Distributional Semantic Models in computational linguistics and cognitive science

Stefan Evert[1], Alessandro Lenci[2]

[1]University of Osnabrück
[2]University of Pisa

Bordeaux, July 29 2009

Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

Distributional similarity as semantic similarity

- DSMs interpret semantic similarity as a quantitative notion
  - if $\vec{a}$ is closer to $\vec{b}$ in the distributional vector space, than *a* is more semantically similar to *b*

rhino	fall	rock
woodpecker	rise	lava
rhinoceros	increase	sand
swan	fluctuation	boulder
whale	drop	ice
ivory	decrease	jazz
plover	reduction	slab
elephant	logarithm	cliff
bear	decline	pop
satin	cut	basalt
sweatshirt	hike	crevice

Types of semantic relations in DSMs

- Neighbors in DSMs have different types of semantic relations with the target

car (InfomapNLP on BNC; n = 2)

- van co-hyponym
- vehicle hyperonym
- truck co-hyponym
- motorcycle co-hyponym
- driver related entity
- motor part
- lorry co-hyponym
- motorist related entity
- cavalier hyponym
- bike co-hyponym

car (InfomapNLP on BNC; n = 30)

- drive function
- park typical action
- bonnet part
- windscreen part
- hatchback part
- headlight part
- jaguar hyponym
- garage location
- cavalier hyponym
- tyre part

Semantic similarity and relatedness

- Semantic similarity - two words sharing a high number of salient features (attributes)
  - synonymy (*car/automobile*)
  - hyperonymy (*car/vehicle*)
  - co-hyponymy (*car/van/truck*)
- Semantic relatedness (Budanitsky & Hirst 2006) - two words semantically associated without being necessarily similar
  - function (*car/drive*)
  - meronymy (*car/tyre*)
  - location (*car/road*)
  - attribute (*car/fast*)

Analyzing a distributional semantic space

Baroni & Lenci (2008)

- We selected the 10 nearest neighbors in a DSM for a set of concrete nouns
  - corpus BNC
  - T 2K most frequent N, V, A and ADVs
  - R 5-word symmetric window
  - d SVD (Infomap NLP)
- Each neighbor was classified with respect to the type of semantic relation with the target concept
  - Wu & Barsalou (2009) taxonomy of property types
- We compared the neighbors in DSMs with human-generated properties extracted from McRae et al. (2005) semantic norms
  - comparison between the human-generated properties and the neighbors generated by the DSM was carried out at the level of their semantic type

Concrete nouns

ESSLLI 2008 dataset

Word	Semantic Category	Word	Semantic Category
chicken	bird-animal-natural	*onion*	green-vegetable-natural
duck	bird-animal-natural	*potato*	green-vegetable-natural
eagle	bird-animal-natural	*bottle*	tool-artifact
owl	bird-animal-natural	*bowl*	tool-artifact
peacock	bird-animal-natural	*chisel*	tool-artifact
penguin	bird-animal-natural	*cup*	tool-artifact
swan	bird-animal-natural	*hammer*	tool-artifact
cat	groundAnimal-animal-natural	*kettle*	tool-artifact
cow	groundAnimal-animal-natural	*knife*	tool-artifact
dog	groundAnimal-animal-natural	*pen*	tool-artifact
elephant	groundAnimal-animal-natural	*pencil*	tool-artifact
lion	groundAnimal-animal-natural	*scissors*	tool-artifact
pig	groundAnimal-animal-natural	*screwdriver*	tool-artifact
snail	groundAnimal-animal-natural	*spoon*	tool-artifact
turtle	groundAnimal-animal-natural	*telephone*	tool-artifact
banana	fruit-vegetable-natural	*boat*	vehicle-artifact
cherry	fruit-vegetable-natural	*car*	vehicle-artifact
pear	fruit-vegetable-natural	*helicopter*	vehicle-artifact
pineapple	fruit-vegetable-natural	*motorcycle*	vehicle-artifact
corn	green-vegetable-natural	*rocket*	vehicle-artifact
lettuce	green-vegetable-natural	*ship*	vehicle-artifact
mushroom	green-vegetable-natural	*truck*	vehicle-artifact

Semantic norms

McRae et al. (2005)

