TR-IT-0293

Topic Management for Spoken Dialogue Systems

クリスティーナ・ヨキネン
Kristiina Jokinen

February, 1999

This report concerns dialogue topics and the usefulness of topic information in language modelling for speech recognition. I describe a linguistically motivated topic-model and the Predict-Support algorithm to recognize topics in task-oriented dialogues, and also report results of the accuracy of the algorithm.

Contents

1 **Introduction** 1

2 **Information Structure of Utterances** 2
- 2.1 Topic and Comment . . . 2
- 2.2 Focus and Background . . . 4
- 2.3 Central Concept and NewInfo . . . 5

3 **Tagging corpus with topics** 6

4 **Topic trees** 9

5 **Topic model and spoken language systems** 10
- 5.1 The Predict-Support Algorithm . . . 11
- 5.2 Speech Recognition . . . 13

6 **Future Work** 14

Acknowledgements 14

References 15

Figure Contents

1 CCs and NewInfos in a small world model. . . . 7

2 Information structure of utterances. The underlines denote old information, and the boxes denote new information. . . . 7

3 A partial topic tree for room reservation domain. . . . 9

4 A partial topic tree for flight reservation domain. . . . 10

5 Scheme of the Predict-Support Algorithm. . . . 11

Table Contents

1 Dialogue statistics. . . . 6

2 Topic tags for the experiment. . . . 9

3 Accuracy results of the first predictions. . . . 12

4 Accuracy and precision results for different topic types. . . . 13

5 Trigram perplexity in the experiments. . . . 14

1 Introduction

Rational agents act coherently. When communicating with others, they produce utterances which are both intentionally and thematically linked to previous utterances: their intentions as well as the content of the utterances contribute to some logical organisation of the related events and propositions. Coherence facilitates understanding and is an integral part of the rational agent's cognitive processing, hence also its importance to dialogue modelling and speech recognition.

Speakers' intentions are usually modelled in terms of dialogue acts (speech acts, communicative acts), and the information content of their utterances with the help of the notion of topic (focus). The set of necessary dialogue acts is by no means agreed upon, but the act classification can be assumed to be domain-independent thus providing a suitable basis for statistical coherence measures, e.g. [1]. The content of the utterances, however, is related to the exchange of domain information, and a similar domain-independent classification for topics is impossible. Domain-independent dialogue models thus tend to discard topic information, although such general models also tend to be less specific and hence less accurate.

In AI-based dialogue modelling, the use of topic (focus) has been mainly supported by arguments regarding processing effort (search space limits) and anaphora resolution, and they are associated with a particular discourse entity, *focus* or *backward-looking center*, which is currently in the centre of attention and which the participants want to focus their actions on [2, 3]. However, the goal in this research is to use thematic information in predicting likely content of the next utterance, and thus the interest lies in the topic *types* that describe new information in the utterance than in the actual topic entity. Consequently, instead of tracing foci or backward-looking centers, we seek a formalisation of the utterance's information structure in terms of the new information that is exchanged in the utterances.

On the other hand, the purpose of the topic model is to assist speech processing, and thus extensive and elaborated reasoning about plans and world knowledge is not possible. Rather, a model that relies on observed facts and uses statistical information is preferred, as it provides a quick mechanism to process input utterances and can be combined with a speech recognizer. The model should also be general and extendable, so that if more factors are to be taken into account, it could easily adapt to these changes. For instance, sentential stress and pitch accent are important factors in speech processing, and they closely interact with the information structure of utterances. Although prosody is not discussed in detail in this report, we require that topic modelling should anticipate an account of these characteristics as well.

The guidelines for topic modelling can be summarised as follows. The topic model should be

1. linguistically motivated: based on the information structure of the utterances
2. surface syntax oriented: no deep analysis of the meaning of the sentence nor world model available,
3. operational: possible to recognise automatically.

In this report, I propose such a topic model and report results of applying the model in spoken language system. First, a *topic tree* which describes possible dependencies between different topic types for the particular domain is extracted from the dialogue corpus. The tree is based on shallow domain modelling, and it provides top-down information for the prediction of the likely next topic. Possible topic sequences are modelled by trigram probabilities, calculated on the basis of the training corpus and smoothened by the backoff method. Then, bottom-up analysis of the information structure of the utterances is used to identify new information conveyed by particular words of the utterance, and these words are matched to possible topic types via *topic-vectors* which encode the mutual information between particular words and the topic types. Finally, the topic is assigned to an unseen input utterance by the *Predict-Support Algorithm* as the best candidate out of the possible topic types proposed by the utterance's topic words, using the mutual information encoded in the topic vectors and the likelihood information encoded into the topic tree about the relative probabilities of the shifts from the previous topic to the candidates.

The report is organised as follows. Section 2 gives linguistic background for the topic model: two different information structures for utterances, the topic-comment structure and the focus-ground structure, are reviewed from the point of view of our research goals. Our own approach which combines the two

structures is also introduced. Section 3 gives an overview of the topic tree and briefly describes topic tagging and tagging principles. Section 4 discusses the usefulness of the topic model in speech recognition and reports on the results of applying the model in speech processing. Finally, Section 5 gives conclusions and points to future work.

2 Information Structure of Utterances

The bottom-up approach to topic modelling is a data-driven study on how information is conveyed in individual utterances: what kind of syntactic-semantic constructions convey information about the topic and how the new information exchanged in the utterances is recognised. Two basic approaches to describe the information content of utterances have been suggested[1]: the topic-comment structure (what is talked about *vs.* what is said about it) and the focus-ground structure (new *vs.* old information). In what follows, the "aboutness" and "newness" approaches are briefly reviewed in regard to our research goals on dialogue modelling, and then our own approach which combines the two information structures is presented.

2.1 Topic and Comment

A note on terminology: in this section, I follow linguistic conventions and use *topic* as a specific technical term which refers to the particular constituent that the utterance talks about. This usage is more restricted than the generic use of the term elsewhere in the paper where it describes the thematic structure of utterances in general. If there is a danger of ambiguity, I will make it clear which meaning is meant.

