A CONDITIONAL MODEL FOR TRIGGERING UNDERSTANDING ACTIONS IN A SPEECH UNDERSTANDING SYSTEM

Frederic Duvert¹ and Renato De Mori¹,²

¹ LIA - University of Avignon, FRANCE
² Mc Gill University, Canada

(frederic.duvert@univ-avignon.fr, rdemori@cs.mcgill.ca)

ABSTRACT
A conditional model is introduced for triggering understanding actions that correct errors of frame hypothesization and composition. Experimental evidence is provided using the French MEDIA corpus that these models trained with automatic speech recognition hypotheses trigger effective corrections of more than half of the errors. The overall frame recall increases from 0.76 to 0.84 while precision increases from 0.78 to 0.85. The number of fully corrected dialog turns increases of 8.8% making about half of the turns fully correct.

Index Terms — Speech understanding, Frames, Semantic composition, Interpretation error correction.

1. INTRODUCTION
Exponential models have been recently used for probabilistic frame semantic parsing. A noticeable example for obtaining resources for a FrameNet lexicon can be found in [1] where target identification is performed by deciding which word sequence evokes frames in a written sentence. The problem considered in this paper is to use an automatic speech recognition (ASR) system for obtaining word hypotheses from which interpretations represented by frame structures are generated. The problem is known to be difficult because the generation of word hypotheses is error prone and linguistic models for interpreting spoken language have to take into account specific events not present in written text. For these reasons, efforts are concentrated on application domains described by specific ontologies.

The interpretation models used in the Spoken Language Understanding (SLU) process should deal with possibly ungrammatical sentences that may contain incorrect word hypotheses. This suggests to use specific methods for processing chunks of words to hypothesize knowledge chunks rather than attempting to fully parse a sentence as in the case of text processing for which a recent, elegant solution can be found in [2].

Knowledge chunks can be represented by descriptions in a frame language as suggested in [3]. An SLU process is proposed in this paper in which semantic structure hypotheses are composed from knowledge chunks. These hypotheses are then verified and possibly corrected using a new set of understanding actions triggered by automatically trained conditional models. Experiments with the French MEDIA corpus [4] on a baseline described in [5] are reported in this paper. They show a considerable interpretation error reduction when training is performed by assigning action labels to sequences of ASR word hypotheses.

Section 2 of this paper introduces semantic structures. Section 3 describes the SLU process. Section 4 introduces the conditional model for performing SLU actions and Section 5 reports experimental evidence on the benefits of the proposed approach.

2. SEMANTIC STRUCTURES
Semantic structures are now briefly introduced. A more comprehensive introduction of them can be found in [6]. A semantic structure is a data model for meaning. In order to be used in a computer program, models for structured concepts are described with computational schemata called frames. Theoretical foundations of frame languages are presented in [7]. A frame has a name and a set of properties. For example a PERIOD frame may have start, date and duration as properties. Let $F_d$ be the name of a frame. The $i^{th}$ property of $F_d$ is represented by a slot indicated as $s_{d,i}$. Frame structures were proposed in [3] for representing fragments of real world knowledge (chunked knowledge).

The understanding process instantiates frames based on the analysis of spoken utterances. This process extracts predicates and values that describe a particular realization of a semantic structure. Some values can be frame instances with slots having frame instances as values. An instance fragment $\Gamma_{d,a}$ is a semantic structure made of a frame name and a slot list and is represented as: $\Gamma_{d,a} = F_d, s_{ld,a}$. A slot list $s_{ld,a}$ is represented as follows as a set of slot names $G_{d,t}$ and slot values $v_{t,d}$:

$\ldots s_{ld,a} = [G_{d,1}v_{d,a}, \ldots, G_{d,t}v_{d,t}, \ldots, G_{d,L}v_{d,L}]$.

