FRAME BASED INTERPRETATION OF CONVERSATIONAL SPEECH

Frédéric Béchet[1], *Christian Raymond*[2], *Frédéric Duvert*[3], *Renato De Mori*[3]

[1] Aix Marseille Université, LIF-CNRS, France
[2] UEB, INSA/IRISA, UMR 6074, Rennes, France
[3] Université d'Avignon, LIA-CERI, France , Mc Gill University, Canada

ABSTRACT

Two approaches to Spoken Language Understanding based on frames describing chunked knowledge are described. They are applied to the MEDIA corpus annotated in terms of concepts expressing chunks of spoken sentences. General rules of knowledge composition and inference appear to be adequate to effectively applying the application ontology for obtaining frame based representations of dialogue turns. The main difficulty appears to be the characterization of the syntactic knowledge expressing semantic links between knowledge chunks. This knowledge can be hand-crafted or automatically learned from examples. It is shown that the latter approach outperforms the former if applied to ASR error prone transcriptions.

Index Terms— Spoken Language Understanding

1. INTRODUCTION

Spoken Language Understanding (SLU) systems attempt to obtain semantic interpretations in a meaning representation language (MRL) from automatic transcriptions generated by an Automatic Speech Recognition (ASR) component. Various approaches have been obtained for this purpose using various types of syntactic analysis. Application dependent semantic non-terminal symbols have been introduced in context-free or context-sensitive grammars so that semantic interpretations can be obtained as parsing results. Non probabilistic context-sensitive grammars [1] as well as probabilistic context-free grammars [2, 3, 4, 5] were proposed for this purpose. Semantic information obtained during parsing is then used for generating the MRL description of a sentence using a frame language. Frame instances are useful for performing system actions such as data-base queries as in the ATIS project. They are also useful, among other things, to characterize dialog acts, to distinguish among multiple instances of or make reference to objects under discussion in a negotiation dialogs.

THIS WORK IS SUPPORTED BY THE 6TH FRAMEWORK RESEARCH PROGRAMME OF THE EUROPEAN UNION (EU), PROJECT LUNA, IST CONTRACT NO 33549.

Application independent mildly context-sensitive lexicalized grammars were also proposed [6] in which syntactic and semantic knowledge is associated to each word and a semantic process is executed to compose semantic information associated to each word when a sentence is parsed. Sentence interpretation is obtained in this way by constructing logic expressions describing the sentence meaning. Probabilities have been introduced in these models to account for knowledge incompleteness and imprecision as well as for errors introduced by the ASR component. Estimation of probabilities in these models is difficult due to data sparseness. Various approximations have been introduced to alleviate this difficulty. Some of them consist in using finite-state approximations of complex grammars or various types of classifiers using features to summarize complex statistical dependencies.

Various approaches have been considered and evaluated in the European LUNA project for obtaining frame-based semantic interpretations of spoken dialogue turns as introduced in section 2. Some of them implement a progressive interpretation process in which fragments of knowledge that can be characterized by robust finite-state models are hypothesized first and then composed using other models. As in early SLU systems [7] these models for chunked knowledge [8] attempt to match template patterns with ASR hypotheses. The approaches described in section 4 of this paper use recently developed and powerful machine learning methods for building the models and compare them with a classical rule-based approach described in section 3. Furthermore, new process steps are introduced for composing fragments into complex frame instances. Experimental results are presented in section 5.

2. A FRAME-BASED REPRESENTATION OF CHUNKED KNOWLEDGE

The interpretation phase of the SLU process is made difficult by possible errors contained in the ASR transcription of a sentence and by disfluencies such as self-corrections and repetitions that may be part of the sentence. In order to increase interpretation robustness, an initial step could be introduced in the process for generating hypotheses of semantic knowledge chunks to be further modified, removed or composed into more complex semantic structures. These hypothe-

ses can be obtained by using partial or shallow parsing using, for example, finite-state approximations of generative grammars, discriminatively trained classifiers or other models. In order to increase robustness, it would be interesting to have language models specific for each type of knowledge chunks that can be used in the decoding process of the ASR module.

