SIGDIAL'02, 3rd SIGdial workshop on Discourse and Dialogue Management.
July 11-12, Philadelphia, USA, 2002. To appear.

Topic Identification In Natural Language Dialogues Using Neural Networks

Krista Lagus and Jukka Kuusisto
Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 9800, FIN-02015 HUT, Finland
krista.lagus@hut.fi

Abstract

In human–computer interaction systems using natural language, the recognition of the topic from user's utterances is an important task. We examine two different perspectives to the problem of topic analysis needed for carrying out a successful dialogue. First, we apply self-organized document maps for modeling the broader subject of discourse based on the occurrence of content words in the dialogue context. On a Finnish corpus of 57 dialogues the method is shown to work well for recognizing subjects of longer dialogue segments, whereas for individual utterances the subject recognition history should perhaps be taken into account. Second, we attempt to identify topically relevant words in the utterances and thus locate the old information ('topic words') and new information ('focus words'). For this we define a probabilistic model and compare different methods for model parameter estimation on a corpus of 189 dialogues. Moreover, the utilization of information regarding the position of the word in the utterance is found to improve the results.

1 Introduction

The analysis of the topic of a sentence or a document is an important task for many natural language applications. For example, in interactive dialogue systems that attempt to carry out and answer requests made by customers, the response strategy employed may depend on the topic of the request (Jokinen et al., 2002). In large vocabulary speech recognition knowledge of the topic can, in general, be utilized for adjusting the language model used (see, e.g., (Iyer and Ostendorf, 1999)).

We describe two approaches to analyzing the topical information, namely the use of *topically ordered document maps* for analyzing the overall topic of dialogue segments, and *identification of topic and focus words in an utterance* for sentence-level analysis and identification of topically relevant specific information in short contexts.

1.1 Document map as a topically ordered semantic space

The Self-Organizing Map (Kohonen, 1982; Kohonen, 1995) is an unsupervised neural network method suitable for ordering and visualization of complex data sets. It has been shown that very large document collections can be meaningfully organized onto maps that are topically ordered: documents with similar content are found near each other on the map (Lin, 1992; Honkela et al., 1996; Lin, 1997; Kohonen et al., 2000).

The document map can be considered to form an ordered representation of possible topics, i.e., a topical semantic space. Each set of map coordinates specifies a point in the semantic space, and additionally, corresponds to a subset of the corpus, forming a kind of associative topical-semantic memory.

Document maps have been found useful in text mining and in improving information re-

trieval (Lagus, 2000). Recent experiments indicate that the document maps ordered using the SOM algorithm can be useful in focusing the language model to the current active vocabulary (Kurimo and Lagus, 2002).

In this article we examine the usefulness of document maps for analyzing the topics of transcripts of natural spoken dialogues. The topic identification from both individual utterances and longer segments is studied.

1.2 Conceptual analysis of individual utterances

Within a single utterance or sentence the speaker may provide several details that specify the request further or provide additional information that specifies something said earlier. Automatic extraction of the relevant words and the concepts they relate to may be useful, e.g., for a system filling out the fields of a database query intended to answer the user's request.

If a small set of relevant semantic concepts can be defined, and if the sentence structures allowed are strictly limited, the semantic concept identification problem can be solved, at least to some degree, by manually designed rule-based systems (Jokinen et al., 2002).

However, if the goal is the analysis of free-form dialogues, one cannot count on hearing full sentences. It is therefore important to try to formulate the task as a learning problem into which adaptive, statistical methods can be applied.

The major challenge in adaptive language modeling is the complexity of the learning problem, caused by large vocabularies and large amount of variation in sentence structures, compared to the amount of learning data available. For English there already exist various tagged and analyzed corpora. In contrast, for many smaller languages no tagged corpora generally exist. Yet the methods that are developed for English cannot as such be applied for many other languages, such as Finnish.

