

DutchSemCor

Building a semantically annotated corpus for Dutch

Piek Vossen, Attila Görög, VU University Amsterdam

Fons Laan, ISLA, University of Amsterdam

Rubén Izquierdo, Tilburg University

Antal van den Bosch, Maarten van Gompel, Radboud University Nijmegen



Overview

- Project goals and planning
- Current progress
- Word-sense-disambiguation results
- Active learning phase

Goals and planning

- Funded by NWO, 2009-2012
- Create a large semantically tagged corpus for Dutch:
 - Sense-tags from the Cornetto database (includes Dutch wordnet)
 - Domain labels from Wordnet Domains
 - Named entities mapped to Wikipedia

Global procedure

- **Phase-1:**
 - 25 examples per meaning for 3,000 most polysemous and frequent nouns, verbs and adjectives (average nr. of meanings = 3)
 - Annotated by two student assistants
 - Minimal IAA 80%
- **Phase-2:**
 - Word-sense-disambiguation (WSD) systems trained with the data of phase-1
 - Active learning: add examples for low performing words and meanings until we reach accuracy of 80% or no progress
- **Phase-3:**
 - Apply WSD to rest of the full corpus

Corpora

- SoNaR: 500M tokens written Dutch
- CGN: 1M tokens spoken Dutch
- Web snippets mediated through WebCorp.co.uk (<http://www.webcorp.org.uk/>)
 - In case no or insufficient examples are found for particular senses in SoNaR and CGN
 - Students select snippets (target word and context) which are added to the corpus in the SoNaR annotation format

Annotation tool

Browser: dutchsemcor

URL: http://dutchsemcor.uvt.nl/sat/

Mode: Free | List | Buddy | Lemma: muis | Category: adj noun verb ? | Context: 75 chars | Source: All | SoNaR | Snippets | CGN | DB Search

#	Examples	Morphosyntax	Resume/Def	Domain	SUMOntology	Synonyms	Relations	Tagged
1		n-de-t	klein knaagdier	biol	Mouse		knaagdier	
2	de muis van de hand	n-de-t	onderdeel v.d. hand bij de duim	biol	AnatomicalStructure	duimmuis	deel gedeelte part	
3	de muis aanklikken	n-de-t	randapparatuur voor de computer	comp	Device		toestel apparaat	
4	Ze is een grijze muis die zich nooit ergens over uit laat.	n-t	spichtig, verlegen persoon	psy alg	SubjectiveAssessme		persoon mens figuur	
5	Voor dit gerecht is een muis een uitstekende aardappel.	n-t	langwerpig soort aardappel	voed	FruitOrVegetable		aardappel pieper	

1 of 5 rows

Tag Co-oc L: M: R: Clear Filter UnTag Usage: Normal 1 ... 50 of 4121

#	tfeL	Left	Sense	Word	Right
24	nee . x'	en uh daar hadden ze 't de hele dag nog over over*x . een ...		muis	xxx ja was wel aardig dat . ja oh ja dat uh ... xxx . hè ? en uh nou ja
30	eleusiv	dan kreeg ie een laptop en een visuele muis . en een uh ... een visuele		muis	wat is dat ? zo'n ding waar geen bolletje onder zit maar een laseroog . oh
41	tad gor	en toch ... d*a die die dat gif dat werkt ook niet . we hebben toen nog dat	1	muisje	wat hier achter de kast zat dat had je ook aan elke kant van de kast een bo
25	ed . ho	toen met dieren . ja zijn uh hele ouwe . xxx . oh . de		muis	was de spion en de olifant was de maarschalk . ggg . oh ja . oh ja . dus ee
46	eid tna	r op en neer totdat je natuurlijk heel die jas kapot gebeten had . want die		muis	waart ook niet gek . mmm . nou en toen was ie weg . ggg . ggg . ja
4	nee ret	x xxx natuurlijk maar ... xxx . tien naar links . ja . kunnen ze beter een*		muis	voor je voet of zo kunnen o*a ontwerpen xxx . ja . waarom nou niet voor je
26	nee lew	t was de maarschalk . ggg . oh ja . oh ja . dus een een olifant kan wel een		muis	vertrappen of nee ... maar die muis kan die olifant ook laten schrikken . j
49		ken boven zijn uh		muis	van z'n van z'n hand zeg maar . ja . ja maar to*a ... oh ja . als een tromm
43		eren . weet je die		muis	van de hand dat zijn wel spieren maar weinig diertjes hebben pijnlijke spie
42		ren dat is dan de		muis	van de hand . want da 's spieren . maar pijnlijk spieren . nou als je nou z
7		het juist van die		muis	toch RSI ? ja . ja . en*x heleboel*x ben*x je*x ... heel veel wel . maar
2		bijna niet te doen		muis	sturen met je met je ... ja nee dat is xxx . in tekenpakketten en zo . als
34		en zat er dus een		muis	op die dorpel . weet je nog die spin die ggg in de hal . en in de
47		... ik denk dat dit		muisje	nog een lelijk staartje heeft*z voor de ... ja . uh ja tot nog toe vind ik
3		zo . als je dan die		muis	met je stem moet sturen dan ben je ... mm-hu ja . en dan houdt 't op . pixe
45		mouw racete die		muis	maar op en neer totdat je natuurlijk heel die jas kapot gebeten had . want
40		gif weg maar de		muis	liep nog vrolijk rond . ggg . xxx . nou dat bordje met gif stond er toch al