Based on the above motivations, an approach is proposed with an early detection of semantic knowledge chunks expressed by local syntactic relations. Concepts expressed by different chunks hypothesized in a sentence are often semantically related. In this case, they can be considered as semantic constituents with each constituent represented by a concept tag. Semantic relations, such as spatial relations are language independent, while relations between their conceptual constituents and words in a sentence are language dependent. Semantic constituent hypotheses, represented by concept tags, can then be composed into semantic structures. The approach is based on the identification of possible conceptual dependencies among constituents according to the application ontology and the validation of the hypothesized semantic dependency relations with supporting syntactic relations. A frame language is used for representing semantic relations. Let V be a vocabulary of words that can be hypothesized by an ASR system. Let $C_1^K = C_1, \ldots, C_k, \ldots, C_K$, with $C_k \in V_C$ be a sequence of concept tags, belonging to the concept vocabulary V_C, that can be hypothesized from a sequence of word hypotheses in an initial interpretation step. Let σ_k indicate the generic *support* of C_k. The following example shows the concept tags and their supporting word sequences for a sentence of the French MEDIA corpus [9].

```
<I would like to make a reservation)>
[command(reservation)]
<a room> [object(room)]
<with an internet access>
[room_facility(internet)]
```

Concept tags are represented by attribute-value pairs between brackets. Word sequences between the brackets <> represent the sentence segment, called *concept support*, that expresses the concept.

A frame describes a general conceptual entity. An instance of it describes a particular realization of the frame. For the sake of simplicity, a frame will be indicated with its name. Let F_i be a frame introduced for describing the semantic knowledge of an application domain. Frame F_i has a name and a set of properties. The j-th property is represented by a slot $S_{i,j}$. The slot can be empty or filled with a value. A frame instance is characterized by the association of values with some or all its slots. Some values can be frame instances with slots having frame instances as values. Structures of this type with partially filled slots are *instance fragments*. An *instance fragment* $\Gamma_{i,a}$ is a semantic structure made of a frame name and a slot list represented as: $\Gamma_{i,a} = F_i.sl_{i,a}$ A slot list $sl_{i,a}$ is represented as follows as a set of slot names and slot values: $sl_{i,a} = [G_{i,1}v_{i,a,1}, \ldots, G_{i,L}v_{i,a,l}]$ where $G_{i,l}$ is a slot name and $v_{i,a,l}$ is its value. Each slot $s_{i,l}$ is associated with a *facet* which is a list of value types. Value types can be names of frames whose instance may be a possible slot value or another kind of value (date, digits, named entities,...).

Frame F_i is the *head* of the instance fragment. All the unfilled slots of the instance fragment are elements of the *fragment tail* list. Some slot values can be the results of methods associated to slots. They are executed by an interpretation process with a specific strategy. Each concept tag C_k characterizes a knowledge chunk and may correspond to one or more fragments of the application ontology, depending on some features of the words of its support and its context. Such a correspondence is established by the designer of the application ontology. The concept tags in V_C are part of the specifications for the annotation of the application corpora. The following example shows a fragment corresponding to a concept tag. The support of a fragment is the support of the corresponding concept hypothesis.

- Concept tag C_k
  `room_facility(internet)`
- Fragment Γ_k

```
ROOM.[r_facility
      FACILITY.[facility_type internet]]
```

Fragment Γ_k is a linear representation of a frame structure in which frame names are in capital letters, the list of slots for a frame is represented between brackets, its elements are separated by a comma and the value of a slot follows the slot name after a blank. When a value is a constant and not a frame instance, it is represented in italics. The relation between a frame name and its slot list is represented by a dot.

3. RULE-BASED COMPOSITION OF FRAME FRAGMENTS

This method is based on handcrafted knowledge: a set of logical inference rules has been defined that take as input all the words output by an ASR process and the semantic concepts C_k presented in the previous section. These concepts can be obtained with different methods, using grammars or a tagging approach. A set of different methods that can be used for concept tagging have been compared within the LUNA project, on the same corpus as the one used in this study [10].