In the analysis of natural language dialogues, theories of information structure (Sgall et al., 1986; Halliday, 1985) concern the semantic concepts and their structural properties within an utterance. Such concepts include the attitudes, prior knowledge, beliefs and intentions of the speaker, as well as concepts identifying information that is shared between the speakers. The terms 'topic' and 'focus' may be defined as follows: 'topic' is the general subject of which the user is talking about, and 'focus' refers to the specific additional information that the user now introduces about the topic. An alternative way of describing these terms is that 'topic' constitutes of the *old* information shared by both dialogue participants and 'focus' contains the *new* information which is communicated regarding the topic.

A traditional way of finding the old and new information is the 'question test' (see (Vilkuna, 1989) about using it for Finnish). For any declarative sentence, a question is composed so that the sentence would be a natural answer to that question. Then the items of the sentence that are repeated in the question belong to the topic and the new items to the focus.

A usual approach for topic–focus identification is to use parsed data. The sentence, or it's semantic or syntactic-semantic representation, is divided into two segments, usually at the location of the main verb, and the words or semantical concepts in the first segment are regarded as 'topic' words/concepts and those in the second as 'focus' words/concepts. For example in (Meteer and Iyer, 1996), the division point is placed before the first strong verb, or, in the absence of such a verb, behind the last weak verb of the sentence. Similar division is also the starting point for the algorithm for topic–focus identification introduced in (Hajiĉová et al., 1995). The initial division is then modified according to the verb's position and meaning, the subject's definiteness or indefiniteness and the number, type and order of the other sentence constituents.

In language modeling for speech recognition improvements in perplexity and word error rate have been observed on English corpora when using language models trained sep-

arately for the topic and the focus part of the sentence (Meteer and Iyer, 1996; Ma et al., 1998). Identification of these concepts is likely to be important also for sentence comprehension and dialogue strategy selection.

In this article we examine the application of a number of statistical approaches for identification of these concepts. In particular, we apply the notions of topic and focus in information structure (Sgall et al., 1986) to tagging a set of natural dialogues in Finnish. We then try several approaches for learning to identify the occurrences of these concepts from new data based on the statistical properties of the old instances.

2 Experiments on recognizing the dialogue topic of a dialogue turn

The ordered document map can be utilized in the analysis of dialogue topics as follows: encode a dialogue turn, i.e., an utterance u (or an utterance combined with its recent history) as a document vector. Locate the best-matching map unit, or several such units. Utilize the identities of the best units as a semantic representation of the topic of the u. In effect, this is a latent semantic representation of the topical content of the utterance. Evaluation of such a latent representation directly amounts to asking whether the dialogue manager can benefit from the representation, and must therefore be carried out by the dialogue manager. This direct evaluation has not yet been done.

Instead, we have utilized the following approach for evaluating the ordering of the maps and the generalization to new, unseen dialogues: An intermediate set of named semantic concepts has been defined in an attempt to approximate what is considered to be interesting for the dialogue manager. The latent semantic representation of the map is then labeled or *calibrated* to reflect these named concepts. In effect, each dialogue segment is categorized to a prior topical category. The organized map is labeled using part of the data ('training data'), and the remaining part is used to evaluate the map ('test data')[1].

Furthermore, a statistical model for document classification can be defined on top of the map. The probability model used for topic estimation is

$$P(A_i|S) = P(X_N|S)P(A_i|X_N), \qquad (1)$$

where A_i is the topic category, S denotes the text transcription of the spoken sentence and X_N is the set of N best map vectors used for the classification. We approximate the probability $P(X_N|S)$ to be equal for each map vector in X_N. We assume that X_N conveys all information about S. The terms $P(A_i|X_N)$ are calculated as the relative frequencies of the topics of the document vectors in the training data that were mapped to the nodes that correspond to X_N.

2.1 Corpus: transcripts of 57 spoken dialogues

The data used in the experiments were Finnish dialogues, recorded from the customer service phone line of Helsinki City Transport. The dialogues, provided by the Interact project (Jokinen et al., 2002), had been transcribed into text by a person listening to the tapes.