The topic-comment structure emphasises the "aboutness"-aspect of the utterances: the speaker announces a topic and says something about it [6, 7], cf. also [8] who uses the term *theme*. The topic is usually the leftmost constituent of the utterance and often this happens to be the grammatical subject denoting the actor (1).

(1) *[I]$_T$ would like to reserve a room for August 23rd.*
Talking about the speaker: what would you like to do?

To assign topicality to some other element than the default subject, syntactic marking can be used as in (2)-(3).[2] The order in which different sentential elements can be considered topics is captured in the topicality hierarchy [10], akin to the availability hierarchy of different foci in AI focus stacks.

(2) *[The room]$_T$ is reserved by me for August 23rd.*
Talking about the room: what is the situation with it?

(3) *[August 23rd,]$_T$ that's when I have my room reservation for.*
Talking about the time: what's so special about August 23rd?

The strict topic-comment structure has several drawbacks. First, utterances may be topicless. Presentational sentences (neutral descriptions, [11]) provide information about a general setting for the discussion but are not about a particular referent. For instance, in the beginning of the dialogue the utterance

(4) *I'd like to make a reservation for a single room.*

introduces the task and the global theme for the dialogue, but is not especially about "I" or reservation or single rooms. Elliptical utterances, however, like the answer in (5), have no explicit surface topic.

(5) *How would you like to pay? - By Master Card.*

The topic of the elliptical answer becomes available if I augment the answer into the full sentence *I would like to pay by Master Card*, but then the problem arises whether the topic is actually the speaker *I*, or the paying method which the exchange is about.

Second, utterances may have the same meaning as far as the aboutness is concerned, but differ in the ways in which the comment updates information about the topic. For instance, the utterances in (6) talk about the speaker, but (a) focuses on what she would like to do, while (b) focuses on how she would like to pay:

[1] The distinction has been independently described in [4] and in [5].

[2] Some languages, like Japanese, grammaticalize topic marking with the help of a special topic marker, while others, like Finnish, use discourse configuration [9].

(6) a. *[I]$_T$ would like to [pay by Master Card.]$_{Focus}$*
b. *[I]$_T$ would like to pay [by Master Card.]$_{Focus}$*

These differences are related to the new information that the comment conveys about the topic, and cannot be captured by the flat topic-comment structure, see Section 2.2.

The biggest concern in regard to topic-comment structure is based on the fact that in task-oriented dialogues, most topics deal with the speaker and provide no help for the task identification. It can be argued that the utterances (7) convey information about the speakers' attitudes and actions related to the task, but usually this sort of information is modelled on the intentional structure of the dialogues and encoded into the speakers' beliefs and intentions (dialogue acts), so it is separate from the thematic organisation of task topics.

(7) *[I]$_T$ would like a single room, please.*
May [I]$_T$ have your name, please?
Then, [I]$_T$ will take a twin
[I]$_T$ am staying at the Washington Hotel right now
[We]$_T$ have singles, and twins and also Japanese rooms available

In her thorough study of the topic structures in information seeking dialogues, [7] points out that information exchanges in fact operate on two levels: they provide both meta-level information about the speakers and task-level information about the task objects. Each utterance has either a meta-level or task-level topic explicitly present (while the other topic is implicit), and the utterance topic can shift between the two levels. However, considering the thematic structure of the dialogues in this study, the speaker is seldom the main issue but rather, the progression of the underlying task[3]. It thus seems somewhat superfluous to postulate another topical line to track the meta-level topics for the utterances like (7). Furthermore, each utterance always carries information about the speaker (about her knowledge and intentions in regard to the task and the partner), so the special character of the utterances (7) is that their syntactic structure makes the attitudinal information explicit. On the task-level, however, their most informative part is the comment which introduces new task-level elements (a single/twin room, the name, as in the examples 7). As in (8) subsequent utterances usually shift topic to this newly introduced task entity (in fact, 71% of the time according to [7]), suggesting that the topics of the utterances (7) serve as grammatically required starting points of the sentences, but that the locus of what the utterance is about is encoded in the new information carried by the comment.

(8) customer: [I]$_T$ would like a single room, please.
clerk: All right, just a moment, please.
[The single rooms with a bath]$_T$ are all full.

We use both the speaker's intentions and the content information to model utterances, and thus the "speaker-topics" in (7), dealing with the speaker's attitudes can be encoded into the speaker's intentions (modelled by dialogue acts), while the important task information embedded into the comment part is available as the new information of the utterance.

[7] also observes a strong tendency toward topic continuation in her dialogues: 76 % of the topic-comment structures continue the previous topic. In fact, of her six different topic-comment structures, defined with respect to the different logical possibilities of how topic-comment structures can be attached to each other, only the comment thematization (the comment is taken as the topic of the next utterance like in (8)), can be interpreted as causing real topic shifts; in other structures the same topic continues either explicitly or by being embedded in the topic-comment structure. This kind of topic continuity seems to support our goal to rely on the new information of the utterances rather than their "aboutness": while the topic entities form topical chains through the dialogue and contribute to the whole dialogue hanging together, they do not model what kind of information is exchanged in the utterances. Thus identification of the topic entity is not regarded as of primary importance in our coherence studies. However, the topic is important in anaphora

[3]Utterances like (7) give information about the speaker, but they do not talk about the speaker in the same way as e.g. the following utterances: *What can I tell about myself? Well, I'm a handsome young chap who likes motor-bikes and old-fashioned girls.*

resolution and pronoun generation, and it must be emphasised that an adequate dialogue model should incorporate knowledge of the topic as well.

2.2 Focus and Background

I now turn to the "newness"-approach which seems a useful basis for our topic model[4]: the speakers exchange new information, this is always realised in utterances, and the locus of new information is related to the sentential nuclear stress, which makes it important for speech processing.

In the linguistic literature, new information is called *focus*, while the known or expected information is *old* or *(back)ground* [12, 13, 14, 5]. An unfortunate terminological confusion is caused by the use of the term *focus* in the AI-literature where it refers to the most salient element activated in the course of the dialogue. The AI-focus can be referred to by a pronoun in subsequent utterances, and it is thus related to the linguistic "topic" (aboutness) rather than to "focus" (newness), see more of the differences in [15, 4, 5]. I avoid the use of "focus", and will use the clearer NewInfo when referring to the new and informative part of an utterance, see section 2.3.

