Topics and Spoken Language Recognition

Kristiina Jokinen, Hideki Tanaka and Akio Yokoo
ATR Interpreting Telecommunications Research Laboratories

1 Introduction

This paper concerns the use of discourse information in spoken dialogue systems, and argues that knowledge of the domain, modelled with the help of topic trees, is important in improving the recognition accuracy of spoken utterances.

The use of discourse information in spoken language systems is usually motivated by arguments concerning the system's *robustness*: (1) support for speech recognition and (2) providing necessary context for higher level dialogue management. Previous research has taken discourse information into account mainly in the form of speech acts [8, 10, 7], which seem to provide a suitable basis for dialogue models: they have been widely studied, and dealing with the speaker's intentions, they can be considered domain-independent. However, speakers do not only produce intentionally linked utterances: they also aim at maintaining thematic coherence. Thus knowledge of the information focus of the utterances is important too. In practical applications (information providing systems, speech-to-speech translation, etc), analysis of the semantic content of the utterances, even if the aim is a shallow analysis only, is required, and knowledge of what is talked about can be used for disambiguation and as an approximation of the content.

Recently interest has thus shifted to the information content of the utterances. In the JANUS-system [5], translation is guided by Semantic Dialogue Units, semantically coherent pieces of information, while [9] report that the Jumping Context approach, which uses syntactic and semantic information of the input utterance, gives consistently better accuracy results than approaches relying strictly on act sequences. [4] present preliminary results of using topic information in spoken language systems.

Our goal is to to combine the speaker's intentions with the semantic focus of the utterance and study how speech recognition can be improved by taking the information content of the utterances into account. The most pertinent task in this is the recognition of dialogue topics, which encode what is being talked about in the dialogue. In the rest of the paper we describe a topic model for spoken dialogue systems and present results of assigning topics to utterances using the Predict-Support Algorithm.

2 The Topic Model

In AI-based dialogue modelling, topics are associated with a particular discourse entity, *focus*, which is currently in the centre of attention and which the participants want to focus their actions on [3]. The topic (focus) is mainly used in anaphora resolution and constraining search space. Our goal is to predict likely content of the next utterance, and thus we are more interested in the topic *types* that describe the information conveyed by utterances than the actual topic entity. Consequently, instead of tracing salient entities in the dialogue and providing heuristics for different shifts of attention, we seek a formalisation of the *information structure* of utterances in terms of the new information exchanged in the course of the dialogue. Furthermore, we rely on observed facts (= word tokens) and use statistical methods instead of elaborated reasoning about plans and world knowledge. Our topic model consists of the following:

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.

2.1 Topic trees

Originally "focus trees" were proposed by [6] to trace foci in NL generation systems. The tree is a subgraph of the world knowledge, and it both constrains and enables prediction of what is likely to be talked about next: random jumps from one branch to another are not very likely, and if they do, they should be appropriately marked.

Our topic tree is an organisation of the domain knowledge in terms of topic types, cf. [1]. The nodes of the tree[1] correspond to topic types which represent clusters of the words expected to occur at a particular point of the dialogue. For our experiments, topic trees were hand-coded from the dialogue corpus, but an automatic clustering program has been developed [11]. Figure 1 shows a partial topic tree in the hotel reservation domain.

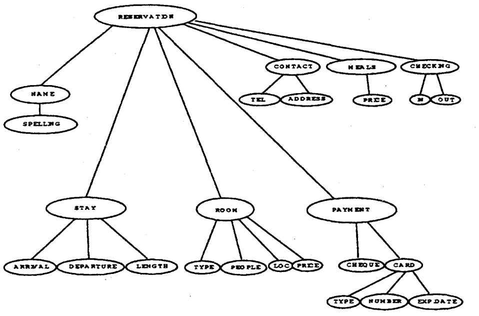

Figure 1: A partial topic tree.

Our corpus contains 80 dialogues from the bilingual ATR Spoken Language Dialogue Database. The dialogues deal with hotel reservation and tourist information, and the total number of utterances is 4228. Segmentation is based on the information structure so that one utterance contains only one piece of new information (more of the segmentation and tagging is reported in [4]).

The number of topic types in the corpus is 62. The types were pruned on the basis of the topic tree, so that only the topmost nodes were taken into account and the subtopics merged into appropriate mother topics. Figure 2 lists the pruned topic types and their frequencies in the corpus.