- Semantic properties collected from approximately 725 participants for 541 living (*dog*) and nonliving (*car*) basic-level concepts
  - property salience estimated with its production frequency
    - number of subjects (out of 20) that have produced the property for a given concept

car	property type	production freq
has_wheels	external_component	19
used_for_transportation	function	19
has_4_wheels	external_component	18
has_an_engine	internal_component	13
has_doors	external_component	13
has_a_steering_wheel	internal_component	12
requires_gasoline	contingency	12
is_expensive	systemic_property	11
a_vehicle	superordinate	9
is_fast	systemic_property	9
used_for_passengers	participant	9
causes_pollution	contingency	8
requires_drivers	contingency	7
different_colours	external_surface_property	6

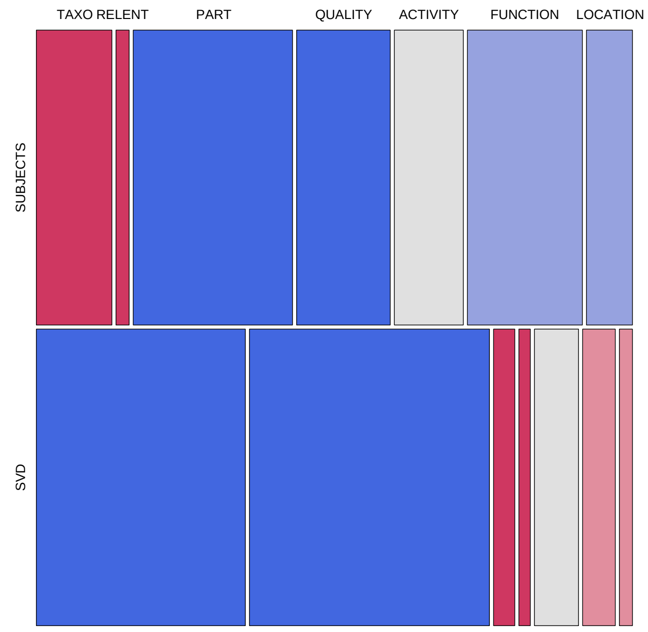
Property types

Wu & Barsalou (2009)

Class	*Property Type*	*Code*	*Example*
Taxonomy (c)	Coordinate	cc	*cat*–dog
	Superordinate	ch	*cat*–animal
Entity (e)	Associated abstract entity	eae	*telephone*–information
	Entity behavior	eb	*lion*–roar
	External component	ece	*truck*–wheel
	External surface property	ese	*banana*–yellow
	Internal component	eci	*car*–engine
	Internal surface property	esi	*pineapple*–crunchy
	Larger whole	ew	*cow*–cattle
	Made–of	em	*bottle*-glass
	Quantity	eq	*pear*–slice
	Systemic feature	esys	*elephant*–wild
Situation (s)	Associated entity	se	*spoon*–bowl
	Associated event	sev	*watermelon*–picnic
	Function	sf	*scissors*–cut
	Action	sa	*banana*–eat
	Location	sl	*ship*–port
	Participant	sp	*boat*–fisherman
	Time	st	*pineapple*–summer
Introspective (i)	Cognitive operation	io	*snail*–like_a_slug
	Evaluation	ie	*pineapple*–delicious
	Negation	in	*penguin*–cannot_fly

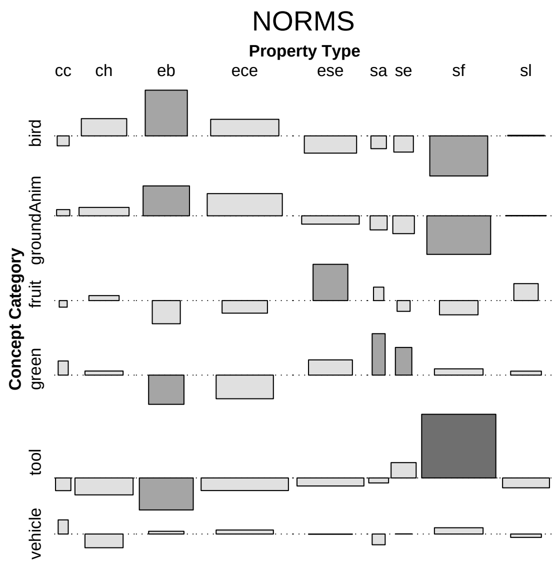
Other semantic relations

Baroni & Lenci (2008)



