Automatische verwerving van fonotactische modellen

Erik F. Tjong Kim Sang
Centrum voor Nederlandse Taal en Spraak
Universitaire Instelling Antwerpen
België
erikt@uia.ac.be

overheadsheets: http://pcger34.uia.ac.be/~erikt/talks/

Onderwerp

Het gebruik van computationele leeralgoritmen voor het simuleren van de verwerving van taalkennis.

Doel

1. ontdekken welke leeralgoritmen de beste taalmodellen kunnen genereren,
2. bepalen wat de invloed is van kennisrepresentatie is op het leerresultaat, en
3. nagaan of de beschikbaarheid van enige initiële taalkennis de gegenereerde modellen verbetert.

Leerprobleem

Het afleiden van modellen voor de fonotactische structuur van woorden.

Voorbeelden

bad, crwth, mloda, Nwanko, ptata

Extra beperkingen

1. gebruik alleen Nederlandse data,
2. gebruik alleen woorden van één lettergreep, en
3. gebruik alleen positieve leerdata.

Werkmethode

1. Laat het leeralgoritme een model bouwen op basis van positieve data.
2. Test het model met positieve data.
3. Test het model met negatieve data.

Leerdata en testdata

De data is beschikbaar in twee representaties: orthografische representatie en fonetische representatie. De data komt uit de CELEX Lexical Database.

Leerdata: 5577 unieke orthografische woorden en 5084 unieke fonetische woorden.

Testdata: vier verzamelingen van 600 letterreeksen uit de vier testcategorieën.

Datarepresentatie

Woorden zijn gerepresenteerd als reeksen van bigrammen van letters:

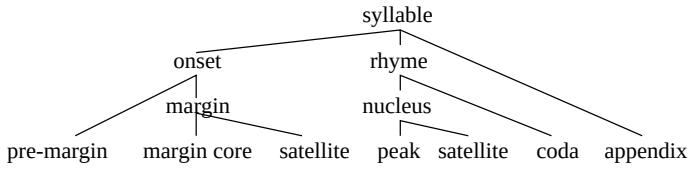
tsjirp = ^t+ts+sj+ji+ir+rp+p*

Datacomplexiteit

1. Bigramentropie
   H(W) = - $\Sigma_{c_i} P(c_i) \Sigma_{c_j} (P(c_j|c_i) * log_2(P(c_j|c_i))$
   orthografisch: 3.298; fonetisch: 3.370
2. Chomskyhierarchie: beide dataverzamelingen bevatten reeksen uit een eindige taal
3. Baselinemethode: accepteer alle letterreeksen die uitsluitend bestaan uit bigrammen die voorkomen in de leerdata.

representatie-formaat	geaccepteerde positieve data	afgewezen negatieve data
orthografisch	99.2%	60.2%
fonetisch	99.2%	71.6%

Het syllabemodel van Cairns en Feinstein



Mogelijke leerparadigma's

1. statistisch leren
2. connectionistisch leren
3. regelgebaseerd leren
4. genetisch leren

Gekozen leeralgoritmen

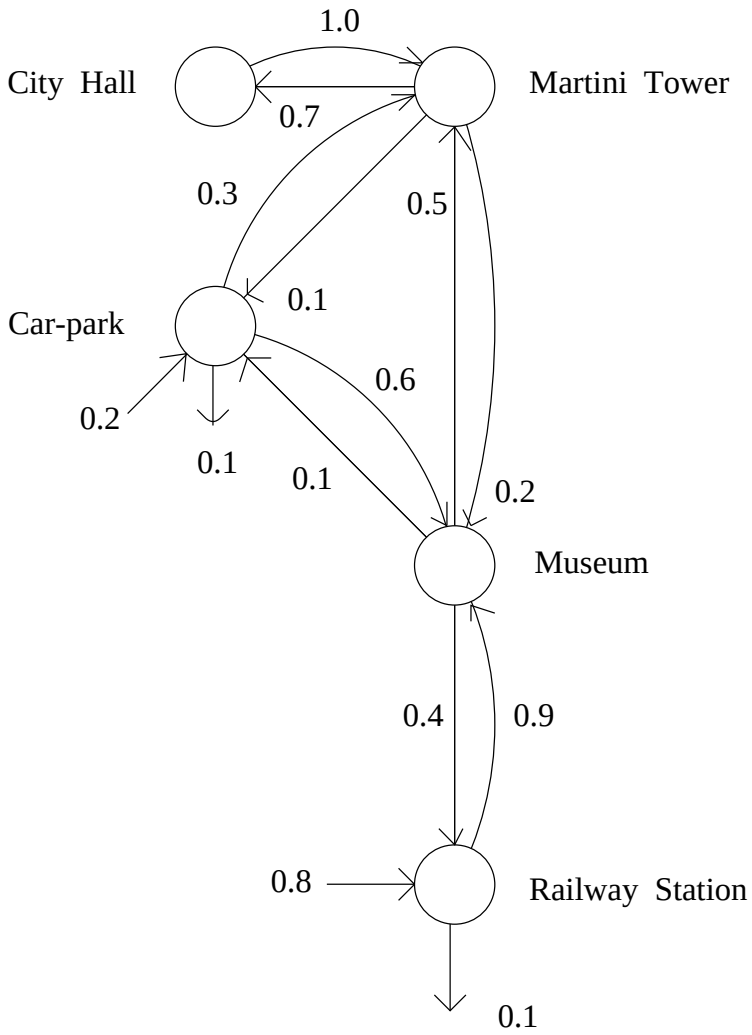
1. Hidden Markov Models (Rabiner)
2. Simple Recurrent Networks (Elman, Cleeremans)
3. Inductive Logic Programming (Muggleton)

