

Semantic Tagging: 'Supersense' and Word Sense Tagging

Diana McCarthy and Siva Reddy

Lexical Computing Ltd

16th March 2011

Two parts:

Semantic Class Tagging

Dante

Word Sense Disambiguation

Frequency Population

Motivation

- ▶ to see behaviour of words, in terms of semantics as well as syntax
- ▶ linking an inventory/dictionary to concordance lines
- ▶ possible applications, for example
 - ▶ language learning resources
 - ▶ linguistics research

Semantic Class Tagging

Tag the corpus with semantic classes. Which classes?

- ▶ USAS (<http://ucrel.lancs.ac.uk/usas/>) Paul Rayson and colleagues
- ▶ Super Sense Tagger (SST) [Ciaramita and Altun, 2006] (<http://sourceforge.net/projects/supersensetag/>)

Super Sense Tagger

- ▶ semantic tags are WordNet lexicographer classes
- ▶ supervised (requires training data) HMM system (tag sequences)
- ▶ SemCor used as training data

Noun Classes

act	acts or actions
object	natural objects (not man-made)
animal	animals
quantity	quantities and units of measure
artifact	man-made objects
phenomenon	natural phenomena
attribute	attributes of people and objects plant plants
...	...

Verb Classes

body	grooming, dressing and bodily care
emotion	feeling
change	size, temperature change, intensifying
motion	walking, flying, swimming
cognition	thinking, judging, analyzing, doubting
perception	seeing, hearing, feeling
communication	telling, asking, ordering, singing
possession	buying, selling, owning
...	...

New Model 'Super Sensed'

Demo (or slides)

Semantic Tags in the Concordance

gen00015 antelope, a steenbuck, a lizard, a striped **mouse** /animal.n , mantis, baobab tree, yellow-crested cobra
gen00087 real world that are then deemed `mickey **mouse** /animal.n . What ever such populist criticism of
gen00087 flags probably come from South Korea, Mickey **Mouse** /other.n T-shirts from Morocco. Emphasising that
gen00111 merchandise now available IYM Scotland T shirts, **mouse** /artifact.n mats, clocks and mugs are now available
gen00111 is: Adult T shirt £10 Kids T shirt £8 **Mouse** /artifact.n mat £5 Small clock £8 Large clock £10
gen00284 were involved in an endless game of cat and **mouse** /animal.n . - - - - - AS MY
gen00325 alpha-1, 3-galactosyltransferase knockout **mouse** /animal.n - implications for xenotransplantation,
gen00337 Chipmunk, Grey Squirrel, and White-footed **Mouse** /animal.n rarely ventured onto road surfaces when
gen00394 could resist? ``Click `` (with the computer **mouse** /artifact.n): a page of description appeared, full
gen00431 allowed entry and searched the house. A **mouse** /animal.n was discovered and removed. Benefits of
gen00522 between two artefacts by simple use of the **mouse** /animal.n . Fig. 7 Plot database The obvious advantages
gen00543 Sprawl report from the Sierra Club. ``A **mouse** /animal.n , and other naturalists, reclaim Paris streets
gen00546 more than 50 %, as did CO emissions. ``A **mouse** /animal.n , and other naturalists, reclaim Paris streets
gen00555 interactive, because you get to click with a **mouse** /animal.n and fill in forms! I have mentioned that
gen00591 grown then in the trifle of her days, a **mouse** /animal.n , a mere tittle, trots offwith the whole
gen00609 knowing the precise focal length.'Mickey **Mouse** /other.n ' - one, usually AF 35-mm Canon. A point
gen00654 social critic Kirkpatrick Sale writes, ``If a **mouse** /animal.n were to be as big as an elephant, it would
gen00669 ``. These pins have a picture of Mickey **Mouse** /other.n on the front, riding an old-fashioned big-wheel
gen00682 follow. Back to top Change the world, one **mouse** /act.n click at a time My friends at Anti-Apathy
gen00682 call it changing the world, one lazy ass **mouse** /act.n click at a time. www. thenag. net / join

Semantic Tags in the Word Sketch

sing_(verb)

