

2^e Deeltentamen Spraakherkenning en -synthese

Rob van Son

10-13 uur, 20 december 2007
GEBP/P2.27

Vermeld op iedere pagina je naam, je studentnummer en het volgnummer per pagina. Gebruik voor elke opgave (1-3) een apart vel. Als je voor 13.00 uur klaar bent, lever dan je tentamen in bij de surveillant en verlaat rustig de zaal.

Het cijfer voor dit schriftelijk deeltentamen bepaalt de helft van je eindcijfer voor dit college. Dat eindcijfer wordt echter pas aan het OWI doorgegeven wanneer je ook alle practicumverslagen voor dit onderdeel hebt voltooid en ze zijn goedgekeurd. Nadere informatie hierover vind je op de Blackboard site.

Beantwoord onderstaande vragen zo nauwkeurig, compleet en gedetailleerd mogelijk, zonder in herhalingen te vervallen. De vragen zijn zo gesteld dat een kort antwoord volstaat. Reken op 5 minuten gemiddeld per sub-onderdeel als je het antwoord weet. Een erg lang antwoord is bijna zeker een verspilling van tijd. Het kan geen kwaad relevante ervaringen vanuit het hoorcollege en de practicumopdrachten in je antwoorden ter verwerken. Bij de beoordeling van dit tentamen wordt meer gelet op begrip dan op feitenkennis. Besteed niet te veel tijd aan afzonderlijke opgaven. Als een opgave te veel tijd vergt, probeer dan eerst een andere opgave.

Veel succes!

Rob van Son

1) Automatische Spraakherkenning

1.a) Bayesian inference

$$\hat{W} = \underset{Words}{\operatorname{argmax}} P(Words|Observation) = \underset{Words}{\operatorname{argmax}} \frac{P(Observation|Words) \cdot P(Words)}{P(Observation)}$$

Centraal in Automatische spraakherkenning staat de bovenstaande formule waarbij $\hat{W}$ de meest waarschijnlijke zin is, $Words$ een kandidaat zin en $Observation$ het waargenomen geluid. Leg uit wat de verschillende componenten van deze formule betekenen, en wat de hele formule beoogt.

1.b) Taalmodel

Leg uit wat een Finite-State en een N-gram taalmodel zijn. Wat zijn de verschillen?

Waarvoor worden taalmodellen gebruikt in de Automatische Spraakherkenning? Hoe kun je deze taalmodellen trainen en wat voor soort corpora heb je daarvoor nodig?

1.c) Tweekaps decodering



Figuur 1.c): Tweekaps N-beste decoder

In de figuur 1.c) zie je het schema van een tweekaps decodering. Hoe werkt dit wanneer je het vergelijkt met de standaard, eentraps, herkenner? Wat wordt bedoeld met de term “N-best”?

Wat zijn de extra mogelijkheden ten opzichte van een éentraps herkenner?

1.d) Trainen van akoestische modellen

Bij het maken van een spraakherkenner voor in de auto worden de sprekers en de geluidstechnici met de auto de straat opgestuurd voor het maken van opnamen. Waarom? Waar kunnen ze het beste gaan rijden?

1.e) Akoestische features

Bij de spraakherkenning wordt intern niet de gewone golfvorm van het geluid of het spectrum gebruikt. Wat voor een representatie wordt er wel gebruikt?

2) Dialoogsystemen

C₁: ... I need to travel in May.
A₁: And, what day in May did you want to travel?
C₂: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A₂: And you're flying into what city?
C₃: Seattle.
A₃: And what time would you like to leave Pittsburgh?
C₄: Uh hmm I don't think there's many options for non-stop.
A₄: Right. There's three non-stops today.
C₅: What are they?
A₅: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C₆: OK I'll take the 5ish flight on the night before on the 11th.
A₆: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C₇: OK.

Figure 19.4 Part of a conversation between a travel agent (A) and client (C).

2.a) Wat voor een type dialoog is dit?

Waarin onderscheidt dit type dialoog zich van gewone conversatie? Wat is het belang van dit onderscheid voor de spraaktechnologie?

2.b) Speech acts

Wat zijn *Speech Acts*? Geef een paar voorbeelden aan de hand van de bovenstaande dialoog.

2.c) Basic control tasks

Wat wordt bedoeld met een *Minimal Response*? En wat met *Grounding*? Geef van beiden voorbeelden. Worden ze gebruikt in de bovenstaande dialoog? Zo ja, hoe?

2.d) Inferences

Grice's Maxims kunnen worden gebruikt om ambigue of onduidelijke uitingen te begrijpen. Geef twee voorbeelden uit deze dialoog waar een deelnemer Grice's Maxims zou kunnen hebben gebruikt om de bedoeling van de andere spreker te bepalen. Geef ook aan om welke Maxims het gaat.

2.e) OVIS training

Bij de constructie van OVIS (Openbaar Vervoer Informatie Systeem) werd begonnen met een bestaand corpus van over de telefoon opgenomen spraak (het Polyphone corpus). Geef aan hoe de OVIS herkenner werd getraind voor zijn taak. Dat wil zeggen, beschrijf welk spraakmateriaal werd gebruikt, hoe men het verzamelde en hoe men voorbeelden van interacties verzamelde voor een taalmodel.

2.f) Beurtwisselingen

We kennen de volgende regels voor wie als volgende mag spreken in een conversatie (geciteerd uit het Engels):

- If during this turn the current speaker has selected *A* as the next speaker then *A* must speak next
- If the current speaker does not select the next speaker, *any* other speaker may take the next turn
- If no one else takes the next turn, the *current* speaker may take the next turn

Dialogsystemen kunnen deze regels gebruiken om een conversatie te sturen door middel *Controlled Turn Switches*,. Leg uit wat dit zijn en hoe ze gebruikt worden. Geef voorbeelden uit de bovenstaande dialoog.

3) Template based speech recognition

3.a) *Structure-based vs Template-based*

Waarin verschillen de *Structure-based* en *Template-based approach* van de spraakherkenning van elkaar?

3.b) Herkenning

Welke techniek wordt in *Template-based speech recognition* gebruikt om de eigenlijke spraakherkenning uit te voeren? Dwz, hoe worden geluidsfragmenten met elkaar vergeleken? Is dit anders dan in HMM spraakherkenning?

3.c) Wat is een template in *Template-based speech recognition*?

Welke informatie bevat een template en hoe wordt die verzameld?

3.d) Hoe construeer je een corpus voor *Template-based speech recognition*

Geef een "kookboek recept".

3.e) Spreekstijlen

Standaard HMM based recognizers kunnen moeilijk overweg met veranderingen in spreekstijl of spreektempo. Leg in het kort uit hoe *Template-based speech recognition* dit oplost.

3.f) Wat is de functie van *clustering* in *Template based speech recognition*?

Geef de relatie tussen het clusteren van templates en het zoekalgoritme dat het best passende template moet vinden dat kan volgen op het laatst gevonden template.