Extracting Structured Semantic Spaces from Corpora

Marco Baroni

Center for Mind/Brain Sciences
University of Trento

National Institute for Japanese Language
July 26, 2007

Collaborators

- Brian Murphy, Massimo Poesio, Eduard Barbu (Trento)
- Alessandro Lenci (CNR, Pisa): ongoing analysis of traditional Word Space Models
- Building on earlier work by Abdulrahman Almuhareb (KACS, Riyadh) and Massimo Poesio

Introduction

- Corpora: large collections of text/transcribed speech produced in natural settings
- Had revolutionary impact on language technologies (speech recognition, machine translation…) and (pedagogical) lexicography

Introduction

- Corpora: large collections of text/transcribed speech produced in natural settings
- Had revolutionary impact on language technologies (speech recognition, machine translation. . .) and (pedagogical) lexicography
- Corpora and cognition: computer seen as statistics-driven agent that “learns” from its environment (distributional patterns in text)
- Can it teach us something about human learning?

Introduction

- Corpora: large collections of text/transcribed speech produced in natural settings
- Had revolutionary impact on language technologies (speech recognition, machine translation…) and (pedagogical) lexicography
- Corpora and cognition: computer seen as statistics-driven agent that “learns” from its environment (distributional patterns in text)
- Can it teach us something about human learning?
- Convergence with probabilistic models of cognition (see, e.g., *Trends in Cognitive Sciences* July 2006 issue)

Outline

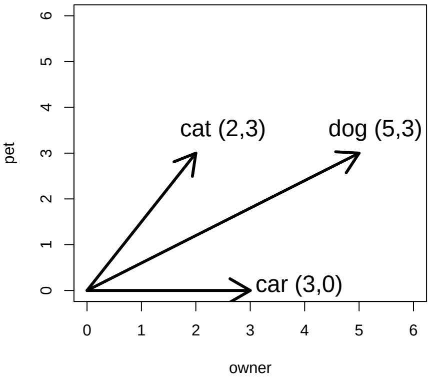
The Word Space Model
Sahlgren 2006

- Meaning of words defined by *set of contexts* in which word occurs
- Similarity of words represented as *geometric distance* among *context vectors*

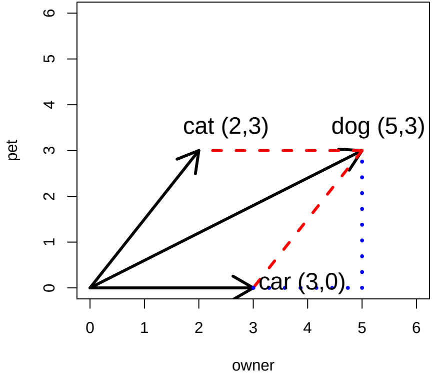
Contextual view of meaning

	leash	walk	run	owner	pet
dog	3	5	2	5	3
cat	0	3	3	2	3
lion	0	3	2	0	1
light	0	0	0	0	0
bark	1	0	0	2	1
car	0	0	1	3	0

Similarity in word space



Euclidean distance in two dimensions



Contextual view of meaning

Theoretical background

- "You should tell a word by the company it keeps" (Firth 1957)
- "[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts" (Cruse 1986)

Corpora as experience

- Of course, humans have access to other contexts as well (vision, interaction, sensory feedback)
- Context vectors can include also non-linguistic information, if encoded appropriately
- At the moment, corpora are only kind of *natural* input that is available to researchers on human-input-like scale
- Given that distribution of linguistic units (and probably other input information) is highly skewed, realistically distributed input is fundamental for plausible simulations

The TOEFL synonym match task

- 80 items

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed, believed, requested, correlated*

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed*, *believed*, *requested*, *correlated*

Human and machine performance on the synonym match task

- Average foreign test taker: 64.5%

Human and machine 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%

Human and machine 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%
- Best reported WSM results (Rapp 2003): 92.5%

Outline

Some problems with traditional Word Space Models

- "Semantic similarity" is multi-faceted notion but a single WSM provides only one way to rank a set of words
- "Representations" produced by models are not interpretable

Multi-faceted semantic similarity
Output of WSM trained on BNC

- Some nearest neighbours of *motorcycle*
  - motor $\rightarrow$ component
  - car $\rightarrow$ co-hyponym
  - diesel $\rightarrow$ component?
  - to race $\rightarrow$ proper function
  - van $\rightarrow$ co-hyponym
  - bmw $\rightarrow$ hyponym
  - to park $\rightarrow$ proper function
  - vehicle $\rightarrow$ hypernym
  - engine $\rightarrow$ component
  - to steal $\rightarrow$ frame?

