知覚情報からの概念構造の抽出にもとづく音声入力による言語獲得

岩橋直人　　田村正統†

ソニーコンピュータサイエンス研究所
Email: {iwahashi,tamura}@csl.sony.co.jp

本報告では、パターン認識により非言語的な知覚情報から概念情報を抽出し、これを用いて音声を通して言語を獲得するアルゴリズムについて述べる。アルゴリズムでは、センサーによる観測信号は知覚特性を考慮した特徴量空間に射影される。語彙項目はこの知覚特徴量空間、文法は許容される規則のパラメータ空間での、確率モデルによりそれぞれ表現される。これらの確率モデルは音声と非言語的知覚情報との関連性にもとづいて学習される。入力音声の構文構造は知覚情報の中の概念間の関係を分析することにより推測する。学習された文法は、概念の確率モデル間の類似性にもとづいて汎化される。ベイジアン学習等の統計的学習手法を用いることで、学習データ中の言語ノイズ、音声処理と知覚情報処理における入力の曖昧性と出力の不確実性、学習データ量の少なさに対して、頑健性を高めている。言語獲得の基本原理として、状況をまたいだ学習、および単語の意味の排他性の原理を部分的に用いている。インプリメントしたアルゴリズムは音声認識処理と動的グラフィックシーンからの概念抽出処理を含むものとなっており、予備的な実験の結果も示される。

Spoken language acquisition based on the conceptual pattern analysis in perceptual information

Naoto Iwahashi and Masatsune Tamura†

Sony Computer Science Labs. Inc.
Email: {iwahashi,tamura}@csl.sony.co.jp

This paper describes a machine learning algorithm for spoken language acquisition by using concepts extracted from nonliguistic perceptual information, based on a pattern recognition technique. The algorithm projects the raw sensor-observed signals into perceptually appropriate feature spaces. Lexicon and grammar are respectively represented in stochastic form in the feature space and in the possible grammar parameter space. Their learning is based on the association between speech and perceptual information. The syntactic structure is inferred from the conceptual structure obtained by analyzing the conceptual patterns in the perceptual information. The grammar is generalized according to the similarity of concept distributions in the feature space. The algorithm is robust against noise, ambiguity, and sparseness in the learning data because it uses statistical learning, such as Bayesian learning In the learning process, cross-situational learning and the principle of exclusivity applied between word meanings are partly implemented in a statistical way. The implemented algorithm includes the processes of the speech recognition and the analysis of graphical scenes containing stationary or moving objects. A preliminary experiment is also described here.

†Masatsune Tamura is also with Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology.

1 Introduction

The computational study of language aquisition has been attracting interest in various research areas, and the recent progress in both linguistics and machine learning theory has made it a very fruitful field of study. Language acquisition algorithms for machines [1, 2] have provided us with an intuitive understanding of and profound insight into the phenomena associated with language acquisition. They have also led to new machine learning methods. On the other hand, the recent progress of such application technologies as computation, telecommunications, sensing, and robotics has made the development of natural language interfaces with machines more important because it has increased the demand for easy and comfortable relationships with machines. The use of language acquisition schemes in interactive machines has been studied in attempts to increase the flexibility of the language interface, and one of the practical applications investigated was on automatic call-routing system using speech recognition [3]. While linguistic theories are inevitably applied in machine learning technology, our resesarch was motivated by this demand for better interfaces.

The language for communication with machines should be grounded in the process dealt with by the machines, which might be either a liguistic or nonlinguistic one. This paper addresses the question of *how a language acquisition algorithm can utilize nonlinguistic perceptual information.* Several pioneering studies exploring this problem have implemented algorithms based on inductive learning using a set of pairs each consisting of a word sequence and its nonlinguistic or semantic information. For instance, algorithms for learning the meanings of English prepositions with surrounding words, using the symbolic attributes for spacial relationship, were presented in [4] and [5]. In addition, Siskind [6] described a word-to-meaning mapping algorithm using the set of pairs each consisting of a sentence and a bunch of its possible whole referencial meanings represented symbolically with Jackendoff style expression. This algorithm is based on cross-situational learning [7] and the principle of exclusivity [8] applied between word meanings, and it succesfully addressed the problems due to homonyms and to noisy learning data. Visual graphical information rather than symbolic nonlinguistic information was used in simple word-to-meanining learning tasks in [9] and [10]. The judgement of whether or not the system's response is appropriate has also been used as nonlinguistic information in [3, 11], where the meaning of a word was defined as the action the system should take. And spoken-word acquisition algorithms based on unsupervised clustering of speech tokens were presented in [9, 12, 13]. There have been some studies [14, 15] on the use of semantic information in learning of syntactic rules. An algorithm for the learning of stochastic regular grammar in a visually-grounded way was presented in [9], in which the linear order of words in utterances was dealt with.

