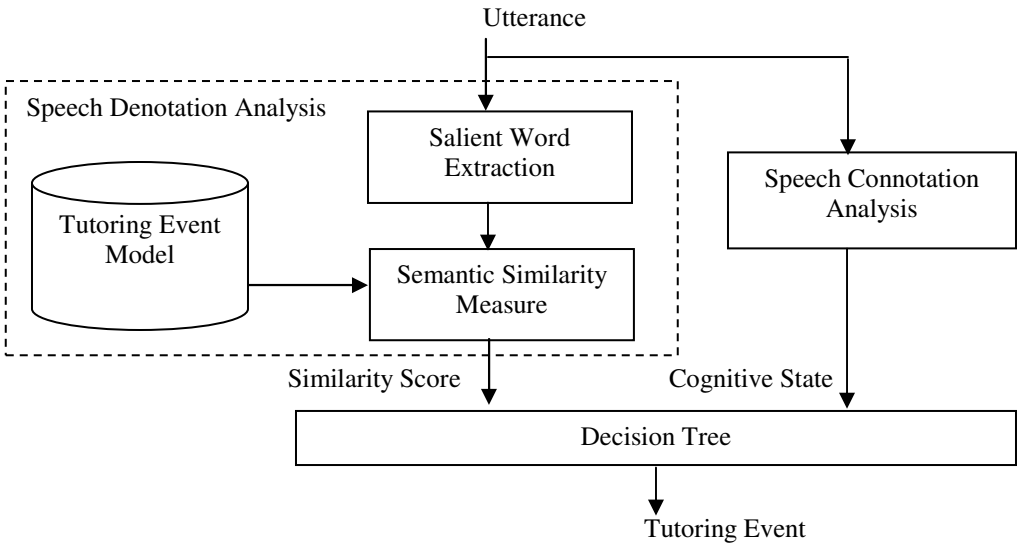

Figure 1: Architecture of the tutoring event classification system.

modeling, and semantic similarity measure. We describe the three components in detail in the next section.

Speech Denotation Analysis

Salient Word Extraction

In this study, we intend to automatically detect the pragmatically and semantically salient words that contain novel information neither presupposed by the tutor nor denoted in the precedent part of the utterance. For example (the salient words are marked with **bold**),

T: What happens to the different gears as you spin the one at the end?
*U: They **move** with the single gear that I'm spinning.*
T: Oh, are you having fun?
*U: **Yeah**, it's kind of interesting.*

The highlighted or unexpected constituents within an utterance are often marked by pitch accents (Kadmon, 2001). The prosodic features closely associated with pitch accent are duration, pitch, energy, and spectral balance cepstral coefficients (Ren, et al., 2004). In addition, part-of-speech (POS) is used as an information source since function words usually are not salient by not containing novel information. POS of all words in the ITS corpus is first tagged using an automatic POS tagger (Munoz, et al., 1999), and then is manually checked against the tagging standard in Treebank-3 (Santorini, 1990).

The semantic meaning of words is the direct indicator of novelty. Let N_i denote the novelty of word w_i given the dialogue context. We compute N_i by the minimum of dissimilarity between w_i and other words in the set S, where S consists of those words appearing in the tutor's presupposition and those precedent of w_i in the utterance, i.e.

$$N_i = \min_{w_j \in S} dis(w_i, w_j). \quad (1)$$

Research in computational linguistics has developed various methods to compute the degree of *semantic similarity* between a pair of words. The methods are basically divided into two strategies: (1) *ontology hierarchies* (e.g., Lee, et al., 1993; Sussana, 1993; Resnik, 1995; Jiang and Conrath, 1997)—ontology is a structural system of categories or semantic types, so that knowledge about a certain domain can be organized through the categorization of the entities of the domain in terms of semantic types; and (2) *corpus statistics* (e.g., Lin, 1998; Pantel and Lin, 2002; Thelen and Riloff, 2002; Terra and Clarke, 2003)—three statistics are commonly employed to model the similarity of words (Higgens, 2004):

Topicality assumption: similar words tend to have the same neighboring content words.
Proximity assumption: similar words tend to occur near each other. Word senses are ultimately grouped according to proximity of meaning.
Parallelism assumption: similar words tend to be found in similar grammatical structures.