1 of 50 rows

SoNaR context of row 3 (41)

hebben toen toch ... d*a die die dat gif dat werkt ook niet . we hebben toen nog dat **muisje** wat hier achter de kast zat dat had je ook aan elke kant van de kast een bordje gif staan 's ochtends was 't bordje gif weg maar de muis liep nog vrolijk rond . ggg . xxx . nou dat bordje met gif stond er toch al een

Current results Phase-1

- PoS: nouns, verbs and adjectives
- number of annotated lemmas: 2,589
- number of word senses: 10,172
- number of overlapping annotations: 255,625
(67% SoNaR, 5% CGN, 28% Snippets)
- Inter Annotator Agreement: 93%
- Coverage of senses with 25 examples: 77%
- Coverage of annotations for words: 86%

WSD systems

- Knowledge-based WSD that employs the relations from Cornetto and English WordNet.
- Supervised machine learning-based WSD that creates word experts from annotated examples
- Named Entity recognition and Wikification
- Domain classification of paragraphs

Knowledge-based WSD

- UKB (Agirre and Soroa 2009):
<http://ixa2.si.ehu.es/ukb/>
- Wordnet considered as a graph
- Personalized PageRank algorithm: meanings of the context words activate the weights in the graphs which are propagated to the target meanings

Semantic relations

- Dutch wordnet relations
- Dutch wordnet to English wordnet
- English wordnet relations
- Wordnet domain relations
 - *tennis player, tennis ball* => tennis => sport
 - *football player, football* => soccer => sport
- Annotation co-occurrence relations
 - *koning* (king), *koningin* (queen), *paard* (horse), *loper* (bishop), *toren* (tower), *stuk* (chess piece) and *slaan* (take a chess piece)