The algorithm we wanted to develop, in contrast, is one that acquires hierarchical characterisitcs of syntax from speech by utilizing the nonsymbolic perceptual information from visual and kinematic sensors, and so on. It projects the raw sensor-observed signals into perceptually appropriate feature spaces. Lexicon and grammar are respectively represented in stochastic form in the feature space and in the possible grammar parameter space. Their learning is based on the association between speech and perceptual information. The syntactic structure is inferred from the conceptual structure obtained by analyzing the conceptual patterns in the perceptual information. It could be considered an algorithmic implentation of the part of the semantic bootstrapping scheme [16] in grammar acquisition. In order to obtain the semantic information from the sensor input, many intermediary processes have to be considered on the feature extraction from sensor input, conceptual representation, and the analysis of semantic structure. The algorithm uses pattern recognition and analysis techniques for implementing these processes. On the whole, it considers many ambiguities due to dealing with contiuous perceptual and speech signal, in a statistical way.

In addition, the word meanings grounded in perception are expected to play an important role in generalizing the grammar. Although the use of hand-written symbolic conceptual attributes have been studied for grammar generalization, we have found no reports of algorithms based on conceptual relationships in perceptual signal space. Particularly, the role of conceptual similarity in perceptual signal space appears to be esssential, and its algorithmic implementation is considered.

The algorithm described in this paper deals with graphical scenes as perceptual information. It includes the process of speech recognition and the conceptual pattern analysis of graphical scenes. The parameters in initial stochastic grammar are gradually changed adaptively in a statistical way. As in the previous studies, cross-situational learning and the principle of exclusivity are used in the learning process.

2 Learning Task

We set up the following spoken language acquisition task. A human and a machine see the same scenes on a display on which some graphical objects are shown. Each graphical object is either stationary or moving. The human participant's possible action is the combination of speaking about the scene into a microphone, poiting and moving one of the graphical objects in the scene by using a pointing device. The machine learns language through a sequence of such strokes, which provides the set of pairs consisting of the operation in a scene and the speech describing that operation. During the course of the task, the human may, to confirm how well machine has learned, ask the machine to speak about a given scene and may also ask the machine to move objects in response to the speech input.

3 Algorithm

Outline

The system initially has a simple and neutral stochastic grammar with no lexicon. The algorithm is based on inductive learning using the set of pairs made up of the operations in scenes and the speech describing those operations. Sequences of spoken words in the speech are recognized. Possible individual concepts are extracted from each operation, and then a possible structural relation among them as a whole is constructed, by scene analysis. This conceptual structure is the chandidate of meaning of the given speech, and possible association between the recognized spoken words and the individual concepts extracted is obtained. The learning of both the lexicon and the grammar are based on this association. Speech recognition is carried out as a part of the unsupervised clustering process for lexicon acquisition. The system generates sysnthetic speech according to the learned grammar, and graphical scene based on maximum likelihood criterion[17]. No text is dealt with in either input or output of the system.

Features for concept representation

The raw speech and graphical data are projected into the appropriate feature spaces. It appears to be important that the perceptual characteristics of a human and the system are shared in the feature space. The features for the speech and the scene are, respectively, decided by the requirements from speech recognition and conceptual scene analysis. In speech recognition there are many proposed features which adopt auditory characteristics, such as Mel frequency scaling and time-frequency masking effect. In scene analysis, we have found few studies on the perceptually appropriate features that are physically grounded in the scene. While simply the use of perceptually uniform color space L*a*b* and egde information may be effective, the perceptually grounded primitives [18] for the concept representation should be explored. While each feature may be either discrete or continuous, depending on the task, the algorithm described here deals with only continuous ones.

