Word Categorization

A comparison of bag of words and syntax-based approaches

Tim Van de Cruys

University of Groningen

ESSLLI lexical semantics workshop
August 6, 2008
Hamburg

Semantic similarity

- Most work on semantic similarity relies on the DISTRIBUTIONAL HYPOTHESIS (Harris 1954)
  - Take a word and its contexts:
    - tasty *spätzlen*
    - greasy *spätzlen*
    - a plate of *spätzlen*
    - *spätzlen* with cheese
- By looking at a word's context, one can infer its meaning

Semantic similarity

- Most work on semantic similarity relies on the DISTRIBUTIONAL HYPOTHESIS (Harris 1954)
  - Take a word and its contexts:
    - tasty *spätzlen*
    - greasy *spätzlen*
    - a plate of *spätzlen*
    - *spätzlen* with cheese

  $\Rightarrow$ **FOOD**

- By looking at a word's context, one can infer its meaning

Semantic similarity

- Most work on semantic similarity relies on the DISTRIBUTIONAL HYPOTHESIS (Harris 1954)

  - Take a word and its contexts:
    - tasty *spätzlen*
    - greasy *spätzlen*
    - a plate of *spätzlen*
    - *spätzlen* with cheese

- By looking at a word's context, one can infer its meaning

Two kinds of context

1. 'Bag of words' context
   - a window around the word is used as context
   - e.g. a fixed numbers of words, the paragraph in which a word appears, ...
2. Syntactic context
   - a corpus is parsed, dependency triples are extracted
   - e.g. <apple, obj, eat>, <apple, adj, red>

Different kinds of similarity

1. Bag of words context
   - thematic similarity
   - **muziek** 'music': *gitaar* 'guitar', *jazz* 'jazz', *cd* 'cd', *rock* 'rock', *bas* 'bass', *song* 'song', *muzikant* 'musician', *musicus* 'musician', *drum* 'drum', *slagwerker* 'drummer'
2. Syntactic context
   - tighter, synonym-like similarity
   - **muziek** 'music': *dans* 'dance', *kunst* 'art', *klank* 'sound', *liedje* 'song', *geluid* 'sound', *poëzie* 'poetry', *literatuur* 'literature', *popmuziek* 'pop music', *lied* 'song', *melodie* 'melody'

Introduction

- Three word categorization tasks
  1. concrete noun categorization
  2. concrete/abstract noun discrimination
  3. verb categorization
- Two contexts
  - bag of words context
  - syntactic context
- Research question: is the difference between both contexts present in the clustering results?

Methodology

- Approach carried out for Dutch (translations)
- TWNC parsed with Dutch dependency parser ALPINO
- paragraph as window in bag of words method
- dependency triples in syntax-based method (e.g. <*apple*,*obj1*,*eat*>)
- Frequency matrices adapted with POINTWISE MUTUAL INFORMATION

Evaluation

- ENTROPY: distribution of various semantic classes within each cluster
- PURITY: extent to which each cluster primarily contains one class
- Both measures: $0 \leq x \leq 1$
- Low entropy and high purity indicates successful clustering

Task

- Clustering of 44 nouns into a number of classes:
  - 2-way clustering: *natural*, *artefact*
  - 3-way clustering: *animal*, *vegetable* and *artefact*
  - 6-way clustering: *bird*, *groundAnimal*, *fruitTree*, *green*, *tool*, *vehicle*

Bag of Words

n-way	entropy	purity
2	.983	.545
3	.539	.705
6	.334	.682

Example clustering

car	banana	boat	bottle	cat	chisel
motorcycle	bowl	helicopter	cup	cow	hammer
telephone	cherry	rocket	kettle	dog	knife
	chicken	ship		duck	pen
	corn	truck		eagle	pencil
	lettuce			elephant	scissors
	mushroom			lion	screwdriver
	onion			owl	
	pear			peacock	
	pineapple			penguin	
	potato			pig	
	spoon			snail	
				swan	
				turtle	

Example clustering

car	banana	boat	bottle	cat	chisel
motorcycle	bowl	helicopter	cup	cow	hammer
telephone	cherry	rocket	kettle	dog	knife
	chicken	ship		duck	pen
	corn	truck		eagle	pencil
	lettuce			elephant	scissors
	mushroom			lion	screwdriver
	onion			owl	
	pear			peacock	
	pineapple			penguin	
	potato			pig	
	spoon			snail	
				swan	
				turtle	