The first steps in the processing of a speech utterance are the speech-to-text transcription process producing a string or a lattice of words and the word-to-concept process translating the sequences of words into sequences of concepts C_k. After these initial steps we map these concepts into Frame fragments as presented in section 2. These fragments correspond to a higher semantic representation than the concepts but they don't provide a structured semantic representation of what is said in a dialog turn. For example the concept `hotel-brand` can be mapped into the fragment: `LODGING.type.HOTEL.hotel_brand` where `LODGING` and `HOTEL` are frames and `type` and `hotel_brand` are Frame elements. This fragment expresses the fact that `HOTEL` is a `LODGING`. In this case and

according to the ontology we can associate more information to the HOTEL and start building relations between this fragment and other fragments occurring in the same turn.

Instance fragments are hypothesized from the sequence C_1^K by a table look-up procedure. When multiple choices are possible for a concept C_k, a finite set of instance fragments is associated to it and disambiguation is based on the features of the support σ_k and its context. Composition of instance fragments into semantic structures is performed by three generative rules for meaning representation in the frame language.

1. Composition by fusion.
   Two instance fragments with the same head can be composed into a single structure with each slot filled by a first-order logic expression of the values filling the constituent slots with the same name if there are no conflicts in the values and there is a syntactic support.
2. Composition by attachment.
   An instance fragment can become the value of another instance fragment if:
   - the latter has an empty slot whose facet contains the head of the former,
   - there is a syntactic support.
3. Composition by attachment and inference.
   The rule defines a composition by attachment in which the link between a slot facet and the head of a fragment is established by inference performed on the domain ontology.

Syntactic supports are expressed by template patterns involving classes of supports for fragments and features extracted in the window of words between pairs of them. For the MEDIA corpus, fragment supports are annotated in the corpus. Very simple template patterns were used in the rule-based approach. They describe relations between fragment supports using only distance and function words as features. The fragment composition process consists of the following steps: generation of the 1 best sequence of word hypotheses with an ASR system, generation of concept tag hypotheses, generation of instance fragments from concept tag hypotheses, generation of possible links between pairs of semantic fragments using the three composition rules, assembly of all compatible compositions into structures.

4. SEMANTIC COMPOSITION WITH CRFS

This approach is based on a 2-step process: the first step consists in tagging the words and the basic semantic concepts C_k with all the semantic information needed to remove most of the composition ambiguities; the second step composes these enriched concepts thanks to a set of straightforward composition rules. We enriched the word and concept strings with

word support	attibute	value
je veux réserver	command-task	reservation
une chambre	amount-room	1
à marseille	location-city	marseille
du deux	date	02/??/??
au cinq août	date	05/08
à paris	location-city	paris
du six	date	06/??/??
au huit	date	08/??/??

Table 1. Example of concept annotation with concept strings and values

4 different levels of information: *semantic specifiers*, *mode*, *connectors*, *reference link*. They are briefly presented in the next section.

4.1. Additional semantic information

Semantic specification
Each concept C_k is made of an attribute and a value. For example, the annotation of the sentence: "*je veux réserver une chambre à marseille du deux au cinq août et une autre à paris du six au huit*" ("*I want to book a room in marseille from the second to the fifth of august and another one in paris from the sixth to the eight*") is displayed in table 1, representing the entities contained in the sentence.

This concept annotation string contains ambiguities. Among them we can note:

- The city ambiguity: is Paris a correction or an addition to Marseille?
- The date ambiguities: we have four dates, which ones are the beginning and the end of a reservation?
- Is this all one or two reservations?

The goal of the semantic specification process is to add to these attribute/value tags a label called a *specifier* that can help removing the interpretation ambiguities. The main idea is that a combination of the specifiers and the attribute names can directly produce a hierarchical representation of a query from its flat annotation. For example, on the concepts of the previous example, the attribute location-city is going to be specified with the label -hotel and the dates with the double specifier begin-reservation or end-reservation.

Three other kinds of annotation are added in order to remove the remaining composition ambiguities: connectors, references and mode. ***Connectors***
The connectors are useful for obtaining the structure of a request. In the previous example the word "*et*" ("*and*") is a connector between the two reservation requests. There are two kinds of connectors: "connectAttr" for a connection

between attributes and "connectProp" for a connection between propositions. The value of a connector can be either "addition", "alternative" or "opposition".