As mentioned in 2.1, utterances may have the same topic-comment structure but differ with respect to the new information that the comment carries. For instance, the utterances in (9) have *I* as the linguistic topic, but they are not interchangable in any given context, since the new information (marked with the subscript *New*) is different in each case. (Capital letters mark sentential stress.)

(9) (a) *[I'd like to pay by MASTER Card]$_{New}$*

(b) *I'd like to [pay by MASTER Card]$_{New}$*

(c) *I'd like to pay [by MASTER Card]$_{New}$*

(d) *[I]$_{New}$ 'd like to pay by Master Card*

(e) *I'd like to pay by [MASTER]$_{New}$ Card*

(f) *I'd like to [PAY]$_{New}$ by Master Card*

[14] defines new and old information as follows:

> If a sentence $S = XBY$ is addressed to a sentence $S' = XAY$, then
>
> string B is *old* if B repeats A, and
>
> string B is *new* or *focus* if B replaces A.

The sentence S addresses the sentence S' if S can be interpreted in the context of S'. The context is usually set up by appropriate questions. For instance, following the two-level question method of [5], contexts for the utterances (9) can be set up as follows:

(10) (a) Hello. How can I help you?
[I'd like to pay by MASTER Card]$_{New}$

(b) What about you and your doings? What would you like to do next?
I'd like to [pay by MASTER Card]$_{New}$

(c) What about paying? How would you like to pay?
I'd like to pay [by MASTER Card]$_{New}$

(d) What about paying? Who would like to pay by Master Card?
[I]$_{New}$ 'd like to pay by Master Card

(e) What about paying? By which card would you like to pay?
I'd like to pay by [MASTER]$_{New}$ Card

(f) Sorry, did you say that you'd like to play with a Master Card?
(No,) I'd like to [PAY]$_{New}$ by Master Card

[4] From now on I again use "topic" in a generic sense referring to the thematic structure of dialogues, and do not use it as a technical term of the topic-comment structure.

(g) What about paying? You will pay somehow, but...
[HOW]$_{New}$ would you like to pay?

The question method works fine for declarative sentences but is less obvious for the interpretation of interrogatives. Traditionally, the context for questions is the set of presuppositions which the question is based on, exemplified by the three dots in (10g), and consequently new information deals with what is asked: the parameter whose value or instantiation is requested, or the proposition whose truth value is questioned.

The problem with the question method is to find the appropriate question(s) that the utterance addresses. In dialogues, the context is often provided by the explicit questions asked, but the speakers also use statements for which no previous question is available. Moreover, the question method allows subjective interpretations and the chosen contexts sound arbitrary. However, the purpose of the question method is not to find the single best question that an utterance addresses, but to clarify the information structure of the utterance in terms of new and old information. The overall dialogue context usually contains enough constraints for this, and although the individual questions attached to an utterance may vary, consensus on what is the appropriate information structure of the utterance can emerge among researchers during repeated revisions. In this, the question method resembles the paraphrase method used in [16] to specify communicative act labels according to the meaning equivalencies of cue patterns in contexts: reliability of judgements is based on observed objects and shared conventions.

It is also true that the question method is difficult to automatize, and the main purpose for using it has mainly been to offer a bootstrapping method for topic tagging and guidelines for linguistically motivated topic modelling. However, by marking the syntactic constructions that carry new information, I observed that they usually contain the main verb and its complement. Typically, the main verb acts as the *pivot* of the utterance, as stated in [17]. The first step in automatizing the question method is thus to divide the utterance in two parts by extracting the main verb and its complement(s) and assigning the new information status to them (see [18] for using the pivot-approach in building speech recognition). I envisage that if the parse also contains information about the word semantics (i.e. the structural part-of-speech tags simultaneously represent the classification of the word along semantic categories as in the ATR tagger [19], or along some topic hierarchy), the topic tags can be hypothetised from the parser output. Moreover, if there is access to prosodic information, new information can be equated with the phrase which carries the sentential stress.

2.3 Central Concept and NewInfo

To combine aboutness and newness approaches in response planning, [4] uses the following notions based on a contextual model which consists of a rich world model (static knowledge of concepts and their conceptual hierarchy) and a dynamically modified context (instantiated discourse referents):

Central Concept, CC: a discourse referent which the utterance is about.

NewInfo: a concept or a property value of a concept which is new with respect to some CC.

NewInfo is the information centre of the utterance, and is always explicitly realised. It can be singled out by the question method. CC fixes the view-point from which NewInfo is presented, and its realisation depends on the context: if recoverable from the context, it need not be explicitly realised (elliptical utterances), while object-type CCs may be realised as pronouns. CC often corresponds to the topic of the topic-comment structure, but as it is defined in terms of concepts rather than syntactic realisation, it may appear in other positions than the first element as well (see example (11c-g). The concepts related to CC form the background for the utterance: the concepts already instantiated in the course of the dialogue belong to old information, while those pending to be realised are potential new information, likely to be talked about next.

Incidentally, the same kind of distinction between aboutness and newness is made by Vallduví [5]. However, his starting point is not in dialogue management, but in cross-linguistic realisation of information packaging, and thus he operates on the level of syntactic pharses rather than discourse referents. Vallduví divides utterances into *focus* and *ground*, and the ground further into *link* and *tail*. The link refers to the topic-entity, roughly the Central Concept, whereas the tail, for which there is no immediate correlate in my

terminology, contains further information about how the focus updates the link. The combinations of the tripartite division then instruct the hearer on the ways in which she should update her information state.

I use my own terminology but follow Vallduví in presenting the information structure of the utterances on the two levels: NewInfo and background, with the (back)ground G containing the central discourse referent, Central Concept, CC. The full information structure of utterances can thus be represented as in (11), and the locus of information that we are interested in our topic modelling is the NewInfo marked with the subscript New.

