SPECOM 1490
15 August 2005 Disk Used
ARTICLE IN PRESS
No. of Pages 26, DTD = 5.0.1

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Speech Communication xxx (2005) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Extraction of pragmatic and semantic salience from spontaneous spoken English

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

Department of Electrical and Computer Engineering, Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews Avenue, Urbana, IL 61801, USA

Received 1 January 2005; received in revised form 16 July 2005; accepted 19 July 2005

Abstract

This paper computationalizes two linguistic concepts, *contrast* and *focus*, for the extraction of pragmatic and semantic salience from spontaneous speech. Contrast and focus have been widely investigated in modern linguistics, as categories that link intonation and information/discourse structure. This paper demonstrates the automatic tagging of contrast and focus for the purpose of robust spontaneous speech understanding in a tutorial dialogue system. In particular, we propose two new transcription tasks, and demonstrate automatic replication of human labels in both tasks. First, we define *focus kernel* to represent those words that contain novel information neither presupposed by the interlocutor nor contained in the precedent words of the utterance. We propose detecting the focus kernel based on a word dissimilarity measure, part-of-speech tagging, and prosodic measurements including duration, pitch, energy, and our proposed spectral balance cepstral coefficients. In order to measure the word dissimilarity, we test a linear combination of ontological and statistical dissimilarity measures previously published in the computational linguistics literature. Second, we propose identifying *symmetric contrast*, which consists of a set of words that are parallel or symmetric in linguistic structure but distinct or contrastive in meaning. The symmetric contrast identification is performed in a way similar to the focus kernel detection. The effectiveness of the proposed extraction of symmetric contrast and focus kernel has been tested on a *Wizard-of-Oz* corpus collected in the tutoring dialogue scenario. The corpus consists of 630 non-single word/phrase utterances, containing approximately 5700 words and 48 min of speech. The tests used speech waveforms together with manual orthographic transcriptions, and yielded an accuracy of 83.8% for focus kernel detection and 92.8% for symmetric contrast detection. Our tests also demonstrated that the spectral balance cepstral coefficients, the semantic dissimilarity measure, and part-of-speech played important roles in the symmetric contrast and focus kernel detections.

* Corresponding author. Tel.: +1 217 328 1542; fax: +1 217 244 8371.
E-mail addresses: tzhang1@ifp.uiuc.edu (T. Zhang), hasegawa@ifp.uiuc.edu (M. Hasegawa-Johnson), sel@ifp.uiuc.edu (S.E. Levinson).

0167-6393/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.specom.2005.07.007

Keywords: Spoken language understanding; Spoken dialogue systems; Computational linguistics; Information extraction

1. Introduction

Words are tools; in real speech, every word is deployed for the purpose of achieving a human goal. The fields of computational semantics and pragmatics study quantifiable goal variables—variables that encode quantifiable aspects of the goals served by a word in context—and their semantic and contextual correlates. This paper describes the computation of two semantic and pragmatic goal variables, *focus* and *contrast*, from spontaneous speech.

The paper is organized as follows. The remainder of Section 1 explains why we are interested in annotating focus and contrast, defines the aspects of focus and contrast that are under study with examples from an intelligent tutoring system (ITS) corpus, and then puts forward the objectives of our study in this paper. Section 2 provides some background in support of our work and describes related work in modern linguistics and computational linguistics. Section 3 describes the ITS corpus in detail, with particular attention paid to annotations and corpus statistics of the proposed focus and contrast variables. Sections 4 and 5 describe the algorithms implemented for the purpose of detecting the proposed focus and contrast variables: Section 4 describes prosodic analysis, and Section 5 describes the measurement of word semantic similarities. Section 6 describes system integration and results of experimental evaluation using the ITS corpus. Section 7 discusses and concludes our work.

1.1. Motivation

The motivation of this study is to achieve robust spontaneous spoken language understanding (SSLU) in an intelligent tutoring dialogue system. The system intends to provide a computer-based environment for education in math and physics, using the *Lego* construction set, with children of primary and early middle school ages (9–12 years old). Due to the characteristics of the dialogue scenario as we describe in Section 3.1, the children users' spontaneous utterances are often dysfluent, ungrammatical, and even incoherent. Our robust speech understanding system design under these circumstances basically involves two steps: (1) Classification of each utterance into one of a list of 30 tutoring events. Similar to *call types* in an automatic call center or call router (Gorin et al., 2002; Chu-Carroll and Carpenter, 1999), the tutoring events are used to summarize the content meaning of utterances in the tutoring dialogue scenario in a broad way. For example, the tutoring event *AskForPlayInstruction* means that the user asks a question requesting the instruction on how to play the Legos; *SpinSpeed* means that the user is talking about the spinning speed of the Lego gears; and *ExplainAction* means that the user explains what is being done with the Legos. (2) Sometimes the tutoring event itself cannot provide sufficient information for the computer to pop up proper response. For example, when the tutoring event *ArithmeticComputation* is detected, sometimes the computer needs to know what the type of the arithmetic computation is; if it is division, then the computer needs to know what the dividend and divisor are for proper response. Such detailed information needs syntactic/semantic structure parsing or named entity recognition (Zhang, 2004).

To analyze the content meaning of an utterance, we are interested in extracting a small set of words, from the utterance, that encode pragmatically and semantically salient information. We investigate the computerization of two linguistic concepts, focus and contrast, that are assumed to be useful for content summarization and structure parsing of spoken messages. Both of the concepts have reasonably clear published definitions. We wish to adapt the published definitions as necessary in order to define a corpus transcription experiment, and to train and test algorithms that automatically detect these two categories of salience based on cues measured in the speech waveform and in its orthographic transcription.

1.2. Focus

The information structure of a sentence can be partitioned into *presupposition* and *focus*: presupposition is what the interlocutor assumes to be true when the sentence is elicited in a conversation, and focus is the non-presupposed part of the sentence (Chomsky, 1971; Zubizarreta, 1998). For example (T represents tutor and U represents user. Focus is marked by $[\ ]_F$),

(1) *T: What are you exploring there?*
U: [Seeing if the small gears move the big gears.]$_F$

(2) *T: How many times does the small gear spin until they line up again?*
U: I think it goes around [one and a half]$_F$ times.

By definition focus is indicative of pragmatically and semantically new information not presupposed by the interlocutor. If focus can be reliably detected, it should be possible to use the distinction between focus and presupposition to detect new information embedded in an utterance. Speakers will often signal focus of a sentence by the use of *pitch accent* (we use pitch accent to mean prosodic prominence marked by *F*0 extrusion; the same word is usually also marked by the other acoustic correlates of prominence, including duration, energy, and spectral balance). Pitch accent marks the constituents within an utterance as highlighted or unexpected; it has been argued that constituents outside focus are expected, and hence tend to be unaccented (Kadmon, 2001; Zubizarreta, 1998; Hedberg and Sosa, 2001). For example (pitch accented words are marked with subscript *a*),

(3) *T: Which gear are you counting?*
U: I am counting the small$_a$ gear.

(4) *T: Which gear do you think is the strongest?*
U: Probably the large$_a$ gear.

However, the phonological manifestation is not straightforward: pitch accent can only approximate the location of focus in a sentence. For example, in the sentence *They turn in the opposite$_a$ direction,* the accented word 'opposite' is focus for the question *What can you tell me about the directions they turn* but not for the question *What else do you notice?* The latter question requires the sentence 'they turn in the opposite direction' to be focus for interpretation. Such ambiguity in pragmatic interpretation of single accent has been known traditionally as the *focus projection* phenomenon, demonstrating that focus expressed by a single accent can project to a larger linguistic constituent than just the word with pitch accent.

Since focus is a syntactic constituent, the boundaries of focus need to be determined to identify focus. A sentence may have multiple foci, and the size of a focus may vary from a single word to a phrase or even a sentence. It is difficult to automatically extract syntactic constituents containing novel information without making use of a complete parse tree for the sentence in question. Even with a parse tree available, automatically selecting the right constituents would be difficult; for example, in the following exchange,

(5) *T: Which gear are you counting?*
U: I am counting the [small]$_F$ gear [in my hand]$_F$,

it would be difficult for an automatic algorithm to determine that focus consists of a single word and a prepositional phrase; it would be nearly impossible without access to a correct parse of the sentence. It is even harder to extract focus from spontaneous speech, since spontaneous speech often has loose grammar structure, dysfluency, and inconsistency between linguistic segments and acoustic segments. For example ('...' represents silence):

(6) *T: What happens when you spin the left gears?*
U: 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.]$_F$

To robustly understand spontaneous speech, we propose labeling individual words containing new

information neither presupposed by the interlocutor nor contained in the preceding part of the utterance. Such a word is usually a content word because of the information content requirement. We hypothesize that words matching this definition will typically be the semantically salient part of focus. Therefore, we call each of these words a *focus kernel*. In the following examples, focus kernels are marked with **bold:**

(7) *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.*
(8) *T: Oh, are you having fun?*
*U: **Yeah**, it's kind of interesting.*
(9) *T: How many times would it take for the reds to come back on top?*
*U: It would take **three** times to have the red be back on top.*

1.3. Contrast

Contrast is a concept having multiple senses: (1) In logic, two propositions are defined to be contrastive if it is impossible for them to be true simultaneously. For example, in the sentence *Bach was an organ mechanic; Mozart knew little about organs*, the two propositions are not contrastive, whereas they become contrastive when 'Mozart' is replaced by 'Bach' at the beginning of the second sentence (Bosch and van der Sandt, 1999). (2) The discourse relation called contrast is induced by 'but,' and constitutes a pair (or pairs) of contrasted alternatives, which can be predicates (e.g., *John cleaned up the room, but he didn't wash the dishes*), individual words (e.g., *John cleaned up the room, but Bill didn't*), or propositions (e.g., *It is raining, but we go out for a walk*) (Umbach, 2004). (3) Some linguists use contrast to denote the mutually exclusive disjunction between the words contributing to a fact and other alternatives made available by context (Vallduví and Vilkuna, 1998). It has been argued that focus in general establishes a contrast since novel information usually conveys contrast between a fact and the potential alternatives (Bolinger, 1961; Kruijff-Korbayova and Steedman, 2003). For example, in the sentence *Last night they had a party*, there is a contrast between the focus 'party' and any other alternative activities of the group. (4) Symmetric contrast consists of a set of words that are parallel or symmetric in linguistic structure but mutually exclusive in meaning; the stress on one word is motivated by its distinction from the others, e.g., 'American' and 'Canadian' in *An American farmer was talking to a Canadian farmer* (Rooth, 1992; Umbach, 2004).