Multi-faceted semantic similarity

- Different ways in which other words can be similar to a target word/concept:
  - Taxonomic relations (*motorcycle* and *car*)
  - Properties and parts of concept (*motorcycle* and *engine*)
  - Proper functions (*motorcycle* and *to race*)
  - Frame relations (*motorcycle* and *to steal*)

Multi-faceted semantic similarity

- Different ways in which other words can be similar to a target word/concept:
  - Taxonomic relations (*motorcycle* and *car*)
  - Properties and parts of concept (*motorcycle* and *engine*)
  - Proper functions (*motorcycle* and *to race*)
  - Frame relations (*motorcycle* and *to steal*)
- Impossible to distinguish in WSM

Multi-faceted semantic similarity

- Different ways in which other words can be similar to a target word/concept:
  - Taxonomic relations (*motorcycle* and *car*)
  - Properties and parts of concept (*motorcycle* and *engine*)
  - Proper functions (*motorcycle* and *to race*)
  - Frame relations (*motorcycle* and *to steal*)
- Impossible to distinguish in WSM
- Different status of different relations:
  - Properties, parts, proper functions constitute *representation* of word/concept
  - Ontological relations are product of overlapping representations in terms of properties etc.

Multi-faceted semantic similarity

- Different ways in which other words can be similar to a target word/concept:
  - Taxonomic relations (*motorcycle* and *car*)
  - Properties and parts of concept (*motorcycle* and *engine*)
  - Proper functions (*motorcycle* and *to race*)
  - Frame relations (*motorcycle* and *to steal*)
- Impossible to distinguish in WSM
- Different status of different relations:
  - Properties, parts, proper functions constitute *representation* of word/concept
  - Ontological relations are product of overlapping representations in terms of properties etc.
- For example:
  - A motorcycle is a motorcycle because it has an engine, two wheels, it is used for racing. . .
  - A car is similar to a motorcycle because they share a number of crucial properties and functions (engine and wheels, driving)

Multi-faceted semantic similarity

- Different ways in which other words can be similar to a target word/concept:
  - Taxonomic relations (*motorcycle* and *car*)
  - Properties and parts of concept (*motorcycle* and *engine*)
  - Proper functions (*motorcycle* and *to race*)
  - Frame relations (*motorcycle* and *to steal*)
- Impossible to distinguish in WSM
- Different status of different relations:
  - Properties, parts, proper functions constitute *representation* of word/concept
  - Ontological relations are product of overlapping representations in terms of properties etc.
- For example:
  - A motorcycle is a motorcycle because it has an engine, two wheels, it is used for racing...
  - A car is similar to a motorcycle because they share a number of crucial properties and functions (engine and wheels, driving)
- This is not captured in WSM representation

Semantic representations

- In WSM, word meaning is represented by co-occurrence vector:
  - long and sparse
  - or, if dimensionality reduction technique is applied, with denser dimensions corresponding to “latent” factors
- In either case, dimensions are hard/impossible to interpret

Semantic representations

- In WSM, word meaning is represented by co-occurrence vector:
  - long and sparse
  - or, if dimensionality reduction technique is applied, with denser dimensions corresponding to “latent” factors
- In either case, dimensions are hard/impossible to interpret
- However, converging evidence suggests rich semantic representation in terms of properties and activities

Semantic representations

- In WSM, word meaning is represented by co-occurrence vector:
  - long and sparse
  - or, if dimensionality reduction technique is applied, with denser dimensions corresponding to "latent" factors
- In either case, dimensions are hard/impossible to interpret
- However, converging evidence suggests rich semantic representation in terms of properties and activities
  - Rich lexical representations needed for semantic interpretation:
    - to finish a book (reading it) vs. an ice-cream (eating it) (Pustejovsky 1995)
    - a zebra pot is a pot with stripes

