Navorsings- en oorsigartikels

Neurale netwerke as moontlike woordafkappings-tegniek vir Afrikaans

Machteld Fick
Departement Kwantitatiewe Bestuur, Universiteit van Suid-Afrika
E-pos: fickm@unisa.ac.za

Ontvang Julie 2002; aanvaar September 2002

UITTREKSEL

In Afrikaans word saamgestelde woorde aanmekaar geskryf en nuwe woorde word voortdurend gevorm. Aangesien daar dus nie 'n statiese verwysingsbron bestaan nie, is die proses van woordafkapping tydens teksprosessering 'n probleem – veral waar smal kolomme gebruik word, soos in tydskrifte en koerante. 'n Neurale netwerk (vorentoevoer-terugpropagering) is vir die afkappingsprobleem ontwikkel en met sowat 5 000 Afrikaanse woorde met korrekte lettergreepverdeling afgerig. Die neurale netwerk het gemiddeld 97,56% van moontlike posisies in 5 000 willekeurig gekose woorde korrek as óf geldige óf ongeldige afkappingspunte geklassifiseer. Tydens 'n toets met woorde uit 'n Afrikaanse tydskrif het die neurale netwerk 98,75% van woordposisies korrek geklassifiseer. Hieruit is die gevolgtrekking gemaak dat neurale netwerke wel suksesvol as afkappingstegniek vir Afrikaans gebruik kan word.

ABSTRACT

Neural networks as possible hyphenation technique for Afrikaans

In Afrikaans compound words are written as one word. New words are therefore created by simply joining words. Word hyphenation during typesetting by computer is often a problem, because the source of reference changes all the time. A neural network (feedforward backpropagation) was trained with about 5 000 Afrikaans words with correct syllabification. The neural network classified 97,56% of possible points in 5 000 randomly chosen words correctly as either valid or invalid hyphenation points. In a test with 510 words from an Afrikaans magazine the neural network classified 98,75% of possible positions correctly. We came to the conclusion that neural networks can be used successfully as hyphenation technique for Afrikaans.

INLEIDING

Programmatuur wat vir rekenaarteksprosessering gebruik word, bevat gewoonlik ingeboude afkappingsfunksies wat hoofsaaklik aangewend word om visuele reëlmatigheid in gedrukte teks te bevorder. Om te voorkom dat groot gapings tussen woorde ontstaan (dubbelgeskouerde teks) of dat regterkante te veel varieer (linksgeskouerde teks) word woorde wat nie aan die einde van 'n reël inpas nie verdeel, eerder as om dit in geheel na die volgende reël oor te dra.

In Afrikaans, soos in Nederlands en Duits, waar nuwe woorde geskep kan word deur woorde aanmekaar te haak, is afkapping 'n groot probleem. Bestaande afkappingsfunksies wat op 'n woordeboekbenadering of 'n reëlbenadering of 'n kombinasie van die twee gebaseer is, lewer gewoonlik onbevredigende resultate, aangesien dit vir 'n rekenaar onmoontlik is om die konteks van saamgestelde woorde te ontleed en korrekte afkappingsposisies te bepaal.

Neurale netwerke, wat op biologiese senuweestelsels gemodelleer is, is by uitstek geskik vir patroonherkennings-probleme. Afkapping is op patroonherkenning gebaseer en neurale netwerke het die vermoë om te veralgemeen. Die moontlikheid word dus ondersoek of 'n neurale netwerk wat met korrekte voorbeelde afgerig word, in staat sal wees om patrone in onbekende woorde te herken en uiteindelik as bruikbare afkappingsinstrument aangewend kan word.

Alhoewel dit 'n tydrowende taak is om data voor te berei en 'n neurale netwerk af te rig, is neurale netwerke vinnig en effektief tydens gebruik.