In this study, we seek to make use of the knowledge about contrast from the pragmatics and prosody literature, for the purpose of detecting pairs of symmetrically contrasted words that are assumed to be useful for spontaneous speech understanding. Symmetric contrast can occur within a sentence, e.g. (contrasted words are marked with **bold**),

(10) *U: The **large** gear has five times as many teeth as the **small** ones.*
(11) *U: How about **small** and **big** and **medium**?*

Topics and/or foci of conjunct phrases or coordinated sentences (by 'and', 'but', etc.) can also constitute symmetric contrast, e.g.

(12) *T: Where are the gears?*
*U: The **red** gear is on the **bottom** and the **yellow** gear is on the **top**.*
(13) *T: How are the gears spinning?*
*U: The two **outside** ones spin in the **same** direction and the **middle** one spins in the **opposite** direction.*

The words participating in a symmetric contrast satisfy semantic parallelism, which has two implications: (a) the conjunct alternatives have to be semantically independent of each other in the sense that neither of them subsumes the other; and (b) there has to be a "common integrator," i.e., a concept subsuming both conjunct alternatives (Umbach, 2004).

1.4. Study objective

As its primary technical goal, the study intends to test whether the proposed word tags, i.e., focus kernel and symmetric contrast, can be reliably

annotated in a spontaneous speech corpus using both manual and automatic annotation. As part of this evaluation, this study tests the relationship of focus kernel and symmetric contrast with the following prosodic and pragmatic variables: (1) prosodic prominence—experiments described in this paper test the reliability of prosodic prominence in the automatic identification of focus kernel and symmetric contrast; (2) novelty and semantic parallelism, the semantic attributes of focus kernel and symmetric contrast. Information theoretic measures of novelty and semantic parallelism are implemented, based on algorithms proposed in the computational linguistics literature. Implemented algorithms are tested for the purpose of automatically identifying focus kernel and symmetric contrast; and (3) part-of-speech. In addition, this study discusses the usefulness of focus kernel and symmetric contrast to spontaneous speech understanding.

2. Background and related work

Focus and contrast in modern linguistics are used to "account for the correlation between certain prosodic patterns and certain pragmatic and semantic effects" (Kadmon, 2001). Sections 2.1 and 2.2 describe related work on contrast and focus published in the linguistics and computational linguistics literature. Section 2.3 describes previous work on the word dissimilarity measure (given a pair of words, how much novel information a word contains with respect to the other word) in natural language processing (NLP).

2.1. Focus

The information structure of sentences can be defined in various ways, e.g., presupposition-focus (Chomsky, 1971; Jackendoff, 1972; Zubizarreta, 1998), topic-comment (Dahl, 1969), theme–rheme (Firbas, 1964, 1966; Bolinger, 1965; Steedman, 2000), given-new (Halliday, 1967; Kay, 1975), background-focus (Dahl, 1969; Steedman, 2000), and background-kontrast (Vallduví and Vilkuna, 1998). Although the information structure definitions are diverse, overlapping and even conflicting, they can be categorized into two dimensions: (1) topic-comment or theme–rheme, in which one part describes what the discourse is talking about, and the other part advances the discourse; and (2) given-new or background-kontrast, in which one part conveys information that has been known, and the other part conveys novel information distinguishing actual occurrence from potential alternatives triggered by context (Kruijff-Korbayova and Steedman, 2003; Lee, 2003). In defining the information structure of sentences in the given-new or background-kontrast dimension, we follow the partition of Chomsky (1971), Jackendoff (1972) and Zubizarreta (1998), i.e., presupposition and focus (see Section 1.2).

Contrastive focus is a special kind of focus that expresses *exhaustive identification* of an element (or subset) given a set of candidates. Exhaustive identification means that the selection of a candidate excludes all other candidates. Ordinary focus introduces new, non-presupposed information. Contrastive focus exhaustively selects one or more candidates from a set of candidates that are presupposed by the interlocutor (Umbach, 2004; Lee, 2003; Hedberg and Sosa, 2001). For example

It was $[John]_{CF}$ *that baked bread for our breakfast.*

When the answer to an alternative question is to choose a disjunct from the disjunctive alternatives, the choice is thought to be exhaustive. For example, the choice of 'money' or 'pen' for *Did the baby pick the money first, or did she pick the pen first?* is contrastive focus (Lee, 2003). Another case of exhaustiveness is *correction* in dialogues: contrastive focus corrects the explicit or implicit assumption made by the interlocutor.

Much of the literature studying focus seems to be motivated by the pitch accent correlate of focus (e.g., Hedberg and Sosa, 2001; Pierrehumbert and Hirschberg, 1990): semantic or pragmatic non-presupposed new information is usually signaled by a language-dependent accentual *F*0 contour. Human detection of focus is very sensitive to the presence or absence of accenting of the focused item and the deaccenting of the post-focus items (Welby, 2003; Xu et al., 2004; Gussenhoven, 2002). Nuclear pitch accent (the last pitch accent in an intermediate phrase) and prenuclear pitch accent have different effects on the listener's interpretation of focus

(Welby, 2003). However, as we have described, there is no decisive information source to clearly mark the focused syntactic constituent. For this reason, automatic detection of focus has received little study. In one relevant study, Heldner et al. (1999) tried to locate *narrow focus* (focus consisting of a single accented word) in Swedish speech: they used energy and high frequency emphasis to automatically detect narrow focus within three-word phrases. Their automatic focus detector was designed based on the assumption that there was only one focused word per phrase and the focused word was accented. About two thirds of the focused words were correctly detected.

In general, pitch accent in spoken English marks subconstituents within an utterance as highlighted. A central reason for an item to be highlighted is novelty. Focus is marked by pitch accent because it expresses novel information. However, pitch accent may also signal something else; there are a variety of factors affecting the placement of pitch accent. As Kadmon (2001) pointed, pitch accent can be assigned to a given item with special importance, while novelty may not be marked with pitch accent. Therefore, pitch accent is an important condition, but neither a necessary nor a sufficient condition for novelty.

2.2. Contrast

Contrast is widely investigated in modern linguistics with special attention given to *contrastive topic* and *contrastive focus* (the latter has been addressed in Section 2.1). Contrastive topic can be defined as a syntactic constituent that is both topic-marked and contains a focused item. That means, the constituent on one hand forms a topic, and on the other hand contains novel information in contrast with other alternatives triggered by the context (Lee, 1999). In the following example

(14) *T: What happens to the gears?*
U: [*The large gear*]$_{CT}$ *only spins once and it spins slower,*

'gears' is the topic that the speaker is addressing, but the choice of 'large' is new and in contrast with 'the other gears.' In addition, the topics of a sequence of independent contrastive answers to a question also form contrastive topics (Krifka, 1999), e.g.

(15) *T: What about the large gear and the medium gear?*
U: The large $_{CT}$ *gear spins left, and the medium*$_{CT}$ *gear spins right.*

In (15), 'large' and 'medium' are identified to be contrastive topic and symmetric contrast simultaneously.

In the literature, much investigation of contrast has focused on the conceptual issue of some problematic cases and specification of the types of pitch accent correlate (e.g., Gundel and Fretheim, 2001; Pierrehumbert and Hirschberg, 1990; Kadmon, 2001; Lee, 1999). In general, contrast is well and clearly defined in linguistics. However, a small set of problematic cases receive a great deal of attention. For example, authors are divided on whether to identify a negated presupposition as contrastive focus or contrastive topic (e.g., the words 'anything extraordinary' in *He did the only thing you could do. He hasn't done anything extraordinary*). In addition, published studies agree that contrast is typically marked with pitch accent, but there is considerable disagreement on the particular patterns of pitch accent used to mark contrastive topic and/or contrastive focus (e.g., Lee, 1999; Hedberg and Sosa, 2001). These debates are usually framed in terms of the pitch accent categories specified by the *ToBI* (tones and break indices) prosodic annotation standard (Beckman and Ayers, 1994), e.g., whether contrastive topic will be marked by H*+!H or not. However, the disputes in the literature generally do not affect the integrity of our task, because: (1) the problematic cases requiring conceptual discrimination are rare in our corpus; and (2) our study does not require the discrimination of pitch accent patterns.

2.3. Word semantic similarity

The computational modeling of novelty and semantic parallelism involves semantic analysis of words, more exactly, the comparison between

words in terms of semantic meaning. Strictly speaking, every word is different from every other word. "Absolute *synonymy*, if it exists, is quite rare. Usually, words that are close in meaning are almost synonyms, but not quite; very similar, but not identical, in meaning; not fully intersubstitutable, but instead varying in their shade of denotation, connotation, or emphasis" (Edmonds and Hirst, 2002). So how can we formally specify similarity and dissimilarity? A word often has multiple senses. What counts as a central trait to compare the meaning of a pair of words? Researchers in computational linguistics have developed various measures to compute the degree of *semantic similarity* between a pair of words. The measures are basically divided into two categories: (1) *ontology hierarchies* (e.g., Lee et al., 1993; Sussna, 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 the types in the ontology. The length of the path between a pair of words is a measure of their semantic dissimilarity. The well-known *edge-based* method reflects the fact that in a hierarchical semantic network, the simplest measure of the distance between two elemental concept nodes, A and B, is the shortest path that links A and B, i.e., the minimum number of edges that separate A and B (Rada et al., 1989). (2) *Corpus statistics* empirically model the context-dependence characteristics of word meaning in text (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 (Higgins, 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.

3. Corpus description, annotations and analyses

3.1. Tutoring dialogue scenario

The intelligent tutoring system helps students learn basic math and physics concepts by playing with Lego gears, with the objective of helping students develop a physical understanding of abstract concepts. For example, one question about the relationship between gear size and spinning speed is *Line up a 24-tooth gear and a 40-tooth gear. If the 24-tooth gear spins 5 times, then how many times must the 40-tooth gear spin for them to line up again? Why?* Children can answer this question by spinning the gears and counting the cycles. Similarly, a physics question about interactive force is *Put one hand on the 40-tooth gear axle, and put another hand on the 8-tooth gear axle. What happens if you hold one of them steady, and try to turn the other one? Why?* Children usually think that the big gear is stronger before they do the experiment. However, it turns out that the small gear is stronger.