Each slot $s_{d,t}$ is associated with a set $\Phi_{d,t}$ called facet and represented as follows:
\[ \Phi_{d,\ell} = \{ \phi_{d,\ell,m} \}\ \phi_{d,\ell,m} = F_q \text{ or } V_q. \]

\[ F_q \text{ is a frame name while } V_q \text{ is another type of value.} \]

Frame \( F_d \) is the head of the instance fragment. All the unfilled slots of the instance fragment are elements of the fragment tail list. If an instance \( \Gamma_{q,z} \) of a frame \( F_q \) is the value of \( G_{d,\ell} \) of \( \Gamma_d \), and \( s_{q,n} \) is a slot of \( \Gamma_{q,z} \), then there exists a relation between \( F_q \) and \( s_{q,n} \) represented by a slot chain indicated as:

\[ \delta_n = F_d \cdot G_{d,\ell} - F_q \cdot s_{q,n}. \]

The first frame name is the head of the chain and the last slot name is the chain tail. An instance fragment is a set of slot chains.

3. SLU PROCESS

3.1. Knowledge chunks

Knowledge chunks are initially hypothesized with a process described in [8] from non-overlapping sequences of word hypotheses generated by an ASR system. Let \( V \) be the word vocabulary and \( V_C \) be the vocabulary of concept tags describing knowledge chunks. Let \( \sigma_k \) indicate the sequence of words, called support, expressing chunk \( C_k \) in a sentence. Supports defined here evoke instance fragments and are similar to targets that evoke frames in FrameNet semantic parsing [1].

A concept tag \( C_k \) hypothesized in a sentence corresponds to an instance fragment. Such a correspondence is established by selecting an element of a list associated with \( C_k \), based on the support of \( C_k \) and its context. The following example from the MEDIA [4] corpus shows a fragment corresponding to a concept tag.

Concept tag \( C_k : \text{night number(y).} \]

Fragment \( \Gamma_q : \text{PERIOD}[\text{dur DURATION}.[\text{nb night NUMBER}],[\text{type cardinal, value y}]] \]

Fragment \( \Gamma_q \) is a linear representation of a frame structure in which frame names are in capital letters and the list of slots for a frame is represented between brackets. A frame instance represents an n-ary semantic relation between a frame name, its slots and their values. Semantic structures establishing n-ary relations between semantic constituents can be composed using binary relations. Composition of instance fragments into semantic structures is performed by identifying possible semantic binary relations between pairs of fragments and validating each possibility with supports in a spoken sentence. Supports for these relations are represented by templates containing features of the support constituents and their links. Links are often represented by local syntactic relations connecting, for example, a noun phrase expressing a tag with another noun phrase expressing another tag. The connection can be expressed, for example, by a preposition.

3.2. Semantic composition

Three types of instance frame composition have been considered, namely composition by fusion of two fragments with the same head, composition by attachment of a fragment as a value of a slot, and composition of inference. Details can be found in [6]. For the sake of brevity, only composition by attachment will be described here.

Given an instance fragment \( \Gamma_d \) with head \( F_d \). Let us consider the set \( \Delta_d \) of all its slot chains having \( F_d \) as a head and ending with an unfilled slot. Let \( \delta_z \in \Delta_d \) be one of these chains.

Let assume that two instances \( \Gamma_d \) and \( \Gamma_q \) have been hypothesized. Let \( F_q \) be the frame name which is the head of \( \Gamma_q \). Let the tail of \( \delta_z \) be an unfilled slot having name \( B_{q,z} \) and facet \( \Phi_{q,z} \). Let \( \Phi_{q,z} \) include the frame name \( F_q \). A binary semantic relation involving \( B_{q,z} \) can be asserted between \( \Gamma_d \) and \( \Gamma_q \) if there is a support for their link in the data from which the instances \( \Gamma_q \) and \( \Gamma_q \) have been hypothesized. The predicate \( \text{supp}[\text{link}(B_{q,z},\Gamma_q)] \) is true when such a support is detected. This semantic relation assigns \( \Gamma_d \) as a value for \( B_{q,z} \). The composition action is described with the predicate \( \text{link}(B_{q,z},\Gamma_q) \) asserted by the following inference rule:

\[ \text{contains}(\Phi_{q,z},F_q)\land \text{supp}[\text{link}(B_{q,z},\Gamma_q)] \Rightarrow \text{link}(B_{q,z},\Gamma_q) \]

The predicate \( \text{contains}(\Phi_{q,z},F_q) \) is true when the facet \( \Phi_{q,z} \) contains \( F_q \). The predicate \( \text{supp}(x) \) is true when a template pattern \( \pi_k(x) \) associated with it matches with the available hypotheses.