New Model super sensed freq = 5675

NP	1543	3.8	subj_NP	1297	4.0	NP_ne
communication	872	4.26	animal	107	4.85	TI
all	28	2.84	person	646	2.33	ANIM
relation	17	2.0	plant	7	1.67	WORK_OF_ART
animal	13	1.81	group	169	1.12	ORGANIZATION
creation	20	1.38	object	12	1.02	ORG_DESC:OT
time	53	1.26	communication	70	0.62	ORGANIZATION
attribute	42	1.17	body	10	0.44	DATE:DA
quantity	10	1.06	phenomenon	6	0.34	PERS
feeling	7	0.93	artifact	46	0.14	
event	15	0.81	quantity	5	0.06	ORG_DESC:CO
object	8	0.43	time	23	0.06	PER_DE
food	5	0.36				
person	158	0.3				
substance	9	0.24				
act	54	0.06				
			subj_NP_ne	1297	4.0	
			ANIM	109	5.61	
			PERS	250	3.43	
			PER_DE	379	3.02	
			ORGANIZATION:CORPORATI	135	2.17	

Semantic Word Sketch, for example

*DUAL

=N_mod/N_premod

*COLLOC "%(2.sense)"

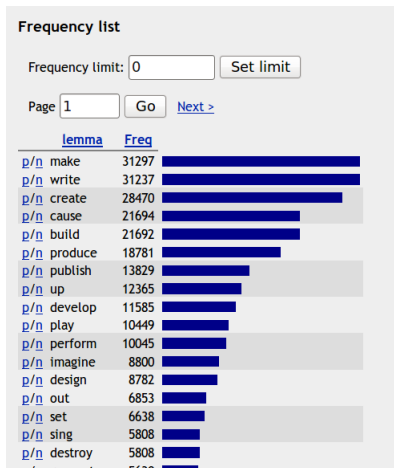
2:any_noun any_noun{0,3} 1:any_noun

Semantic Tags in the Word Sketch

eat (*verb*) New Model super sensed freq = 14232

NP	4956	5.5	subj NP	1301	1.8	NP_ne	4956	5.5	subj NP_ne	1301	1.8
food	1989	8.96	animal	160	5.43	SUBSTANCE:FO	1353	9.06	ANIM	134	5.91
plant	248	6.74	person	575	2.16	ANIM	235	6.65	NORP:NATIONALI	20	3.91
animal	258	6.08	food	13	1.74	QUANTITY:WEIG	8	5.27	PER_DE	383	3.03
all	212	5.73	object	17	1.52	PLA	21	4.98	PERS	142	2.61
quantity	122	4.64	substance	20	1.4	SUBSTANCE:OTH	63	4.89	SUBSTANCE:FO	10	2.04
shape	23	4.61	feeling	7	0.94	DATE:DA	233	4.74	ORGANIZATION:CORPORATI	97	1.69
substance	146	4.24	group	140	0.85	TI	34	4.24	SUBSTANCE:OTH	5	1.32
consumption	51	3.74	artifact	63	0.6	PRODUCT:OTH	10	3.99	ORG_DESC:CORPORATI	17	1.21
body	82	3.45	quantity	7	0.55	SUBSTANCE:CHEMIC	11	3.89	FAC_DESC:BUILDI	6	0.89
time	153	2.78	state	30	0.51	SUBSTANCE:DR	8	3.29	GPE:COUNT	6	0.09
artifact	259	2.63	body	9	0.29	CARDIN	7	3.18			
state	127	2.58	time	26	0.23	DATE:DURATI	17	3.17			
object	33	2.45				FAC_DESC:OTH	7	2.98			
relation	19	2.13				PRODUCT:VEHIC	5	2.9			

Semantic Word Lists



DANTE: Database of Analysed Texts of English

- ▶ lexical resource: a monolingual analysis of English
- ▶ commissioned by Foras na Gaeilge for New English Irish Dictionary
- ▶ Lexicographers led by Sue Atkins
- ▶ corpus based (1.7 billion words) using word sketches
- ▶ concordance sorted with 'GDEX'
- ▶ entries for:
 - ▶ 42,000 headwords (6,300+ verbs)
 - ▶ 27,000 idioms and phrases
 - ▶ 20,500 compounds
 - ▶ just under 3,000 phrasal verbs