AFRIGTINGSDATA

In Afrikaans word afkapping op grond van lettergreepverdeling gedoen.[10] 'n Neurale netwerk word dus ontwikkel en afgerig om woorde in lettergrepe te verdeel. Afrikaanse woorde met lettergreepaanduidings is uit die elektroniese weergawe van die *Verklarende Handwoordeboek van die Afrikaanse Taal*[8] (ELHAT) onttrek en in 'n datalêer geplaas (52 167 woorde). Aanvanklik is 'n deelversameling van 1 238 woorde hieruit gebruik om die neurale netwerk mee af te rig.

ENKODERING

Data wat aan die neurale netwerk gevoer word, bestaan uit *vensters* van agt opeenvolgende letters uit woorde wat vir afrigting gebruik word. Die neurale netwerk moet uitslag lewer oor die posisie tussen die vierde en vyfde letter in die venster, naamlik *ja* (1), dit is 'n geldige afkappingsposisie of *nee* (0), dit is nie.

Daar is op 'n venstergrootte van 8 besluit aangesien lettergrepe met vier letters algemeen in Afrikaans voorkom. Die vier letters weerskante van die beslissingsposisie verskaf die konteks waarbinne besluit moet word of dit 'n afkappingsposisie is of nie.

Die opeenvolgende vensters vir die woord *geletterd* met ooreenstemmende teikens verskyn in figuur 1. Afrikaanse woorde kan met 'n lettergreep wat uit 'n enkele letter bestaan, begin of eindig, soos in *A-fri-ka* en *ge-heu-e*. Daar kan dus drie spasies voor die eerste letter en drie spasies na die laaste letter in die venster voorkom.

Venster								Teiken
			g	e	l	e	t	0
		g	e	l	e	t	t	1
	g	e	l	e	t	t	e	0
g	e	l	e	t	t	e	r	0
e	l	e	t	t	e	r	d	1
l	e	t	t	e	r	d		0
e	t	t	e	r	d			0
t	t	e	r	d				0

Figuur 1 *Opeenvolgende vensters met ooreenstemmende teikens*

Die volgende 39 karakters kom algemeen in Afrikaanse woorde voor:

- Die 26 letters van die alfabet.
- Letters met kappies, deeltekens en aksente (ä, ë, ê, é, è, ï, ö, ô, ü en û). (Daar word aangeneem dat aksente soos á, í, ò en ú geen invloed op lettergreepverdeling het nie en word nie ingesluit nie.)
- Die afkappingsteken (') wat in woorde soos *pa'tjie*, *wag-'n-bietjie*, *foto's*, ens. voorkom.
- Lettergreepaanduidings (-) en koppeltekenaanduidings (=).

Aan elk van hierdie karakters word 'n syferwaarde toegeken. Vokale en konsonante vervul verskillende fonologiese funksies en resultate uit vorige navorsing[9] toon dat neurale netwerke beter presteer indien 'n duidelike onderskeid tussen die twee groepe gemaak word. In tabel 1 verskyn die koderingstabel wat gebruik is. Die vokale en konsonante word apart gegroepeer, met "y" – 'n konsonant wat soos 'n vokaal funksioneer – tussenin. Die lettergreep- en koppeltekenaanduidings is soortgelyk, maar word apart geplaas om later weer tussen hulle te kan onderskei. Verder is 'n niewoordkarakter (#) bygevoeg om die skikking te voltooi.

Tabel 1 Koderingstabel vir Afrikaanse woorde

-	a	ä	e	ë	ê	é	è	i	ï
1	2	3	4	5	6	7	8	9	10
o	ö	ô	u	ü	û	y	b	c	d
11	12	13	14	15	16	17	18	19	20
f	g	h	j	k	l	m	n	p	q
21	22	23	24	25	26	27	28	29	30
r	s	t	v	w	x	z	'	#	=
31	32	33	34	35	36	37	38	39	1

In die proses om die onverwerkte data na bruikbare invoer vir die neurale netwerk om te skakel, word elk van die woorde in die afrigtingsdata geënkodeer, die koppeltekenaanduidings word verwyder en 'n teikenvektor word geskep wat die posisies van koppeltekens aandui. 'n Venster van agt karakters word oor elke woord geskuif om 'n vierletter-konteks weerskante van elke moontlike afkappingsposisie te kry. Elke venster word na 'n 40×8 matriks van nulle en ene omgeskakel waar elke ry een van die karakters in tabel 1 verteenwoordig. Hierdie matriks word na 'n kolomvektor met 320 elemente omgeskakel sodat elke venster deur 'n enkele kolom weergegee word.

