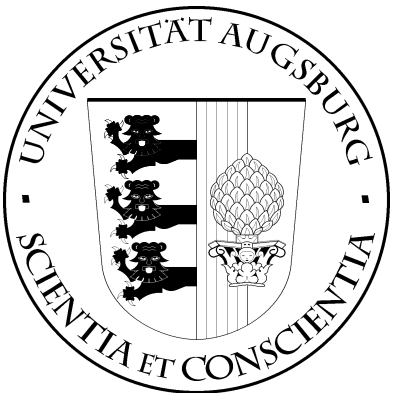
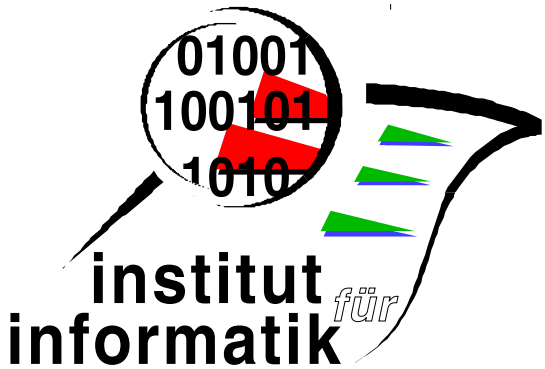
UNIVERSITÄT AUGSBURG



Language Modeling with Utterance-Meaning Pairs

Günther J. Wirsching and Christian Kölbl

Report 2011-12 May 12, 2011



INSTITUT FÜR INFORMATIK

D-86135 AUGSBURG

Introduction

In linguistics or computer science, a *language* usually is defined as set of finite word-strings, where the words are drawn from a given *dictionary*. *Language modeling* is, amongst others, the presentation of a language to a computer, i.e., in an algorithmically accessible way. As pointed out by Jurafsky and Martin [5], most known methods are mathematically equivalent to weighted finite state transducers.

When designing a speech dialog system, we are not satisfied with an arbitrary language model—what we need is a possibility to extract *meaning* from an *utterance* of the user. In order to come to terms with this need, there are two extremal strategies:

(A) Use a statistical language model (trained on a very large corpus) for recognition, and then try to extract meaning by a robust parser.

Disadvantage: When the statical language model is trained on a general corpus, parsing for special meanings often doesn't give results.

(B) Fix the set of phrases which can be understood by giving a linear grammar.

Disadvantage: An utterance not fitting the grammar cannot be understood by the system.

In this paper, we develop an argument from behavioristic psychology leading to the definition of a mathematical object which we call *utterance-meaning pair*. With this concept, the combination of a language model with a method to associate one or more meanings to an utterance can be mathematically described as set of utterance-meaning pairs. In particular, strategies (A) and (B) are included as special cases. As an application of this concept, we introduce the notion *horizon of comprehension* to indicate how to deal with the disadvantage of strategy (B) by keeping the probability of an utterance not fitting the grammar low but not zero.

Note that we do not give specific algorithms for representing language model and meaning extraction; rather, we concentrate on the functional structure.

1 Utterance-meaning pairs

An *utterance-meaning pair* will be a mathematical model for the information recognized by the system in a dialog turn, i.e., when the user says something which should be understood by the dialog system. An *utterance* is modeled as a string of *words*, which can be defined arbitrarily as smaller or larger phonetic units, and a *meaning* is a feature-values relation [3].

For getting an idea in which way these two elements are connected, we assume that the possible utterances to a specified set of possible meanings are learned *verbal operants* in the sense of behavioristic psychology.

In order to be more explicit about this, we start by giving a short account of Skinner's [7] notion *verbal operant*, leading to an argument why this can be applied in dialog modeling. With this preparation, we then proceed to mathematical definitions.