The complete system has not been finished yet; the database used in this study was collected by Wizard-of-Oz simulations of the finished system. In the experiments, the user and the tutor (human wizard) were sitting in separate rooms. The user orally communicated with a computerized talking head shown on the computer screen ahead of him using a head-set microphone. The lip movement of the talking head was coincident with speech synthesized from text that was typed by the tutor. The user's speech was transmitted through the microphone to a digital camera placed opposite the user, and then was transmitted to the earphone of the tutor sitting in another room. Both the tutor and the user had Lego gearsets on the table in front of them. A video of the tutor's gearset was displayed on the user's computer, and vice versa. The WoZ experiments allowed data to be collected that was similar in most respects to the data the final spoken language system would need to understand: because children felt that they were communicating with a computer tutor instead of a human tutor, they behaved as they

would in a real computer-interacting environment. The number of experimental sessions in which each child participated varied from one to three, depending on the interest and cooperative attitude of the child. The tutor adjusted the content of experiments according to the intelligence, cooperation, and learning progress of each child subject.

3.2. Characteristics of the ITS corpus

Unlike the users of a telephone weather or flight ticket system, who are often expert users interested in achieving known goals using known tools in the shortest time possible, the users of our intelligent tutoring system are perpetually naïve with respect to the future content of the dialogue. Each child participates in at most three experimental sessions and each session has different tutorial content, because we do not ask children to re-learn knowledge that they have mastered. Therefore, although children who participate in more than one experiment may gain some expertise in the use of the computer interface, the children are never able to memorize menu prompts, re-use conversation content, or otherwise become expert users of the dialogue system.

In addition, we encourage children to participate in the experiments, and we instigate their interest in scientific learning by asking them open-ended questions rather than questions with an absolute answer. 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 single-choice questions such as *In which direction are the gears turning?* Instead, he would ask questions whose answers are not absolute, e.g., *What are you noticing*? Compared with the single-choice questions, the open-ended questions open a wider space for children and arouse children's enthusiasm to use their imagination, knowledge, and observation to solve problems.

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 single-choice questions, their utterances are even more incoherent and dysfluent than is typical in interpersonal conversations. The utterances usually include loose grammar structure, fragments, restarts, repairs, meaningless speech (e.g., *That if the…*), and repetitions. Moreover, sentence boundaries in spontaneous speech are ambiguous because of mismatch between acoustic segmentation and linguistic segmentation. The following example illustrates the characteristics of the ITS speech data:

(16) *U: Big gears move in different ways and… uhm…with the first when you push one of the first gears, the other gear the last gear moves you know, and the gear after that moves.*

3.3. Focus kernel annotation

Three annotators worked independently on identification of focus kernel, if any, in each utterance of the ITS corpus. Annotation was based on perception, text transcriptions, and dialogue context. A vast majority of the ITS corpus are question-answer pairs between the tutor and user, in which the tutor initiates local dialogue topics by asking questions or providing suggestions (e.g., *Can you make them look like this?*). In this case, presupposition of an utterance lies in the question or suggestion of the tutor (Kadmon, 2001). Students sometimes initiate local dialogue topics by issuing commands, asking questions or simply explaining what they are doing. In this case, presuppositions for the students' utterances do not exist. Annotators were given the following criteria to use in their labeling of focus kernels in an utterance:

1. Mark the content words that contain information not already available in presupposition, if any, nor in the preceding words of the utterance. For example,

   (17) *T: What if we now add a gear?*
   U: The ***third*** *one* ***moves*** *or spins the* ***same direction*** *that the* ***first*** *one does.*

   Mark contrastive focus: (a) If the tutor asks an alternative disjunctive question and the user responds by choosing a single disjunct, then

the disjunct is characterized by contrastive focus, and thus, is marked as focus kernel. For example,

(18) *T: Does the top or the bottom do the pushing?*
U: I think it's the ***top***.

(b) If the user's utterance is to correct the assumption of the tutor, then the corrective word is characterized by contrastive focus, and thus, is marked as focus kernel. For example,

(19) *T: Take that big gear, please.*
U: I thought you said ***this*** *gear.*

2. Focus defined by Chomsky (1971) and Zubizarreta (1998) is confined to the rhyme of a sentence (Steedman, 2000). However, we are also interested in marking the new item in contrastive topic, e.g., 'large' in example dialogue (14), since it conveys novel information. We label such item as focus kernel although it is not in the rhyme, because: in our corpus this case is very rare and focus kernel are generally in the focus (rhyme), so the inclusion of this rare case basically does not affect the fact that we can call the words containing novel information as the kernel of focus.
3. In the case of dysfluency, discourse marker and repetition should not be marked, while repair should be marked. For example, in the dysfluent constituent $[\ ]_{dys}$ of the utterance *If I spin 3 times it end up, [all yellows at the bottom, and all yellows at the... I mean, all reds at the bottom,]$_{dys}$ and all yellows at the top,* the discourse marker 'I mean' should not be marked. As correction of 'yellows', 'red' should be marked.
4. Function words are not marked unless they carry novel information, and are thus stressed by the speaker, e.g., 'this' in example (19).

Each word in an utterance is labeled either focus kernel or nonfocus kernel. We used the kappa statistics to evaluate consistency among the three annotators. The kappa statistics is a chance-corrected inter-transcriber agreement rate $\kappa = (P_O - P_C)/(1 - P_C)$, where P_O is the rate of inter-transcriber agreement, and P_C is the average rate that would be achieved by chance (Flammia, 1998). Comparison of the three transcriptions yielded a kappa score of 0.79, indicating reasonably good agreement among the transcribers. We used majority voting to resolve the annotation difference among the three annotators. That is, for each word, if two or more annotators had the same focus kernel/nonfocus kernel label, then we assigned that label to the word as the target label.

3.4. Symmetric contrast annotation

The labeling of symmetric contrast was based on pairs of words. For example, 'small' and 'big' are labeled to be a symmetric contrast. If more than two words were symmetrically contrastive with one another in an utterance, then we labeled them by pairs. For example, in *I'm counting the small, medium, and big*, we labeled 'small', 'medium' and 'big' by three symmetric contrasts: 'small' and 'medium', 'small' and 'big', and 'medium' and 'big'. Moreover, in the case of repair (dysfluency), e.g., *The* ***small*** *one, no, the* ***big*** *one is turning left*, the repairing word (e.g., 'big') and the repaired word (e.g., 'small') were also labelled as a symmetric contrast.

Two transcribers worked together to identify instances of symmetric contrast in the ITS corpus. The annotation was performed based on speech perception, transcription, and dialogue context. Because the transcribers worked together, no inter-transcriber agreement statistics were computed.

3.5. Corpus analysis

To date 29 WoZ experiments with 17 subjects have been carried out and transcribed, and we have collected 11.7 h of audio-visual data. In each experiment, the child often spent most time silently playing with the Legos. Therefore, the collected audio data corpus was small. Recordings of the students were manually transcribed and segmented into conversation sides. Each of the user's conversation sides was considered an *utterance*. Some

speech data had to be discarded because of Lego block noise, heavy breathing, etc. This process resulted in a total of 714 transcribed user utterances, containing a total of approximately 50 minutes of relatively clean speech. On average each utterance had 4.2 s speech and 8.1 words. The vast majority of the utterances contained 1–20 words, while the longest utterance had 57 words. Fig. 1 shows the histogram of utterance lengths in the corpus.

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 merely repeat the words of the tutor, 19 utterances that are not semantically meaningful (e.g., *That if the...*), and 589 multi-word, multi-phrase utterances containing meaningful and novel information. We use the 630 multi-word, multi-phrase utterances (589 meaningful utterances, 22 repetitions of the tutor, and 19 semantically void utterances) for experiments.

Short utterances tend not to contain symmetric contrast. We choose the average number of words in a sentence, which is 8, as the threshold to distinguish long utterances from short utterances. Utterances containing at least 8 words occupy 40% of the 630-utterance corpus. Table 1 shows the relationship between utterance length and occurrences of symmetric contrast. The table shows that 45% of the long utterances contain symmetric contrast, while 5% of the short utterances contain symmetric contrast. Phrased in another way, more than 85% of symmetric contrast instances occur in utterances of 8 words or more. Fig. 2 shows the histogram of focus kernels per utterance in the 630-utterance corpus. An utterance can has maximally 10 focus kernels, while on average each utterance has 2 focus kernels.

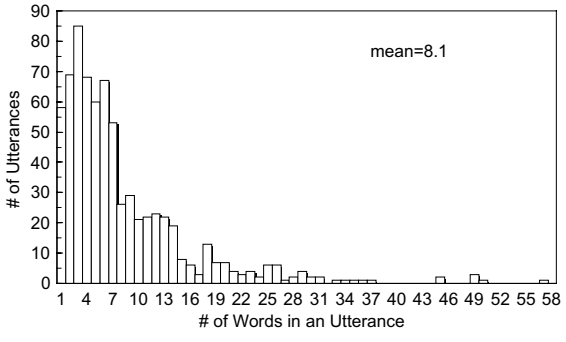

Fig. 1. Histogram of utterance lengths in the ITS corpus.

Table 1
Relationship between utterance length and occurrence of symmetric contrast

	Utterance length $\geqslant 8$	Utterance length < 8
# of utterances containing symmetric contrast	113	19
# of utterances not containing symmetric contrast	140	357

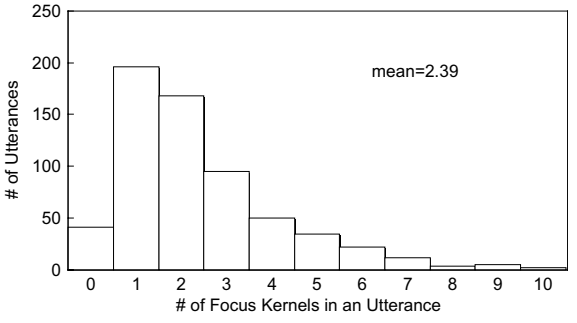

Fig. 2. Histogram of focus kernels per utterance.

Part-of-speech of all words in the ITS corpus is first tagged using an automatic part-of-speech tagger (Munoz et al., 1999), and then manually checked against the tagging standard in Treebank-3 (Santorini, 1990). The part-of-speech tags are used in the automatic detection of symmetric contrast and focus kernel, since part-of-speech can be used to distinguish content words from function words.