Hidden Markov Models (HMMs)

- een verzameling cellen
- iedere cel kan tekens produceren met een bepaalde waarschijnlijkheid
- er kan van de ene cel naar een andere worden gesprongen

Na het leerproces kan een HMM waarden toekennen aan tekenreeksen. De waarden zijn getallen tussen 0 en 1.

Een voorbeeld-HMM



Onderzoeksvragen voor HMMs

1. Hoe kan de context van een teken worden meegenomen in het verwerkingsproces?
2. Hoe kan worden voorkomen dat per definitie korte letterreeksen een hogere waarde krijgen toegewezen dan lange letterreeksen?

Antwoorden

1. Beschouw letterreeksen als reeksen van bigrammen.
2. Voeg een extra startteken en eindteken toe.
   Normaliseer waarden voor de lengte van de reeksen

HMM-resultaten

De HMMs bevatten acht cellen. Elke cel kan een deelverzameling van de 29 orthografische tekens of een deelverzameling van de 43 fonetische tekens verwerken.

Orthografische data

	geaccepteerde positieve data	afgewezen negatieve data
zonder kennis	98.9%	91.0%
met kennis	98.9%	94.5%

Fonetische data

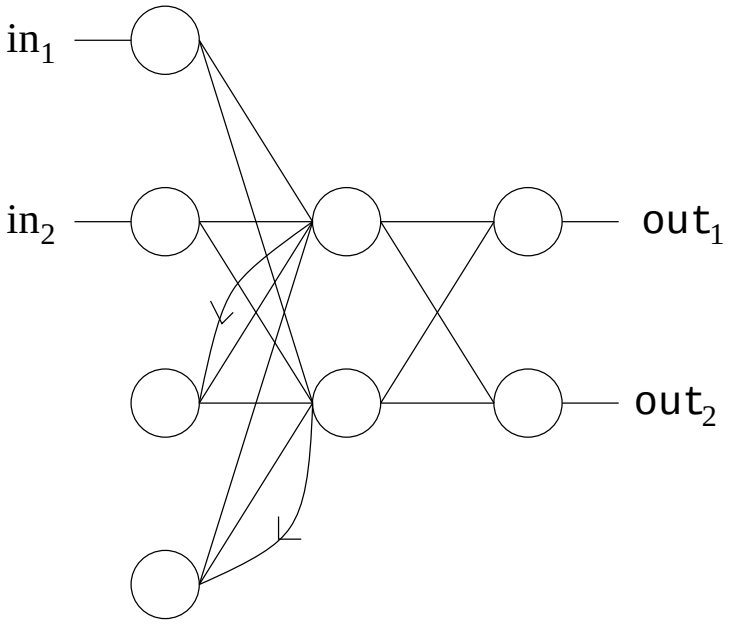
	geaccepteerde positieve data	afgewezen negatieve data
zonder kennis	99.1%	98.3%
met kennis	99.1%	99.1%

Simple Recurrent Networks (SRNs)

- een verzameling cellen
- iedere cel krijgt één of meer numerieke invoersignalen
- iedere cel genereert precies één numeriek uitvoersignaal
- cellen sturen signalen naar elkaar toe via verbindingen met gewichten

Na het leerproces kan een SRN een uitvoer-reeks produceren voor een invoerreeks. De reeksen bestaan uit getallen.

Voorbeeld-SRN



SRN-experimentvragen

1. Hoe kunnen we letters representeren met reeksen getallen?
2. Hoe kunnen we een SRN iets laten leren van positieve data?

Antwoorden

1. Gebruik de representatietechniek *local-list mapping*: voor elke letter wordt één binair cijfer gereserveerd.
2. Laat het netwerk de volgende letter van een reeks voorspellen.

SRN-resultaten

De SRNs bevatten 29 cellen in de invoerlaag en 29 cellen in de uitvoerlaag. De beste resultaten zijn behaald met een netwerk met vier cellen in de verborgen laag.