1.1 Verbal stimulus-response associations

The famous first scientifically recorded example for a learned stimulus-response association is Pavlov's dogs [6]. During his research on the physiology of digestion in dogs, Pavlov endowed some dogs with a surgically implanted cannula to measure salivation. Soon he noticed that, rather than simply salivating in the presence of food, the dogs began to salivate in the presence of the lab technician who normally fed them. Motivated by this observation, Pavlov started experimental research. In his initial experiment, Pavlov used a bell to call the dogs to their food, and, after a few repetitions, he noticed that the dogs started to salivate in response to the bell. The idea is that the physiological behavior *salivation* is learned through reinforcement provided by repeated feeding after ringing the bell. In general, the situation can be describe by the following scheme:

$$\text{Stimulus} \longrightarrow \text{Response} \longrightarrow \text{Consequences}\,. \tag{1}$$

For learning a wished response, the experimentator adjusts stimulus and consequences in such a way that the wished response is reinforced.

Some years later, Skinner [7] argues that 'verbal behavior' is learned through reinforcement like other behavior, where he calls a behavior *verbal*, if its consequences are mediated by another organism. By this definition, the purpose of *verbal behavior* is to have consequences. This is expressed in the terminus *verbal operant* defined by Skinner [7, p. 20] to refer to a unit of verbal behavior. Note that, by Skinner's definition, most spoken utterances are to be considered as *verbal behavior*, but also non-verbal communication may contain *verbal operants*. In addition, the focus on the communicative aspect of language relates Skinner's approach to philosophical concepts of Wittgenstein [8], and to Bühlers *Sprachtheorie* [1].

We note in passing that Skinner has been criticized by many authors, e.g., Chomsky [2] or Hörmann [4, S. 101ff.], and a major point of criticism being the open question how a child could receive sufficient reinforcement to learn a language with all its complexity. For our purpose, it suffices to restrict attention to learned situations which we assume to be stable—we need not dwell on the details of the learning process.

1.2 Modeling meaning

What is the 'meaning' of a verbal behavior? Let us look in Skinner's book:

> *Technically, meanings are to be found among the independent variables in a functional account, rather than as properties of the dependent variable. When someone says that he can see the meaning of a response, he means that he can infer some of the variables of which the response is usually a function.* [7, p. 14]

In order to see what this means, we analyze the situation as if it were a setting in experimental psychology. Then we have to distinguish between *independent* and *dependent* variables. The *independent* variables are set by the experimentator, they include the stimulus, the possible consequences, and possibly other elements of the environment. The *dependent* variables are to be found in the behavior of the probands, i.e., in the responses.

In this section, we concentrate on the *independent* variables. Returning to the scheme (1) above, we see that these are to be found in the *stimulus* and in the *consequences*. Restricting attention to speech dialog systems, we see that in a single turn of the system, the *stimulus* is provided by

(i) a prompt provided by the system,

(ii) the dialog turns performed up to now, and

(iii) the knowledge of the user about the residual current context.

For describing the possible *consequences*, observe that the task of a speech dialog system can be described as performing operations on a given data base. Hence, the possible *consequences* of an utterance of the user are corresponding changes in the state of information of the system.In [3] we assume that the structure of the data base is given by an entity-relationship model and a corresponding relational database, and give an algorithm which constructs feature-values relations needed to represent the state of information of the system. The state of information itself is given as a weighted feature-values relation, and changes implied by what is understood by the system will be modeled as changes in the weights.

Let us now turn to mathematical definitions. The basic mathematical objects we use are *sets*, as defined axiomatically in Zermelo-Fraenkel set theory, and *relations* on a given set V, which are, by definition, subsets of the cartesian product $V \times V$. A relation $E \subseteq V \times V$ is called *acyclic*, if there is no finite sequence $v_0, \ldots, v_n \in V$ such that $v_0 = v_n$ and, for each $i \in \{1, \ldots, n\}$, $(v_{i-1}, v_i) \in E$. A pair $(v_1, v_2) \in E$ is called an *edge* of the relation. We call a triple $R = (V, E, \ell)$ a *feature-values relation*, if it consists of a set V, an acyclic

relation $E \subseteq V \times V$, and a labeling ℓ associating to each $v \in V$ an identifier $\ell(v)$. The first component v_1 of an edge $(v_1, v_2) \in E$ is called *feature*, and the second component v_2 is called the *value* of feature v_1. The set $F(V) \subseteq V$ of elements $v \in V$ which occur as features generates the *feature structure* of the feature-values relation:

$$R_f := \Big(F(V), E \cap (F(V) \times F(V)), \ell|_{F(V)}\Big) .$$

A feature-values relation is a mathematical object designed to describe 'meaning' in a speech dialog system. Consider the following very simple example: Assume that the a small portion of the feature-values relation derived from the data-base is the triple $\Gamma = (V, E, \ell)$ with four vertices with vertex labels

$$\ell(V) = \{\text{name}, \text{Elisabeth}, \text{Xaver}, \text{Anton}\}$$

and edges

$$E = \{(\text{name}, \text{Elisabeth}), (\text{name}, \text{Xaver}), (\text{name}, \text{Anton})\}.$$