Tydens afrigting bestaan die data wat aan die neurale netwerk gevoer word uit 'n invoermatriks (**P**) waarvan elke kolom 'n venster uit die afrigtingsdata verteenwoordig en 'n teikenvektor (**t**) wat die korrekte uitvoer vir elke ooreenstemmende venster bevat.

DIE NEURALE NETWERK

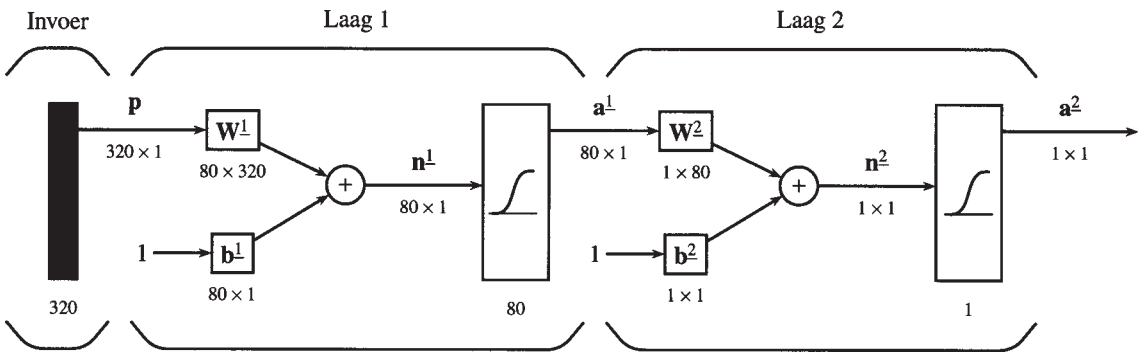
'n Vorentoevoer-terugpropageringsnetwerk (*feedforward back-propagation network*) is met MATLAB se neurale-netwerk-pakket afgerig. Die argitektuur van die tweelaagnetwerk wat die beste resultate gelewer het, verskyn in figuur 2 en het die volgende eienskappe:

- Die invoer bestaan telkens uit 'n kolomvektor **p** met 320 elemente uit die invoermatriks **P**. Die invoer word nie as 'n laag van die netwerk beskou nie.

- Laag 1
  - Die gewigte verskyn in $\mathbf{W}^1$ ('n 80 × 320 matriks).
  - Die beladings verskyn in $\mathbf{b}^1$ ('n kolomvektor met 80 elemente).
  - Die netto invoer word gegee deur $\mathbf{n}^1 = \mathbf{W}^1\mathbf{p} + \mathbf{b}^1$.
  - Die uitvoer $\mathbf{a}^1$ word deur die *log-sigmoidale* funksie uit $\mathbf{n}^1$ bepaal ('n kolomvektor met 80 elemente). Dit is ook Laag 2 se invoer.

- Laag 2
  - Die gewigte verskyn in $\mathbf{W}^2$ ('n ryvektor met 80 elemente).
  - 'n Enkele belading verskyn in $\mathbf{b}^2$.
  - Die netto invoer word gegee deur $\mathbf{n}^2 = \mathbf{W}^2\mathbf{a}^1 + \mathbf{b}^2$ ('n enkele waarde).
  - Die uitvoer $\mathbf{a}^2$ is 'n enkele waarde en word deur die *log-sigmoidale* funksie uit $\mathbf{n}^2$ bepaal.

- Die uitvoer van die netwerk is die afgeronde waarde van $\mathbf{a}^2$. Dit dui vir elke venster (**p**) aan *ja* (1), dit is 'n geldige afkappingsposisie of *nee* (0), dit is nie.