Nr. relations for UKB-graph

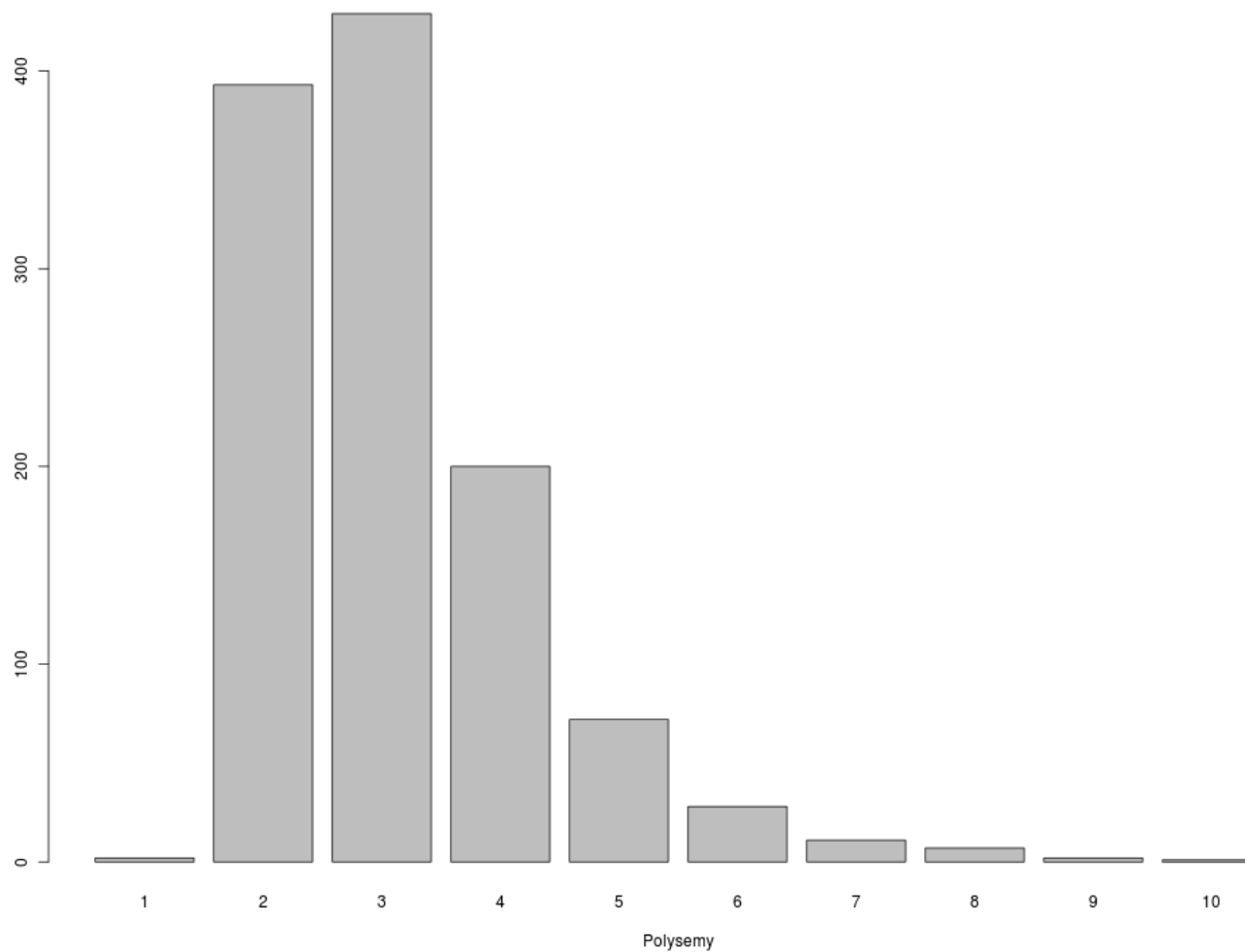
Type of relation	Relations
Dutch_synset/Dutch_synset	140,219
Domain/Domain	125
Dutch_synset/Domain	86,798
Dutch_synset/English_synset	73,935
English_synset/English_synset	252,392
English_synset/English_gloss_synset	419,387
Annotation co-occurrences polysemous	17,152
Annotation co-occurrences monosemous	151,598
Total	1,266,481

Different UKB graphs

- UKB1: D. Synset:D. Synset & Dom.:Dom. & D. Synset:Dom.
- UKB2: UKB1 & D. synset:E. synset
- UKB3: UKB2 & E. synset:E. Synset & E. synset:E. gloss
- UKB4: UKB1 & annotation co-occurrences
- UKB5: UKB3 & annotation co-occurrences

Test set

35,269 tokens (nouns and verbs)



Evaluation

	Precision	Recall	F-measure
UKB1	0.4557	0.4491	0.4523
UKB2	0.4557	0.4491	0.4524
UKB3	0.4560	0.4493	0.4526
UKB4	0.6360	0.6272	0.6316
UKB5	0.6411	0.6322	0.6366

For comparison SemEval2010 Task on WSD in specific domain, all-words-task:

- UKB3 **52.6** precision
- English UKB **48.1** precision

- UKB5 & UKB4 gained 9 points on UKB3 due to co-occurrence relations

Memory-based word-experts

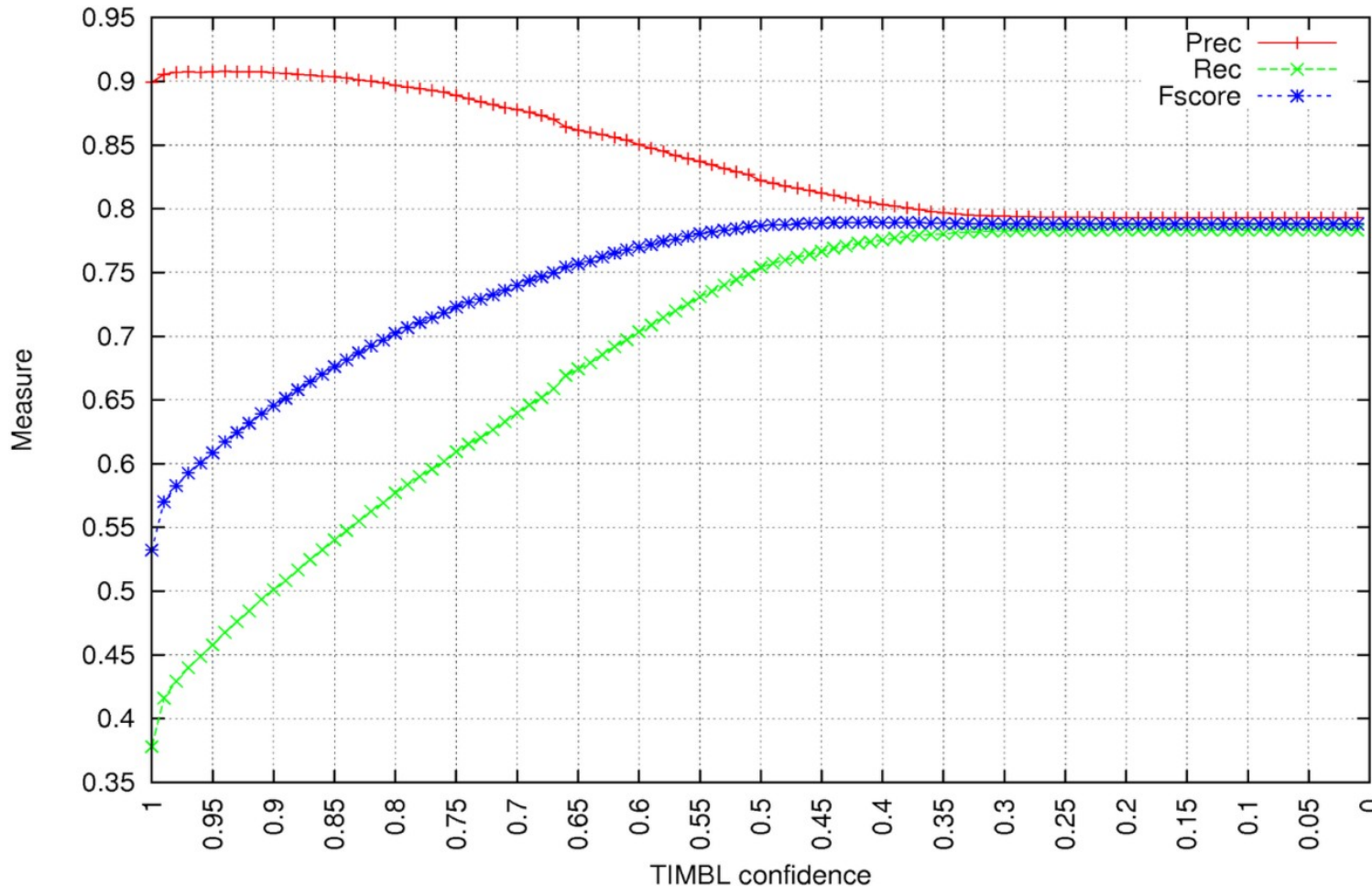
- TiMBL (Daelemans et al 2007)
- Word expert for each word (3,000), employing k-Nearest Neighbour classification
- Feature-vector:
 - Local context (word + words left and right)
 - Global context (binary bag-of-predictor-words in same sentence)
 - Domain label
 - Parameter optimisation
- 10 test examples & 15 train examples (random)

TiMBL results

Feature set	Token accuracy
Words ₁	0.6462
Words ₁ + Bag-of-words	0.7259
Words ₁ + PoS ₁ + Bag-of-words	0.7226
Words ₁ + Bag-of-words + PS	0.7931