Application-Oriented Ontology. Designing ontology actually means to determine the set of semantic categories which properly reflects the particular conceptual organization of the domain. In this study, we adapt *WordNet*, a general linguistic resource, to our ontology construction in an application domain. Partial of the ontology is shown in Figure 2 using a tree structure. Then we employ the *edge-based* method, in which an edge represents a direct association between two concepts, to compute the distance between a pair of words in the

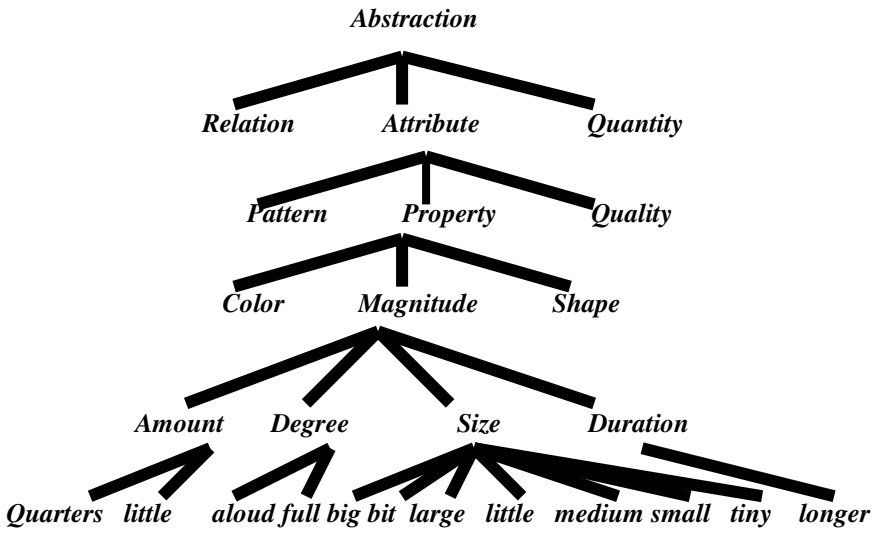

Figure 2: Partial hierarchy of the application-oriented ontology.

ontology. Generally, the distance shrinks as one descends the hierarchy, since differentiation is based on finer and finer details (Jiang and Conrath, 1997). Therefore, we propose the following weighted edge distance measure:

$$dis(w_1, w_2) = \min_{c_1 \in sem(w_1), c_2 \in sem(w_2)} \left(\frac{d_{c_1,c_2} + 1}{d_{c_1,c_2}}\right)^{\alpha} len(c_1, c_2), \quad (2)$$

where $sen(w)$ denotes the set of possible senses for word w in case w has multiple senses; d_{c_1,c_2} is the mean depth of nodes c_1 and c_2 in the hierarchy; $len(c_1, c_2)$ is the smallest number of edges connecting c_1 and c_2; α is constant and we choose $\alpha = 3.0$ here.

Corpus Statistics. We use *GigaWord*, a billion-word archive of English newswire text and distributed by the Linguistic Data Consortium, as the text database for corpus statistics. Given a context $C = \{w'_1, w'_2, ..., w'_n\}$, a pair of words w_1 and w_2 are more semantically similar if they are more likely to co-occur with the words in the context. Therefore, the similarity between w_1 and w_2 can be computed by the cosine value between the two partial mutual information (PMI) vectors corresponding to w_1 and w_2 (Pantel and Lin, 2002). Then we have our dissimilarity measure:

$$dis(w_1, w_2) = 1 - \frac{\sum_{w' \in C} PMI(w', w_1) PMI(w', w_2)}{\sqrt{\sum_{w' \in C} PMI(w', w_1)^2} \sqrt{\sum_{w' \in C} PMI(w', w_2)^2}}, \quad (3)$$

where $C = C(w_1) \cup C(w_2)$, $C(w_1)$ and $C(w_2)$ are the context of w_1 and w_2, respectively, and

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1) p(w_2)}. \quad (4)$$

Knowledge Combination. The computation of word semantic dissimilarity is a combination of lexical semantics and corpus statistics by linear interpolation

$$dis(w_1, w_2) = \lambda dis_1(w_1, w_2) + (1 - \lambda) dis_2(w_1, w_2), \quad (5)$$

where dis_1 and dis_2 are the peak-normalized word distances based on the ontological method and the statistical method, respectively; λ is constant. The prerequisite to use corpus statistics is that there must be at least two contextual words for a given pair of words (w_1, w_2). Otherwise, dis_2 will be 1 (# of contextual words is 0) or 0 (# of contextual words is 1). If the prerequisite cannot be satisfied, then dissimilarity is computed using only the ontology-based method.