4. Prosodic analysis

The literature in both prosody and pragmatics reports the pitch accent correlate of contrast and focus. Therefore, pitch accent is a reasonable first step in the automatic classification of focus kernel and symmetric contrast. Because of the man power involved in manual labeling of pitch accent in the ITS corpus, we try to use an automatic system to label pitch accent. To date pitch accent automatic

detection concentrates on Radio Speech, in which half of all words may be pitch accented (Kim et al., 2003), whereas in conversational telephone speech, typically about 20% of words are pitch accented (Yoon et al., 2004). Pitch accent can be automatically detected in the Boston Radio Speech Corpus with reasonable accuracy: a gender-dependent and speaker-independent system achieved a 10% pitch-accent detection equal error rate (Kim et al., 2003). This section describes the derivation of the prosodic measurements used for pitch accent labeling.

4.1. Duration, pitch and energy

English speakers tend to signal prosodic prominence using extruded pitch, increased energy, and longer phoneme durations. In order to match durations and energy with individual words, we first train a speaker-independent mixture Gaussian HMM speech recognizer, and then adapt the recognizer to children's speech (Zhang et al., 2004). The known word-level transcriptions of each utterance are expanded into phoneme transcriptions, and forcibly aligned to the speech waveform using the automatic speech recognizer, resulting in an automatic estimate of the duration and time alignment of each word. Pitch and probability of voicing are extracted using the FORMANT program in Entropic XWAVES. Pitch measurements in frames with low probability of voicing are discarded. Extremely high $F0$ values (doubling) and extremely low $F0$ values (halving) are also discarded. Valid pitch measurements of an utterance are then normalized by the highest $F0$ in the utterance to compensate for inter-speaker $F0$ differences. Kim et al. (2003) found that 95% of the pitch accents in their corpus were high-$F0$ pitch accents (in ToBI notation, these were variants of H* or !H*), and similar statistics have been reported for other corpora (e.g., Yoon et al., 2004). Therefore, the normalized pitches in the higher half pitch region of a word are averaged in order to obtain an "average peak pitch" measurement for the word. The averaging scheme is given by

$$D_m = \frac{1}{N_{\psi_m}} \sum_{f \in \psi_m} F_m[f], \tag{1}$$

where D_m is the pitch value of word W_m, $F_m[f]$ is the pitch value of frame f in W_m, ψ_m is a subset of frames in W_m, N_{ψ_m} is the total number of frames in ψ_m, and

$$\psi_m = \left\{ f | F_m[f] \geqslant \frac{1}{2} T_m \right\}, \tag{2}$$

$$T_m = \max_{F_m[n] \in W_m} \{F_m[f]\}. \tag{3}$$

Energy of a word is computed using the same strategy as described by Eqs. (1)–(3).

4.2. Spectral balance cepstral coefficients

Spectral balance is intensity emphasis in the higher frequency region. Sluijter et al. (1997) showed that in lexically stressed syllables, speech spectral intensity at higher frequencies (above 500 Hz) increased more than the intensity at lower frequencies. Sluijter et al. also found that manipulating the high-frequency intensity of speech resulted in stronger stress cues than manipulating the entire band. These studies demonstrated that high frequency intensity is a stronger cue for stress than overall intensity. Since pitch accent in spoken sentences normally falls on the lexically stressed syllables, we speculate that intensity of the spectrum above 500 Hz may be a useful acoustic cue for pitch accent detection.

We propose spectral balance cepstral coefficients (SBCC) to encode the intensity and shape of the speech spectrum in the range between 500 Hz and 5000 Hz. The speech waveform is first pre-emphasized and windowed by 30 ms windows with an inter-frame window shift of 10 ms. Speech samples in each frame are decomposed into a series of bands between 0 and 5000 Hz through multiresolution analysis (see Fig. 3). The multiresolution analysis is achieved by iterative application of a filterbank, which consists of quadrature mirror filters (a low-pass smoothing filter and a high-pass differencing filter) with Daubechies-4 orthogonal coefficients (Daubechies, 1990); each time the filterbank decomposes speech signals into a higher band and a lower band. The Daubechies-4 time domain filterbank has better out-of-band rejection than a filter constructed by adding DFT coefficients (like

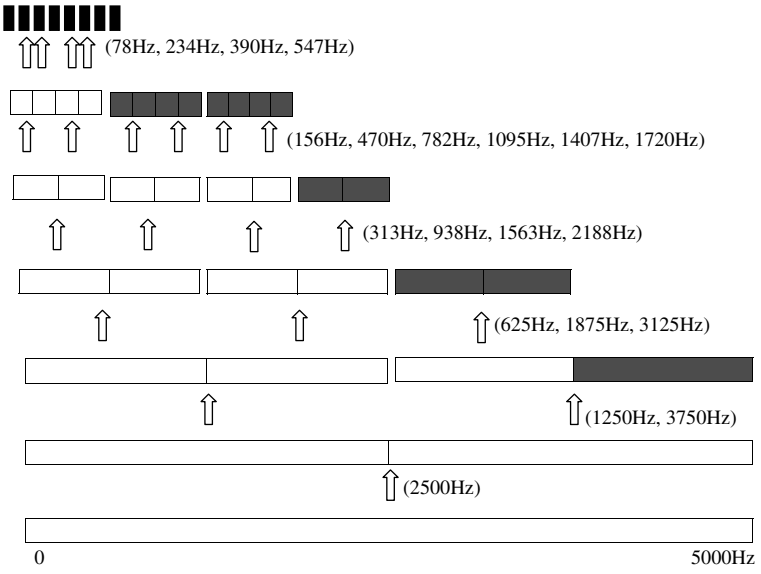

Fig. 3. Multiresolution decomposition of 0–5000 Hz. An arrow splits the band from which it originates into two equal subbands. The frequencies of the splits as arrows indicate are labeled at the right side. The subbands finally obtained from the decomposition are dark marked. The subbands over 500 Hz are the 14 rightmost bands.

the filters used in MFCC or PLP analysis), but retains the desirable properties of MFCC analysis: semi-logarithmic frequency scaling, and the flexibility necessary to select desired sub-bands, e.g., in our case, the sub-band above 500 Hz. After multiresolution analysis, each word is characterized by its average peak signal intensity in each band, using equations similar to those used for pitch:

$$e_m = \frac{1}{N_{\phi_m}} \sum_{n \in \phi_m} y_m^2[n], \quad 1 \leqslant m \leqslant M, \tag{4}$$

where $y_m[n]$ is sample n in band B_m, e_m is the intensity of band B_m, M is the total number of bands spanning from 500 Hz to 5000 Hz, N_{ψ_m} is the total number of speech samples in set Φ_m, Φ_m is a subset of B_m and

$$\phi_m = \left\{ n \middle| |y_m[n]| \geqslant \frac{1}{2} T_m \right\}, \tag{5}$$

with

$$T_m = \max_{y_m[n] \in B_m} \{|y_m[n]|\}. \tag{6}$$

The log energies in each band (logarithms of Eqs. (4)–(6)) are transformed using an inverse discrete cosine transform to compute the cepstral coefficients:

$$E_l = \sum_{m=1}^{M} \log(e_m) \cos\left[\frac{l(m-0.5)\pi}{M}\right], \quad \leqslant l \leqslant L, \tag{7}$$

where L is the desired length of the cepstrum. Cepstral mean subtraction is then applied to compensate for disturbances caused by the transmission channel.

5. Word semantics analysis

We use a word dissimilarity measure to model the degree of novelty that a word has in comparison with other words. We use N_i to denote the novelty of word w_i given dialogue context. According to the definition of focus kernel, we compute N_i by the minimum of the dissimilarity between w_i and the words in set S, where S consists of those words appearing in the interlocutor's presupposition and those precedent of w_i in the utterance, i.e.,

$$N_i = \min_{w_j \in S} \mathrm{dis}(w_i, w_j). \tag{8}$$

Semantic parallelism is the semantic attribute characterizing symmetric contrast. Given a couple of words, the quantification of their semantic parallelism should emphasize: (1) word dissimilarity, i.e., a word cannot be parallel to itself or its synonym; (2) word similarity, e.g., 'blue' and 'white' are more likely than 'drink' and 'pier' to become a symmetric contrast; and (3) non-hypernym relation, i.e., a word cannot form symmetric contrast with its hypernym. For example, 'dog' and 'cat' can form symmetric contrast, while 'dog' and 'animal' cannot.

Similarity and dissimilarity of a pair of words was computed using measures based on methods from the computational linguistics literature. Specifically, as suggested in Section 2.3, two methods were used: an ontology-based method and a method based on corpus statistics. Neither of these two methods was found to be an adequate dissimilarity measure by itself, but the linear interpolation of the two measures was found to be reasonably adequate.

5.1. Application-oriented ontology

Ontology design determines the set of semantic categories that properly reflects the particular conceptual organization of a target domain (Lenci, 2001). Designing a completely new ontology is comparably difficult to the design of a completely new dictionary. An attractive solution is to adapt a general linguistic resource to the application domain. In this study, we use *WordNet* as the universal background linguistic source. WordNet is a hierarchical semantic database of English words (Miller and Fellbaum, 2002). In WordNet, the main relations between words are: synonym, antonym, and hypernym (representing the '*is–a*' relationship) for nouns, verbs and adjectives; and synonym, antonym for adverbs.

The ontology in the ITS domain is defined by $O = \{C,R,H\}$, where C is the set of concepts, R is the set of relations, and H is the concept hierarchy. Each content word in the ITS lexicon is a concept. Concepts should interconnect with each other in the ontology. The relationship between concepts is represented by links. In this study, the ontology encodes the information of synonym, antonym, and hypernym for the content words. The hierarchy of word concepts is constructed following the procedures in Appendix A. The major semantic categories used in the hierarchical structures are listed in Appendix B. A subset of the ontology is shown schematically in Fig. 4 using a tree structure. In addition, function words are categorized by their part-of-speech under the main semantic class 'function' (in contrast with 'content', the main semantic class for content words).