The transcribed data is extremely colloquial. Both the customers and the customer service personnel use a lot of expletive words, such as 'nii' ('so', 'yea') and 'tota' ('hum', 'er', 'like'), often the words appear in reduced or otherwise non-standard forms. The word order does not always follow grammatical rules and quite frequently there is considerable overlap between the dialogue turns. For example, the utterance of speaker A may be interjected by a confirmation from speaker B. This had currently been transcribed as three separate utterances: A1 B A2.

2.2 Tagging and segmentation of dialogues

The data set was split into training and test data so that the first 33 dialogues were used for organization and calibration of the map

[1]Note that even in this case the map is ordered in an unsupervised manner, although it is applied for the classification of new instances based on old ones.

Table 1: Proportions of customer utterances in each topic category in the data sets.

	Training data	Test data
Beginnings	0.08	0.11
Endings	0.12	0.14
Timetables	0.49	0.59
Tickets	0.16	0.11
OOD	0.15	0.06

and the 24 dialogues collected later for testing.

A small number of broad topic categories were selected so that they comprehensively encompass the subjects of discussion occurring in the data. The categories were 'timetables', 'beginnings', 'tickets', 'endings', and 'out of domain'.

The dialogues were then manually tagged and segmented, so that each continuous dialogue segment of several utterances that belonged to one general topic category formed a single document. This resulted in a total of 196 segments, 115 and 81 in training and test sets, respectively. Each segment contained data from both the customer and the assistant.

Of particular interest is the analysis of the topics of individual customer utterances. The data was therefore split further into utterances, resulting in 450 and 189 customer utterances in the training and test set, respectively. The relative frequencies of utterances belonging to each topic category for both training and test data are shown in Table 1. Each individual utterance was labeled with the topic category of the segment it belonged to.

2.3 Creation of the document map

The documents, whether segments or utterances, were encoded as vectors using the methods described in detail in (Kohonen et al., 2000). In short, the encoding was as follows. Stopwords (function words etc.) and words that appeared fewer than 2 times in the training data were removed. The remaining words were weighted using their entropy over document classes. The documents were encoded using the vector space model by Salton (Salton et al., 1975) with word weights. Furthermore, sparse random projection of was applied to reduce the dimensionality of the document vectors from the original 1738 to 500 (for details of the method, see, e.g., (Kohonen et al., 2000)).

In organizing the map each longer dialogue segment was considered as a document. The use of longer segments is likely to make the organization of the map more robust. The inclusion of the utterances by the assistant is particularly important given the small amount of data—all information must be utilized. The document vectors were then organized on a SOM of $6 \times 4 = 24$ units.

2.4 Experiments and results

We carried out three tests where the length of dialogue segments was varied. In each case, different values of N were tried. In the first case, longer dialogue segments in the training data were used to estimate the term $P(A_i|X_N)$ whereas recognition accuracy was calculated on customer utterances only. Next, individual customer utterances were used also in estimating the model term. The best recognition accuracy in both cases were obtained using the value $N = 3$, namely 60.3% for the first case and 65.1% for the second case. In the third case we used the longer dialogue segments both for estimating the model and for evaluation, to examine the effect of longer context on the recognition accuracy. The recognition accuracy was now 87.7%, i.e., clearly better for the longer dialogue segments than for the utterances.

It seems that many utterances taken out of context are too short or nondescript to provide reliable cues regarding the topical category. An example of such an utterance is 'Onks sinne mitää muuta?' (lit. 'Is to there anything else?', the intended meaning probably being 'Does any other bus go there?'). In this case it is the surrounding dialogue (or perhaps the Finnish morpheme corresponding to 'to') that would identify the correct category, namely 'timetables'.

Moreover, results on comparing a docu-

ment map to Independent Component Analysis on the same corpus are reported in (Bingham et al., 2002). The slightly higher percentages in that paper are due to evaluating longer segments and to reporting the results on the whole data set instead of a separate test set.