Prosodic observations, POS tagging, and semantic novelty measure are integrated using a time-delay recurrent neural network (TDRNN). TDRNN is a neural network that models the dynamic context using a combination of delayed input nodes and delayed recurrent nodes (Kim, 1998).

Tutoring Event Modeling

Each event model consists of a set of salient words extracted from the component utterances of that event. We notice that some salient words share common attributes, and the clusters of these salient words are more robust representative than the salient words themselves. We call these clusters salient concepts, some of which are listed in Table 2. In addition, the multiple tenses of a word are

clustered, e.g., 'push,' 'pushes,' and 'pushing' are clustered together.

Table 2: Some semantic concepts and their component salient words

Semantic Concepts	*Component Words*
Color	blue, color, colors, grey, red, red's, reds, white, yellow, yellow's yellows
Fraction	a-half, a-quarter, half, quarter
Meet	Even, evens, line, lined, meet, meeting, touch, touching
TurnTimes	once, twice, X-times (X denotes digit)

Each event model is represented by $E = \{w_1, c_1, w_2, c_2 \ldots w_n, c_n\}$, where $c_1 \ldots c_n$ are salient words/concepts extracted from the component utterances of event E, and $w_1 \ldots w_n$ are the corresponding weights derived by the *tf* • *idf* method (Sparck Jones, 1972), i.e. for c_i and E_j,

$$w_{ij} = tf_{ij} \bullet \log_2(idf_i), \quad (6)$$

where tf_{ij} is the percentage of utterances containing c_i in event E_j, idf_i is the frequency of events containing c_i, and

$$idf_i = \frac{total\ \ number\ \ of\ \ events}{number\ \ of\ \ events\ \ containing\ \ c_i}. \quad (7)$$

Semantic Similarity Measure

The semantic similarity between an incoming utterance and each of the tutoring event candidates is measured by *lexical similarity*, which promotes an unstructured approach better reflecting the unconstrained nature of human language. In computing the event-utterance lexical similarities, we consider three factors: (1) the length of the utterance—a long utterance tends to contain more salient words/concepts than a short utterance; (2) the type of events--different events tend to contain different amount of salient words/concepts. For example, event *ExplainAction* tends to contain more salient words/concepts than event *Accept*; and (3) the discrepancy among salient words/concepts in their contribution to the similarity measure—the salient words/concepts that are more representative of the event meaning should contribute more significantly to the similarity measure than the others. Therefore, we propose a weighted counting of common feature for $SS(u|E_j)$, the similarity between utterance u and event E_j:

$$SS(u \mid E_j) = \frac{\sum_{c_i \in (u \cap E_j)} w_i(E_j)}{len(u \cup E_j)}, \quad (8)$$

where c_i is the i^{th} salient word/concept contained by both utterance u and event E_j, w_i (E_j) is the corresponding weight of c_i in event E_j, and $len(u \cup E_j)$ is the number of salient words contained by either utterance u or event E_j.

System Evaluation

Cognitive State Classification

We tested cognitive state classification using the manual transcriptions and automatic recognition of the utterances, and yielded 96.6% and 95.7% accuracies, respectively. Table 3 lists the featue vector accuracy ranking in terms of average *F*-score.

Table 3: Feature vector accuracy ranking for cognitive state classification. Averaged *F*-score (sum of three one-class *F*-score divided by 3), based on classification using the manual transcription. Classification based on automatic transcription results in the same relative ranking of feature vectors.

Average F-score	*Feature Vector*
0.96	spectrum + prosody + lexicon + POS
0.95	Spectrum + prosody
0.94	spectrum
0.76	prosody
0.72	lexicon
0.49	POS

We also compared the robustness of different features to speech recognition errors, and present the results in Table 4. The table shows that the spectrum-based classification and prosody-based classification were robust to speech recognition errors, much more than the lexicon-based classification. As for recognized speech, the part-of-speech-based classification could not converge during learning, so we did not obtain the classification result. When spectrum and prosody were combined, classification correctness for recognized speech and transcribed speech were almost identical.