We employ the edge-based method, in which an edge represents a direct association between a pair of semantic concepts, to compute the distance between the pair of words. In a more realistic scenario, the distances between any two adjacent nodes are not necessarily equal. Generally, the distance shrinks as one descends the hierarchy, since differentiation is based on finer and finer details (Jiang and Conrath, 1997), e.g., the distance between 'abstraction' and 'entity' (at the top level of the hierarchy) is much bigger than the distance between 'many' and 'some' (at a lower level). It is therefore necessary to consider that the edge connecting the two nodes should be weighted by their depth in the hierarchy. We propose the following weighted edge-distance:

$$\mathrm{dis}(w_1,w_2) = \min_{c_1 \in \mathrm{sen}(w_1) c_2 \in \mathrm{sen}(w_2)} \left(\frac{d_{c_1,c_2}+1}{d_{c_1,c_2}}\right)^{\alpha} \mathrm{len}(c_1,c_2), \quad (9)$$

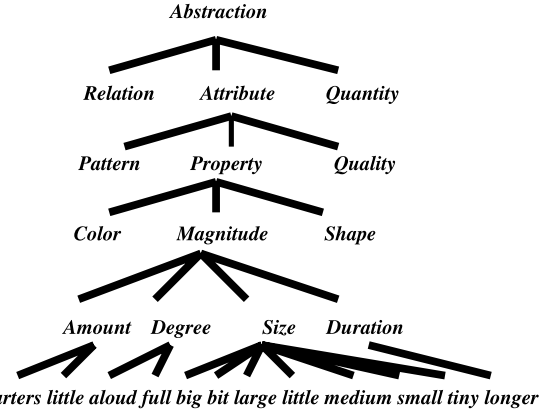

Fig. 4. Partial hierarchy of the application-oriented ontology.

where sen(w) denotes the set of possible senses for word w if 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 number of edges connecting c_1 and c_2; and α is a constant. Here we choose $\alpha = 3.0$. In addition, we set the distance score as 0 for synonym, and 15.0 for antonym based on the maximum distance between a pair of words in the ontology.

5.2. Corpus statistics

Statistical methods attempt to measure dependence between words by using statistics taken from a large corpus. Here we use the English *GigaWord* corpus, a billion-word archive of English newswire text distributed by the Linguistic Data Consortium. The entire GigaWord corpus consists of 314 files, nearly 11.7 GB. All texts are presented in SGML form, using a simple markup structure. We remove the SGML tags, leaving only text content. The raw database consists of four different types of documents: *Story*, *Multi*, *Advis*, and *Other*. *Story* has uniform format, each consisting of a few paragraphs describing a same topic. The other three types of documents do not have long paragraphs of text. Therefore, we only choose the story session for experiments; the story session (with paragraph breaks and some other control characters removed) contains 8.57 GB of text.

In this study, similarity between a pair of words w_1 and w_2 is measured based on the proximity and topicality assumptions (Section 2.3). First, the proximity assumption tells us that we can model word similarity by the probability of a pair of words occurring together in a sentence, a paragraph, a topic, or even a document. We measure the degree of association between a pair of words w_1 and w_2 by mutual information

$$\mathrm{PMI}(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

Let the co-occurrence frequency of w_1 and w_2 be denoted by $f_{w1,w2}$. Because GigaWord is a very large text corpus, the co-occurrence frequency is roughly estimated by the number of topics (stories) in which w_1 and w_2 co-occur. Let N be the size of the corpus in terms of topics. The maximum likelihood estimate of the co-occurrence probability is given by $p(w_1, w_2) = f_{w1,w2}/N$. To compensate for the sparsity of the training data, we compute $p(\mathrm{w}_1, \mathrm{w}_2)$ by

$$p(w_1, w_2) = \frac{f_{w1,w2} + 1}{N + 1} \tag{10}$$

The computations of $p(w_1)$ and $p(w_2)$ adopt the same smoothing strategy. Second, the topicality assumption tells us that if we have decided that two words are similar, then we may infer that they have similar mutual information with some other word, w. Given a context $C = \{w'_1, w'_2, \ldots, w'_n\}$, w_1 and w_2 are considered semantically similar if they are both likely to co-occur with the words in C. We apply a method (Pantel and Lin, 2002) that computes the cosine distance between the two partial mutual information (PMI) vectors corresponding to w_1 and w_2:

$$\mathrm{dis}(w_1, w_2) = 1 - \frac{\sum_{w' \in C} \mathrm{PMI}(w', w_1)\mathrm{PMI}(w', w_2)}{\sqrt{\sum_{w' \in C} \mathrm{PMI}(w', w_1)^2}\sqrt{\sum_{w' \in C} \mathrm{PMI}(w', w_2)^2}}, \tag{11}$$

where $C = C(w_1) \cup C(w_2)$, $C(w_1)$ and $C(w_2)$ are context of w_1 and w_2, respectively. Given a word w, its context $C(w)$ can be the simple collection of all (content) words appearing within the window of w in the ITS corpus. To more properly represent those words semantically associated with w, we modify the method proposed by Dagan et al. (1995) to determine $C(w)$:

1. Define a pair of words (w, v) to be strong neighbors if $f(w, v) > t_f$, where $f(w, v)$ is the count of (w, v) in a window of size d, t_f is threshold and
$$t_f = \begin{cases} \lambda_1 \times \min(N_w, N_v) & \min(N_w, N_v) > \gamma_1 \\ \tau_1 & \min(N_w, N_v) \leqslant \gamma_1 \end{cases}, \tag{12}$$
where N_w and N_v are the number of occurrences of w and v, respectively, and λ_1, γ_1, and τ_1 are constant.
2. Collect all the strong neighbors of w as potential candidates and denote as $C_1(w)$
$$C_1(w) = \{v | f(w, v) > t_f\}. \tag{13}$$

3. Collect the strong neighbors of all words in $C_1(w)$ as potential candidates, and denote as $C_2(w)$

$$C_2(w) = \{v | v \in C_1(C_1(w))\}. \quad (14)$$

4. Collect those words which share at least t_N strong neighbors with w in the lexicon and denote as $C_3(w)$

$$C_3(w) = \{v | C_1(w) \cap C_1(v) > t_N\}, \quad (15)$$

where

$$t_N = \begin{cases} \lambda_2 \times \min(C_1(w), C_1(v)) & \min(C_1(w), C_1(v)) > \gamma_2 \\ \tau_2 & \min(C_1(w), C_1(v)) \leqslant \gamma_2 \end{cases}, \quad (16)$$

N_w and N_v are the number of occurrences of w and v, respectively, λ_2, γ_2, and τ_2 are constant.

5. $C(w) = C_1(w) \cup C_2(w) \cup C_3(w)$. (17)

Examples of corpus statistics are given in Appendix C.

5.3. Knowledge combination

The computation of word semantic dissimilarity is a combination of lexical ontology and corpus statistics. First, the manually built pseudo-knowledge base has advantages in efficient paraphrasing, inference and reasoning. The ontology is especially useful when the dissimilarities of some words are dependent on topic and pragmatic context. For example, 'play' and 'build' are dissimilar to each other under a basketball game topic, but similar under the ITS topic. However, the top-down organization of lexical knowledge cannot incorporate the dynamic nature of word meanings. Word meanings change dramatically depending on the linguistic context. This context-dependent characteristic is one of the main empirical arguments for real text data (Lenci, 2001). Second, the statistical model provides computational evidence from distributional analysis of corpora data. The statistical model provides a quantification of the semantic space. However, corpus statistics has some limitations, as it requires the exact correspondence between terms (word or character n-grams). In unrestricted language many reasonable co-occurrences may fail to occur in the training corpus.

In order to combine information from both the ontology-based dissimilarity measure and the corpus-based dissimilarity measure, we use linear interpolation as

$$\mathrm{dis}(w_1, w_2) = \lambda \mathrm{dis}_1(w_1, w_2) + (1 - \lambda)\mathrm{dis}_2(w_1, w_2), \quad (18)$$

where dis_1 and dis_2 are the peak-normalized word distances based on the ontological and the statistical methods, respectively; and parameters λ is constant. Here we arbitrarily choose $\lambda = 0.5$. 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 in Eq. (11) will be 1 (# of contextual words is 0) or 0 (# of contextual words is 1). If the prerequisite fails to be satisfied, then the dissimilarity is computed using only the ontology-based dissimilarity measure. Examples of feature combination are listed in Appendix D.

6. System evaluation

The corpus for focus kernel classification consisted of 630 multi-word, multi-phrase utterances, containing approximately 5700 words and 48 min of speech. In the experiments of extracting focus kernel and symmetric contrast, training and test data included different utterances from the same set of talkers, so the experiments were multi-speaker speaker-dependent.

6.1. Focus kernel detection

Prosodic observations (duration, average peak pitch, average peak energy, and spectral balance cepstral coefficients) and part-of-speech tagging were integrated with each word's semantic novelty measurement (Eq. (8)) using a time-delay recurrent neural network (TDRNN) for the purpose of automatically detecting focus kernels. The TDRNN is a neural network that encodes and integrates dynamic signal context using a combination of delayed input nodes (for temporal loca-

tion of important information in the input sequence) and delayed recurrent nodes (for temporal context information) (Kim, 1998). This study used a modified TDRNN: we added a recurrent layer between the output layer and the hidden layer to feed back the delayed values from the output layer; and the time indices were synchronized to successive words rather than centisecond speech frames.

We have tried using the non/pitch accent automatic label as a feature for focus kernel detection: we employed a model trained from Boston Radio News Corpus (Ren et al., 2004a) for our automatic pitch accent labeling task; then the pitch accent/nonpitch accent labels were combined with semantic similarity measure to input to the TDRNN. However, the performance was not satisfying, apparently because of: (1) the application of a model trained on radio news speech to spontaneous speech; (2) the noisy recording environment; and (3) the fact that erroneous pitch accent detection would lead to erroneous extraction of focus kernel. Therefore, we used the acoustic correlates of pitch accent, rather than pitch accent itself, for the purpose of detecting focus kernel.

We used 90% of the corpus for training and the remaining 10% for test. Our experiments yielded an accuracy of 83.8% on the test set, which consisted of 536 words in total. The confusion matrices summarizing the classification performance in terms of utterances and words are shown in Tables 2 and 3, respectively. In Table 2, *novel* and *nonnovel* denote whether the utterances contain focus kernel or not. Precision, recall, and *F*-score ($f = \frac{1}{0.5/p+0.5/r}$) of non/focus kernel automatic labeling is presented in Table 4. We further present the *F*-scores of non/focus kernel labeling based on different features in Fig. 5. The figure shows that pitch and energy yielded better performance for focus kernel labeling than nonfocus kernel labeling, while the other features were the visa versa, especially part-of-speech and spectral balance cepstral coefficients.