Learning concepts and spoken words

Individual concepts of graphical objects and spoken words are acquired as membership functions on the respective feature spaces, and are represented by a probability density functions. These membership functions are used as discriminant functions in speech recognition and scene analysis. Hidden Markov models (HMM) [26] are used to represent the dynamic characteristics of graphical objects and speech sounds. In each HMM, the output probalitity density function (*p.d.f.*) at each state is given by a multivariate normal distribution. A normal distribution is used to represent the static characteristics of graphical objects. The values of the parameters in these *p.d.f.*s are calculated by inductive learning methods. Particularly, Bayesian learning [19] is used for the learning of the *p.d.f.*s on static graphical characteristics. Bayesian learning reduces the severity of the problem of data sparseness and the so-called *curse of dimensionality*, and makes it easy to use high-dimentional features.

Lexicon acquisition

Each lexical item in the lexicon includes concept and spoken word representations. The set of pairs of the features of the spoken word and graphical scene is divided into clusters, each of which corresponds to a lexical item. Using tokens in each cluster, each the *p.d.f.* for graphical concept and spoken word is estimated as described above. The lexicon is built up by using an unsupervised clustering method [20] in an incremental manner with regard to token input. When a new word (one not in the lexicon) is spoken, a new cluster should be generated. Any phonemic units are not assumed as prior knowledge to exclude language dependency. The algorithm is implemented as follows:

1. Get a new speech sample o, and add o to the sample set O.

2. Select the HMM h_c associated with cluster c which gives the highest value of likelihood with regard to o, and add sample o to cluster c.

3. If c is to be split, split it into new two clusters and estimate new HMMs, otherwise reestimate HMM h_c.

4. Resplit O into clusters by selecting the HMM h_i that gives the highest likelihood for each $o_i \in O$.

5. For all clusters, estimate HMM h_{c_j} for each c_j. Then go to step 1.

The determination of the number of HMM states is based on cross validation in a cluster. Although the splitting decision in step 3 is the problem of recognition verification [21], likelihood gain is used as threshold criterion as in the conventional regression tree[22]. Splitting the set of time sequential signals of speech is not as straightforward as the conventional cluster splitting, and is carried out by splitting HMM based on expectation-maximization algorithm [23]. In the algorithm described here the clustering is based only on speech information, but the algorithm can be extended to also use perceptual information.

Conceptual scene analysis

Each scene is analyzed in order to obtain its conceptual expression (CE), which is defined by a rudimentary form consisting of the individual concepts in the lexicon with semantic attributes[1], such as [object], [action], and [to]. The scene analysis assigns the semantic attributes to each extracted graphical concept. In the process, the system judges whether the object movement is spontaneous or forced by the human. For instance, if the concept *rotate* is extracted as the state of an object, the attribute [object] is assigned. And if it is extracted as the action to an object, the attribute [action] is assigned. When the operation is

to put a rotating red ball on the blue block on the right-hand side,

the CE might be

$$\begin{bmatrix} \text{[action]} & : & \textit{put} \\ \text{[object]} & : & \textit{rotate, red, ball} \\ \text{[to]} & : & \textit{blue, block, right} \end{bmatrix}$$

[1] These can be defined dependently on the task. We may be able to use the semantic primitives described in [24] and [25], although how to extract such semantic primitives from a scene is a problem.

where written in the right-hand column are the concepts in the lexicon. The *CE* is constructed using the individual concepts of possible words in a speech such that the likelihood of the membership fuction made by the composition of the individual concept membership functions is maximized for the operation in the scene.

Grammar learning

Grammar is learned through the adaptation of the initial grammar. Learnable grammar is rather restricted so far: functional words such as prepositions and articles, for example, are not treated. Let a constituent of a sentence be defined as a word group that describes a concept, which may be a structural combination of multiple concepts. Each constituent is characterized by the semantic attribute assigned to the concept that the constituent describes. The initial grammar consists of the following three parts:

1. The set A of semantic attributes a_i ($i = 1, 2, ...$).

2. A constituent formation rule:
   If two words are used to describe a concept, which may be constructed using multiple concepts, the word between these two words must be used to describe same concept.