The special topic type IAM is assigned to utterances which deal with InterAction Management: short fixed phrases (*Let me confirm; thank you; good-bye; ok; yes*) and temporizers (*well, ah, uhm*) which do not request or provide information about the domain, but control the overall flow of the dialogue in terms of time management requests or conventionalised dialogue acts (feedback-acknowledgements, thanks, greetings, closings, etc.).

tag	count	%	interpretation
iam	1747	41.3	Interaction Management
room	826	19.5	Room, its properties
stay	332	7.9	Staying period
name	320	7.6	Name, spelling
res	310	7.3	Make/change/extend/ cancel reservation
paym	250	5.9	Payment method
contact	237	5.6	Contact Info
meals	135	3.2	Meals (breakfast,dinner)
mix	71	1.7	Single unique topics

Figure 2: Topic tags for the experiment.

The topic type MIX is reserved for utterances which contain information not directly related to the domain (safety of the downtown area, business taking longer than expected, a friend coming for a visit etc.), thus marking out-of-domain utterances.

2.2 Topic shifts

Probabilities of different topic shifts were estimated by a trigram backoff model [2], where conditional probabilities are calculated as follows:

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

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

2.3 Topic vectors

Each word may support several topics. E.g. the occurrence of *room* in *I'd like to make a room reservation.* supports the topic MAKERESERVATION, but in *We have only twin rooms available on the 15th.* it supports the topic ROOM. To estimate how well the words supports the different topic types, we measured *mutual information* between each word w and the topic type t (ln is log base two, $p(t|w)$ the conditional probability of t given w, and $p(t)$ the probability of t):

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

[1] We will continue talking about a topic *tree*, although in statistical modelling, the tree becomes a topic *network*.

Each word is then associated with a *topic vector*, which describes how much information the word carries about each possible topic type. E.g. the topic vector of the word *room* is:

```
topvector(room,[mi(0.21409750769169117,contact),
                mi(-5.5258041314543815,iam),
                mi(-3.831955835588453,meals),
                mi(0,mix),
                mi(-1.26971341136738,name),
                mi(-2.720924523199709,paym),
                mi(0.9687353561881407,res),
                mi(1.9035899442740105,room),
                mi(-4.130179669884547,stay)]).
```

The word supports the topics ROOM and MAKE-RESERVATION (res), but gives no information about MIX (out-of-domain) topics, and its presence is highly indicative that the utterance is not at least IAM or STAY. It also supports CONTACT because the corpus contains utterances like *we can reach you at the Hotel New Tokyo room 803* which give information about how to contact the customer who is staying at a hotel.

2.4 The Predict-Support Algorithm

Topics are assigned to utterances given the previous topic sequence (what has been talked about) and the words carrying new information (what is said). The Predict-Support Algorithm goes as follows (schematically presented in Figure 3):

1. Prediction: get the set of likely next topics in regard to the previous topic sequence using the topic shift model.
2. Support: link each NewInfo word w_j to the possible topics types by retrieving its topic vector. For each topic type t_i, add up the amounts of mutual information $mi(w_j; t_i)$ with which the type is supported by the words w_j, and rank the topics in the descending order of mutual information.
3. Selection:
   (a) Default: From the set of predicted topics, select the most supported one as the current topic.
   (b) What-is-said heuristics: If the predicted topics do not include the supported topics, rely on what is said and select the most supported topic as the current topic (cf. the Jumping Context method in [9]).
   (c) What-is-talked-about heuristics: If the input words do not support any topic (unknown or out-of-domain words), rely on what is predicted and select the most likely topic as the current topic.

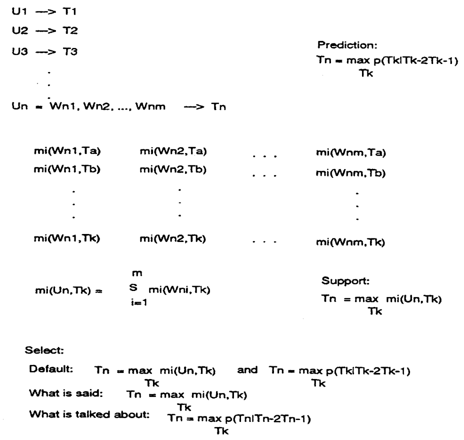
U1 —> T1
U2 —> T2
U3 —> T3
Prediction:
Tn = max p(Tk|Tk-2Tk-1)
Tk
Un = Wn1, Wn2, ..., Wnm —> Tn
mi(Wn1,Ta)
mi(Wn1,Tb)
mi(Wn1,Tk)
mi(Wn2,Ta)
mi(Wn2,Tb)
mi(Wn2,Tk)
. . .
mi(Wnm,Ta)
mi(Wnm,Tb)
mi(Wnm,Tk)
mi(Un,Tk) = S mi(Wni,Tk) m i=1
Support:
Tn = max mi(Un,Tk)
Tk
Select:
Default: Tn = max mi(Un,Tk) Tk and Tn = max p(Tk|Tk-2Tk-1) Tk
What is said: Tn = max mi(Un,Tk) Tk
What is talked about: Tn = max p(Tn|Tn-2Tn-1) Tk

Figure 3: The Predict-Support Algorithm.

