

Cornetto: a combinatoric and relational network for language technology

Piek Vossen, Irion Technologies

International Colloquium on Word structure and lexical system: models and applications.

Paiva, December, 16-17th, 2004

Overview

- Natural language clichés
- Cornetto database: two-layered relational semantic network;
- Experiments to improve document classification using multiword units;

Natural language clichés

- Noun-phrase index is not feasible due to combinatorial explosion (Evans 1995);
- Most language usage is **not** creative “*hot coffee*”:
 - a relatively limited set of expressions (clichés);
 - a limited set of compositional relations;
- We can still decompose expressions to combinations of concepts provided that:
 - We can exactly detect concepts;
 - We can value the ‘appropriate’ expression of concept combinations ;

Cumulative combinations

market

marketplace
grocery
grocery store

rise

arise
be lifted
climb
climb up
get out of bed
get to one's feet
get up
go up
heighten
increase in volume
jump
lift

mount
move up
prove
rebel
renounce allegiance
resist forcefully
rise up
stand up
straighten
surface
turn out
wax

Knowledge based approaches to WSD

- Wordnet Domains work for larger and global contexts but have a low recall for small context;
 - Senseval-2, all-words: 75% precision, 35% recall
- Calculation of semantic distance or conceptual density, based on the wordnet structure and even Machine Learning approaches;
 - Senseval-2/3 all-words: at most 65% precision & recall
- Wordnets:
 - Mostly contain vertical relations: bread -> food
 - Hardly any horizontal relations: bread -> eat
- Wordnets do not contain sufficient multiword expressions for compositional clichés

Cornetto

- Initiative from Irion Technologies and the Vrije Universiteit Amsterdam
- Two-layered relational structure
 - Conceptual relations, like Wordnet
 - Lexical expressions, more like Melčuk
- Conceptual relations: massive storage of horizontal relations: *roles, properties, manners*;
- Combinatoric possibilities constrained by encoded multiword expressions (clichés) for conceptual combinations

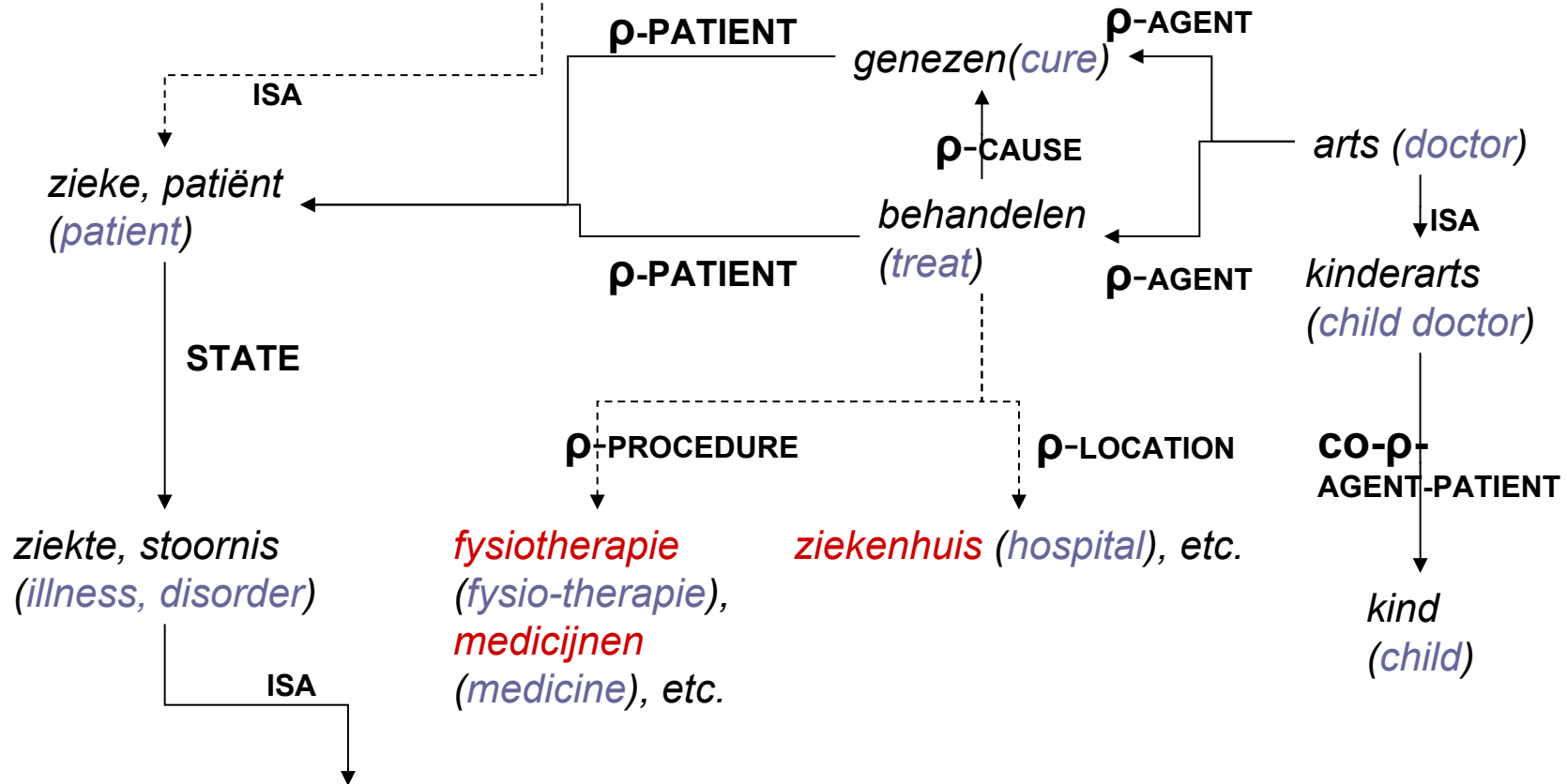
Dutch Cornetto database

- Dutch wordnet:
 - Concepts: 44,000
 - Vertical relations (hyponymy): 46,021
 - Horizontal relations (roles, meronymy, etc.): 5,843
- RBN (Referentie Bestand Nederlands):
 - Concepts: 40,000
 - Relations: 60,000 (mostly horizontal)
 - Typical multiword expressions with various degrees of freedom, encoded at various levels