References

Detecting references is useful for semantic composition, even if the reference link is not solved. In the previous example the reference "*une autre*" ("*another one*") indicates that the room we want to book is different from the first one and therefore these new entities "city-name" and "dates" are not corrections of the first ones but rather another request expressed by the caller. In this case the reference detected must be associated with a specifier "exclusion" indicating that the object referred is different from the one introduced.

Mode

The mode indicates if a concept is expressed in an affirmative, negative or interrogative way. In the previous example all the concepts are affirmative. However in some sentences this information is crucial in order to build a correct interpretation. For example the sentence: "*pas l'Hotel du Centre*" ("*not the Hotel du Centre*") contains the entity "hotel-du-centre" with a negative mode and in the sentence: "*y a t'il une piscine*" ("*is there a swimming pool*") the attribute "piscine" has an interrogative mode. In both cases the mode information influences the final Frame interpretation that can be associated with the sentences.

This additional semantic information is added to the concept annotation with a tagging approach presented in the next section.

4.2. Adding semantic specifiers to the concept strings and obtaining semantic frames

Adding this additional semantic information to a word and concept sequence is seen here as a tagging problem. We consider three different levels of label to predict: connectors and reference links, seen as additional concepts to predict; semantic specifiers, added to each concept predicted; and mode, also added to each concept. We use the Conditional Random Fields (CRF) [11] approach for our tagging process thanks to the *CRF++* tool[1]. CRF have been widely used for various word labelling tasks such as Part-Of-Speech tagging or Named Entity detection. CRF is a discriminant approach; it has proven to give better results on these tasks than generative HMM-based approaches. The main advantage of CRF is the ability to predict a word label according to a whole set of features related to the entire message, and not just the short history of the word to tag. This is very important for the task of adding specifier and mode labels to concepts as this information depends on features that can be far away from the concept to tag in the message.

We train one CRF for each level of labels to predict. The training corpus of these CRFs is obtained on a dialogue corpus semantically annotated with Frames, following the model presented in section 2. Each message is a sequence of features (words, attributes, values), labelled with a specifier label extracted from the Frame structure or the symbol "NULL". At decoding time each word/concept sequence hypothesis of the structured n-best list of concept strings output by the concept tagger is processed by the CRF in order to add these specifier labels. The same strategy is deployed for adding mode labels, and detecting connectors and reference links. Once the concepts are enriched with this additional semantic information, a simple semantic parser is in charge of building the final frame representation. Since the semantic ambiguities have been removed by all these additional semantic labels, this step is straightforward: a set of handcrafted composition rules is used for projecting the semantic annotations into the frame representation and linking the different Frames, Frame elements and sub-Frames together.

5. EXPERIMENTAL RESULTS WITH THE MEDIA CORPUS

The application used in this study is the hotel booking application defined in the MEDIA project [9]. The task is the reservation of hotel rooms with tourist information, using information obtained from a web based database. The MEDIA dialog corpus was recorded using a WOZ system simulating a vocal tourist information phone server. In this way, each user/caller believes she or he is talking to a machine whereas she or he is talking to a human being (a wizard) who simulates the behaviour of a tourist information server. Variations in the dialogue strategy in MEDIA are due in part to the behaviour of the wizard. 1250 dialogs were recorded, from 250 different speakers where each caller carried out 5 different hotel reservation scenarios. Several starting points were possible for the dialogs i.e. choice of town, itinerary, tourist event, festival, price, date etc. Eight scenario categories were defined each with a different level of complexity. The final corpus is about 70 hours of dialogs, which have been transcribed and annotated at the concept level by ELDA. This corpus is available through ELDA. During the LUNA project[2] we have added to the MEDIA corpus Frame annotations following the model presented in section 2. The MEDIA corpus is split in three sets of dialogues: a training, development and test sets.