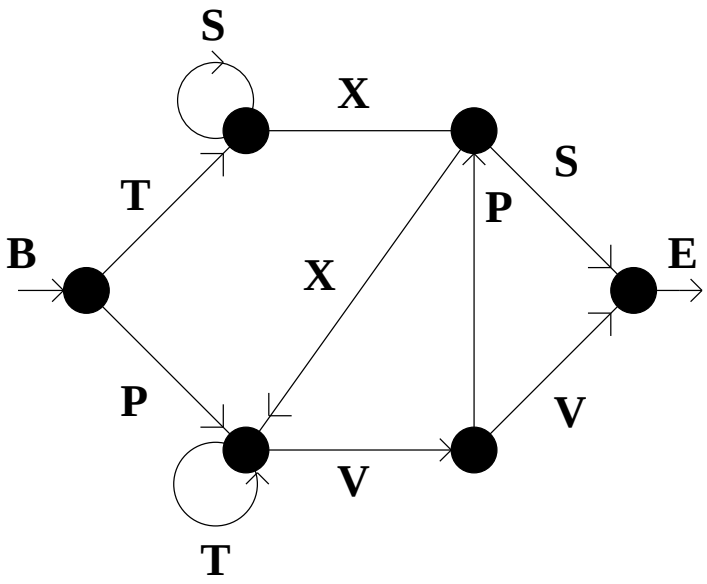
Orthografische data

	geaccepteerde positieve data	afgewezen negatieve data
zonder kennis	100%	8.3%
met kennis	100%	4.8%

Voor de fonetische data zijn geen experimenten uitgevoerd.

Verklaring SRN-resultaat (1)

Bron van inspiratie: Cleeremans et.al. 1993: in experimenten met letterreeksen gegenereerd door een Rebergrammatica konden SRNs na een leerproces 100% van de reeksen correct classificeren.



Verklaring SRN-resultaat (2)

Tjong Kim Sang 1995: de prestatie van SRNs op letterreeksen gegenereerd door een Rebergrammatica zal afnemen als de complexiteit van de grammatica toeneemt

aantal mogelijke vervolgsymbolen	geaccepteerde positieve data	afgewezen negatieve data
2	100%	100%
3	100%	92.2%
4	100%	82.2%

Oorzaak: een afnemende signaal-ruisverhouding voor complexere grammatica's.

Inductive Logic Programming (ILP)

Een methode voor het afleiden van hypotheses van observaties en achtergrondkennis.

Voorbeeld

P_1	Alle mannen zijn sterfelijk.
P_2	Socrates is sterfelijk.
IC	Socrates is een man.

Onderzoeksvragen bij ILP

1. Hoe beperken we het aantal afgeleide hypotheses?
2. Welke achtergrondkennis moeten we gebruiken?

Antwoorden

1. Gebruik een beperkte verzameling van afleidregels en verwijder hypotheses die kunnen worden afgeleid uit andere hypothese.
2. Voorbeeld: SUFFIXREGEL: Neem aan dat het woord W=$w_1 ... w_{n-1} w_n$ ($w_1 ... w_n$ is een reeks van n letters) en dat de suffixhypothese SH(w_{n-1},w_n) bestaat. Dan volgt uit het feit dat W een goed woord is dat $w_1 ... w_{n-1}$ een goed woord is en omgekeerd.

ILP-resultaten

De door ILP gegenereerde modellen bevatten tussen de 600 en 1200 regels.

Orthografische data

	geaccepteerde positieve data	afgewezen negatieve data
zonder kennis	99.2%	60.0%
met kennis	97.8%	97.7%

Fonetische data

	geaccepteerde positieve data	afgewezen negatieve data
zonder kennis	99.2%	70.6%
met kennis	99.2%	98.3%

Conclusies

Doel

1. ontdekken welke leeralgoritmen de beste taalmodellen kunnen genereren,
2. bepalen wat de invloed is van kennis-representatie is op het leerresultaat, en
3. nagaan of de beschikbaarheid van enige initiële taalkennis de gegenereerde modellen verbetert.

Resultaten

1. HMMs en ILP genereren betere modellen dan SRNs.
2. De modellen gegenereerd voor fonetische data zijn vrijwel altijd beter dan de modellen voor orthografische data.
3. Het beschikbaar zijn van enige initiële taalkennis verbetert de gegenereerde modellen, in het bijzonder die van ILP.

Adres

Adres:	Centrum voor Nederlandse Taal en Spraak Universitaire Instelling Antwerpen Universiteitsplein 1 B-2610 Wilrijk België
telefoon:	+31 3 8202765
WWW:	http://ger-www.uia.ac.be/u/erikt/

Literatuur

- Erik F. Tjong Kim Sang, *Machine Learning of Phonotactics*, PhD thesis University of Groningen, The Netherlands, 1998.
  http://stp.ling.uu.se/~erikt/mlp/

Overheadsheets

http://pcger34.uia.ac.be/~erikt/talks/