In the MEDIA corpus, semantic entities called specifiers have been used for indicating possible composition links. As an example, let us consider the following sentence: "Un hotel Marseille pour trois nuits" (A hotel in Marseille for 3 nights). The hypothesized fragments are:

- \( \Gamma_d : \text{HOTEL}.[\text{at loc ADDRESS}.[\text{adr city Marseille}]] \]
- \( \Gamma_q : \text{PERIOD}.[\text{dur DURATION}.[\text{nb night NUMBER}],[\text{type cardinal, value 3}]] \]
- \( \text{Slot chain} : \delta_q : \text{HOTEL}.\text{for time}; \)
- \( \text{tail}(\delta_q) : B_{q,z}=\text{for time}; \Phi_{q,z} = \{ \text{PERIOD, DATE} \}; \)
- \( \text{head}_q = \text{PERIOD}; \text{head}_q \in \Phi_{q,z} \)
- \( \text{Fragment supports} : \sigma_1: [\text{HOTEL}], \sigma_2: [\text{PERIOD}] \)
- \( \text{supp}[\text{link}(B_{q,z},\Gamma_q)] = \text{match}<\sigma [\text{HOTEL}], \sigma [\text{PERIOD}]>\text{, data} \)

Relation support is generalized by considering \( \sigma[\text{HOTEL}] \) as an abstraction of all sequences of words expressing a fragment having \( \text{HOTEL} \) as head and \( \sigma[\text{PERIOD}] \) as an abstraction of all supports for \( \text{PERIOD} \). The semantic relation is simply supported by the contiguity of the supports of the constituents. The result of composition is:

\( \text{HOTEL}.[\text{at loc ADDRESS}.[\text{adr city Marseille}], \text{for time PERIOD}.[\text{dur DURATION}.[\text{nb night NUMBER}],[\text{type cardinal, value 3}]]] \)
4. UNDERSTANDING ACTIONS

Composition by attachment is performed by an understanding action triggered by an inference rule that is application independent. Two similar rules are used for the other two types of composition, namely by fusion and by inference. Possible compositions are generated by using simple matching patterns for asserting $\text{supp}(x)$. These hypotheses are verified by a conditional model that trigger understanding actions based on hypothesized frame sequences and the sequence of words from which the hypotheses have been generated. The effects of these actions are modifications of the hypothesized structures by removing insertions or performing further compositions.

Let $W = w_1, ..., w_n, ..., w_N$ be a sequence of word hypotheses and $A = a_1, ..., a_n, ..., a_N$ be a sequence of understanding action symbols. Let $a_n$ indicate the understanding action involving the instance fragment whose support contains $w_n$. An example of action is to identify the head of a frame instance that has been inserted or could have been composed with already hypothesized structures but a support for the composition was not found. Let $\alpha_1$ be the symbol of this action and $\alpha_0$ be the symbol for a NULL action. Let $\Gamma$ be the set of hypotheses generated from $W$.

The conditional model uses conditional random fields [9] for computing probabilities of actions given a sequence of word and frame hypotheses as follows:

$$P(a_1^N | \Gamma, W) = \frac{1}{Z} \exp \left\{ \sum_{n=1}^{N} \sum_{q=1}^{Q} \gamma_q g_q(a_n, v_{n+2}^n) \right\}$$