3 Identification of old and new information in utterances

We define this task as the identification of 'topic words' and 'focus words' from utterances of natural Finnish dialogues. There are thus no restrictions regarding the vocabulary or the grammar. By observing previous, marked instances of these concepts we try to recognize the instances in new dialogues. It should be noted that this task definition differs somewhat from those discussed in Section 1.2 in that we do not construct any conceptual representation of the utterances, nor do we segment them into a 'topic' part and a 'focus' part. This choice is due to utilizing natural utterances in which the sentence borders do not always coincide with the turn-taking of the speakers—a turn may consist of several sentences or a partial one (when interrupted by a comment from the other speaker). In other words, we try to identify the central words that communicate the topic and focus in an utterance. We assume that they can appear in any part of the sentence and between them there may be other words that are not relevant to the topic or focus. Whether these central words form a single topic or focus or several such concepts is left open.

3.1 Corpus and tagging

The corpus used includes the same data as in section 2 with additional 133 dialogues collected from the same source. Basically each dialogue turn was treated as an utterance, with the exception that long turns were segmented into sentence-like segments, which were then considered to be utterances[2]. Utterances consisting of only one word were removed from the data. The training data contained 11464 words in 1704 utterances. Of the words 17 % were tagged as topic, and 28 % as focus. The test data consisted of 11750 words in 1415 utterances, with 14 % tagged as topic and 25 % as focus.

In tagging the topic and focus words in the corpus, the following definitions were employed: In interrogative clauses focus consists of those words that form the exact entity that is being asked and all the other words that define the subject are tagged as belonging to the topic. In declarative sentences that function as answers words that form the core of the answer are tagged as 'focus', and other words that merely provide context for the specific answer are tagged as 'topic'. In other declarative sentences 'topics' are words that define the subject matter and 'focus' is applied to words that communicate what is being said about the topic. Regardless, the tagging task was in many cases quite difficult, and the resulting choice of tags often debatable.

As is charasteristic of spoken language, the data contained a noticeable percentage (35 %) of elliptic utterances, which didn't contain any topic words. Multiple topic constructs, on the other hand, were quite rare: more than one topic concept occurred in only 1 % of the utterances. The pronouns were quite evenly distributed with regard to position in the utterances: 32 % were in medial and 36 % in final position[3].

3.2 The probabilistic model

The probability of a word belonging to the class topic, focus or other is modeled as

$$P(T_i|W,S) = \frac{P(T_i|W)P(T_i|S)}{P(T_i)}, \qquad (2)$$

where W denotes the word, S its position in an utterance, and $T_i \in \{\text{topic, focus, other}\}$ stands for the class. The model thus assumes that being a topic or a focus word is dependent on the properties of that particular word as well as its position in the utterance. Due

[2]Non-textual cues such as silences within turns could not be considered for segmenting because they were not marked in the data.

[3]We interpreted 'medial' to mean the middle third of the sentence, and 'final' to be the last third of the sentence.

to computational reasons we made the simplifying assumption that these two effects are independent, i.e., $P(W, S) = P(W)P(S)$.

Maximum likelihood estimates are used for the terms $P(T_i|W)$ for already seen words. Moreover, for unseen words we use the average of the models of words seen only rarely (once or twice) in the training data.

For the term $P(T_i|S)$ that describes the effect of the position of a word we use a softmax model, namely

$$P(T_i|S_j) = \frac{e^{q_i(x_j)}}{\sum_i e^{q_i(x_j)}}, \tag{3}$$

where the index j identifies the word and x_j is the position of the word j. The functions q_i are defined as simple linear functions

$$q_i(x_j) = a_i x_j + b_i \tag{4}$$

The parameters a_i and b_i are estimated from the training data. For the class T_3 (other), these parameters are set to a constant value of zero.

3.2.1 ML estimation

When evaluating the rest of the model parameters we use two methods, first Maximum Likelihood estimation and then Bayesian variational analysis.