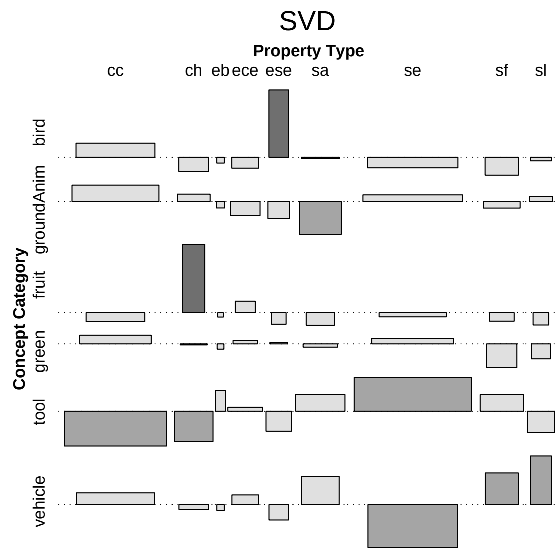
Property spaces

NORMS (McRae et al. 2005)



Property spaces

SVD



Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

DSMs and semantic similarity

- These models emphasize paradigmatic similarity
  - words that tend to occur in the same contexts
- Words that share many contexts will correspond to concepts that share many attributes (attributional similarity), i.e. concepts that are taxonomically/ontologically similar
  - synonyms (*rhino/rhinoceros*)
  - antonyms and values on a scale (*good/bad*)
  - co-hyponyms (*rock/jazz*)
  - hyper- and hyponyms (*rock/basalt*)
- Taxonomic similarity is seen as the fundamental semantic relation, allowing categorization, generalization, inheritance

DSMs for attributional similarity

- Synonym identification
  - TOEFL test
- Modeling semantic similarity judgments
  - the Rubenstein/Goodenough norms
- Noun categorization
  - the ESSLLI 2008 dataset
- Semantic priming
  - the Hodgson dataset

Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *imposed*, *believed*, *requested*, *correlated*
- DSMs and TOEFL
  1. take vectors of the target ($\vec{t}$) and of the candidates ($\vec{c_1} \dots \vec{c_n}$)
  2. measure the distance between $\vec{t}$ and $\vec{c_i}$, with $1 \leq i \leq n$
  3. select $\vec{c_i}$ with the shortest distance in space from $\vec{t}$

Human performance on the synonym match task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%

DSMs take the TOEFL

- Humans
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%
- Machines
  - Classic LSA: 64.4%
  - Padó and Lapata's dependency-based model: 73%
  - Rapp's 2003 SVD-based model trained on lemmatized BNC: 92.5%

Semantic similarity judgments

Dataset Rubenstein and Goodenough (1965) (R&G)
65 noun pairs rated by 51 subjects on a 0-4 similarity scale

car	*automobile*	3.9
food	*fruit*	2.7
cord	*smile*	0.0

- DSMs and R&G
  1. for each test pair $< w_1, w_2 >$, take vectors $\vec{w_1}$ and $\vec{w_2}$
  2. measure the distance (e.g. cosine) between $\vec{w_1}$ and $\vec{w_2}$
  3. measure (with Pearson's r) the correlation between vector distances and R&G average judgments (Padó and Lapata 2007)

model	*r*
dep-filtered+SVD	0.8
dep-filtered	0.7
dep-linked (DM)	0.64
window	0.63

Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

Categorization

- In categorization tasks, subjects are typically asked to assign experimental items - objects, images, words - to a given category or to group together items belonging to the same category
  - categorization presupposes an understanding of the relationship between the items in a category
- Categorization is a basic cognitive operation presupposed by further semantic tasks
  - inference
    - if X is a CAR then X is a VEHICLE
  - compositionality
    - $\lambda y : FOOD \lambda x : ANIMATE(eat, x, y)$
- "Chicken-and-egg" conundrum in the relationship between categorization and similarity (cf. Goodman 1972, Medin et al. 1993)

Noun categorization

Dataset 44 concrete nouns (ESSLLI 2008 Distributional Semantics shared task)
- 24 natural entities
  - 15 animals: 7 birds (*eagle*), 8 ground animals (*lion*)
  - 9 plants: 4 fruits (*banana*), 5 greens (*onion*)
- 20 artifacts
  - 13 tools (*hammer*), 7 vehicles (*car*)