3. The stochastic grammar *SG* consisting of probabilities $P(a_i \to a_j a_k)$ that the constituent with semantic attribute a_i consists of two constituents each with semantic attribute a_j and a_k in this order.

The constituent formation rule is fixed, but the *SG* is adapted by Bayesian learning. The adaptation is based on the association between the speech and the scene. The following is an example of the procedure of determining such an association. Ideally, the speech recognizer recognizes the word seqence *'put'-'rotate'-'ball'-'blue'-'block'* perfectly. Then the scene analyzer produces a *CE* by using the concepts of these words under the constraints of the constituent formation rule. That is, the *CE* is produced so that the likelihood of the overall conceptual structure based on each concept pattern's likelihood is maximized under the constraints. The comparison of the recognized word sequence with the *CE* results in the sentence being divided into the following three constituents:

(([action], *'put'*), ([object], *'rotate'-'ball'*),
([target], *'blue'-'block'*)).

The *SG* is adapted using the information of the order [action]-[object]-[target] of the constituents' attributes. Although any adaptation schemes for stochastic language models [27] can be used, we should consider that the algorithm does not always produce the proper constituents and their semantic attributes because there are many uncertainties in the whole process. The input utterance itself could be ungrammatical or could not even describe the scene correctly. The algorithm therefore utilizes Bayesian learning, by which the values of *SG* probabilities are adapted robustly. That is, a small number of improper samples will not influence the grammar adaptation much.

Grammar generalization

The generalization of the grammar is considered in the speech generation process. The system generates synthetic speech describing a given scene on demand, and the word set associated with the scene is selected by using concept membership functions. The word order in the synthetic speech is determined according to the learned grammar. If a grammar rule including the selected word set exists, the word order is determined by this rule. If such a rule does not exist, the word order is determined using the grammar rule including the word set most similar to the selected word set. The calculation of the similarity of words is based on the distance between the corresponding concept *p.d.f.*s, such as the Bhattacharyya distance or the KL divergence. To focus on the shape of *p.d.f.* rather than the their location, we use the similarity measure R_w as follows:

$$R_w(w_1, w_2) = -\ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}, \qquad (1)$$

where the term related to mean vectors has been eliminated from Bhattacharyya distance. Σ_1 and Σ_2 denote the covariance matrices of concept *p.d.f.*s for words w_1 and w_2, respectively. The word sequence is determined by searching an appropriate grammar rule so that the sum of word similarities between the selected word set and the word set in the rule. For instance, when the word set { *'red'*, *'right'*, *'ball'* } is selected in the association with a object in a scene, the rule ([object], *'yellow'-'left'-'ball'*) is chosen as a rule which has the word set most similar to the selected word set, where the distributions of the concepts *right* and *left* are far apart in the feature space but the their shapes are similar.

4 Experiments

The algorithm was tested using the following setup. The position on the display (two-dimensional: horizontal and vertical coordinates) and the color information (three-dimensional: L*a*b* parameters) were used as the features of the concept representation of graphical objects. Mel-scale cepstrum coefficients [28] and their delta parameters (thirty-two dimensional) were used as the features of speech. A normal-Wishart distribution was used as prior distributions of the parameters in normal distribution for each concept representation. The parameters in the prior distribution were set empirically.

A male participant taught language to the machine, according to the task described in Section 2, under acoustic conditions typical of an office environment. In the first step, fifteen lexical items consisting of concepts about position, color, and movement were taught (Table. 1). Note that for convenience each concept in Table. 1 is denoted by a text word which has a concept similar to it. These lexical items were taught in sixty learning strokes. In each stroke, either concepts about the static characteristics of objects (static position and color) were taught by uttering one or more words and pointing to a stationary object, or else concepts about dynamic characteristics (movement) were taught by uttering one word and moving an object.