Table 4: Comparison of classification correctness between transcribed speech and recognized speech using different features

	transcribed speech	*recognized speech*	*recognized-transcribed (absolute)*
spectrum	94.6%	90.0%	-4.6%
prosody	85.4%	87.2%	1.8%
lexicon	81.6%	70.9%	-10.7%
part-of-speech	73.6%	-	-
spectrum + prosody	96.0%	95.5%	-0.5%

Speech Denotation Analysis

The ITS corpus has many question-answer pairs, in which the tutor initiates dialogue topics by asking questions or providing suggestions. In this case, presupposition of an utterance lies in the questions or suggestions of the tutor. Students sometimes initiate dialogue topics by making commands, asking questions or simply explaining on what they are doing. In this case, presuppositions for the students' utterances do not exist. Three annotators independently identified the salient words that we have defined based on perception, text transcription, and dialogue context. The consistency among the annotators was (Kappa score) $\kappa = 0.79$. Then we used majority voting to resolve the annotation inconsistency among the three annotators. We used 90% of the corpus for training and the remaining 10% for test. Our experiment yielded an accuracy of 83.8% on the test set, which consisted of 536 words. We also compared the features according to the accuracy of salient word extraction (see Table 5).

Table 5: Feature accuracy ranking for salient word extraction

Average F-score	*Feature Vector*
0.849	feature combination
0.656	word dissimilarity
0.593	duration
0.577	part-of-speech
0.546	spectral balance cepstral coeff.
0.345	energy
0.339	pitch

We can see that pitch played unexpectedly low efficiency in salient word extraction. Because of the noisy recording environment, it was hard to discriminate voiced regions from unvoiced regions. The energy of unvoiced regions carried information irrelevant to pitch estimate. So the automatically extracted pitch showed too many pitch tracking errors to be an efficient feature. Spectral balance cepstral coefficients showed better performance, possibly because the band-pass filters (Ren, et al., 2004) eliminated the disturbance of low frequency noise that adversely affected the pitch and energy estimates.

Tutoring Event Classification

We applied *See5* (Rulequest Research, 2004), a decision tree classifier, to combine the two distinct information sources, semantic similarity score and cognitive state, for tutoring event classification. The reason that we used decision tree was that it provided reliable performance when the amount of training data was small, and/or when the features were high dimensional. The tutoring event classification was independent on the dialogue state. Assume there are N event candidates ($N = 30$ here). Then the feature vector of a given utterance u is $f(u) = \{SS_1, SS_2, \ldots, SS_N, Cs\}$, where SS_i is the score of the semantic similarity of u to E_i, and Cs is the cognitive state of u. The features used for information fusion are listed in Table 6.

Table 6: List of the features for tutoring event classification

Feature	*Size*	*Description*
semantic similarity measure	30	Continuous
cognitive state	1	Character: confidence, puzzlement, hesitation
target	1	Discrete: 1, 2,…30 are labeled for E_1, E_2, …E_{30}, respectively.

We used the 10-fold cross-validation to evaluate our tutoring event classification system. Each time we randomly chose 397 (63%) samples for training and the remaining 233 (37%) samples for test. We first used the manually annotated salient words for the experiment, and the classification achieved 75.5% correctness. Then we used the automatically detected salient words for the experiment, and the classification accuracy reduced by 15.4% relative.

The information structure of an utterance can be partitioned into the presupposed part(s) and the non-presupposed part(s). But in our experiments, the literal meaning of an utterance was modeled by its salient words that excluded the presupposed information. Therefore, it seemed that the information representation of an utterance was not complete. However, our classification performance was satisfying. This was because: the inclusion of presupposition for information representation sometimes played positive function, sometimes played negative function, and sometimes played zero function for the identification of tutoring events. For example, if the tutor asked how much bigger a big gear was than a small gear, and the user answered "5 times", then we knew that "5 times" was about the *GearSize*. If the tutor asked how much faster the small gear was turning than the big gear, and the user answered "5 times", then we knew '5 times' was about the *SpinningSpeed*. In this case, presupposition would help the tutoring event identification. However, for example, if the target tutoring event was *AskToRepeat*, the tutor could say various kinds of things before the user asked him to repeat. In this case, if we brought the tutor's utterance (presupposition) into the utterance meaning representation, then it would bring complexity and difficulty for the target event identification. When the user initiated a dialogue topic by asking a question or making a command, then there was no presupposition, which meant that the function of presupposition was zero. If we wish to incorporate presupposition into the information representation of utterances' meaning in the future, we might need to distinguish which tutoring events need the presupposed information for identification, and which do not.