(11) (a) Hello. How can I help you?
[I'd like to pay by MASTER Card]$_{New}$

(b) What about you and your doings? What would you like to do next?
As for me,
[[I]$_{CC}$ would like to]$_G$ [pay by MASTER Card]$_{New}$

(c) What about paying? How would you like to pay?
As for how to pay,
[I would like to [pay]$_{CC}$]$_G$ [by MASTER Card]$_{New}$

(d) What about paying? Who would like to pay by Master Card?
As for paying by Master Card,
[I]$_{New}$ [would like to [pay]$_{CC}$ by Master Card]$_G$

(e) What about a card? By which card would you like to pay?
As for paying by a card,
[I'd like to pay by]$_G$ [MASTER]$_{New}$ [[Card]$_{CC}$]$_G$

(f) You'd just like to show off your Master Card.
As for showing off Master Card,
No, [I'd like to]$_G$ [PAY]$_{New}$ [by [Master Card]$_{CC}$]$_G$

(g) What about paying? You will pay somehow, but...
[HOW]$_{New}$ [would you like to [pay]$_{CC}$]$_G$

Given a world model and a conceptual hierarchy on which the discourse referents are based as unique instantiations of the concepts in the world model, the relation between the CC and NewInfo in each case can be depicted as like that in Figure 1. The characters refer to the particular examples, and the arrow points from CC to NewInfo.

3 Tagging corpus with topics

To enable computational testing with different topic types, I tagged 80 transcribed English dialogues of the spoken bilingual ATR Travel Dialogue Corpus according to the information structure outlined above. The dialogues deal with hotel reservations and tourist information between a clerk and a customer, and their basic statistics are given in Table 1.

The speakers' turns are segmented into utterances according to written language markers (periods, commas) already in the transcripts. Hesitations, filled pauses, etc. marked as [uhh], are analysed as temporizers (time management acts), and considered separate acts unless they occur in the middle of a sentence (production errors). Complex utterances are divided into their constituent clauses (12), except for conditional

Table 1: Dialogue statistics.

speaker	turns	percentage	utterances	percentage	word forms	percentage
clerk	881	53 %	2486	59 %	845	84 %
customer	788	47 %	1736	41 %	663	66 %
total	1669	100 %	4222	100 %	1008	

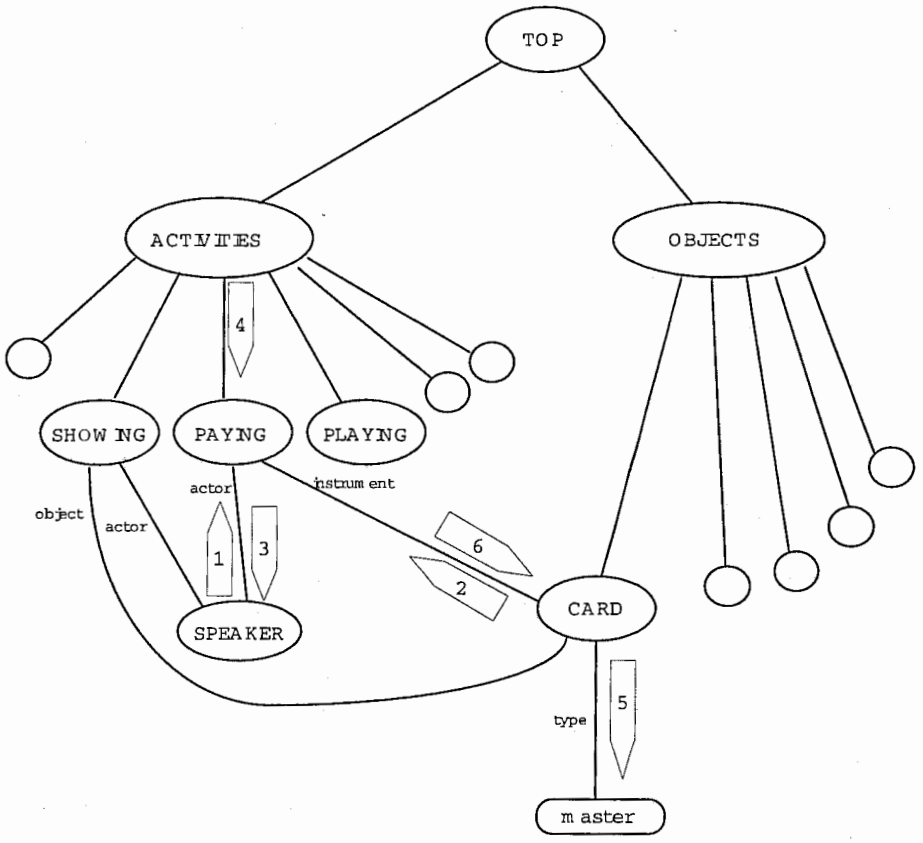

Figure 1: CCs and NewInfos in a small world model.

and temporal clauses (13) and coordinations of reasons and conclusions (14), which are regarded as single utterances. (In the examples, topic tags are enclosed in asterisks.)

(12) *My name is Kazuo Suzuki* *Name*
and I have a VISA card *CardType*
and the number is 4883 5800 4088 1718. *CardNumber*

(13) *Please wait while I check availability.* *iam* (waiting)
if I can help you with anything else, please feel free to give me a call. *iam* (calling)
I'd like a Japanese-style room if possible. *RoomType*

(14) *We really love Kyoto and would like to stay for an extra night.* *ExtendReservation*
I'd like to relax so I definitely need a room with a bath. *Room*

The tags are abstractions of the NewInfo which is exchanged on a particular CC. The examples in Figure 3 show how the tags are assigned to utterances with different information structures.

Immediate Context: Hello. How can I help you?
Utterance: I'd like to [pay by Master Card]

Immediate Context: About paying, yes, I know you'll pay somehow, but....
(= presupposition)
Utterance: [How] would you like [to pay]?

Figure 2: Information structure of utterances. The underlines denote old information, and the boxes denote new information.

One topic tag is assigned to each utterance. If the utterance contains multiple items of new information, we go up in the topic tree (p. 9), and a tag which subsumes the different subtopics is used.