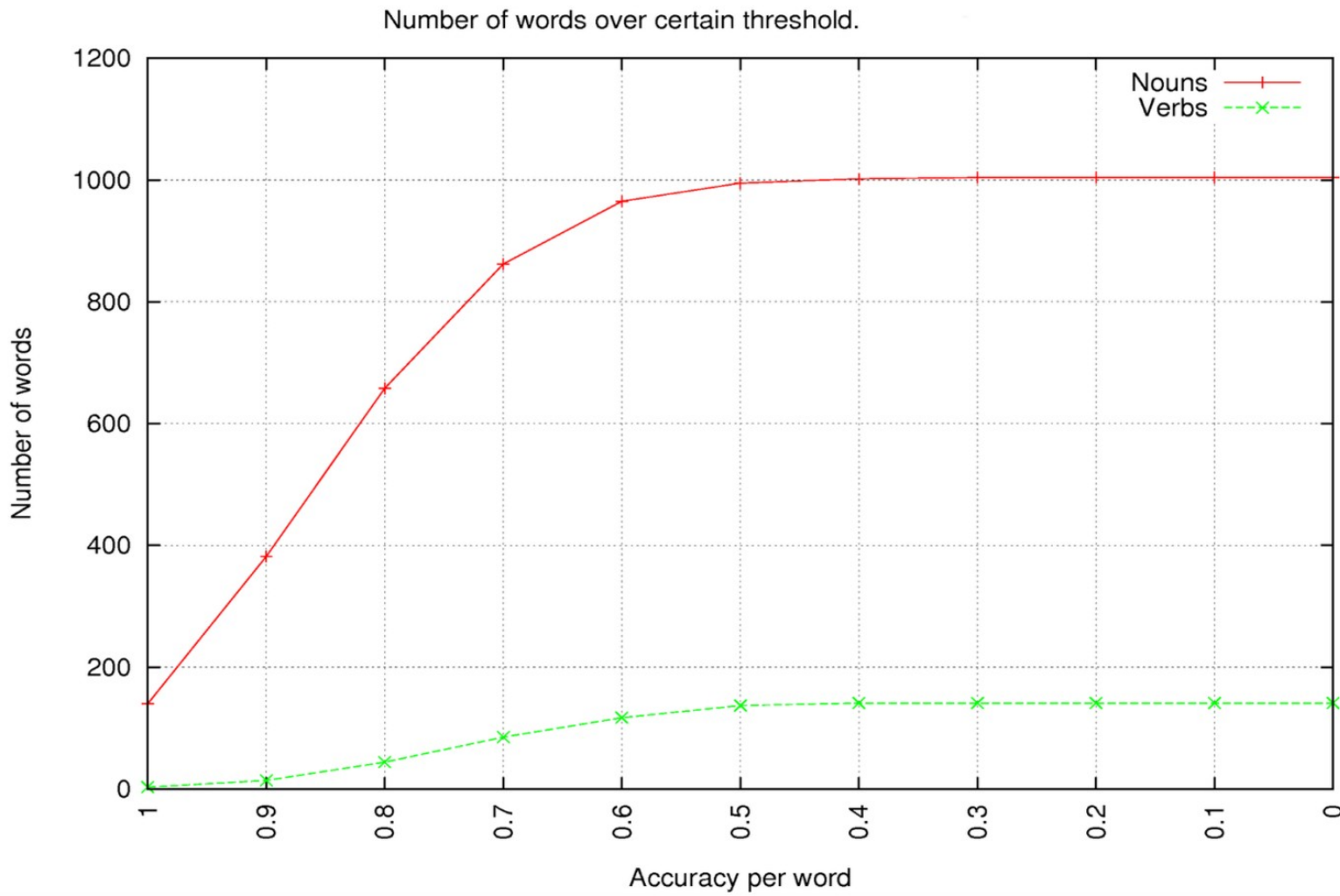
- Bag-of-predictor-words improvement of 8%
- Parameter optimisation (PS) improvement of another 7%

TiMBL results



confidence 0.55 & precision 0.84 (+0.44 compared to no filtering) and Fscore 0.78 (only -0.03 less than no filtering)

Accuracy per word



Phase-2: active learning

- 1) Train WSD with current data (minus the test set) & determine word accuracies and word meaning F-measures.
- 2) Select words with accuracy below 80%. Words that already perform well are ignored.
- 3) Apply WSD to all tokens of selected words not annotated.
- 4) Select tokens where WSD assigned meaning with F-measure below 80%. Tokens with good performing meanings are ignored.
- 5) Active-learning score: tokens of which WSD is sure it is the weak sense but that have different features from training set):
 - 1) $(\text{TiMBLE confidence} + \text{feature-distance}) / 2$
- 6) Annotate 50 examples per weak meaning in two weeks and return to step 1)

Future work

- Fine tune the active learning;
- Optimise the WSD systems
- Combine named-entity recognition with WSD
- Combine different WSD systems
- Test on independent texts in all-words task
- Apply optimal system to full corpora (over 500K tokens)

Thanks to

- Anneleen Schoen
- Charlotte van Tongeren
- Daphne van Kessel
- Dieke Janssen
- Elizabeth van Zutphen
- Gratia Bruining
- Jonica Kaagman
- Laura Kipp
- Lisanne Ranzijn
- Marlisa Hommel
- Wilma van Velzen
- Milou Kerkhof
- Sam Vossen
- Niqee Vossen
- Rosa Scheffer
- Chantal van Son