Figuur 2 *Argitektuur van die neurale netwerk*

Daar is met die volgende veranderlike komponente geëksperimenteer ten einde die beste neurale netwerk vir die probleem te verkry:

- Oordragfunksies
  Die invoer en uitvoer vir die afkappingsprobleem bestaan uit nulle en ene. Op grond hiervan is besluit om die *log-sigmoidale*[5] oordragfunksie te gebruik.

- Aantal neurone in die verborge laag
  Tydens 'n empiriese ondersoek is bepaal dat 'n neurale netwerk met om en by 80 neurone in die verborge laag ('n 320-80-1 netwerk) die beste resultate lewer.

- Aantal verborge lae
  'n Tweede verborge laag is ingevoer en die aantal neurone in elke laag is gevarieer. Met min afrigtingsdata (1 238 woorde) het 'n 320-50-10-1 netwerk beter resultate as 'n 320-80-1 netwerk gelewer. Dit was egter nie die geval met meer afrigtingsdata (4 000 woorde) nie.

- Afrigtingsalgoritmes
  Deur verskillende afrigtingsalgoritmes te vergelyk, is bepaal dat veerkragtige terugpropagering (*resilient backpropagation*[5]) die geskikste algoritme vir die afkappingsprobleem is.

- Vroeë beëindiging
  Die resultate het getoon dat afrigtingsdata nie gememoriseer word nie. Vroeë beëindiging is dus onnodig vir die afkappingsprobleem.

Die data is stelselmatig verbeter soos probleemareas opgeduik het. Die volgende stappe is deurloop:

- Beter afrigtingsdata
  Die datalêer waaruit die afrigtingsdata kom, is skoongemaak deur woorde wat meer as een keer voorkom en woorde uit ander tale te verwyder, verkeerde lettergreepverdelings reg te stel en ontbrekende koppeltekens in te voeg. Die neurale netwerk is met 'n ewekansige deelversameling van 1 000 woorde uit die verbeterde datalêer van 51 080 woorde afgerig. Die prestasie van die neurale netwerk het heelwat verbeter.

- Meer afrigtingsdata
  Die afrigtingsdata is stelselmatig meer gemaak deur telkens die invoermatriks en teikenvektor wat 'n volgende 1 000 woorde verteenwoordig, aan die bestaandes te haak. Die resultate toon dat die prestasie van die neurale netwerk beslis verbeter soos die afrigtingsdata toeneem. As gevolg van beperkte rekenaarkapasiteit kon nie meer as 40 000 afrigtingspare gebruik word nie.

- Kort woorde bygevoeg
  'n Datalêer wat bestaan uit woorde wat nie in lettergrepe verdeel kan word nie, is uit ELHAT onttrek. Die neurale netwerk is met 'n samestelling van die bestaande data en hierdie kort woorde afgerig. Die resultate toon dat die netwerk se prestasie afneem met die byvoeging van kort woorde.

- Vreemde letterkombinasies
  Ontleding van die uitvoer het getoon dat woorde verkeerd afgekap word indien die tweede deel van 'n saamgestelde woord met 'n vokaal begin (soos in *verenig*, *eienaardig*) en indien ongewone letterkombinasies wat nie in die afrigtingsdata ingesluit was nie, voorkom (soos in *kalkswael*). Sulke woorde is per hand uit die datalêer gesoek en by die afrigtingsdata gevoeg. Resultate het verbeter.

RESULTAAT

Die beste neurale netwerk is met 39 302 afrigtingspare (uit 4 777 woorde) afgerig totdat 'n gemiddelde kwadraatfout van minder as 0,01 na 118 epogge bereik is. Die afrigtingsproses het nagenoeg 8 uur op 'n Pentium 4 rekenaar met 1,4 gigahertz verwerkerspoed geneem.

Die neurale netwerk klassifiseer gemiddeld 97,56% van moontlike afkappingsposisies in 5 000 woorde (nie die afrigtingsdata nie) reg. Dit klassifiseer 99,50% van moontlike afkappingsposisies in die afrigtingsdata reg.