- DSMs and noun categorization
  - categorization can be operationalized as a clustering task
    1. for each noun w_i in the dataset, take its vector $\vec{w_i}$
    2. use a clustering method to group close vectors $\vec{w_i}$
    3. evaluate whether clusters correspond to gold-standard semantic classes

Noun categorization

- Clustering experiments with CLUTO (Karypis 2003)
  - repeated bisection algorithm
  - 6-way (birds, ground animals, fruits, greens, tools and vehicles), 3-way (animals, plants and artifacts) and 2-way (natural and artificial entities) clusterings
- Clusters evaluation
  - entropy – whether words from different classes are represented in the same cluster (best = 0)
  - purity – degree to which a cluster contains words from one class only (best = 1)
  - global score across the three clustering experiments

$$\sum_{i=1}^{3} Purity_i - \sum_{i=1}^{3} Entropy_i$$

Noun categorization
results

model	6-way		3-way		2-way		global
	P	E	P	E	P	E	
Katrenko	89	13	100	0	80	59	197
Peirsman+	82	23	84	34	86	55	140
dep-typed (DM)	77	24	79	38	59	97	56
dep-filtered	80	28	75	51	61	95	42
window	75	27	68	51	68	89	44
Peirsman–	73	28	71	54	61	96	27
Shaoul	41	77	52	84	55	93	-106

`Katrenko`, `Peirsman+/–`, `Shaoul`: ESSLLI 2008 Shared Task
`DM`: Baroni & Lenci (2009)

Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

Semantic priming

- Hearing/reading a “related” prime facilitates access to a target in various lexical tasks (naming, lexical decision, reading)
  - the word *pear* is recognized/accessed faster if it is heard/read after *apple*
- Hodgson (1991) single word lexical decision task, 136 prime-target pairs (cf. Padó & Lapata 2007)
  - similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):
    - synonyms (synonym): *to dread/to fear*
    - antonyms (antonym): *short/tall*
    - coordinates (coord): *train/truck*
    - super- and subordinate pairs (supersub): *container/bottle*
    - free association pairs (freeass): *dove/peace*
    - phrasal associates (phrasacc): *vacant/building*

Simulating semantic priming
McDonald & Brew (2004), Padó & Lapata (2007)

- DSMs and semantic priming
  1. for each related prime-target pair, measure cosine-based similarity between pair items (e.g., *to dread/to fear*)
  2. to estimate unrelated primes, take average of cosine-based similarity of target with other primes from same relation data-set (e.g., *value/to fear*)
  3. similarity between related items should be significantly higher than average similarity between unrelated items

Semantic priming results
Padó & Lapata (2007)

Mean distance values for Related and Unrelated prime–target pairs; Prime Effect size (= Related − Unrelated) for the dependency model and ICE.

Lexical Relation	*N*	Related	Unrelated	Effect (dependency)	Effect (ICE)
Synonymy	23	0.267	0.102	0.165**	0.063
Superordination	21	0.227	0.121	0.106**	0.067
Category coordination	23	0.256	0.119	0.137**	0.074
Antonymy	24	0.292	0.127	0.165**	0.097
Conceptual association	23	0.204	0.121	0.083**	0.086
Phrasal association	22	0.146	0.103	0.043**	0.058

**$p < 0.01$ (2-tailed)

Outline

Finding and distinguishing semantic relations

- Classic distributional semantic models are based on attributional similarity
  - single words/concepts that tend to share contexts/attributes are similar
- Attributional similarity can be modeled with DSMs that have single words as matrix raws and contexts as matrix columns
  - the contexts are the attributes shared by similar words

	die	kill	gun
teacher	109.4	0.0	0.0
victim	1335.2	22.4	0.0
soldier	4547.5	1306.9	105.9
policeman	68.6	38.2	30.5

Attributional and relational similarity
Turney (2006)

- *Policeman* is attributionally similar to *soldier*
  - they both occur in contexts such as: *kill X*, *with gun*, *for security*
- The pair *policeman-gun* is relationally similar to *teacher-book*
  - they both occur in contexts in which they are connected by *with*, *use*, *of*
- It is not always possible to reduce relational similarity to attributional similarity
  - *mason*:*stone*::*carpenter*:*wood* vs. *traffic*:*street*::*water*:*riverbed*
    - *mason* - *carpenter* and *stone* - *wood* are attributionally similar
    - *traffic* - *water* and *street* - *riverbed* are not attributionally similar