The rule-based Frame parser of section 3 has been developped on the training set, with the reference transcriptions and concept annotations already available in the corpus. Then, this parser has been applied to the whole MEDIA corpus (train+dev+test) in order to obtain a corpus annotated with semantic frames. The development and test corpora were manually corrected with the Frame editor SALTO[3] in order to have a gold standard annotation at the Frame level for evaluating our approaches. An example of such an annotation is given in figure 1. The CRF models of the approach presented

[1] http://crfpp.sourceforge.net/

[2] www.ist.luna.org

[3] http://www.coli.uni-saarland.de/projects/salsa

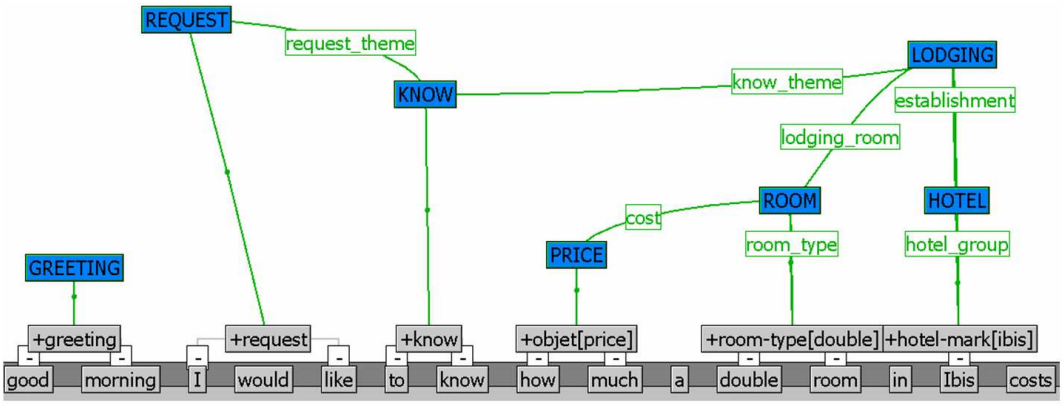

Fig. 1. Frame representation, projection from FrameNet to Media

in section 4 have been trained on the MEDIA train corpus automatically annotated with Frames. Experiments presented here have been conducted on the test set described in table 2.

# dialogs	208
# dialog turns	3005
# words	25.5K
# concept/value occurrences	9131
# different concepts	73
# different values	603
# frame occurrences	9500
# different frames	38

Table 2. Description of the MEDIA test corpus annotated with semantic frames

To compare the different approaches proposed in this study, we use the precision, recall and F-measure on the following evaluation measures:

1. Frame Recognition. Rate of individual correct frames (disregarding their related arguments).
2. Argument Boundary Detection. Of exact word-level boundaries for individual arguments, disregarding their labels.
3. Argument Labeling Detecting of the correct argument labels, disregarding their boundaries.
4. Argument Recognition: evaluates the composition of the tasks defined in 2 and 3,
5. Frame Realization: composition of 1+2+3.
6. Frame Composition Frame-to-Frame links/relations.
7. Turn Analysis: composition of 5+6.

The two frame annotation systems presented in this paper have been evaluated on the speech files of the MEDIA test corpus with the following protocol: firstly the speech files of each dialog turn are automatically transcribed with the ASR system developped at the University of Avignon (*SPEERAL*); secondly the CRF concept tagger presented in [10] is applied on the automatic transcriptions in order to produce the concept sequences; finally we apply the frame annotation systems directly on these automatic transcrptions and concept sequences. The word error rate (WER) of the ASR module is 27.4% and the concept error rate (CER) of the sequences of concepts is 31.3%.