Table 1: The concepts taught in the experiments

position	color	movement
right	*red*	*up*
left	*blue*	*down*
top	*yellow*	*rotate*
bottom	*green*	*put*
middle	*gray*	*slide*

In the next step, grammar was taught by uttering one or more words while pointing or moving a object which is stationary or moving. Utterances in the experiments were rather simple. If, for example, the operation was '*to lift a red rotating object at the bottom*' and the utterance was *"up rotate red bottom "* , the extracted *CE* was:

$$\left[\begin{array}{lcl} \text{[action]} & : & \textit{up} \\ \text{[object]} & : & \textit{rotate, red, bottom} \end{array}\right]$$

The following probability was learned:

$$P([\mathrm{S}] \to [\mathrm{action}]\ [\mathrm{object}]) = \\ 1-P([\mathrm{S}] \to [\mathrm{object}]\ [\mathrm{action}])$$

A beta distribution ($\alpha = \beta = 5.$) was used as a prior distribution of the value of this probability in Bayesian learning.

As a result, the lexicon was learned after sixty strokes. The concepts could be acquired with such a small number of learning strokes because of the Bayesian learning. When maximum likelihood learning was used instead, the concept *p.d.f.* often became too sharp due to small amount of samples. This often led to unintuitive errors in scene analysis. For example, a position near the edge of the left-hand side of the display was recognized as *right* because the *p.d.f.* so far obtained for the concept *left* was much sharper than the *p.d.f.* for the concept *right.* Bayesian learning reduced the severity of this problem by making the *p.d.f.* less sensitive to distributions of small amounts of learning samples.

The stochastic grammar was also learned robustly. The change of the value of probability $P([\mathrm{S}] \to [\mathrm{action}][\mathrm{object}])$ is shown in Figure 1, where it is obvious that the learing was robust against early errors in the learning process.

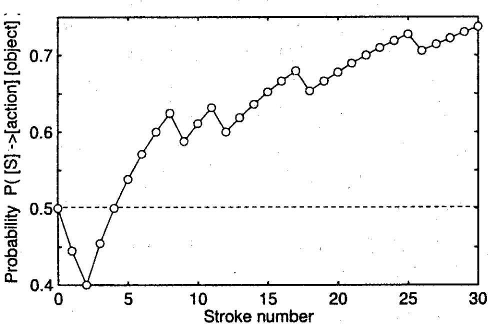

Figure 1: The change of probability in the *SG* during the course of learning

The grammar generalizaion process was also tested. After *'right'-'top'-'red'* was uttered in the previous stroke, a red object at the bottom of the left side of the display was pointed to and the system was asked to describe it. Then the phrase *'left'-'bottom'-'red'* was generated using the rule ([object], *'right'-'top'-'red'*).

5 Discussion

Although algorithm tested was a preliminary version and the spoken language used was not natural, the experimental results are promising. We found that statistically robust learning provides human participants with an intuitive understanding of how well the machine is learning. Such information appears to be important when considering what strategy should be used in later teaching. The learning efficiency depended on learning samples. The distribution of learning samples the teacher showed, for example, greatly influenced on the learning performance. Bayesian learning provides robustness in learning individual concepts but does not improve directly the efficiency of the cross-situational learning. A mechanism for improving the cross-situational learning is necessary.

Robustness is quite important in the learning of *SG* because the speech recognition process uses the *SG* which has been learned so far. If the *SG* is learned wrongly because of errors in the early stage, the speech recognition becomes likely to be erroneous. This would make it difficult to get the correct grammar by adapting the *SG*.

The scheme for learning the *SG* probabilities can be expanded to deal with other kinds of grammars, such as one based on the principles-and-parameters (P&P) approach [29]. Indeed, the *SG* probability obtained in the experiments can be taken as the parameter which determines the order of the head and the complement in a verb phrase. The robust statistical learning might be part of the solution to *poverty of stimulus* problem: "how can children learn language using only a small amount of learing samples?".

Although the presented formulation of the word similarity is rudimentary, it shows a concrete method of measuring the relationship of concepts represented in continuous feature space. In dealing with metaphor with anologies between concepts, sucn an approach based on geometrical relation in the feature space is worth pursuing.

6 Conclusion

The preliminary framework for the acquisition of spoken language grounded in perception, that does not use symbolic imformation, was presented. The method described is a statistical method based on the conceptual analysis of the perceptual information. Its algorithm recognizes words in an utterance and infers their conceptual structure by analyzing the scene associated with the speech. The grammar generalization method was implemented using a measure of concept similarity based on the relation of the concept *p.d.f.*s in the feature space. Simple experiments demonstrated that the lexicon and the grammar were learned robustly. The natural expansion of the range of learnable language, and the development of the perceptually plausible features are future works.