RBN information

- een patiënt voor een nieraandoening/keelpijn behandelen
(lit: a patient for a kidney disorder/throat affection treat)
- iemand aan zijn verwondingen behandelen
(lit: somebody at his injuries treat)
- iemand met fysiotherapie/medicijnen behandelen
(lit: somebody with physiotherapy/medicines treat)
- twee kinderen moesten in het ziekenhuis behandeld
worden
(two children had to be treated in the hospital)

chronisch zieke (*chronical patient*), *langdurig zieke* (*long-term patient*),
psychisch/geestelijk zieke (*mental patient*)



maagaandoening (*stomach disorder*)
nieraandoening (*kidney disorder*), *keelpijn* (*sour throat*).

Disambiguating combinations

slag & leveren (fight)
slag & vrije (free style swimming)

punt & delen (share points)
punt & pakken (take a point)
punt & scoren (score a point)
punt & hard (firm point)

trap & geven (to kick)
trap & krijgen (be kicked)
trap & vrij (free kick)
trap & hoge (high stairs)

band & lekke (flat tire)
band & lopende (converter belt)
band & losse (loose relation)
band & spelen (music band plays)
band & starten (start a tape or belt)
band & sterke (strong relation)

bui & flinke (heavy shower)
bui & plaatselijk (local shower)
bui & vallen (there can be shower)
bui & slechte (bad mood)

Prototypical contexts

functie & bekleden (take a position)
functie & hoge (high position)
functie & neerleggen (put down position)
functie & ontheffen (discharge from position)
functie & openen (open up a position)
functie & treden (start a position)
functie & uitoefenen (carry out position)
functie & vervullen (fulfill position)

werk & aannemen (take on work)
werk & bieden (offer work)
werk & gaan (go to work)
werk & half (incomplete work)
werk & staken (strike)
werk & zetten (to put to work)
werk & zoeken (search for work)
werk & zwaar (heavy work)

First experiments with multiword units

- Document Classification using statistical techniques: vector space model and idf*tf weighing;
- Every normalized form is an index item: vector in a multi-dimensional space;
- Compounds are decomposed into smaller elements and stored as separate index items:
mensenrechtenactivistenleider -> mensenrechten (human rights) activisten (activists) leider (leader);
- Multiword expressions are also stored as combinations and as separate index items: *werk (work) aannemen (take on)-> werk_aannemen*

Combination words for nouns from RBN

		Combination word		
		Nouns	Verbs	Adjectives
Noun kernel	8442	771	6071	3489
Combinations	15418	898	9398	5316
Average	1,83	1,16	1,55	1,52

Dutch news items

	Test	Training
Nr of files	278	2.647
Words	57.266	520.411
RBN Nouns	20.050	180.287
Proportion	0,35	0,35
Detected multiwords (25 word window)	1.500	14.366
Proportion of detected nouns	0,07	0,08
Single word types	8367	37035
Multiword types	655	2661

Test files

Test	5	15	25	35	45	55
Nouns	20.050	20.050	20.050	20.050	20.028	20.000
Single	19.401	18.962	18.655	18.372	18.084	17.822
Combinations	685	1.160	1.500	1.822	2.139	2.414
Average words	205	205	205	205	205	205
Average nouns	72	72	72	72	72	71
Average single	69	68	67	66	65	64
Average combinations	2	4	5	6	7	8
Proportion of single	0,97	0,95	0,93	0,92	0,90	0,89
Proportion of combinations	0,03	0,06	0,07	0,09	0,11	0,12

Training files

Training	5	15	25	35	45	55
Nouns	180.287	180.284	180.274	180.256	180.218	179.978
Single	174.103	169.824	166.870	164.287	162.010	159.753
Combinations	6.404	10.993	14.366	17.398	20.136	22.636
Average words	196	196	196	196	196	197
Average nouns	68	68	68	68	68	67
Average single	65	64	63	62	61	60
Average combinations	2	4	5	6	7	8
Proportion of single	0,97	0,94	0,93	0,91	0,90	0,89
Proportion of combinations	0,04	0,06	0,08	0,10	0,01	0,13

Results

		Multiword window					
75	base line	5	15	25	35	45	55
RECALL	86,4	86,4	87,4	88,1	88,1	86,5	85,2
PRECISION	71,2	71,9	72,8	74,1	73,6	73,3	72,1
COVERAGE	95,3	95,6	94,6	94,2	93,8	93,8	94,9

Future plans

- Combine Dutch wordnet with RBN;
- Expand multiword expressions to (co)-hyponyms;
- Train domain disambiguator with multiword expressions;
- Build classifiers for for each polyemeous word based on combinatoric and horizontal relations;
- Automatic acquisition of multiword expressions per domain;

Thank you for your attention!