Conclusion

We have addressed how to robustly understand spontaneous speech by combining speech connotation and denotation analyses. We used an ITS dialogue scenario, which was collected by the WoZ simulations and full of dysfluencies and loose grammars, as the target application platform. We interpreted utterances by classifying them into 30 tutoring events, which summarized the content meaning of the ITS utterances in a broad and shallow way. We used cognitive state classification for speech connotation analysis. We defined and then automatically extracted salient words that encode the literal meaning of spoken messages. We used speech and transcribed text for experiments. The experiments achieved 75.5% accuracy when the salient words were manually annotated. The accuracy reduced by 15.4% relative when the salient words were automatically detected.

Acknowledgement

This work is supported by NSF grant number 0085980. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

References

Arai, K., Wright, J. H., Riccardi, G., and Gorin, A. L. 1998. Grammar fragment acquisition using syntactic and semantic clustering. *Proc. of ICSLP*.

Chu-Carroll, J. and Carpenter, B. 1999. Vector-based natural language call routing. *Computational Linguistics*, 25(3): 361-388.

Gorin, A. L., Abella, A., Alonso, T., Riccardi, G., and Wright, J. H. 2002. Natural spoken dialog. *IEEE Computer Magazine*, 35(4): 51-56.

Higgins, D. 2004. Which statistics reflect semantics? Rethinking synonymy and word similarity. *Intl. Conf. on Linguistic Evidence.*

Jiang, J. J. and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proc. of Intl. Conf. Research on Computational Linguistics.*

Kadmon, N. 2001. Formal Pragmatics. Malden, MA: Blackwell Publishers.

Kim, S.-S. 1998. Time-delay recurrent neural network for temporal correlations and prediction. *Neurocomputing*, 20:253-263/

Lee, J. H., Kim, M. H., and Lee, Y. J. 1993. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation, 49(2), 188-207.

Lin, D. 1998. Automatic retrieval and clustering of similar words. *Proc. of COLING-ACL*, Montreal, Canada.

Munoz, M., Punyakanok, V., Roth, D., and Zimak, D. 1999. A learning approach to shallow parsing. *EMNLP-WVLC*.

Narayanan, S. and Potamianos, A. 2002. Creating conversational interfaces for children. *IEEE Trans. on Speech and Audio Processing*, 10(2), 65-78.

Pantel, P. and Lin, D. 2002. Discovering word senses from text. *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*.

Ren, Y., Kim, S.-S., Hasegawa-Johnson, M., and Cole, J. 2004. Speaker-independent automatic detection of pitch accent. *ISCA Intl. Conf. on Speech Prosody.*

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proc. of the 14th Intl. Conf. on Artificial Intelligence*, 1, 448-453.

Rulequest Research. 2004. Data mining tools. *http://www.rulequest.com/see5-info.html*.

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank project. *Linguistic Data Consortium*.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application retrieval. *Journal of Documentation*, 28(1), 11-20.

Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. *Proc. of the 2nd Intl. Conf. on Information and Knowledge Management*.

Terra, E. and Clarke, C. L. A. 2003. Frequency estimates for statistical word similarity measures. *Proc. of the HLT-NAACL.*

Thelen, M. and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Proc. of the EMNLP.*

Wutiwiwatchai, C. and Furui, S. 2004. Hybrid statistical and structural semantic modeling for Thai multi-stage spoken language understanding. *Proc. of the HLT-NAACL Workshop*.

Zhang, T., Hasegawa-Johnson, M., and Levinson, S. E. 2004. Children's emotion recognition in an intelligent tutoring scenario. *Proc. of ICSLP.*