Here the vertex with label "name" is a feature admitting three possible values "Elisabeth, Xaver, Anton". In this case, we would call a subrelation $\Gamma' \subseteq \Gamma$ a *meaning*, if it specifies exactly one of these values, which we consider mutually exclusive. We also call this a *specifying* feature-values graph.

In the general case, the situation is more complicate, as it may happen that a feature is split into subfeatures, each of them having values, e.g., we could have edges

$$\left\{ \begin{array}{c} (\text{name}, \text{first name}), (\text{name}, \text{last name}), \\ (\text{first name}, \text{Elisabeth}), (\text{last name}, \text{Meyer}), \\ \dots \end{array} \right\} .$$

In order to get a precise definition of 'meaning' in terms of feature-values relations, we distinguish between features with possibly more than one value, and features whose values are considered as mutually exclusive. In our example, the features "first name" and "last name" would be considered as *features admitting mutually exclusive values* (abbreviated MEV-features), but the feature "name" would be considered as a feature with two non-exclusive values "first name" and "last name".

Now we are ready to give a formal definition:

Definition 1 A feature-values relation (V, E, ℓ) with a given subset $V_e \subseteq V$ of MEV-features is called a *meaning*, if, for each feature $f \in V_e$, there is at most one edge $(f, v) \in E$.

We have now the ingredients for a mathematical model for 'meaning' in the situation, when the system has generated a prompt to ask the user for some information, and the answer of the user is to be evaluated: The class of all possible meanings that can be understood by the system should be given as feature-values relation Γ, the prompt should ask for the values of a specified part of the features of Γ, and the recognition results should be given as a weighted list of meanings of Γ, i.e., of subrelations $\Gamma' \subseteq \Gamma$ which are meanings.

1.3 Modeling the utterance

Now we turn to the *dependent* variable connected to a single user input to a dialog system. In a scheme (1), Skinner identifies the dependent variable as follows:

> *The probability that a verbal response of given form will occur at a given time is the basic datum to be predicted and controlled. It is the "dependent variable" in a functional analysis.* [7, p. 28]

What is the connection to dialog modeling? When designing a dialog system, we would like to know, at a given dialog situation with a given stimulus, the set of possible responses of the user, preferably with their probabilities of occurrence. So our goal seems to coincide with the goal of the experimental psychologist as described by Skinner. But this is not exactly true: there are many aspects of a 'verbal response' which can be observed by a psychologist but which usually are hidden to the speech recognizer, despite the fact that each of these aspects could carry meaning. In most cases, the recognizer is not able to perceive gesture or mimikry, and even prosody is not modeled and therefore not perceivable—and we have to live temporarily with possible losses of meaning caused by these circumstances.

In this paper, we only take into account phonetic descriptions, with or without prosody, of utterances. Assuming that a user when speaking to the system intends to utter a certain sequence of phones, we suppose that we are given a *phonetic alphabet* $\mathbb{A}$ containing sufficiently many characters to encode any reasonable phonetic intention of a user as string of characters from $\mathbb{A}$. For example, if we design an english language dialog system, and expect the user to speak english, a phonetic alphabet covering all english phonemes will suffice to describe what our speech recognizer could receive from an utterance.

Moreover, we assume that the user of our system intends to convey some information to the system, and that this information fits into the possible 'meanings' given by stimulus and possible consequences. Recall that

in the preceding section, we modeled these possible meanings as set of specifying feature-values graphs. Given a specifying feature-values graph M, we associate a set

$$\mathcal{U}(M) \subset \mathbb{A}* \tag{2}$$

of utterances conveying just the information given by M. Note that we do not *a priori* exclude the empty string, because, according to Skinner, an "empty" answer can also have consequences.