In ML estimation the cost function is the log likelihood of the training data D given the model M, i.e,

$$\ln P(D|M) = \ln \prod_w P(T_i|S_w) \tag{5}$$

$$= \sum_{w \in T_1} q_1 + \sum_{w \in T_2} q_2 + \sum_w (-\ln(1 + e^{q_1} + e^{q_2})). \tag{6}$$

The logarithmic term is approximated by a Taylor series of first degree and the parameters can then be solved as usual, by setting the partial derivatives of $\ln P(D|M)$ to zero with regard to each parameter. The parameters b_i can be solved analytically and the parameters a_i are solved using Newton iteration.

3.2.2 Bayesian estimation

The ML estimation is known to be prone to overlearning the properties of the training data. In contrast, in the Bayesian approach, also the model cost is included in the cost function and can be used to avoid overlearning. For comparison, we thus tried also the Bayesian approach utilizing the software and methodology introduced in (Valpola et al., 2001). The method is based on variational analysis and uses ensemble learning for estimating the model parameters. The methodology and the software allows for the optimization of the model structure with roughly linear computational complexity without the risk of over-fitting the model. However, in these experiments the model structure was not optimized.

3.2.3 Disregarding position information

Furthermore, to study the importance of the position information, we calculated the probabilities using only ML estimates for $P(T|W)$, i.e., disregarding the position of the word.

3.2.4 Tf×idf

As a comparison, we applied the tf×idf weighting scheme, which is commonly used in information retrieval for weighting content words. This method does not benefit from the labeling of the training data. For this reason, it does not differentiate between 'topic' and 'focus' words.

3.3 Experiments and results

The following experiment was performed using each described method: For each utterance in the test data, n words were tagged as topic, and likewise for the focus category. Further, n was varied from 1 to 8 to produce the results depicted in Figure 1.

As can be seen, the Bayesian variational analysis and the maximum likelihood estimation produce nearly identical performances. This is perhaps due to the use of very smooth model family, namely first-order polynomials, for taking into account the effect of the position of the word. For this reason, overlearn-

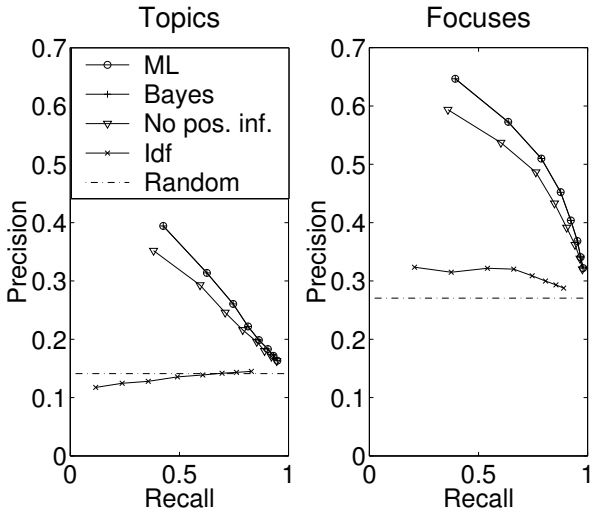

Figure 1: The precision–recall curves for topic–focus estimation. (ML = maximum likelihood, Bayes = Bayesian variational analysis, No pos. inf. = without position information, Idf = tf×idf weighting, Random = the average precision with random selection.)

ing is not problem even for the ML estimation. However, since the nearly identical results were obtained using two completely different implementations of quite similar methods, this can be considered as a validation experiment on either implementation and optimization method. In total, it seems that the full statistical model designed works rather well especially in focus identification.

When compared to the full model, disregarding position information altoghether results in inferior performance. The difference is statistically significant ($p \leq 0.05$) in focus identification for all values of n and in topic identification for small values of n. Moreover, the performance of the tf×idf scheme is clearly inferior in either task. However, it seems that the tf×idf definition of word importance corresponds more closely with the definition of 'focus' than that of 'topic'.

4 Discussion and conclusions

We examined two different viewpoints for the topic identification problem in natural language understanding. In experiments utilizing document maps it was found that longer dialogue segments are reliably modeled, but especially for short segments the history of the utterance must be consulted. A perhaps more interesting idea would be to also look at morphological features, such as cases, and include them in the encoding of the utterances. We plan to study this possibility in further work.

