Taal- en spraaktechnologie 2011

Grafeem-foneem omzetting

Michael Moortgat

Samenvatting

Het teamwerk van Week 3 is het omzetten van geschreven tekst naar een foneemrepresentatie. Die foneemrepresentatie op haar beurt wordt de input voor de difoonsynthese later in de cursus.

Op de cursuspagina vind je de vertaalsleutel tussen schriftsymbolen en het gebruikte foneemalfabet. Hieronder enkele tips voor de te bouwen transducer.

1. Determinisme

De taak komt erop neer dat je de lettergreep herkenner van Week 1 ombouwt tot een transducer: een machine die op zijn invoerband de eenlettergrepige woorden herkent, en op de uitvoerband hun vertaling in het foneemalfabet geeft.

Determinisme Je wil dat de vertaling voor een gegeven inputwoord een unieke omzetting produceert. Anders dan bij FSA is het niet zo dat elke FST in een deterministische machine omgezet kan worden: je zal zelf moeten checken of je code deterministisch is voor zijn input.

- Het testscript testtd geeft feedback over meervoudige output.
- Als je machine aan de voorwaarden voldoet, kan je je code optimaliseren met t_minimize en t_determinize (in FSA Utilities).

2. Context

In een aantal gevallen is de omzetting afhankelijk van de context waarin een invoerletter zich bevindt. Bijvoorbeeld:

- woordbegin/midden versus woordeinde
- open versus gesloten lettergreep
- gekleurde klinkers

Sommige afhankelijkheden laten zich niet uitdrukken in het schema

$$\text{onset} + (\text{nucleus} + \text{coda})$$

omdat ze bijvoorbeeld informatie over de overgang van onset naar nucleus, of van nucleus naar coda nodig hebben. Voor zulke gevallen zal je een 'platte' lettergreepstructuur willen gebruiken, of markeerders op tussenliggende representatieniveau's (volgende slide).

3. Compositie

In het SLT hoofdstuk heb je gezien hoe je een transductietaak op kan splitsen in deeltaken, die je dan in compositie schakelt: de uitvoer van transducer T_i wordt doorgegeven als invoer voor T_{i+1}.

$$w_0(T_1 \circ \cdots \circ T_n)w_n \quad = \quad w_0 T_1 w_1 \ldots w_{n-1} T_n w_n$$

Markeerders Een veelgebruikte techniek bij dit soort cascades is het invoegen van markeerdersymbolen op tussenliggende niveau's. Die markeerders kunnen dan gebruikt worden om de transducties te sturen. Als ze hun werk gedaan hebben ruim je ze op.

4. Voorkeurskeuze (priority union)

Gegeven transducers P en Q is de operatie `priority_union` als volgt gedefinieerd:

```
macro(priority_union(P,Q),{P,~domain(P) o Q}).
```

- als de invoer matcht met P wordt die omzetting uitgevoerd
- voor invoer die niet matcht met P wordt de transducer Q aangeroepen

Neem deze definitie in je macro bestand op, als je deze operatie gaat gebruiken.

Mogelijke toepassingen

- leenwoorden: vang ze af voor je aan de regelmatige NL patronen begint
- veel symbolen blijven ongewijzigd: specificeer met p.u. de patronen die veranderen; invoerpatronen waarvoor je geen omzetting specificeert kunnen naar de identiteitsomzetter `?+`