Semantic representations

- In WSM, word meaning is represented by co-occurrence vector:
  - long and sparse
  - or, if dimensionality reduction technique is applied, with denser dimensions corresponding to “latent” factors
- In either case, dimensions are hard/impossible to interpret
- However, converging evidence suggests rich semantic representation in terms of properties and activities
  - Rich lexical representations needed for semantic interpretation:
    - to finish a book (reading it) vs. an ice-cream (eating it) (Pustejovsky 1995)
    - a zebra pot is a pot with stripes
  - Strong functional neuro-imaging evidence for property-based activation of sensory and motor systems (Martin 2007)

Semantic representations

- In WSM, word meaning is represented by co-occurrence vector:
  - long and sparse
  - or, if dimensionality reduction technique is applied, with denser dimensions corresponding to “latent” factors
- In either case, dimensions are hard/impossible to interpret
- However, converging evidence suggests rich semantic representation in terms of properties and activities
  - Rich lexical representations needed for semantic interpretation:
    - to finish a book (reading it) vs. an ice-cream (eating it) (Pustejovsky 1995)
    - a zebra pot is a pot with stripes
  - Strong functional neuro-imaging evidence for property-based activation of sensory and motor systems (Martin 2007)
  - From practical point of view: property-based representations more useful in (pedagogical) lexicography

Outline

Structured Word Spaces

- Instead of counting generic co-occurrence, try to extract meaningful *concept-property* relations
- Assign *type* to relation

Ideal output

Target word: *motorcycle*

- for riding
- for racing
- is a vehicle
- has engine
- has two wheels
- ...

Corpus-based extraction of structured word spaces

- Basic idea (from Hearst 1992 and others): in a sufficiently large corpus, interesting relations will be explicitly cued by (noisy) superficial patterns
  - vehicles *such as* motorcycles
  - motorcycles *have* [smaller] engines
  - motorcycles *that are* [not] *used for* racing

Corpus-based extraction of structured word spaces

- Basic idea (from Hearst 1992 and others): in a sufficiently large corpus, interesting relations will be explicitly cued by (noisy) superficial patterns
  - vehicles *such as* motorcycles
  - motorcycles *have* [smaller] engines
  - motorcycles *that are* [not] *used for* racing
- Large body of work on relation extraction using similar techniques
- However, we are not aware of other attempts to extract both properties and relation types in a fully unsupervised manner for a variety of related and unrelated concepts as we do here

The basic steps

- Extract list of potential **concept + pattern + property** tuples
- Rank **concept + property** pairs on the basis of number of *distinct* tuples in which they occur
- Assign type to **concept + property** pair based on analysis of shared parts in patterns that connect them

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle → OK

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle → OK
  - motorcycle *that he got for his* birthday

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle → OK
  - motorcycle *that he got for his* birthday → OK (unfortunately)

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle → OK
  - motorcycle *that he got for his* birthday → OK (unfortunately)
  - birthday *John got a* motorcycle

Pattern extraction

- From enWaC, large Web-based corpus of English (more than 2 billion tokens)
- List of target concepts: provided by experimenter, all nouns
- Potential properties: any noun, verb, adjective within a window of 6 words left or right of a concept
- Potential patterns: can contain only function words, adjectives (converted to JJ) or very frequent content words (a bit more complicated than this, but I will skip the details)
- E.g.,
  - rides *a yellow* motorcycle → rides *a JJ* motorcycle → OK
  - motorcycle *that he got for his* birthday → OK (unfortunately)
  - birthday *John got a* motorcycle → NO

Ranking

- Given list of potential **concept + pattern + property** tuples, *count* how many distinct patterns connect a concept and a property

Ranking

- Given list of potential **concept + pattern + property** tuples, *count* how many distinct patterns connect a concept and a property
- Intuition: *frequent patterns* could simply be (part of) fixed phrases
- True semantic relations are likely to be expressed by a variety of different superficial patterns

Ranking

- Given list of potential **concept + pattern + property** tuples, *count* how many distinct patterns connect a concept and a property
- Intuition: *frequent patterns* could simply be (part of) fixed phrases
- True semantic relations are likely to be expressed by a variety of different superficial patterns
- E.g.:
  - Bad: year *of the* tiger; * year *of some* tigers, * tigers *have* years, . . .
  - Good: tail *of the* tiger, tail *of some* tigers, tigers *have JJ* tails, tiger *with its* tail, . . .