(15) (a) *Tell me [the type] and [expiration date of your card]*
CardType* and *CardExp* ⇒ *Card*

(b) *I'd like to stay [for two nights] [on August 10th]*
StayLength* and *ArrivalDate* ⇒ *StayingTime*

In cases where no special lexical item realises the NewInfo, the tag is assigned on the basis of the context as the previous tag. This kind of utterances occur only after the partner's suggestion, explanation or confirmation request when the speaker continues the topic by accepting (or rejecting) it, and they have a fixed the form like *That's correct; Yes please; That sounds nice; I guess I have no choice.*

Utterances like those in (16), control the dialogue flow in terms of time management requests or conventionalised dialogue acts (feedback-acknowledgements, openings, closings, thanks, etc.), and they have their topics classified as *IAM*, InterAction Management topics. These utterances do not request or provide information about the domain, but rather deal with the dialogue on a meta-level. More than one third of the utterances fall in this class.

(16) (a) *Could you wait for a moment while I check?* *iam* (request to waiting)
(b) *Sorry to keep you waiting.* *iam* (apology and renewed contact)
(c) *Okay.* *iam* (acknowledging what the partner said)
(d) *We will be looking forward to your arrival.* *iam* (closing)
(e) *Thank you for calling the New Washington Hotel.* *iam* (thanking)

There are also utterances which explicitly convey information about the speaker's attitudes, preferences and abilities like those in (17). For these, NewInfo is the actual attitude that the speaker wants to convey. However, as discussed in relation to the utterances (7) in Section 2.1, all utterances carry attitudinal information, and even if this is explicitly expressed, subsequent utterances usually refer to the task-related information locus. Since the speaker's attitudes are modelled by dialogue acts and the information content by topics, the topic tag is assigned to the utterances (17) on the basis of the factual content of the particular attitude.

(17) (a) *I don't care how much it costs as long as the room is on the second or third floor.* *RoomPrice*
(b) *I think we can arrange that.* *RoomLoc*
(c) *We will have the room available for you.* *Room*

The main problem in topic tagging is the level of specificity: how fine do tag-distinctions have to be made to be useful? For instance, talking about a reasonable room price, in six dialogues the customer first introduces the limit by talking about her budget (18). These topics resemble side-sequences brought in as explanations or reasons rather than the main threads of the domain, and the topic *RoomPrice* is generalised to subsume these topics as well.

(18) (a) *We're planning a budget for fifty dollars per person* *RoomPrice*
(b) *That's a little over my budget.* *RoomPrice*

The distinction between single unique topics and topics which can be subsumed by a more general class is not clear, however. The principle that was followed in this study was to classify topics which do not have a direct relation to the main task as unique topics. Examples of such topics are given in (19), and in the corpus, there were 71 such utterances (1.7 %).

(19) (a) *My business is taking longer than I expected.*
(b) *And if it isn't too much trouble, could you please tell me how safe that area is?*
(c) *My wife will be relieved to hear that.*

Altogether 62 topic tags were used to tag the corpus of 80 dialogues. Since the original number of topic tags (62) is too big to be used in successful statistical testing, the topic tree was pruned so that only nine topmost nodes are taken into account, and the subtopics are merged into approriate mother topics. The pruned tag set and the distribution of individual tags (%) is given in Table 2.

Based on this seminal work, topic tagging was extended and refined in the research done within the ATR SLDB Tagging Group on tagging 210 English dialogues with dialogue act and topic tags [20].

4 Topic trees

Topic trees provide a "top-down" approach to dialogue coherence by offering a means to constrain information flow in the dialogues: the branches describe what sort of shifts are cognitively easy to process and thus likely to occur in dialogues. For instance, focusing on the action "make a reservation" highlights what the speaker knows about reservations, and she is likely to move on to discuss the room that she wants to reserve or the dates and length of her stay. Originally, "focus trees" were suggested by [21] to enable more flexible focussing management than a stack, and [22] e.g. use topic trees to structure domain knowledge in order to provide information for their text planner. Partial topic trees are shown in Figure 3.

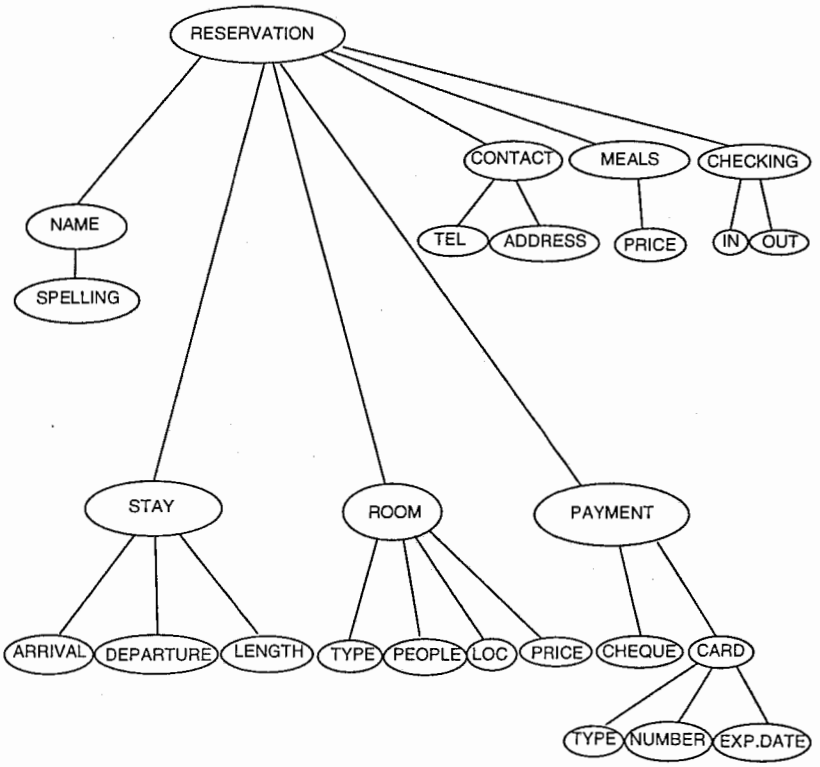

Figure 3: A partial topic tree for room reservation domain.