Syntax-based

n-way	entropy	purity
2	.000	1.000
3	.000	1.000
6	.173	.841

Example clustering

banana	chisel	duck	boat	bottle	cat
cherry	hammer	eagle	car	bowl	chicken
corn	knife	owl	helicopter	cup	cow
lettuce	pen	peacock	motorcycle	kettle	dog
mushroom	pencil	penguin	rocket	spoon	elephant
onion	scissors	swan	ship	telephone	lion
pear	screwdriver		truck		pig
pineapple					snail
potato					turtle

Example clustering

banana	chisel	duck	boat	bottle	cat
cherry	hammer	eagle	car	bowl	chicken
corn	knife	owl	helicopter	cup	cow
lettuce	pen	peacock	motorcycle	kettle	dog
mushroom	pencil	penguin	rocket	spoon	elephant
onion	scissors	swan	ship	telephone	lion
pear	screwdriver		truck		pig
pineapple					snail
potato					turtle

Example clustering

banana	chisel	duck	boat	bottle	cat
cherry	hammer	eagle	car	bowl	chicken
corn	knife	owl	helicopter	cup	cow
lettuce	pen	peacock	motorcycle	kettle	dog
mushroom	pencil	penguin	rocket	spoon	elephant
onion	scissors	swan	ship	telephone	lion
pear	screwdriver		truck		pig
pineapple					snail
potato					turtle

Example clustering

banana	chisel	duck	boat	bottle	cat
cherry	hammer	eagle	car	bowl	chicken
corn	knife	owl	helicopter	cup	cow
lettuce	pen	peacock	motorcycle	kettle	dog
mushroom	pencil	penguin	rocket	spoon	elephant
onion	scissors	swan	ship	telephone	lion
pear	screwdriver		truck		pig
pineapple					snail
potato					turtle

Task

- Part I: 30 nouns (15 HI concrete, 15 LO concrete) clustered in two clusters
- Part II: 10 nouns ($\pm$ concrete) added
- Part III: 3-way clustering of all 40 nouns

Bag of words

part	entropy	purity
part 1	.000	1.000
part 3	.605	.700

Example clustering

banana	pollution	$\leftrightarrow$	ache	belief
bottle	smell		ceremony	concept
bowl	weather		empire	distraction
car			fight	gratitude
chicken			foundation	hope
eagle			invitation	hypothesis
hammer			shape	insight
lion				jealousy
onion				luck
pencil				mercy
potato				mystery
ship				pride
telephone				temptation
truck				truth
turtle				wisdom

Syntax-based

part	entropy	purity
part 1	.000	1.000
part 3	.367	.750

Example clustering

banana	ceremony	↔	ache	belief
bottle	invitation		empire	concept
bowl	weather		fight	distraction
car			foundation	gratitude
chicken			pollution	hope
eagle			shape	hypothesis
hammer			smell	insight
lion				jealousy
onion				luck
pencil				mercy
potato				mystery
ship				pride
telephone				temptation
truck				truth
turtle				wisdom

Task

- Clustering of 45 nouns into a number of classes:
  - 5-way clustering: *cognition*, *motion*, *body*, *exchange*, *changeState*
  - 9-way clustering: *communication*, *mentalState*, *motionManner*,*motionDirection*, *changeLocation*, *bodySense*, *bodyAction*, *exchange*, *changeState*

Bag of words

n-way	entropy	purity
5	.463	.600
9	.442	.556

Example clustering

pay	leave	drive	breathe	break
repair	remember	check	move	kill
buy	cry	evaluate	carry	enter
lend	smile	speak	drink	arrive
notice	read	suggest	push	ride
sell	listen	request	eat	die
acquire	talk	send	look	destroy
	forget		run	fly
	feel		smell	
	know		rise	
			pull	
			walk	

Syntax-based

n-way	entropy	purity
5	.464	.667
9	.408	.556

Example clustering

move	drive	breathe	break	drink
push	pay	remember	kill	eat
look	check	read	evaluate	enter
run	carry	notice	repair	leave
ride	buy	talk	arrive	cry
pull	lend	speak	die	smile
fly	sell	suggest	fall	listen
walk	destroy	forget	request	smell
send	acquire	feel		rise
		know		

Conclusion

- Semantic space models – both bag of words and syntactic – are fruitful models for the induction of semantic classes
- Noun categorization works very well – verbs are more difficult
- Syntax-based models tend to work better (for the workshop's clustering tasks)
- Bag of word topics $\leftrightarrow$ syntax-based 'features'?