Words and frame features like names and possible links can form two parallel streams. For specific tasks, like the triggering of action $\alpha_1$, the two streams can be fused. Let $F_k$ be the name of a frame that is the head of an instance fragment $\Gamma_k \in \Gamma$ hypothesized with support $\sigma_k$ made of a sequence of words in $W$. The input variables $v_n$ associate frame names with words as follows:

$$v_n = \begin{cases} F_k w_n & \text{if } w_n \in \sigma_k, \forall \Gamma_k, \Gamma_k \in \Gamma \\ w_n & \text{otherwise} \end{cases}$$

the exponent functions $g_q(a_n, v_{n+2}^n)$ assert the presence of n-grams of the input variables. When the train set is not large, coding the inputs in this way has been found to be more efficient than having two separate input streams. The result of the execution of an action is a new semantic structure $\Gamma_b$. Assuming that it can be obtained by multiple actions on multiple semantic hypotheses, it is scored by the probability:

$$P(\Gamma_b | W) = \sum_i P(A_i | \Gamma_i, W) P(\Gamma_i | W)$$

where $P(\Gamma_i | W)$ is computed with another exponential model.

5. EXPERIMENTAL RESULTS

A system baseline for the MEDIA corpus has been built [5] in which the predicate $\text{supp}(x)$ is asserted if simple conditions between the supports of the concept tags expressing the components are verified. The objective of the experiments is to reduce errors in the head frames of the hypothesized semantic structures by applying action $\alpha_1$.

The experiment consists in automatically learning from an annotated portion of the MEDIA corpus, when to perform action $\alpha_1$. The annotation consists in association action labels to frame instances that should be removed or composed with existing structures. In order to have enough data, CRF training with the CRF++ tool\(^1\) was performed on a third of the test set containing 823 of the 3005 dialogue turns. The test was performed on the remaining 2182 turns.

Figure 1 shows an example of a semantic structure. Nodes correspond to frame heads, arcs are labeled with slot names and concept tags are listed above words. An instance of DATE has been detected. It is caused by the wrong insertion of a monosyllabic word in French not compatible with the context. The ASR word error rate (WER) on the test set is 27.4% and the concept tag error rate (CER) is 31.3%.

The execution of $\alpha_1$ indicates the presence of frame instances that appear to be inconsistent with the context. They may depend on the limits of patterns for support of composition links. ASR errors inserting words expressing semantic concepts such as numbers and references, poor resolution of repetitions, false starts and self-corrections. Action $\alpha_1$ may trigger new compositions or other operations involving search in the trellis of word hypotheses. Characterizing this operation is the object of future research.

Table 1 shows the results when the triggering of action $\alpha_1$ is automatically learned. ASR word hypotheses are used to produce test inputs. A comparison is made on the effect of $\alpha_1$ on the test data when training is performed with manual word annotations and with the one-best sequence of word hypotheses generated by the ASR component. The number of frames on which action $\alpha_1$ should be applied in the test set is 907. The results clearly show that CRFs trained with ASR hypotheses trigger correction of more than half of the errors, effectively correcting many effects of ASR errors. Training on annotations is not very effective because interpretation of annotations has few errors. The overall frame recall increases from 0.76 to 0.84 while precision increases from 0.78 to 0.85. The number of fully corrected turns increases of 8.8% making about half of the turns fully correct.

\(^{1}\)http://crfpp.sourceforge.net/
Fig. 1. Example of an error consisting in the instance of DATE not compatible with the context

Table 1. Results of the application of action on the test data produced with ASR word hypotheses when training is performed with manual word annotation and with the one-best sequence of word hypotheses generated by the ASR component.

<table>
<thead>
<tr>
<th></th>
<th>training corpus</th>
<th># corrected frames</th>
<th># false alarms</th>
<th># fully corrected turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>annotations</td>
<td>59/907</td>
<td>21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>ASR hyp</td>
<td>580/907</td>
<td>126</td>
<td>101</td>
<td></td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

A sequential SLU process has been described in which understanding actions can be triggered to correct semantic fragments that appear to be not coherent with the semantic context. It is shown that when most of the frame instance errors are caused by ASR errors, a considerable number of corrections can be made by learning from ASR hypotheses when to trigger understanding actions. Future research will investigate when suitable understanding actions should be performed for finding more semantically plausible and syntactically coherent fragment hypotheses in a trellis of word and concept tag hypotheses.

7. REFERENCES