Table 2: Topic tags for the experiment.

tag	count	precentage	interpretation
iam	1743	41.3 %	Interaction Management (no task topic)
room	824	19.5 %	Room and its properties, Availability
stay	332	7.9 %	Staying period and length
name	320	7.6 %	Name and its spelling
res	310	7.3 %	Making/changing/extending/canceling Reservation
paym	250	5.9 %	Method of Payment
contact	237	5.6 %	Contact Information
meals	135	3.2 %	Meals (breakfast, dinner)
mix	71	1.7 %	Mix (single unique topics)

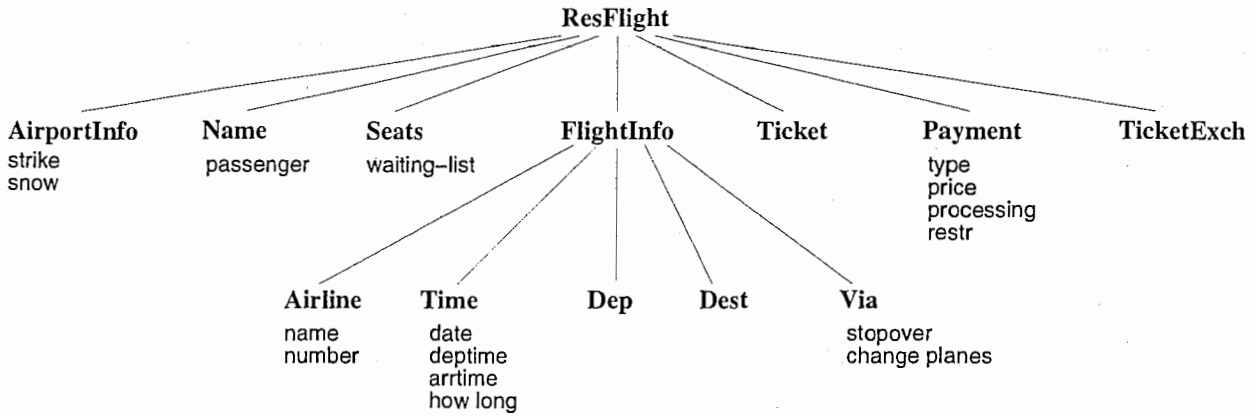

Figure 4: A partial topic tree for flight reservation domain.

Possible traversals of the tree describe possible thematic structures of the domain, i.e. likely sequences of topic tags in the dialogues. The tree can be traversed in whatever order, but in practise some of the transitions are favoured: likely topic shifts correspond to shifts from a node to its daughters or sisters, while shifts to nodes in separate branches are less likely to occur.[5] For instance, after *RoomLoc* it is unlikely that the topic *Card* occurs if the topic *RoomPrice* has not yet been discussed: *RoomLoc* and *Card* are not sisters or subtopics of each other, and switching attention back and forth between them would make their processing difficult.[6] On the other hand, once a topic and its subtopics have been exhaustively discussed, they are not likely to re-occur in the dialogue. For instance, if all necessary information about the room is obtained, the probability of the *Room*-topic drops close to zero. Towards the end of the dialogue it is thus not very likely that topics (or: words related to the topic types) which are already closed will be found, unless the speaker explicitly opens them for confirmation (*I'd like to confirm....*)

At the moment, the topic trees are manually built from the corpus, but on-going research is aimed at making the tagging automatic. Domain models can be constructed using conceptual clustering techniques [23] or word classification [24], and research has also been done on keyword-based topic identification [25].

5 Topic model and spoken language systems

The topic model consists of the following parts:

1. domain knowledge structured into a topic tree
2. prior probabilities of different topic shifts
3. topic vectors describing the mutual information between words and topic types
4. Predict-Support algorithm to measure similarity between the predicted topics and the topics supported by the input utterance.

To test the feasibility and accuracy for the proposed topic model for spoken language applications, different experiments were conducted. The experiments and their results are reported in detail in different conference papers, and this is just an overview of the results.

[5] We will continue talking about a topic *tree*, although in statistical modelling, the tree becomes a topic *network* where the shift probability between nodes which are not daughters or sisters of each other is close to zero.

[6] Awkward shifts are usually syntactically marked: afterthoughts and jumps to distant topics are accompanied by syntactic markers such as *By the way; oh, forgot to say; just to confirm.....*

5.1 The Predict-Support Algorithm

The Predict-Support Algorithm was developed to test the topic model and its applicability to recognize dialogue topics [26]. Topics are assigned to utterances given the previous topic sequence (information about topic shifts: what has been talked about) and the words that carry new information (information about words and topic types: what is actually said). The Predict-Support Algorithm goes as follows:

1. Prediction: get the set of likely next topics in regard to the previous topic sequences using the topic shift model.

2. Support: link each NewInfo word w_j of the input to the possible topics types by retrieving its topic vector. For each topic type t_i, add up the amounts of mutual information $mi(w_j; t_i)$ by which it is supported by the words w_j, and rank the topic types in the descending order of mutual information.

3. Selection:

   (a) Default: From the set of predicted topics, select the most supported topic as the current topic.

   (b) What-is-said heuristics: If the predicted topics do not include the supported topic, rely on what is said, and select the most supported topic as the current topic (cf. the Jumping Context approach in [27]).

   (c) What-is-talked-about heuristics: If the words do not support any topic (e.g. all the words are unknown or out-of-domain), rely on what is predicted and select the most likely topic as the current topic.

Figure 5 shows schematically how the algorithm works.

Figure 5: Scheme of the Predict-Support Algorithm.