In the second viewpoint, individual utterances were analyzed to automatically identify 'topics' (what the user is talking about) and 'focuses' (what is being said about the topic). Each word in an utterance was labeled as 'topic', 'focus' or 'other'.

A statistical model that utilized the identity of the word and its position in the utterance was found to be rather successful, especially for identification of words belonging to the 'focus' category. Without the position information significantly lower performance was observed, which indicates that position information is indeed relevant for the identification. In this case, the Bayesian modeling paradigm and the maximum likelihood estimation produced nearly identical performance. However, this is not the case in general, when less smooth model families and optimization of model structure are applied. In the future we plan to examine other kinds of model structures for this task, perhaps integrating new types of information sources regarding the words, as well. For example, it would be interesting to see whether the addition of prosodic information would provide additional cues to improved solving of this task.

5 Acknowledgements

We thank Harri Valpola for his valuable advice concerning the estimation of the topic-focus identification model and for the possibility to apply the Bayesian software package developed by his group.

This work is part of the collaborative 'Interact' project on natural language interaction in Finnish.

References

Ella Bingham, Jukka Kuusisto, and Krista Lagus. 2002. Ica and som in text document analysis. In *The 25th ACM SIGIR Conference on Research and Development in Information Retrieval,August 11-15, 2002, Tampere, Finland.* Submitted.

Eva Hajiĉová, Petr Sgall, and Hana Skoumalová. 1995. An automatic procedure for topic–focus identification. *Computational Linguistics*, 21(1):81–94.

M. A. Halliday. 1985. *Introduction to Functional Grammar.* Oxford University Press, Oxford, UK.

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. 1996. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

R.M. Iyer and M. Ostendorf. 1999. Modelling long distance dependencies in language: Topic mixtures versus dynamic cache model. *IEEE Trans. Speech and Audio Processing*, 7.

Kristiina Jokinen, Antti Kerminen, Mauri Kaipainen, Tommi Jauhiainen, Markku Turunen, Jaakko Hakulinen, Jukka Kuusisto, and Krista Lagus. 2002. Adaptive dialogue systems — interaction with interact. In *3rd SIGdial Workshop on Discourse and Dialogue, July 11 and 12, 2002.* To appear.

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojrvi, Vesa Paatero, and Antti Saarela. 2000. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585.

Teuvo Kohonen. 1982. Analysis of a simple self-organizing process. *Biological Cybernetics*, 44(2):135–140.

Teuvo Kohonen. 1995. *Self-Organizing Maps.* 3rd, extended edition, 2001. Springer, Berlin.

Mikko Kurimo and Krista Lagus. 2002. An efficiently focusing large vocabulary language model. In *International Conference on Artificial Neural Networks, ICANN'02.* To appear.

Krista Lagus. 2000. Text mining with the WEBSOM. *Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 110*, 54 pp. December. D.Sc(Tech) Thesis, Helsinki University of Technology, Finland.

Xia Lin. 1992. Visualization for the document space. In *Proceedings of Visualization '92*, pages 274–81, Los Alamitos, CA, USA. Center for Comput. Legal Res., Pace Univ., White Plains, NY, USA, IEEE Comput. Soc. Press.

Xia Lin. 1997. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48:40–54.

Kristine Ma, George Zavaliagkos, and Marie Meteer. 1998. Sub-sentence discourse models for conversational speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2*, Seattle, Washington, USA.

Marie Meteer and Rukmini Iyer. 1996. Modeling conversational speech for speech recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Petr Sgall, Eva Hajiĉová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* D. Reidel Publishing Company, Dordrecht, Holland.

Harri Valpola, Tapani Raiko, and Juha Karhunen. 2001. Building blocks for hierarchical latent variable models. In *In Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, California, USA.

Maria Vilkuna. 1989. *Free Word Order in Finnish. Its Syntax and discourse functions.* Suomalaisen Kirjallisuuden Seura, Helsinki.