We compared the efficiency of individual features using classification accuracy and the *F*-score-based evaluation measures. Classification accuracy is the fraction of the test set that is correctly classified with respect to focus kernel and nonfocus kernel. *F*-score is a harmonic mean of precision and recall: *F*-score is high only when

Table 2
Confusion matrix for utterance classification: novel vs. nonnovel; novel = utterance containing focus kernel, nonnovel = utterance not containing focus kernel

	Novel	Nonnovel
Novel	53	7
Nonnovel	0	3

Table 3
Confusion matrix for word classification: focus kernel vs. nonfocus kernel

	Focus kernel	Nonfocus kernel
Focus kernel	148	17
Nonfocus kernel	70	301

Table 4
Precision *p*, recall *r*, and *F*-score *f* of non/focus kernel labelling using various features

	Focus kernel			Nonfocus kernel		
	p	*r*	*f*	*p*	*r*	*f*
Dur	0.417	0.727	0.530	0.819	0.547	0.656
Egy	0.313	0.885	0.462	0.725	0.135	0.227
Pit	0.312	0.891	0.462	0.723	0.127	0.216
Pos	0.582	0.236	0.336	0.731	0.925	0.817
SBCC	0.373	0.364	0.368	0.720	0.728	0.724
SDM	0.472	0.873	0.613	0.909	0.566	0.698
Dur+Egy+Pit+Pos +SBCC+SDM	0.809	0.770	0.789	0.900	0.919	0.909

Dur = duration, Egy = energy, Pit = pitch, Pos = part-of-speech, SBCC = spectral balance cepstral coefficients, SDM = semantic dissimilarity measure.

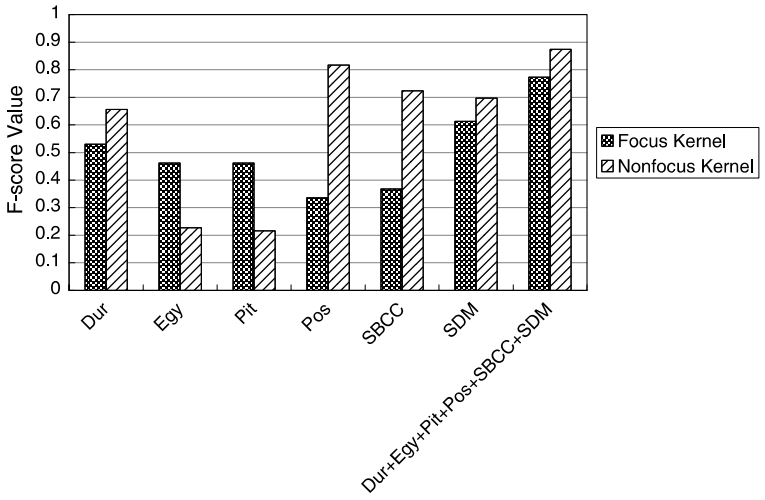

Fig. 5. *F*-score values of focus kernel labeling and nonfocus kernel labeling based on various features. Dur = duration, Egy = energy, Pit = pitch, Pos = part-of-speech, SBCC = spectral balance cepstral coefficients, SDM = semantic dissimilarity measure.

both precision and recall are high. Therefore, *F*-score is a more reliable measure. However, the *F*-score values that we have derived were class-dependent. We were interested in the global performance measure across the classes, so we designed two global *F*-score measures: (1) averaged focus kernel *F*-score and nonfocus kernel *F*-score; and (2) $f_{\text{ave}} = \frac{1}{0.5/p_{\text{ave}}+0.5/r_{\text{ave}}}$, where p_{ave} was the averaged focus kernel precision and nonfocus kernel precision, and r_{ave} was the averaged focus kernel recall and nonfocus kernel recall. The comparison results using classification accuracy and global *F*-score measures are shown in Fig. 6. The figure shows that the combination of features much outperformed the individual features. In addition, comparison based on the global *F*-score measures shows that word dissimilarity measure was the most important feature, followed by duration, part-of-speech, and SBCC.

As for the acoustic correlates of pitch accent, the test results depicted in Fig. 6 show that duration and SBCC played more important roles than energy and pitch for the detection of focus kernels. Both duration and high-frequency intensity have been shown to be reliable acoustic correlates of lexical stress (Sluijter et al., 1997). Pitch played the least significant role in the focus kernel detection, probably by the frequent occurrence of pitch tracking inaccuracies in noisy speech data. It was hard to discriminate voiced regions from unvoiced regions in the noisy recording environment. The energy of unvoiced regions carried information irrelevant to pitch estimate, therefore the automatically extracted pitch might have contained too many pitch tracking errors to be an efficient feature. Spectral balance cepstral coefficients showed better performance than pitch and energy, possibly because: (1) the band-pass filters eliminated the low frequency noise that adversely affected pitch and energy estimates; and (2) spectral balance may be more representative of vocal effort than was overall energy, because the difference in the shape of glottal waveform tended to cause differences in the higher frequency regions (Sluijter et al., 1997).

6.2. Symmetric contrast detection

Symmetric contrast was detected using the algorithm shown in Fig. 7: candidate word pairs were pre-filtered using a series of knowledge-based rules, prior to application of a decision tree program. For the decision tree program, we used *See5*, which is a data mining tool for discovering patterns or relationships in data, assembling them into classifiers that are expressed as decision trees

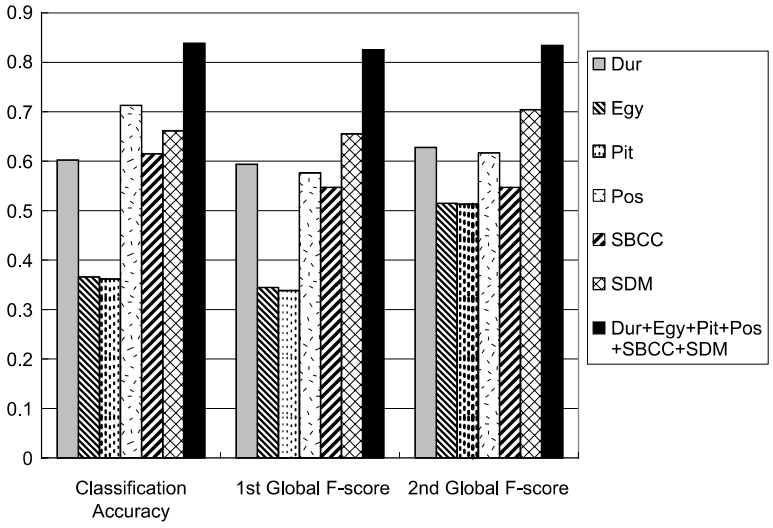

Fig. 6. Feature vector accuracy ranking according to the classification accuracy and two global *F*-score measures: the first measure is the averaged focus kernel *F*-score and nonfocus kernel *F*-score; the second measure is *F*-score computed from the average precision and average recall of the focus kernel and nonfocus kernel labeling. Dur = duration, Egy = energy, Pit = pitch, Pos = part-of-speech, SBCC = spectral balance cepstral coefficients, SDM = semantic dissimilarity measure.

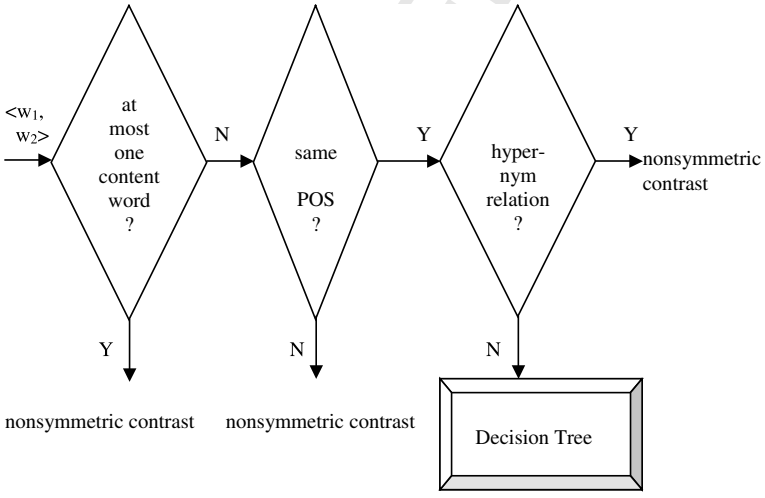

Fig. 7. Architecture of the symmetric contrast detector.

or sets of if-then rules, and using them to make valid predictions (Rulequest Research, 2004).

In experimental tests, we labeled all of word pairs in an utterance as either *symmetric contrast* or *nonsymmetric contrast*. We had 265 symmetrically contrasted pairs. The corpus study showed that 100% of symmetric contrast cases were pairs of content words, and 99.2% of contrasted content-word pairs had the same part-of-speech tags. Therefore, we labeled a pair of words $\langle w_1, w_2 \rangle$ as *nonsymmetric contrast* if at most one of them was a content word, or if they had different part-of-speech tags. We did not model the common integrator requirement for semantic parallelism, but

the semantic independence requirement could be easily represented by requiring that a word may never contrast with its own hypernym.

A pair of words $\langle w_1, w_2 \rangle$ in the utterance was input to the *See5* program, with feature vector being the combination of prosodic measurements and the semantic similarity score, i.e., $f = \{\text{duration, pitch, energy, 13 SBCCs, similarity score}\}$. The similarity score of w_1 and w_2 was

$$\text{sim}(w_1, w_2) = 1 - \text{dis}(w_1, w_2). \quad (19)$$

The prosodic features were the arithmetic mean of w_1 and w_2. The classification target was 1 if $\langle w_1, w_2 \rangle$ was symmetric contrast, and 0 else. The *See5* program was invoked with the options *rule-sets* (collections of *if-then* rules), *ignore costs file* (ignore the penalty for misclassification), and *global pruning* (a large tree is first grown to fit the data closely and is then pruned by removing parts that are predicted to have a relatively high error rate).