TOETS MET ONBEKENDE DATA

Die neurale netwerk is met woorde uit 'n paar artikels in *Sarie* van 9 Januarie 2002 getoets. Die data is aangepas deur herhalende woorde, Engelse woorde en eiename van vreemde herkoms te verwyder om uiteindelik 510 onbekende woorde te kry. Uit hierdie 510 woorde is 2 636 vensters gegenereer. Elk van die vensters verteenwoordig 'n moontlike afkappingsposisie.

In tabel 2 verskyn die 30 woorde waarin die neurale netwerk afkappingsfoute gemaak het.

Tydens ontleding is die volgende soorte afkappingsfoute geïdentifiseer:

1. 'n Koppelteken word tussen geldige lettergrepe weggelaat. (Aangedui deur onderstreping.) Hierdie foute word as *goeie* foute beskou aangesien dit nie verkeerde afkappings tot gevolg sal hê nie, maar bloot nalaat om koppeltekens in te plaas.

Tabel 2 Afkappingsfoute deur neurale netwerk gemaak

ag - tja-ri-ge	leen-tjie - sklip	skoo(-)lu-niform
baie	me(-)teens	stof-ko -r- rel-tjie
be -s- ka-wing - sont-wik-ke-ling	mi - v–vigs	sui-ker -k- lont–feetjie
be - ste	na-vor-sing - sraad	tea-ters
feetjies	po-niestert	vanjaar
ge -s- on-de	pre-si-dent(-)sak-ker	verd - wyn
ge -s- wig	re(-)goor	vin -g- er-tjie
karmo-syn	saal -s- ak-ke	volkspe-le
lae	seepunt	vrag-mo -t- ors
lankal	sek-sue-le	wa - ar's

2. 'n Koppelteken verskyn op 'n ongeoorloofde plek in 'n woord. Hierdie is *slegte* foute aangesien dit tot verkeerde afkappings aanleiding gee. Die volgende soorte slegte foute word onderskei:

   - Die koppelteken verskyn op 'n verkeerde plek in die woord, soos in *ag-tja-ri-ge*, *be-ste*, ens. (Aangedui met 'n grys blokkie.)

   - Waar die eerste deel van 'n saamgestelde woord op 'n konsonant eindig en die tweede deel met 'n vokaal begin, word voor die konsonant afgekap, soos in *me-teens*, *re-goor*, *skoo-lu-ni-form*, ens. (Aangedui met 'n sirkel.)

   - Koppeltekens verskyn weerskante van 'n konsonant, soos in *ge-s-on-de*, *ge-s-wig*, ens. (Aangedui met 'n reghoekige raampie.) Ontleding van die neurale netwerk se uitvoer toon dat die waarde vir die verkeerde koppelteken telkens laer as dié vir die korrekte een is. Hierdie foute kan reggestel word deur die afkappingsposisie met die hoogste waarde as korrek te aanvaar.

Die neurale netwerk het 94,12% (480/510) van die woorde volledig en korrek in lettergrepe verdeel, terwyl dit 98,75% van die moontlike afkappingsposisies korrek geklassifiseer het. Onder die 5,88% woorde waarin foute voorgekom het, was daar 2,15% met *goeie* foute en 3,73% met *slegte* foute, terwyl daar onder die 1,25% verkeerd geklassifiseerde afkappingsposisies 0,49% *goeie* foute en 0,76% *slegte* foute was. Indien die goeie foute nie as foute gereken word nie, het die neurale netwerk 96,27% van die woorde korrek in lettergrepe verdeel en 99,24% van die afkappingsposisies reg geklassifiseer.

GEVOLGTREKKING

Dit is duidelik dat neurale netwerke wel as effektiewe afkappingstegniek vir Afrikaans gebruik kan word, aangesien 'n netwerk wat met minder as 5 000 woorde afgerig is, meer as 98% van die afkappingsposisies in vreemde woorde reg klassifiseer.

Die volgende areas vir moontlike verbetering behoort verder ondersoek te word:

- Datavoorbereiding

  – Deur die afrigtingsdata planmatig op grond van taalkundige kennis te kies, eerder as om willekeurige woorde te gebruik, kan beter resultate moontlik verkry word.