Finding and distinguishing semantic relations with DSMs

- Find non-taxonomic semantic relations
  - look at direct co-occurrences of word pairs in texts (when we talk about a concept, we are likely to also mention its parts, function, etc.)
- Distinguish between different semantic relations
  - use the contexts of pairs to measure pair similarity, and group them into coherent relation types by their contexts
  - *pairs* that occur in similar contexts (i.e. connected by similar words and structures) will tend to be related, with the shared contexts acting as a cue to the nature of their relation, i.e., measuring their *relational* similarity (Turney 2006)

DSMs and relational similarity

T (rows) word pairs

C (columns) syntagmatic links between the word pairs

		in	at	with	use
teacher	school	11894.4	7020.1	28.9	0.0
teacher	handbook	2.5	0.0	3.2	10.1
soldier	gun	2.8	10.3	105.9	41.0

Recognizing SAT analogies

- 374 SAT multiple-choice questions (Turney 2006)
- Each question includes 1 target pair (stem) and 5 answer pairs
- the task is to choose the pair most *analogous* to the stem

mason	stone
teacher	chalk
carpenter	wood
soldier	gun
photograph	camera
book	word

- Relational analogue to the TOEFL task
  1. for each pair p, take its row vector $\vec{p}$
  2. for each stem-pair, select the closest answer-pair (e.g. the one with the highest cosine)

Recognizing SAT analogies

Results

model	*% correct*	*model*	*% correct*
LRA	56.1	KnowBest	43.0
PERT	53.3	DM−	42.3
PairClass	52.1	LSA	42.0
VSM	47.1	AttrMax	35.0
DM+	45.3	AttrAvg	31.0
PairSpace	44.9	AttrMin	27.3
k-means	44.0	Random	20.0

`LRA, PERT, PairClass, VSM, KnowBest, LSA`: ACLWiki

`AttrMax, AttrAvg, AttrMin`: Turney(2006)

`DM+, DM-`: Baroni & Lenci (2009)

Domain analogies

- Turney (2008) extends the relational approach to entire analogical *domains*

solar system	→	atom
sun	→	nucleus
planet	→	electron
mass	→	charge
attracts	→	attracts
revolves	→	revolves
gravity	→	electromagnetism

Outline

Generalizations in DSMs

- DSMs allow us to build contextual representations for specific linguistic items (e.g. words, word pairs, etc.)
- Semantic types are generalizations over specific expressions
  - *swallow*, *sparrow*, *robin*, ...: BIRD
  - *cake*, *apple*, *pork*, ...: FOOD
  - <*car-wheel*>, <*bird-wing*>, ...: MERONYM
- Is it possible to represent in DSMs abstract semantic types?
  - this is like moving towards more general concepts

The case of selectional preferences

- The selectional preferences of a predicate can not be reduced simply to the set of its attested arguments in a corpus
- Selectional preferences specify an abstract semantic type
  - *kill*-obj: LIVING_ENTITY
  - *eat*-obj: FOOD
  - *drink*-obj: LIQUID
- This is necessary to account for the possibility of generalizations to unseen arguments
  - we can discriminate between the different plausibility of the following phrases, despite that fact that we may have never encountered them in any corpus
    - *kill the aardvark* – OK, since *aardvark* is a living entity
    - *kill the serendipity* – BAD, since *serendipity* is not a living entity

Hypothesis

A noun can occur as an argument of a predicate if it is similar to the prototypical filler for that argument

Semantic types and prototypes

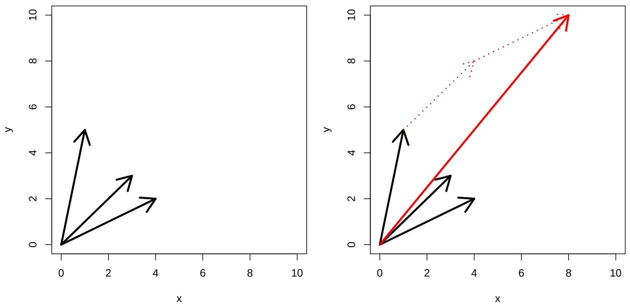
- A semantic type can be represented by a prototype derived from its instances
  - *sparrow, swallow, robin, eagle, penguin, ostrich, ...* $\Rightarrow$ BIRD_PROTOTYPE
  - the class prototype should be more similar to the prototypical class instances (e.g. *robin* and *eagle*), than to the non prototypical ones (e.g. *penguin* and *ostrich)*
  - cf. prototype models of concepts (cf. Rosch 1973 et al.)