After pre-filtering the word pair candidates using the three knowledge-based rules, we obtained 803 word pairs (260 utterances out of the 630 utterance corpus) to input to the *See5* program. Due to the reason that we describe in Section 6.1, we used the pitch accent acoustic correlates (pitch, energy, duration, and SBCC) rather than pitch accent/nonpitch accent labeling from these acoustic correlates for symmetric contrast detection. We used 50% of the filtered word pairs (in 121 utterances) for training the *See5* program. Then we used the symmetric contrast detector depicted in Fig. 7 to test the other 509 ($630 - 121 = 509$) utterances. Table 5 reports the precision, recall, and *F*-score with which utterances were correctly labeled as either containing symmetric contrast (first row in the table) or not containing symmetric contrast (second row in the table). Our test results showed that some individual prosodic features, i.e., pitch, duration and energy, all classified the test word pairs (after pre-filtering) into *nonsymmetric contrast*. In addition, when they were combined with SBCC and semantic dissimilarity measure, they played little role in the symmetric contrast detection: almost all the variables in the rule set used for non/symmetric contrast classification were SBCC variables and the semantic measure. Therefore, pitch, energy and duration were not efficient in the symmetric contrast identification. We further compared the performance of SBCC and semantic similarity measure on the pairs of words after the knowledge rule-based pre-filtering, and listed the results in Table 6. The table shows that SBCC outperformed word similarity measure in the efficiency of non/symmetric contrast classification. We also computed the non/symmetric contrast classification accuracy over all of the word pairs in the 509-utterance test corpus, and the accuracy was 92.8%.

Table 5
Precision *p*, recall *r*, and *F*-score *f* for the correct labeling of utterances containing symmetric contrast (row 1) and not containing symmetric contrast (row 2)

	p	*r*	*f*
Symmetric Contrast	1.0	0.952	0.975
Nonsymmetric Contrast	0.925	1.0	0.961

6.3. Application of focus kernel to a practical SSLU system

Focus kernel has been successfully applied to spontaneous spoken language understanding in the tutoring dialogue scenario by participating in the tutoring event classification (Zhang, 2004). The study corpus was divided into a training set

Table 6
Precision *p*, recall *r*, and *F*-score *f* of symmetric contrast labelling using various features on the word pairs after the three knowledge rule-based pre-filtering

	Symmetric contrast			Nonsymmetric contrast		
	p	*r*	*f*	*p*	*r*	*f*
SBCC	0.789	1.000	0.882	1.000	0.841	0.914
SSM	0.701	1.000	0.824	1.000	0.746	0.855
Dur+Egy+Pit+SBCC+SSM	0.838	1.000	0.912	1.000	0.885	0.939

Dur = duration, Egy = energy, Pit = pitch, SBCC = spectral balance cepstral coefficients, SSM = semantic similarity measure.

and a test set. Focus kernels were extracted from training utterances. A nonparametric model of each tutoring event class was constructed by simply listing all of focus kernels used in training utterances of that tutoring event. The tutoring event classification of test utterances was based on: (1) lexical similarity measure between focus kernels in the test utterance and focus kernels in the nonparametric models of all tutoring event candidates; and (2) cognitive state (*confidence*, *puzzlement*, or *hesitation*) that reflects the students' mental activities during the process of knowledge acquisition. Because the lexical similarity measure was trained using a task-independent corpus (GigaWord) combined with a small task-oriented ontology, it was possible to create this SSLU system using a comparatively small amount of task-specific training data. There were 30 tutoring events in total, and the perplexity of the classification task was 22.5. Accuracy of the tutoring event classification achieved 75.5% when focus kernels and cognitive states were manually annotated, and reduced by 15.4% relative when focus kernels and cognitive states were automatically extracted.

Symmetric contrast labels have not yet been implemented to our speech understanding task, but we expect that automatically labeled symmetric contrast will be useful as a cue for robust semantic parsing.

7. Discussion and conclusions

This paper has computationalized two linguistic concepts, contrast and focus, that were assumed to be useful for robust understanding of spontaneous spoken messages in a dialogue system. Standard and reasonable linguistic definitions of focus are difficult to implement computationally, because the scope of focus is dependent on the syntactic structure of the utterance and highly variable. In order to create a computationally feasible focus detector, this paper has defined focus kernel to be the word in an utterance containing new information neither presupposed by the interlocutor nor contained in the precedent words of the utterance. This paper has also defined symmetric contrast to be a pair of words that are parallel or symmetric in linguistic structure but different or contrastive in meaning. Symmetric contrast marks information about the discourse structure of an utterance that may be useful for robust semantic parsing.

Novelty detection has been widely investigated in natural language processing (NLP) at the sentence or document level. For example, TREC Novelty Track and Topic Detection and Tracking intend to extract sentences or documents that discuss new development of a document or discourse topic. In this study, we have extended the study of informative novelty to an analysis of word-level novelty in spontaneous speech. In future work, it may be possible to use the focus kernel for the purpose of efficiently summarizing the content of an utterance. By definition, the set of words defined to be focus kernel is a maximally informative summary of the utterance in context. In this sense, the use of focus kernels for utterance summarization would be similar to the task of sentence-selection based topic summarization of a document in NLP: topic summarization of a document can be realized by seeking a pool of sentences, each of which expresses information not contained in its precedent sentences in the document (Zechner and Waibel, 2000).

The effectiveness of the proposed symmetric contrast detection and focus kernel detection systems has been evaluated using a transcribed children's spontaneous speech corpus, which was collected using Wizard-of-Oz simulations of an intelligent tutoring dialogue system. The detection of symmetric contrast and focus kernel was based on word dissimilarity measures, part-of-speech tagging, and measurements of the acoustic correlates of prosody including duration, pitch, and energy. We also propose spectral balance cepstral coefficients (intensity and shape of the high-frequency spectrum) as one more acoustic correlate of pitch accent. The word dissimilarity measure combines corpus statistics and application-oriented ontology. Classification achieved accuracies of 83.8% for focus kernel classification and 92.8% for symmetric contrast classification. Our tests showed that spectral balance cepstral coefficients, the dissimilarity semantic measure, and

part-of-speech played important roles in the focus kernel and symmetric contrast detections.

Focus kernel extraction has been applied in a spontaneous spoken language understanding system. Symmetric contrast has not yet been applied in any spoken language understanding (SLU) system, but seems well suited for use as a cue in robust semantic parsing. Our SLU based on the detection of focus kernel and symmetric contrast differs from most existing SLU systems in that it is not keyword-based (does not require the system designer to specify, in advance, a list of "case markers"), and it also does not require predefined grammar (used for syntactic parsing or extraction of finite-state-machine-based semantic concepts). Focus kernel and symmetric contrast begin SSLU by extracting, in a relatively general way, the pragmatic and semantic salient information of the utterance, thus allowing the interpretation of utterance meaning from unconstrained human language.

8. Uncited reference

Ren et al. (2004b).

Acknowledgements

We would like to thank Richard W. Sproat for helpful discussions and providing us with the GigaWords text corpus. We would also like to thank Carla Umbach and Chungmin Lee for their comments and suggestions. 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.

Appendix A

The ontology construction procedure:

1. Noun, verb, adjective, and adverb are categorized into the semantic class *Content*, and function words are categorized into the semantic class *Function*.
2. The content words are clustered according to their semantic meanings, regardless of their part-of-speech:
   (1) For each *noun*
      a. According to the word meaning in the ITS dialogue context, search the WordNet for the most appropriate hypernym of the word. Delete those concepts that contain redundant information for knowledge representation in the ITS scenario. For example, in WordNet we have month → Gregorial month → April (' → ' denotes subordination). However, we delete 'Gregorial month' in constructing ontology of the ITS corpus.
      b. Check WordNet to see if any word in the corpus can be the synonym of the noun.
      c. If the noun has no hypernym, then it becomes a direct subordinate of *Content*.

   (2) After the hierarchy of noun words has been constructed, we
      a. Verify that every member of a concept hierarchy is subordinate of the root concept.
      b. Verify that the direct subordinates of a concept are semantically parallel to each other.
      c. Verify that the hierarchy rooted in a concept has distinct meaning from the hierarchies rooted in other parallel concepts.

   (3) For each *verb*, strip tense, and then check the WordNet to see whether its hypernyms ('is one way to' relationship) are identical or similar to some noun concepts has been defined in the ontology. If yes, then merge the associated hierarchy into the existing ontology. Otherwise, add a new verb concept to the ontology following the step (1). Some verbs have multiple meanings relevant to the ITS scenario, e.g., *make* means

Abstraction
- *Attribute*
  - *Disposition*
  - *Pattern*
  - *Property*
    - *Color*
    - *Magnitude=Amount + Degree + Duration + Size*
    - *Shape=Circles + Crooked + Space*
  - *Quality=Certainty + Correctness + Difference + Excellency + Hardness + Importance + Strength*
- *Quantity*
  - *DefiniteQuantity*
    - *Number=Count + Integer + Fraction*
  - *FundamentalQuantity*
    - *TimePeriod=Unit + Calendar + Time*
  - *IndefiniteQuantity*
- *Relation=Group + Indication + Link + QuantitativeRelation*

Act:
- *Change*
  - *Integrity*
  - *Magnitude*
  - *Ownership*
  - *Position=Lined + Move + Rotation + Step*
  - *State*
- *Do=Exploring + Make + Paint + Put + Show + Try + Use*
- *Duty*
- *Interact*
  - *Communication=Approval + Convey*
  - *Contact*
- *SpeechAct*
  - *Inform*
  - *Request=Ask + Question*
  - *Say=Answer + Repeat*

Be=BeClass + Equal + Have + Stay

Entity
- *Location=Line + Point + Region*
- *Object*
  - *Artifact=Construction + Device + Representation + Transport*
- *Thing=Anything + Body + Life*

Phenomenon
- *Effect*
- *PhysicalPhenomenon=Force + Resistance*

PsychologicalFeature
- *Cognition*
  - *Awareness*
  - *Information=Example + Experience*
  - *Intelligence*
  - *Perception=Find + See*
  - *Prospect*
  - *Think=Calculation + Figure + Guess*
- *Feeling*

Fig. B1. Major ontological categories in the ITS corpus.

'make somebody do something' or 'produce something.' Decode the multiple meanings of a word, if any.

(4) Check the ontology structure:
   a. Search among the ontology for sets of concepts that can be merged together.
   b. Search for more synonym words in the ITS dialogue scenario. The synonyms of a word can be words of different part-of-speech, e.g., a noun can be a synonym of a verb as long as their semantic meanings are similar, such as *rotation* and *spin*.
   c. Go to step (2).