References

[1] M. R. Brent "Advances in the computational study of language acquisition," *Cognition*, **61**, pp.1–61, 1996.

[2] J. L. Elman et al. "Rethinking innateness: a connectionist perspective on development," MIT Press, 1996.

[3] A. L. Gorin et al. "Adaptive acquisition of language," *Computer Speech and Language*, 5, pp.101-132, 1991.

[4] C. Harris "A connectionist approach to the story of 'over'," *Berkeley Linguistic Society*, 15, pp.126–138, 1989.

[5] P. Munro et al. "A network for encoding, decoding and translating locative prepositions," *Connection Science*, 3, pp.225–240, 1991.

[6] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, **61**, pp.39–91, 1996.

[7] S. Pinker, "Learnability and cognition," Cambridge, MA: Haarvard University Press. 1989

[8] E. M. Markman, "Constraints on word learning: Speculations about thier nature, origins, and domain specificity," *Minnesota Symposium on Chile Psychology*, 25, Lawrence Erlbaum Associates, 1992.

[9] S. Nakagawa et al. "An acquisition system of concept and grammar based on combining with visual and auditory information," (in Japanese) *Trans. of Information Society of Japan*, Vol.10, No.4, pp.129–137, 1994.

[10] T. Regier "The Human Semantic Potential," MIT Press, 1997.

[11] S. Fujie and T. Kobayashi "Language acquisition of autonomous robot with action," (in Japanese) *Proc. of 13th Annual Conf. of Japanese Society of Artificial Intelligence*, pp.223-224, 1999.

[12] A. L. Gorin et al. "An experiment in spoken language acquisition," *IEEE Trans. on speech and audio processing*, Vol.2, No.1, pp.224–240, 1994.

[13] D. Roy and A. Pentland "Word learning in a multimodal environment," *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, Vol.6, pp.3761-3764, 1998.

[14] P. Langley, "Language acquisition through error recovery," *Cognition and Brain Theory*, Vol.5, pp.221-225, 1982.

[15] R. C. Berwick, "The acquisition of syntactic knowledge," MIT Press, 1985.

[16] S. Pinker "Language Learnability and Language Development," Harvard niversity Press, 1984.

[17] K. Tokuda, T. Masuko, T. Kobayashi and S. Imai "An algorithm for speech parameter generation from HMM using dynamic features," *The Journal of the Acoustical Society of Japan*, Vol.53, No.3, pp.192-200, 1997.

[18] J. M. Siskind "Grounding Language in Perception," *Artificial Intelligence Review*, 8, 371-391, 1994-5.

[19] M. H. DeGroot "Optimal statistical decisions," McGraw-Hill, 1970.

[20] R. O. Duda et al. "Pattern classification and scene analysis," John Wiley & Sons, 1973.

[21] R. C. Rose "Word spotting from continuous speech utterances," In C-H. Lee et al. (eds), Automatic speech and speaker recognition advanced topics, Kluwer Academic Publisher, 1996.

[22] L. Breiman et al. "Classification and regression tree," Wadsworth & Brooks, 1984.

[23] A. Dempster et al. "Maximum likelihood from incomplete data vie the EM algorithm," J. Roy. Statist. Soc. B, Vol.39, pp.1–38, 1977.

[24] R. C. Shank "Conceptual Dependency: The theory of Natural Language Understanding," *Cognitive Psychology*, 3(4), pp.552–631, 1972.

[25] R. Jackendoff "Semantics Structures," MIT Press, 1990.

[26] F. Jenelik "Statistical methods for speech recognition," MIT Press, 1998.

[27] R. Bod et al. "Data-Oriented Language Processing," *In S. Young et al. (eds), Corpus-based methods in language and speech processing*, Chapter 5, 1997.

[28] K. Tokuda et al. "Spectral estimation of speech based on mel-cepstral representaion," *Trans. IEICE*, vol.J74-A, pp.1240–1248, 1991.

[29] N. Chomsky, "Knowledge of Language: Its Nature, Origin and Use," Praeger, 1986.