Prototypes in DSMs

- both semantic type prototypes and their instances are represented as distributional vectors
- prototype vectors are built from instance vectors through some operation in vector space

Prototype vectors

- Simply sum the vectors of relevant class to obtain prototype (average, centroid) vector
- No need to rescale the resulting vector (e.g., to mean values) as long as we are only going to look at its angle/cosine with other vectors



Outline

1. The shape of semantic spaces
2. Attributional similarity
   - Synonym identification and semantic similarity judgements
   - Noun categorization
   - Semantic priming
3. Relational similarity
4. Representing semantic types in DSMs
   - Selectional preferences
   - Semantic relation classification
5. DSMs meet linguistics
   - Argument alternations
   - Nomina actionis

Verb selectional preferences

Padó et al (2007)

- Correlation with human acceptability judgments of noun-verb pairs from McRae et al. (1997) (100 pairs, 36 raters) and Padó (2007) (211 pairs, ~20 raters per pair):
  - how typical is deer as an object of shoot?
- DSMs and selectional preferences
  1. for each verb, build its prototype subj (obj) argument vector
     - select a set of prototypical noun arguments for each $< verb - dependency >$ pair (prototypical arguments can be chosen by measuring their association score - e.g. information, etc. - with the verb: e.g. the top 50 nouns connected to shoot by an obj link)
     - sum the vectors of the nouns in the prototype set (WARNING: if the noun in the test set is in the prototype set, its vector is subtracted from the prototype)
  2. the prototype-noun cosine space is taken as the model “plausibility judgment” about the noun occurring as the relevant verb argument

Verb selectional preferences

results

- Performance measured with with Spearman ρ correlation coefficient between the average human ratings and the model predictions (Padó et al. 2007)

model	McRae		Padó	
	coverage	ρ	*coverage*	ρ
Padó	56	41	97	51
dep-typed (DM)	96	28	98	50
ParCos	91	21	98	48
dep-filtered	96	21	98	39
window-typed	96	12	98	29
window-filtered	96	12	98	27
Resnik	94	3	98	24

`Padó, ParCos`: Padó et al. (2007)
`Resnik`: Resnik (1996)
`DM`: Baroni & Lenci (2009)

Acceptability of some potential objects of *kill*

object	*cosine*
kangaroo	0.51
person	0.45
robot	0.15
hate	0.11
flower	0.11
stone	0.05
fun	0.05
book	0.04
conversation	0.03
sympathy	0.01

Acceptability of some potential instruments of *kill*

instrument	*cosine*
hammer	0.26
stone	0.25
brick	0.18
smile	0.15
flower	0.12
antibiotic	0.12
person	0.12
heroin	0.12
kindness	0.07
graduation	0.04

Outline

Semantic relation classification

- SemEval-2007 Task 04: Semantic relations between Nominals
- 7 relation types: CAUSE-EFFECT, INSTRUMENT-AGENCY, PRODUCT-PRODUCER, ORIGIN-ENTITY, THEME-TOOL, PART-WHOLE, CONTENT-CONTAINER
- Instances harvested with patterns from the Web, and manually labeled as hits or misses
- For each relation, 140 training examples (about 50% hits), about 80 test cases

Semantic relation classification

Baroni & Lenci (2009)

- For each relation, we build hit and miss prototype vectors by averaging across vectors of training examples
- Prototype vectors are built from the rows in a word pair X context matrix
- For each test pair, we base hit/miss choice on cosine similarity to hit and miss average vectors

Semantic relation classification

Baroni & Lenci (2009)

model	*precision*	*recall*	*F*	*accuracy*
UCD-FC	66.1	66.7	64.8	66.0
UCB	62.7	63.0	62.7	65.4
ILK	60.5	69.5	63.8	63.5
DM+	60.3	62.6	61.1	63.3
UMELB-B	61.5	55.7	57.8	62.7
SemeEval avg	59.2	58.7	58.0	61.1
DM−	56.7	58.2	57.1	59.0
UTH	56.1	57.1	55.9	58.8
majority	81.3	42.9	30.8	57.0
probmatch	48.5	48.5	48.5	51.7
UC3M	48.2	40.3	43.1	49.9
alltrue	48.5	100.0	64.8	48.5