Ranking

- Given list of potential **concept + pattern + property** tuples, *count* how many distinct patterns connect a concept and a property
- Intuition: *frequent patterns* could simply be (part of) fixed phrases
- True semantic relations are likely to be expressed by a variety of different superficial patterns
- E.g.:
  - Bad: year *of the* tiger; * year *of some* tigers, * tigers *have* years, . . .
  - Good: tail *of the* tiger, tail *of some* tigers, tigers *have JJ* tails, tiger *with its* tail, . . .
- (More precisely, ranks are based on statistical association between concepts and properties sampled from the list of distinct tuples – akin to sampling from a dictionary rather than from a corpus)

Type assignment

- Type expressed by single word connector (*in, for, have, . . .*); in the case of verbs and adjectives, “zero” connector also possible

Type assignment

- Type expressed by single word connector (*in, for, have, …*); in the case of verbs and adjectives, “zero” connector also possible
- Type assigned to pair, based on frequency of occurrence of single word connectors in distinct patterns connecting the pair

Type assignment

- Type expressed by single word connector (*in, for, have, . . .*); in the case of verbs and adjectives, “zero” connector also possible
- Type assigned to pair, based on frequency of occurrence of single word connectors in distinct patterns connecting the pair
- For example, *on* chosen as type for *motorcycle+rider* relation on the basis of:
  - rider + *on large* + motorcycles
  - rider + *on the* + motorcycle
  - rider + *on a* + motorcycle
  - motorcycle + *says a lot about the* + rider
  - riders + *use* + motorcycles
  - ...

Type assignment

- Type expressed by single word connector (*in, for, have, . . .*); in the case of verbs and adjectives, “zero” connector also possible
- Type assigned to pair, based on frequency of occurrence of single word connectors in distinct patterns connecting the pair
- For example, *on* chosen as type for *motorcycle+rider* relation on the basis of:
  - rider + *on large* + motorcycles
  - rider + *on the* + motorcycle
  - rider + *on a* + motorcycle
  - motorcycle + *says a lot about the* + rider
  - riders + *use* + motorcycles
  - ...
- (With some complications, and a lot of work remains to be done on this)

Examples (top 10 properties)

Target: book

property	*type*
to read	*verb* _ *concept*
author	*concept* by *noun*
to write	*verb* _ *concept*
reader	*concept* for *noun*
chapter	*noun* in *concept*
library	*concept* in *noun*
publish	*verb* _ *concept*
reading	*noun* from *concept*
publisher	*concept* from *noun*
review	*noun* on *concept*

Examples (top 10 properties)

Target: tiger

property	*type*
jungle	*concept* in *noun*
cat	*noun* as *concept*
species	*noun* as *concept*
stripe	*noun* as *concept*
animal	*noun* as *concept*
to maul	*verb* by *concept*
habitat	*concept* in *noun*
lion	*noun* as *concept*
tame	*verb* _ *concept*
zoo	*concept* in *noun*

Examples (top 10 properties)

Target: motorcycle

property	*type*
ride	*verb* _ *concept*
rider	*noun* on *concept*
vehicle	*noun* as *concept*
moped	*noun* for *concept*
road	*concept* on *noun*
park	*verb* _ *concept*
scooter	*noun* up *concept*
car	*noun* as *concept*
insurance	*noun* for *concept*
bike	*noun* out *concept*

Most frequent property types

All Wu and Barsalou's neurally grounded types are represented

type	*WB classification*
verb _ *concept*	situational
noun in *concept*	situational/taxonomic/entity
concept in *noun*	situational/taxonomic/entity
concept _ *verb*	situational
noun for *concept*	situational
adj _ *concept*	all, including fair amount of introspective
noun as *concept*	taxonomic
concept for *noun*	situational
noun on *concept*	entity
concept on *noun*	entity

Outline

Clustering by shared properties

- As proposed above, we can now use semantic representation in terms of properties and proper functions to identify taxonomic relations
- Moreover, properties used to identify classes are interpretable, and can be seen as emergent semantic representation of abstract classes