On the basis of the tagged dialogue corpus, probabilities of different topic shifts were estimated using the Carnegie Mellon Statistical Language Modeling (CMU SLM) Toolkit [28]. This builds a trigram model where smoothing is done via the backoff method [29]: if the string on one level does not occur, its probability is assigned by backing off to the next lower level and estimating the probability there, i.e. if the trigrams do not occur, the probability is calculated on the basis of bigrams, then on unigrams, multiplied by some estimation of their relative weight. The conditional probabilities are calculated as follows:

$$p(w3|w1,w2) = \begin{cases} p3(w1,w2,w3) & \text{if trigram exists} \\ bo\_wt2(w1,w2) \times p(w3|w2) & \text{if bigram (w1,w2) exists} \\ p(w3|w2) & \text{otherwise.} \end{cases}$$

$$p(w2|w1) = \begin{cases} p2(w1,w2) & \text{if bigram exists} \\ bo\_wt1(w1) \times p1(w2) & \text{otherwise.} \end{cases}$$

To estimate how well different words support the different topic types, *mutual information* between each word and the topic types is calculated. Mutual information describes how much information a word w gives about a topic type t, and is calculated as follows (ln is log base two, $p(t|w)$ the conditional probability of t given w, and $p(t)$ the probability of t):

$$I(w,t) = \ln \frac{p(w,t)}{p(w) \cdot p(t)} = \ln \frac{p(t|w)}{p(t)}$$

Each word is associated with a *topic vector*, which describes how much information the word w carries about each possible topic type t_i:

$$topvector(mi(w,t_1), mi(w,t_2), ..., mi(w,t_n))$$

The Predict-Support algorithm was tested using cross-validation on the corpus with the pruned topic types. 70 randomly picked dialogues were used for training, and the other 10 dialogues for testing in each test cycle (each test file contained about 400-500 test utterances). The accuracy results of the first predictions are given in Table 3. PP is the corpus perplexity which represents the average branching factor of the corpus, or the number of alternatives from which to choose the correct label at a given point. The average accuracy rate, 78.68 % is a satisfactory result. Another set of cross-validation tests were conducted using 75 dialogues for training and 5 dialogues for testing, and as expected, a bigger training corpus gives better recognition results when perplexity stays the same. Finally, a cross-validation was done using the whole set of topic tags

Table 3: Accuracy results of the first predictions.

Test type	PP	PS-algorithm	BO model
Topics = 10 train = 70 files	3.82	78.75	41.30
Topics = 10 train = 75 files	3.74	80.55	40.33
Topics = 62 train = 70 files	5.59	64.96	41.32

Table 4: Accuracy and precision results for different topic types.

Topic type	Correct	Recognised	Hit	Accuracy (Hits/Correct)	Precision (Hits/Recognised)
iam	1951	2267	1787	91.594	78.827
contact	259	395	221	85.328	55.949
name	358	330	290	81.006	87.879
mix	92	138	72	78.261	52.174
paym	263	228	198	75.285	86.842
stay	361	288	251	69.529	87.153
res	342	348	217	63.450	62.356
room	926	606	580	62.635	95.710
meals	159	111	94	59.119	84.685
Average	4711	4711	3710	78.752	78.752

(62), and it's interesting to notice that the results do not drop as drastically as one would expect, given the huge number of tags. This is probably due to the compensatorial effect of support part of the algorithm. The results are compared to the backoff model that only relies on information about probable topic sequences.

The average accuracy and precision of the Predict-Support algorithm is also calculated for each topic type. While accuracy is the ratio of correctly assigned tags to the total number of tags, precision is the ratio of correctly assigned tags to the total number of assigned tags. The results for cross-validation tests where 70 files are used for training and 10 files for testing are given in Table 4 .

The average accuracy and precision for all topic types is of course the same (78.752%), but varies a lot between different topic types. The IAM topic has the highest accuracy but its precision rate is only average, i.e. almost all IAM topics are recognised but this tag was also assigned to some utterances incorrectly. The situation is opposite for the ROOM topic which has the highest precision 95.71% while its accuracy is rather poor. In other words, only about two thirds of the topics are recognised but almost all that are recognised are also correct.

The results of the topic recognition show that the model performs well. Although the rates are somewhat optimistic as the calculations use transcribed dialogues (= the correct recognizer output), it is still safe to conclude that topic information provides a promising starting point in attempts to provide an accurate context for the spoken dialogue systems.

5.2 Speech Recognition

The quality of statistical models can be measured by test set perplexity. The goal is to find models which estimate low perplexity: the direct interpretation of the perplexity in our case is the number of words among which the interpretation of the next word must be chosen.

To study the effect of topic information on speech recognition, trigram perplexities were calculated[7] [30].

In speech recognition, given the utterance $W_1^n = w_1, ..., w_n$ to be recognized, the likelihood of the word string is maximized by maximizing the probability of each word w_i in the context in which it occurs. In the trigram model, the context for a word contains its two previous words, and the conditional probability of a word w_i given the two previous words w_{i-1} and w_{i-2} is calculated as:

$$P(w_i|w_{i-2}w_{i-1}) = \frac{Occ(w_{i-2}w_{i-1}w_i)}{\sum_{w_x} Occ(w_{i-2}w_{i-1}w_x)}$$

The test perplexity was calculated using the normalized formula:

[7] I are grateful to Sabine Deligne (ATR-ITL, Dept 1) for useful discussions on speech recognition and also for doing the trigram perplexity measures using the Carnegie Mellon Statistical Language Modeling (CMU SLM) Toolkit.

Table 5: Trigram perplexity in the experiments.

	General model	Topic-dependent model	
		Random topics	Manually tagged
known words	12.77	23.93	10.31
open word set	14.81	33.71	12.72

$$PP = 2^{-\frac{1}{n}\log L(W_1^n)}$$

where n is the number of words in the utterance $W_1^n = w_1, ..., w_n$ and L is the likelihood of the word string.

Trigram perplexity results are shown in Table 5.

Compared to the general model trained on all dialogues, perplexity decreases by 20 % for a topic-dependent model where topics have been (manually) tagged, and by 14 % if we use an open test with unknown words included as well. Since any consistent classification is likely to improve quality of statistical models, we need to conduct further experiments on automatically tagged topic corpora. However, the results show that a topic model based on linguistically motivated classification provides a good starting point, and at least for the amount of topic-dependent data we used for each topic, it is useful to specialise the language model for speech recognition depending on the topic.

6 Future Work

This paper reports on-going work on topic modelling. The preliminary results of the proposed model show that the Predict-Support algorithm performs well, and the use of topic information in speech recognition is promising when the quality of the model is measured in terms of trigram perplexity.