Dante: Contents

- ▶ meanings with definitions
- ▶ over 622,000 examples from the corpus,
- ▶ argument structure (valency) e.g. NP-Vinf *let him go* (42 frames for verbs, further specified by preposition)
- ▶ attitude e.g. *meddle* (pejorative)
- ▶ regional e.g. *nick* (British) as in *you're nicked*
- ▶ style e.g. *oxidise* (technical) *perambulate* (humorous)
- ▶ register e.g. *ameliorate* (formal) *go ape* (informal)
- ▶ subject e.g. *multiply* (maths)
- ▶ time e.g. *punch* (cattle: dated) or *quoth* (obsolete)
- ▶ inherent grammar e.g. reciprocal
John marries Mary ↔ *Mary and John marry*
- ▶ support verbs e.g. *make an appeal*

see webdante.com

Dante Disambiguation

mouse: (PoS: n)

meaning: a small long-tailed rodent

domain: zoo

example: *The **mouse** was dead in his cage the following day*

...

SCF: N_PREMOD

COLLOC: droppings nest hole cage

example: *Look for signs of **mouse droppings** etc.*

example: **Mouse** cages are available in various stages, sizes and designs.

...

mouse: (PoS: n)

meaning: a computer input device controlled with one hand which moves the cursor on the computer screen

domain: IT

example: *If your mouse runs off the mat edge, lift the **mouse** up, move it back to the mat middle, and put it down*

...

Outline

Semantic Class Tagging

Dante

Word Sense Disambiguation

Frequency Population

Demo

Corpus: DanteDisambiguationDemo

Hits: 31

Page 1 of 2 Go [Next](#) | [Last](#)

1777 objective. Are those the candy-floss , Mickey **Mouse** /

3559 design is , frankly , an abomination . The **mouse** /a computer input device controlled with one hand which moves the cursor

5569 accomplished with a few clicks of a computer **mouse** /a computer input device controlled with one hand which moves the cursor

8376 Europe Researchers are adepts at manipulating **mouse** /a computer input device controlled with one hand which moves the cursor

27499 baited the trap again just to check that the **mouse** /a computer input device controlled with one hand which moves the cursor

30569 's Trust charity . THE fact that Mickey **Mouse** /

36268 than a decade of research on the biology of **mouse** /a small long-tailed rodent

42716 direct human intervention . You can use your **mouse** /a computer input device controlled with one hand which moves the cursor

51526 Click on the printer icon with the right **mouse** /a computer input device controlled with one hand which moves the cursor

52431 intelligence service to an endless game of cat and **mouse** /

53117 logging the call . A) Donald Duck B) Mickey **Mouse** /

58152 consistently entertaining series of cat and **mouse** /

61932 of an earthquake , and through which the **Mouse** /

62432 short of a tax form or loan application . A **mouse** /a computer input device controlled with one hand which moves the cursor

67294 above logo . You answer by clicking with the **mouse** /a computer input device controlled with one hand which moves the cursor

67297 a quick click and it 's all over A right **mouse** /a computer input device controlled with one hand which moves the cursor

79025 and racing consumes their time . Like the **mouse** /

79874 a save dialog , you can right click the **mouse** /a computer input device controlled with one hand which moves the cursor

105734 with that wedge of paper . He fell on a **mouse** /

111287 of mountaiiny men . " When I move my USB **mouse** /a computer input device controlled with one hand which moves the cursor

Page 1 of 2 Go [Next](#) | [Last](#)

Dante Frequency Population

Marrow

Meaning: the substance inside bone cavities that produces blood cells

PoS n

Freq in lexmci: 6026

Structural Information:

GRAMREL: N_premod

Freq in lexmci: 3131

Collocate: *transplant*

Freq in lexmci: 602

Collocate: *donor*

Freq in lexmci: 182

Collocate: *biopsy*

Freq in lexmci: 116

Reference



Ciaramita, M. and Altun, Y. (2006).

Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger.

In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 594–602, Sydney, Australia. Association for Computational Linguistics.