For the construction of the weighted finite state transducer which performs the phonetic pattern recognition, it is essential to combine the elements of $\mathcal{U}(M)$ to their meaning, whence the following definition.

Definition 2 Let $\mathbb{A}$ be phonetic alphabet and M be a specifying feature-values graph. Then a pair (u, M) with $u \in \mathbb{A}*$ is called an *utterance-meaning pair*.

Note that there is no uniqueness, in neither direction: a meaning M may be expressed verbally in more than one different ways, and a given utterance may have more than one meaning.

1.4 Utterance-meaning pairs and the dependent variable

We are now ready to describe the dependent variable connected to a single user input to the dialog system. If the possible meanings are defined by a stimulus (prompt, context, etc.) and possible consequences (operations on a data base), then the dependent variable is a probability function on the union $\bigcup \mathcal{U}(M)$, where M runs over the specifying graphs representing one of the possible meanings.

This probability function is strongly influenced by the prompt given by the system: if the system asks for, e.g., a name, then it is very likely that user response contains a name. For modeling the set of responses at a specific dialog turn, we argue that the set of possible (reasonable) meanings consists of the following five parts:

(i) a set $\mathcal{E}$ of meanings exactly asked for by the prompt,

(ii) a set $\mathcal{U}$ of meanings answering the prompt only partially (underanswering),

(iii) a set $\mathcal{O}$ of meanings containing more information as directly asked for by the prompt (overanswering),

(iv) a set $\mathcal{D}$ of meanings overanswering part of what has been asked for (deviating answer), and

(v) a set $\mathcal{G}$ of meanings generally available, for instance triggering an interruption.

It appears reasonable to take the union of these sets to describe the set of answers which the system should be able to understand.

Definition 3 Given a prompt in some specific dialog context, we call the set $\mathcal{E}$ of meanings exactly asked for the *horizon of expectation*, und the union $\mathcal{C} := \mathcal{E} \cup \mathcal{U} \cup \mathcal{O} \cup \mathcal{D} \cup \mathcal{G}$ the *horizon of comprehension* at the given stimulus.

The probability function could, for instance, always dependent on the current dialog state, highlight the set $\mathcal{E}$ while keeping the other sets at a lower level.

We close this report by formulating some ideas how to get language models for designing a speeach dialogue system.

1. Specify the target of the intended system.
2. Collect the data from an appropriate set of wizard-of-Oz experiments.
3. Associate to each recorded utterance a meaning in the sense defined above.
4. Use combinatorial grammar methods to describe an expanded set $\mathcal{A}$ of possible utterance-meaning pairs.
5. Now local grammars can (automatically) be constructed for each horizon of comprehension $\mathcal{C} \subseteq \mathcal{A}$.

References

[1] Karl Bühler. *Sprachtheorie : die Darstellungsfunktion der Sprache*. Fischer, Stuttgart, second (unchanged) edition, 1965.

[2] Noam Chomsky. A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58, 1959.

[3] Markus Huber, Christian Kölbl, Robert Lorenz, Ronald Römer, and Günther Wirsching. Semantische Dialogmodellierung mit gewichteten Merkmal-Werte-Relationen. In Rüdiger Hoffmann, editor, *Elektronische Sprachsignalverarbeitung 2009. Tagungsband der 20. Konferenz. Dresden, 21. bis 23. September 2009*, volume 53 of *Studientexte zur Sprachkommunikation*, pages 25–32. TUDpress, September 2009.

[4] Hans Hörmann. *Psychologie der Sprache*. Springer, second edition, 1977.

[5] Daniel Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, NJ, 2009.

[6] Ivan P. Pavlov. *Conditioned Reflexes: an Investigation of the Physiological Activities of the Cerebral Cortex*. Oxford University Press, London, 1927. (Translated by G. V. Anrep).

[7] B. F. Skinner. *Verbal Behavior*. Prentice-Hall, Englewood Cliffs, New Jersey, 1957.

[8] Ludwig Wittgenstein. *Philosophische Untersuchungen. Kritisch-genetische Edition*. Suhrkamp, Frankfurt, 2001.