Level	Metrics	Prec.	Recall	F1
1.	Frame Recognition	88.45	82.44	85.34
2.	Argument Boundary Det.	83.44	75.89	79.49
3.	Argument Labeling	86.24	79.54	82.75
4.	Argument Recognition	80.85	74.57	77.58
5.	Frame Realization	74.62	69.55	72.00
6.	Frame Composition	81.46	68.47	74.40
7.	Turn Analysis	54.99	54.33	54.66

Table 3. Results of the rule-based system on the test corpus with reference transcriptions (WER=0)

Level	Metrics	Prec.	Recall	F1
1.	Frame Recognition	85.15	82.98	84.05
2.	Arg. Boundary Det.	96.58	88.19	92.20
3.	Argument Labeling	82.89	77.45	80.08
4.	Argument Recognition	77.48	72.40	74.85
5.	Frame Realization	69.13	67.36	68.23
6.	Frame Composition	80.25	69.06	74.24
7.	Turn Analysis	52.64	52.05	52.35

Table 4. Results of the CRF system on the test corpus with reference transcriptions (WER=0)

Tables 3 and 4 summarize the results obtained with both systems on the reference transcriptions and tables 5 and 6

Level	Metrics	Prec.	Recall	F1
1.	Frame Recognition	78.15	75.90	77.01
2.	Arg. Boundary Det.	75.41	66.22	70.52
3.	Arg Labeling	76.24	70.74	73.39
4.	Arg Recognition	70.06	65.01	67.44
5.	Frame Realization	59.17	57.46	58.30
6.	Frame Composition	72.73	62.01	66.94
7.	Turn Analysis	46.37	45.46	45.91

Table 5. Results of the rule-based system on the ASR test corpus (WER=27.4)

Level	Metrics	Prec.	Recall	F1
1.	Frame Recognition	78.37	77.28	77.82
2.	Arg. Boundary Det.	87.21	76.99	81.78
3.	Argument Labeling	76.78	72.05	74.34
4.	Argument Recognition	70.55	66.20	68.31
5.	Frame Realization	59.87	59.03	59.45
6.	Frame Composition	73.56	63.73	68.29
7.	Turn Analysis	46.77	45.98	46.37

Table 6. Results of the CRF system on the ASR test corpus (WER=27.4)

present the results on the ASR transcriptions. As we can see the rule-based approach is better than the CRF approach on the reference transcriptions. This result is expected as the CRF models are trained on a corpus automatically labeled with the rule-based parser. However, on the ASR transcriptions, the CRF approach appears to be more robust to ASR errors and obtain better F-measure for every evaluation level. The huge gain in the second evaluation level (*argument boundary detection*) obtained with the CRF approach is due to an artifact in the evaluation process (the rule-based parser don't keep the word boundaries of each concept, unlike the CRF tagger, and it's these concept boundaries that are used as *argument boundaries* in the evaluation process).

6. CONCLUSION

Two approaches to SLU based on frames describing chunked knowledge have been described. They have been applied to the MEDIA corpus annotated in terms of concepts expressing chunks of spoken sentences. General rules of knowledge composition and inference appear to be adequate to effectively applying the application ontology for obtaining frame based representations of dialogue turns. The main difficulty appears to be the characterization of the syntactic knowledge expressing semantic links between knowledge chunks. This knowledge can be hand-crafted or automatically learned from examples. It is shown that the latter approach outperforms the former if applied to ASR error prone transcriptions.

7. REFERENCES

[1] W.A. Woods, L. Bates, G. Brown, C.C. Cook, and BC Bruce, "Speech understanding systems," 1976.

[2] S. Seneff, "TINA: a probabilistic syntactic parser for speech understanding systems," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1989, pp. 168–178.

[3] Y. He and S. Young, "Hidden vector state model for hierarchical semantic parsing," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, 2003.

[4] Y.Y. Wang and A. Acero, "Combination of cfg and n-gram modeling in semantic grammar learning," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[5] M. Lease, E. Charniak, and M. Johnson, "Parsing and its applications for conversational speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, 2005.

[6] L.S. Zettlemoyer and M. Collins, "Learning context-dependent mappings from sentences to logical form," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 976–984.

[7] R. Pieraccini, E. Levin, and C.H. Lee, "Stochastic representation of conceptual structure in the ATIS task," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992.

[8] R.E. Fikes and N.J. Nilsson, "STRIPS: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.

[9] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[10] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, "A comparison of various methods for concept tagging for spoken language understanding," *Proceedings of LREC, Marrakech, Morocco*, 2008.

[11] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*. 2001, pp. 282–289, Morgan Kaufmann, San Francisco, CA.