(5) For each *adjective*, check the WordNet to see whether its hypernyms ('is a value of' relationship) are identical or similar to a noun concept. If yes, then merge the associated hierarchy into the existing ontology. Otherwise, use the verb or noun derivation (or root) of the adjective to find its hypernym, following step (1). For example, we use the information about *confuse* to find the hypernym of *confused.* Define its synonym and antonym with reference to the WordNet. If the hypernym of the adjective cannot be found, then we use the hypernym of its noun or verb synonyms. For example, we use the hypernym of the verb *agree* to find the hypernym of *alright*. Otherwise, use the synonyms of the adjective as clues to find its hyponyms. Then go to step (4).

(6) For each *adverb*, WordNet does not define a hypernym. Therefore, we use the stem adjectives as information clues to find the hypernyms. If an adverb has no stem adjective, then use the stem adjectives of its synonyms as information clues. Then go to step (4).

3. Function words typically have little or no semantic content apart from their syntactic use, so we categorize them according to syntactic usage: the words of the same part-of-speech are synonyms to each other, and their part-of-speech is their hypernym.

Appendix B. Fig. B1.

Appendix C

To illustrate the procedure of using corpus statistics for word dissimilarity, consider the utterance "*I lost count let me try again.*" We choose $\lambda_1 = \lambda_2 = 0.6$, $\gamma_1 = 2$, $\tau_1 = 2$, $\gamma_2 = 1$, $\tau_2 = 1$, and present the derived context of the content words in Fig. C1. The PMI scores of some pairs of words are shown in Fig. C2. The listed scores show the problem with mutual information is that it is biased towards infrequent words, such as 'alright.' The dissimilarity scores are shown in Fig. C3. Therefore as defined by Eq. (8), we have the novelty measures for the content words in the utter-

		C_1	C_2	C_3
1	*lost*	Count	let, loud	∅
2	*count*	let, lost, loud	alright, out, start, try	Try
3	*Let*	alright, count, start, try	Figure	figure, loud
4	*Try*	figure, let	alright, count, out, start	count, out
5	*again*	∅	∅	∅

Fig. C1. Context examples.

1	alright count	2.545	8	count figure	2.102
2	alright let	3.848	9	count lost	1.945
3	alright lost	2.567	10	count loud	1.989
4	alright try	2.982	11	count out	1.873
5	figure lost	0.969	12	count start	1.809
6	figure out	1.237	13	count try	1.840
7	figure try	1.100	14	start try	1.903

Fig. C2. PMI examples.

1	count lost	0.021	6	try let	0.006
2	let lost	0.014	7	again lost	0.002
3	let count	0.037	8	again count	0.003
4	try lost	0.004	9	again let	0.011
5	try count	0.027	10	again try	0.003

Fig. C3. Examples of dissimilarity scores.

ance: $N_{\text{count}} = 0.021$, $N_{\text{let}} = 0.014$, $N_{\text{try}} = 0.004$, $N_{\text{again}} = 0.002$.

Appendix D.

Fig. D1 lists the dissimilarity scores of some word pairs obtained by the ontological, statistical, and combinational methods, respectively. The table shows the examples of the four cases in knowledge combination: (1) dissimilar by both ontology and corpus, as in ⟨*any let*⟩ and ⟨*center happen*⟩; (2) dissimilar by ontology but similar by corpus, as in ⟨*already out*⟩ and ⟨*take down*⟩; (3) similar by ontology but dissimilar by corpus, as in ⟨*count figure*⟩ and ⟨*rate speed*⟩; and (4) similar by both ontology

		Ontology	*Corpus*	*Combination*
1	*any let*	14.861	0.052	0.561
2	*center happen*	13.824	0.125	0.771
3	*already out*	16.898	0.003	0.456
4	*take down*	14.857	0.002	0.399
5	*count figure*	0	0.032	0.104
6	*rate speed*	1.536	0.028	0.131
7	*how when*	2.985	0.005	0.095
8	*Ones teeth*	2.985	0.009	0.108

Fig. D1. Examples of word dissimilarities by combined ontology and statistics.

and corpus, as in ⟨*how when*⟩ and ⟨*ones teeth*⟩. The combination occurs when the interpolation parameter $\lambda = 0.5$. The examples show that the combination compromises the errors caused by the specification and the generality of ontology and corpus, respectively.

References

Beckman, M.E., Ayers, G.M., 1994. Guidelines for ToBI Labeling. Available from: <http://www.ling.ohio-state.edu/phonetics/ToBI/main.html>.

Bolinger, D., 1961. Contrastive accent and contrastive stress. Language 37, 83–96.

Bolinger, D., 1965. Forms of English. Harvard University Press, Cambridge, MA.

Bosch, P., van der Sandt, R., 1999. Focus: Linguistic, Cognitive, and Computational Perspective. Cambridge University Press, Cambridge, UK.

Chu-Carroll, J., Carpenter, B., 1999. Vector-based natural language call routing. Comput. Linguistics 25 (3), 361–388.

Chomsky, N., 1971. Deept structure, surface structure and semantic interpretation. In: Steinberg, D., Jakobovits, L. (Eds.), Semantics: An Interdisciplinary Reader in Linguistics, Philosophy and Psychology. Cambridge University Press, Cambridge, UK.

Dagan, I., Marcus, S., Markovitch, S., 1995. Contextual word similarity and estimation from sparse data. Computer Speech Language 9, 123–152.

Dahl, D., 1969. Topic and Focus: a Study in Russian and General Transformational Grammar. Elandres Botryckeri, Göteborg.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Information Theory 36 (5), 961–1005.

Edmonds, P., Hirst, G., 2002. Near-synonymy and lexical choice. Comput. Linguistics 28 (2), 105–144.

Firbas, J., 1964. On defining the theme in functional sentence analysis. Travaux Linguistiques de Prague 1, 267–280.

Firbas, J., 1966. Non-thematic subjects in contemporary English. Travaux Linguistiques de Prague 2, 229–236.

Flammia, G., 1998. Discourse segmentation of spoken dialogue: an empirical approach. Ph.D. Thesis. MIT.

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

Gundel, J.K., Fretheim, T., 2001. Topic and focus. In: Horn, L., Ward, G. (Eds.), The Handbook of Pragmatic Theory. Blackwell Publishers, Malden, MA.

Gussenhoven, C. 2002. Intonation and interpretation: phonetics and phonology. in: Bel, B., Marlien, I. (Eds.), Speech Prosody 2002, Aix-en-Provence.

Halliday, M., 1967. Notes on transitivity and theme in English. Part II. J. Linguistics 3, 199–244.

Hedberg, N., Sosa, J.M., 2001. The prosody of topic and focus in spontaneous English dialogue. LSA Topic and Focus Workshop.

Heldner, M., Strangert, E., Deschamps, T., 1999. A focus detector using overall intensity and high frequency emphasis. Internat. Congress of Phonetic Science.

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

Jackendoff, R., 1972. Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA.

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proc. Internat. Conf. Research on Comput. Linguistics.

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

Kay, M., 1975. Syntactic processing and functional sentence perspective. In: Schank, R., Nash-Webber, B. (Eds.), Theoretical Issues in Natural Language Processing. MIT Press, Cambridge, MA.

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

Kim, S.-S., Hasegawa-Johnson, M., Chen, K., 2003. Automatic recognition of pitch movements using multi-layer perceptron and time-delay recurrent neural network. IEEE Signal Process. Lett. 11 (7), 645–648.

Krifka, M., 1999. Additive particles under stress. Proc. of SALT 8.

Kruijff-Korbayova, I., Steedman, M., 2003. Discourse and information structure. J. Logic Language Inf. 12, 249–259.

Lee, C., 1999. Contrastive topic: a locus of the interface. In: Turner et al. (Eds.), The Semantics/Pragmatics Interface from Different Points of View (CRiSPI 1). Elsevier Science, Amsterdam.

Lee, C., 2003. Contrastive topic and/or contrastive focus. in: McClure, B. (Ed.), Japanese/Korean Linguistics 12, CSLI, Stanford.

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

Lenci, A., 2001. Building an ontology for the lexicon: semantic types and word meaning. In: Jensen and Skadhauge (Eds.) Ontology-based Interpretation of Noun Phrases: Proc. 1st Internat. OntoQuery Workshop.

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

Miller, G., Fellbaum, C., 2002. WordNet. http://www.cogsci.princeton.edu/~wn/.

Munoz, M., Punyakanok, V., Roth, D., Zimak, D., 1999. A learning approach to shallow parsing. In EMNLP-WVLC '99.

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

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M.E. (Eds.), Intentions in Communication, pp. 271–311.

Rada, R., Mili, H., Bicknell, E., Bletner, M., 1989. Development and application of a metric on semantic nets. IEEE Trans. Systems Man Cybernet. 19 (1), 17–30.

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

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

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

Rooth, M., 1992. A theory of focus interpretation. Natural Language Semantics 1, 75–116.

Rulequest Research, 2004. Data Mining Tools. Available from: <http://www.rulequest.com/see5-info.html>.

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

Sluijter, A., van Heuven, V.J., Pacilly, J., 1997. Spectral balance as a cue in the perception of linguistic stress. The J. Acoust. Soc. Amer. 101 (1), 503–513.

Steedman, M., 2000. Information structure and the syntax–phonology interface. Linguistic Inquiry 31 (4), 649–689.

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

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

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

Umbach, C., 2004. On the notion of contrast in information structure and discourse structure. J. Semantics 21 (2), 155–175.

Vallduví, E., Vilkuna, M., 1998. On rheme and contrast. In: Culicover, P., McNally, L. (Eds.), Syntax and Semantics, The Limits of Syntax, Vol. 29. Academic Press, San Diego, CA.

Welby, P., 2003. Effects of pitch accent position, type, and status on focus projection. Language Speech 46 (1), 53–81.

Xu, Y., Xu, C.X., Sun, X., 2004. On the temporal domain of focus. Proc. ISCA Internat. Conf. on Speech Prosody.

Yoon, T.-J., Chavarria, S., Cole, J., Hasegawa-Johnson, M., 2004. Inter-transcriber reliability of prosodic labeling on telephone conversation using ToBI. Proc. ICSLP.

Zechner, K., Waibel, A., 2000. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. Proc. of COLING.

Zhang, T. 2004. Spoken language understanding in an intelligent tutoring scenario. Dissertation, University of Illinois at Urbana-Champaign.

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

Zubizarreta, M.L., 1998. Prosody, Focus, and Word Order. MIT Press, Cambridge, MA.