However, many questions are also left open. First, future work would need to specify the effect and cost of extending the domain to new tasks. The work within the ATR SLDB Tagging Group shows that it is necessary to increase the set of topic tags with the missing tags when the work is extended to new domains. This is a general problem with domain related tags: new domains usually require extension of the tag set. Although limited models are not hard to build, we want to explore find methods for the automatic construction of the tree. Some work has already been done with the decision trees (H. Tanaka), and further ideas can be gotten from conceptual clustering in general. Also, ways to automatize the question method and to automatically recognize the NewInfo in utterances is needed. Something in the line of [18] to distinguish new and old information in terms of the pivot of the utterance (usually the main verb) is promising in this respect.

Furthermore, the relation between topics and dialogue acts need to studied. The two sources of discourse information are independent but also interrelated: the question and its answer have the same topic, but several inform-type utterances can only be recognized on the basis of their content. Future research will also answer the question of how the dialogue model's accuracy can be improved when the two sources are combined.

Acknowledgements

I am grateful to all the people at ATR for providing an inspiring environment in which to conduct this research. I'd especially like to thank all the former and present members of ITL for their help and useful discussions on spoken dialogues, topics and topic modelling for speech recognition.

References

[1] N. Katoh and T. Morimoto. Statistical method of recognizing local cohesion in spoken dialogues. In *Proceedings of the 16th COLING*, pp. 634–639, 1996.

[2] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12, No. 3, pp. 175–204, 1986.

[3] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, 1995.

[4] K. Jokinen. *Response Planning in Information-Seeking Dialogues.* PhD thesis, University of Manchester Institute of Science and Technology, 1994.

[5] E. Vallduví and E. Engdahl. The linguistic realization of information packaging. *Linguistics*, Vol. 34, pp. 459–519, 1996.

[6] J. K. Gundel. Shared knowledge and topicality. *Journal of Pragmatics*, Vol. 9, pp. 83–107, 1985.

[7] M. Rats. *Topic Management in Information Dialogues.* ITK Dissertation Series, Katholieke Universiteit Brabant, Tilburg, 1996.

[8] M. A. K. Halliday. Notes on transitivity and theme in English. *Journal of Linguistics*, Vol. 3, No. 2, pp. 199–244, 1967.

[9] M. Vilkuna. *Free Word Order in Finnish. Its Syntax and Discourse Functions.* Suomalaisen Kirjallisuuden Seura, Helsinki, 1989.

[10] T. Givón. Iconicity, isomorphism and nonarbitrary coding in syntax. In J. Haiman, editor, *Iconicity in Syntax*, pp. 187–219. John Benjamins,Amsterdam, 1985.

[11] S. Kuno. Functional sentence perspective. *Linguistic Inquiry*, Vol. 3, pp. 269–320, 1972.

[12] H. H. Clark and S. E. Haviland. Comprehension and the given-new contract. In R. O. Freedle, editor, *Discourse Production and Comprehension, Vol.1.* Ablex, 1977.

[13] E. Prince. On the given/new distinction. *CLS*, Vol. 15, , 1979.

[14] L. Carlson. *Dialogue Games.* D. Reidel Publishing Company, Dordrecht, 1983.

[15] K. Jokinen. Coherence and cooperation in dialogue management. In K. Jokinen, editor, *Pragmatics in Dialogue Management*, pp. 97–111. Proceedings of The XIVth Scandinavian Conference of Linguistics, University of Göteborg, Göteborg, 1994. Gothenburg Papers in Theoretical Linguistics 71.

[16] M. Seligman, L. Fais, and M. Tomokiyo. A bilingual set of communicative act labels for spontaneous dialogues. Technical Report ATR Technical Report TR-IT-81, ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, 1994.

[17] M. Meteer and R. Iyer. Modeling conversational speech for speech recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

[18] K. W. Ma, G. Zavaliagkos, and M. Meteer. Sub-sentence discourse models for conversational speech recognition. In *ICASSP'98*, pp. 693–696, 1998.

[19] E. Black, S. Eubank, H. Kashioka, D. Magerman, R. Garside, and G. Leech. Beyond skeleton parsing: Producing a comprehensive large-scale general-english treebank with full grammatical analysis. In *Proceedings of the 16th COLING*, pp. 107–112, 1996.

[20] K. Jokinen, H. Tanaka, and H. Iwamoto. Manual for tagging the sldb english dialogues with speech acts and topics. Technical Report TR-IT-, ATR, 1999.

[21] K. McCoy and J. Cheng. Focus of attention: Constraining what can be said next. In C. L. Paris, W. R. Swartout, and W. C. Moore, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pp. 103–124. Kluwer Academic Publishers, Norwell, Massachusetts, 1991.

[22] D. Carcagno and Lidija Iordanskaja. Content determination and text structuring: two interrelated processes. In H. Horacek and M. Zock, editors, *New Concepts in Natural Language Generation*, pp. 10–26. Pinter Publishers, London, 1993.

[23] D. H. Fisher and P. Langley. Approaches to conceptual clustering. In *Proceedings of the IJCAI-85*, pp. 691–697, 1985.

[24] J. G. McMahon and F. J. Smith. Improving statistical language model performance with automatically generated word hierarchies. *Computational Linguistics*, Vol. 22:2, pp. 217–247, 1996.

[25] P. Garner. On topic identification and dialogue move recognition. *Computer Speech and Language*, Vol. 11, pp. 275–306, 1997.

[26] K. Jokinen, H. Tanaka, and A. Yokoo. Context management with topics for spoken dialogue system. In *Proceedings of COLING-ACL'98*, pp. 631–637, 1998.

[27] Y. Qu, B. Di Eugenio, A. Lavie, L. Levin, and C. P. Rosè. Minimizing cumulative error in discourse context. In *Dialogue Processing in Spoken Dialogue Systems*, pp. 60–64. Proceedings of the ECAI'96 Workshop, Budapest, Hungary, 1996.

[28] P. Clarkson and R. Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Eurospeech-97*, pp. 2707–2710, 1997.

[29] S. M. Katz. Estimation of probabilities for sparse data for the language model component of a speech recognizer. In *I.E.E.E. Transactions on Acoustics, Speech and Signal Processing*, Vol. 3, pp. 400–401. ASSP-35, 1987. (in Japanese).

[30] K. Jokinen and T. Morimoto. Topic information and spoken dialogue systems. In *NLPRS-97*, pp. 429–434. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, 1997.