Dialogue systems

Luděk Bártek

Laboratory of Searching and Dialogue, Fakulty of Informatics, Masaryk University, Brno

spring 2023

Speech Recognition

- Continuous speech recognition – transforms continuous speech into the corresponding text.
- Commend/isolated words recognition.
- Speech recognition principles:
  1. Acquire feature vectors using a short-term signal analysis method.
  2. Classify the speech using the feature vectors from previous step.

Continuous Speech Recognition

- The principle differences to the isolated word recognition:
  - the pattern database can not be created
  - the prosodic factors must be taken into the account
  - there is a need to separate the words (find the words borders)
  - the algorithm must deal with filler sounds and speech errors.
- Solution – statistical approach:
  - a language model
  - a user model.
- Example: HMM returns the same probability for Czech words "máma" (mother) and "nána" (not a very smart girl) – mother will be used with higher probability – is used often.

- There are:
  - a sequence of words (utterance) $W = (w_1, \ldots, w_n)$
  - a sequence of acoustic vectors $O = (o_1, \ldots, o_t)$.
- We want to find W^* (set of all utterances), that maximize $P(W|O)$.
- According to the Bayes' rule:

$$P(W^*|O) = max\, P(W|O) = max\, \frac{P(W) * P(O|W)}{P(O)}$$

Continuous Speech Recognition

Language Models – cont.

- We need following to calculate the maximum of probability:
  - a speaker model – $P(O|W)$
  - a language model – $P(W)$.
- The speaker model can be replaced by the probability of generating the W using the corresponding Markov Model
- The trigram model:
  - Experimentally evaluated that:

$$P(w_n|w_1 \dots w_{n-1}) \cong P(w_n|w_{n-2}w_{n-1})$$

.

Continuous Speech Recognition

A Topic Recognition

- The continuous speech recognition is in a range 50 % — 99 % in dependency of a task, a language, ...
- The recognition success rate can be improved by limiting the recognition domain:
  - a topic recognition
  - using speech recognition grammars.
- The well-known topic:
  - a change of the state space and trigrams probabilities:
    - for example in stock market news was recognized either "honey" or "money"?
  - a more accurate language model can be developed.

Speech Recognition Grammars

- A general continuous speech recognition accuracy can drop to 50 % approx.
- The improvement can be reached by the recognition domain restriction – a specification of accepted inputs for example.
- The speech recognition grammars can be used:
  - context-free grammars
- Used grammars notations:
  - a logic programming notation
  - proprietary solutions
  - open standards – JSGF, W3C SRGS, ...

Speech Recognition Grammars

Java Speech Grammar Specification (JSGF)

- Platform and reseller independent textual grammar format.
- Used for a speech recognition.
- Part of the Java Speech API.
- Uses the Java language style and conventions.
- Present version 1.0 (Oct 1998).
- Used in the Použit např. v rozpoznávači Sphinx-4 recognizer, VoiceXML interpreter VoiceGlue, . . .
- More details in the 2nd half of semester – a dialogue interface creation tools.

Speech Recognition Grammar

JSGF example

#JSGF
<root> = I want to go by <what> .|
I want to go by <what> from <where> to <where> .|
I want to go by <what> from <where> to <where> at
<when> .;
<what> = train | bus;
<where> = <city>;
<when> = <dateTime>;

Speech Recognition Grammars

W3C Speech Recognition Grammar Specification (SRGS)

- The W3C standard.
- Current version 1.0 (Mar 2004).
- Defines the notation of rules and their referencing.
- Two types of the notation:
  - XML
  - ABNF (Augmented BNF).
- More details on 2nd half of semester – the topic dialogue interface creating.

W3C SRGS Example


```
#ABNF 1.0 UTF-8
root $greeting;
language en-GB;
mode voice;
$greeting = hi
```

```
<?xml version="1.0" encoding="utf-8" ? >
<grammar root="greating" xml:lang="en-GB" version="1.0" >
<rule id="greating" >
hi
< /rule>
< /grammar>
```

Utterance Semantic Interpretation

- Objective – computer understandable interpretation of the informations entered by user.
- Example:
  I want to buy the Shakespeare's The Taming of the Shrew.
  - action = shopping
  - title = The Taming of the Shrew
  - author = Shakespeare
- Representation – the (attribute, value) pairs.
- General semantic analysis steps:
  1. acquiring the utterance structure (syntactic analysis)
  2. the part of the speech interpretation
  3. deriving the whole utterance interpretation from the parts of the speech interpretations.
- The utterance semantic interpretation $\neq$ the utterance intended sense (pragmatic interpretation).

Utterance Semantic Interpretation

Implementation

- Attributes containing a part of speech semantic interpretation are assigned to the speech recognition grammar rules.
- We can perform operations used to derive the semantic interpretation of the entire utterance from the interpretations of its parts.
  - The ECMAScript (see Semantic Interpretation for Speech Recognition can be used.
- To find the intended utterance meaning we need to process its context as well.
  - The context can be described using finite automaton with output (the Mealy automaton – see later).

Semantic Interpretation Description

- JSGF:
  - Assigned using tags
  - notation – {semantic interpretation}

  $< sentence >=< intro >< title >$ od $< author >$
  $< titlie >=$ Pejska a kočičku
  {Povídání o pejskovi a kočičce}|
  (Zlou ženu|Zkrocení zlé ženy) {Zkrocení zlé ženy}|...
- SRGS – the SISR standard :
  - standard W3C Voice Browser Activity.
  - uses ECMAScript.
  - The semantic interpretation is added to the rules using the *tag* tag or attribute.
  - The semantic interpretation is is return back using the JSON notation.
- ...