Comparison with SemEval 2007 - 04

`Models in group A: WordNet = NO & Query = NO`

Outline

Outline

Argument alternations

Argument alternations

Alternative syntactic realizations of a verb argument structure

- verbs differ with respect to their possible syntactic alternations

- Causative/inchoative alternation (Levin 1993)
  - the patient of *break* can also surface as its (inchoative) subject, whereas this does not happen with *mince*
    - The cook broke the vase → The vase broke
    - The cook minced the meat → *The meat minced
- Measuring similarity between verb slots can be used to study syntactic alterations (Joanis et al. 2008)

A new type of space

T (rows) dependency-typed verbs (= verb slots)
C (columns) : slot fillers

		teacher	victim	soldier	policeman
kill	subj_tr	0.0	22.4	1306.9	38.2
kill	obj	9.9	915.4	8948.3	538.1
die	subj_in	109.4	1335.2	4547.5	68.6

Argument alternations DSMs

Baroni & Lenci (2009)

- For alternating verbs, the direct object vector (`verb + obj`) should be similar to the intransitive subject vector (`verb + subj-in`)
  - the same things that are broken break
- for non-alternating verbs, the two slots should not be similar
  - the things that are minced are different from those that mince them

Argument alternations

experiments (Baroni & Lenci 2009)

- 402 verbs extracted from Levin Classes (Levin 1993)
  - 232 alternating causatives/inchoatives (*break*)
  - 170 non-alternating transitives (*mince)*
- Median per-verb pairwise cosines among slots:

	subj-intr subj-tr	subj-intr obj	subj-tr obj
alternating	0.28	0.31	0.16
non-alternating	0.29	0.09	0.11

Outline

Nomina actionis in Italian

- *Nomina actionis* are event denoting nouns morphologically derived from verbs via different suffixes
  - *violare* "to violate" ⇒ *viola-zione* "violation"
  - *cambiare* "to change" ⇒ *cambia-mento* "change"
  - *passeggiare* "to walk" ⇒ *passeggi-ata* "walk"
  - *rompere* "to break" ⇒ *rott-ura* "breaking"
- Not all the nouns matching the deverbal suffixes are actual *nomina actionis*
  - lots of false positives
    - *condizione* "condition" $*\Leftarrow$ *condire* "to season"
    - *sostanza* "substance" $*\Leftarrow$ *sostare* "to stop"
  - many nomina actionis are morphologically opaque (i.e. they have lost their original event meaning)
    - *parlamento* "parliament" $*\Leftarrow$ *parlare* "to talk"
    - *spazzatura* "trash" $*\Leftarrow$ *spazzare* "to sweep"

Recognizing Italian *Nomina actionis* with a DSM

Quaresima (2008)

- Hypothesis: true nominal actionis should share many contexts with their base verb
  - *The barbarians destroyed the city* vs. *The barbarians' destruction of the city*
- DSMs and nomina actionis
  1. extract from a corpus candidate V-N pairs, by using just suffix matching heuristics (N is a potential event nominal)
     - $*$*zione*, $*$*mento*, $*$*ata*, etc.
  2. for each candidate pair $< V - N >$, collect the nouns that appear as arguments of V (subj and obj) or of N (PP modifiers) in the corpus
  3. build a distributional matrix
     - rows: Vs and Ns in the potential $< V - N >$ pairs
     - columns: the nouns in the corpus appearing as arguments of V and N
  4. measure the distance in vector space between V and N in each potential $< V - N >$ pair
- Prediction: true event nominals should be closer in DSM to their base verb V than false positiives

Recognizing Italian *Nomina actionis* with a DSM

some results (Quaresima 2008)

- The DSM was built with Infomap NLP, trained on *La Repubblica Corpus* (ca. 350 Mega words)

verb	*noun*	cosine
violare “to violate”	*violazione* “violation”	0.619
cambiare “to change”	*cambiamento* “change”	0.378
passeggiare “to walk”	*passeggiata* “walk”	0.704
rompere “to break”	*rottura* “breaking”	0.577
condire “to season”	*condizione* “condition”	-0.178
sostare “to stop”	*sostanza* “substance”	-0.096
parlare “to talk”	*parlamento* “parliament”	-0.101
spazzare “to sweep”	*spazzatura* “trash”	0.070