  – Die probleem van ongewone letterkombinasies wat as gevolg van die samestelling van woorde ontstaan, sal waarskynlik nooit heeltemal uitgeskakel kan word nie, maar kan opgelos word deur soveel moontlik letterkombinasies in die afrigtingsdata in te sluit.

- Afrigting

  – Die koderingsmatriks (tabel 1) kan aangepas word om sekere taalkundige eienskappe te weerspieël. 'n Taalkundige ontleding van die frekwensie waarmee sekere letters saam voorkom, mag daartoe lei dat lettervolgorde en -groepering anders moet wees om beter resultate te kry.

  – Daar kan ook met die toekenning van syferwaardes geëksperimenteer word, soos om byvoorbeeld waardes tussen 0 en 1, eerder as heelgetalle, te gebruik.

  – Deur 'n kragtiger rekenaar en/of 'n ander neurale-netwerkpakket of programmeertaal te gebruik, kan 'n neurale netwerk met meer woorde (sê 10 000) afgerig word. Die prestasie behoort verder te verbeter.

- Uitsetontleding

  Verdere oplossings vir probleemgevalle kan moontlik geïdentifiseer word deur die neurale netwerk se uitvoer te bestudeer, soos in die geval van koppeltekens weerskante van konsonante.

LITERATUURVERWYSINGS

1. Daelemans, W., Van den Bosch, A. (1992). *Generalization performance of backpropagation learning on a syllabification task*. http://ilk.kub.nl/downloads/pub/antalb/twlt3-92.ps.gz.
2. Demuth, H.B., Beale, M. (1992 – 2001). *Neural Network Toolbox: For Use with MATLAB*. (The Math-Works, Inc.).
3. Fick, M. (2002). *Neurale Netwerke as moontlike Woordafkappingstegniek vir Afrikaans*. (Meestersgraadverhandeling, Universiteit van Suid-Afrika).
4. Fritzke, B., Nasahl, C. (1991). 'A neural network that learns to do hyphenation', in *Artificial Neural Networks*. Eds. Kohonen, T., Mäkisara, K., Simula, O., Kangas, J. (North-Holland, Amsterdam, Netherlands), 1375–1378.
5. Hagan, M.T., Demuth, H.B., Beale, M. (1996). *Neural Network Design*. (PWS Publishing Company).
6. Liang, F.M. (1983). *Word hy-phen-a-tion by com-pu-ter*. (Ph.D. thesis, Stanford University).
7. McIntosh, R., Fawthrop, D. (1990). *Hyphenation* (third ed.). http://www.hyphenologist.co.uk/book/book-ed3.htm.
8. Odendal, F.F., Schoonees, P.C., et al. (1983). *Verklarende Handwoordeboek van die Afrikaanse Taal (HAT)*. (Perskor-Boekdrukkery).
9. Smrž, P., Sojka, P. (1996). 'Word hy-phen-a-tion by Neural Networks'. *FI MU Report Series*.
10. Suid-Afrikaanse Akademie vir Wetenskap en Kuns. Taalkommissie (1991). *Afrikaanse Woordelys en Spelreëls*. (Kaapstad: Tafelberg).

MACHTELD FICK

Machteld Fick behaal in 1973 die graad B.Sc. (Wiskunde en Wiskundige Statistiek) aan die Universiteit van Pretoria, waarna sy tot 1977 as aktuariële klerk by Sanlam se hoofkantoor in Bellville werksaam was. Sedert 1984 is sy verbonde aan die Departement Kwantitatiewe Bestuur aan die Universiteit van Suid-Afrika waar sy eers in 'n tydelike hoedanigheid en sedert 1997 as voltydse dosent werksaam is. In 1995 behaal sy die graad B.Sc.Hons. (Operasionele Navorsing) en in 2002 die graad M.Sc. (Operasionele Navorsing) met lof aan die Universiteit van Suid-Afrika.