3 Experiments

The Predict-Support Algorithm is tested using cross-validation on our corpus. The accuracy results of the best predictions are given in Table 4. 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.

In each test cycle we reserved 10 randomly picked dialogues for testing (about 400-500 test utterances), and used the other 70 dialogues for training. For the pruned topic types the average accuracy rate, 78.68 % is a satisfactory result. Another set of cross-validation tests using 75 dialogues for training and 5 dialogues for testing was also conducted, and as expected, a bigger training corpus gives better recognition results when perplexity stays the same.

Test type	PP	PS-algorithm	BO model

Topics = 10
train = 70 files
	3.82	78.68	41.30

Topics = 10
train = 75 files
	3.74	80.55	40.33

Topics = 62
train = 70 files
	5.59	64.96	41.32

Figure 4: Accuracy results of the best predictions.

To estimate the effect of a bigger number of topic tags, we conducted cross-validation tests with the original 62 topic types. A finer set does worsen the accuracy, but not as much as we expected: the Support-part of the algorithm effectively remedies prediction inaccuracies.

As the lower-bound experiments we conducted cross-validation tests using the trigram backoff-model, i.e. relying only on the context which records the history of topic types. For the best ranked predictions the accuracy rate is about 40%.

The rates are somewhat optimistic as we used transcribed dialogues (= the correct recognizer output), but we can still safely conclude that topic information provides a promising starting point in attempts to provide an accurate context for the spoken dialogue systems. This can be further verified in the perplexity measures for *word* recognition: compared to a general language model trained on non-tagged dialogues, perplexity decreases by 20 % for a language model trained on topic-dependent dialogues, and by 14 % if unknown words are included in the test [4].

4 Conclusions

The paper has presented a probabilistic topic model to be used as a context model for spoken dialogue systems. The model combines both top-down and bottom-up approaches to topic modelling: the topic tree, which structures domain knowledge, provides expectations of likely topic shifts, whereas the information structure of the utterances is linked to the topic types via topic vectors which describe mutual information between the words and topic types. The Predict-Support Algorithm assigns topics to utterances, and achieves an accuracy rate of 78.68 %.

Research on statistical topic modelling and combining topic information with spoken language systems is still new and there are several areas for future work. One is the coverage of topic trees. Topic trees can be generalised in regard to world knowledge, but this requires deep analysis of the utterance meaning, and an inference mechanism to reason about conceptual relations. We will explore ways to extract semantic categories from the parse trees and integrate them with the topic knowledge. Another research issue is the relation between topics and speech acts. We will investigate their respective roles in context management for spoken dialogue systems. Finally, statistical modelling is prone to sparse data problems, and we need to consider ways to overcome inaccuracies in calculating mutual information.

References

[1] 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.

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

[3] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

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

[5] A. Lavie, D. Gates, N. Coccaro, and L. Levin. Input segmentation of spontaneous speech in JANUS: a speech-to-speech translation system. In *Dialogue Processing in Spoken Dialogue Systems*, pp. 54–59. Proceedings of the ECAI'96 Workshop, Budapest, Hungary, 1996.

[6] 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 (Eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pp. 103–124. Kluwer Academic Publishers, Dordrecht, 1991.

[7] J-U. Möller. Using DIA-MOLE for unsupervised learning of domain specific dialogue acts from spontaneous language. Technical Report FBI-HH-B-191/96, University of Hamburg, 1996.

[8] M. Nagata and T. Morimoto. An information-theoretic model of discourse for next utterance type prediction. In *Transactions of Information Processing Society of Japan*, volume 35:6, pp. 1050–1061. 1994.

[9] 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.

[10] N. Reithinger and E. Maier. Utilizing statistical dialogue act processing in Verbmobil. In *Proceedings of the 33rd Annual Meeting of the ACL*, pp. 116–121, 1995.

[11] H. Tanaka, K. Jokinen, and A. Yokoo. Automatic dialogue domain profiling: towards efficient topic tagging environment. In *Proceedings of the 4th Annual Meeting of The Association for Natural Language Processing*. University of Kyushu. 1998.