Clustering by shared properties

- As proposed above, we can now use semantic representation in terms of properties and proper functions to identify taxonomic relations
- Moreover, properties used to identify classes are interpretable, and can be seen as emergent semantic representation of abstract classes
- Test set of 402 concepts, 21 categories, developed by Abdulrahman Almuhareb and Massimo Poesio
- Difficult data:
  - Difficult classes: motivation (e.g., compulsion, incentive, superego), legal document, creator...
  - Similar classes: feeling, pain, disease...
  - Rare concepts: icosahedron, hornbeam, zloty...
  - Ambiguous concepts: samba as a tree, divan as a social unit...

Semantic (sub-)spaces

- AAMP: state-of-the-art model proposed by Almuhareb and Poesio, clustering based on properties selected with few hand-picked patterns
- PROP: clustering based on properties that are among top 20 for at least one concept
- TYPED-PROP: clustering using same properties, with types added (e.g., distinguishing *for author* and *by author*)
- COMMON-TYPED-PROP: clustering using typed properties, based on typed properties belonging to one of 10 most common types only (*verb-concept*, *in*, *on*. . .)
- TAXO-PROP: clustering based on two frequently "taxonomic" types only (*in* and *as*)

Clustering

- Using CLUTO toolkit
- No parameter tuning
- Performance measured in terms of cluster *purity*

Results

(sub-)space	*purity*
AAMP	57.7%
PROP	60.6%
TYPED-PROP	65.0%
COMMON-TYPED-PROP	68.4%
TAXO-PROP	60.9%

Emergent abstract concepts

Top typed properties for some cluster

- fruit: it is a fruit, it is eaten, it is tasted, it is sliced, it is a flavour, it is used for juice, it is in bowls, it ripens, it is peeled, it is picked
- animal: it is an animal, it is killed, it is fed, it is bred, it is a mammal, it is in cages, it is a species, it eats stuff, it is in zoos, it is rescued
- illness: it is a disease, treatments have a function for it, it causes stuff, it is pain, it is cured, it is a condition, it is common, it is an infection, it has something to do with dying, it is an ailment
- creator: they are employed, they create stuff, they are asked, they are artists, they are in studios, they build stuff, they are commissioned stuff, cameras have a function for them, they are hired, they sell stuff

Highlighting different types of properties lead to different notions of similarity

- Nearest neighbours of *motorcycle* in the common property space (ordered by decreasing cosine >= .15):
  - bicycle, van, car

Highlighting different types of properties lead to different notions of similarity

- Nearest neighbours of *motorcycle* in the common property space (ordered by decreasing cosine >= .15):
  - bicycle, van, car
- Nearest neighbours of *motorcycle* in “functional” space (defined by properties of type *concept verb*, *verb concept*, *concept for noun*) (ordered by decreasing cosine >= .15):
  - divan, automobile, van, car, bicycle, camel

Highlighting different types of properties lead to different notions of similarity

- Nearest neighbours of *motorcycle* in the common property space (ordered by decreasing cosine >= .15):
  - bicycle, van, car
- Nearest neighbours of *motorcycle* in “functional” space (defined by properties of type *concept verb*, *verb concept*, *concept for noun*) (ordered by decreasing cosine >= .15):
  - divan, automobile, van, car, bicycle, camel
- You sit on divans, use camels for transportation, motorcycles *look* more like bicycles but they are used more like cars…

Outline

Conclusion

- We developed a fully unsupervised model that, given list of target words and corpus, automatically builds a semantic representation in terms of:
  - characteristic properties of the target words
  - type of the relation linking the target and each property
- Good quantitative and qualitative evaluation results

Ongoing and future work

- Smooth rough edges (in particular, property type identification)
- Compare with databases of properties generated by human subjects

Ongoing and future work

- Smooth rough edges (in particular, property type identification)
- Compare with databases of properties generated by human subjects
- Test predictive power of model in psycholinguistic experiments and linguistic tasks
- Integrate with other data sources (“visual” information from image labeling databases)
- Pedagogical lexicography application (project with EurAc research institute, to start this fall)
- More languages (Japanese!)

Some references

A. Almuhareb and M. Poesio (2004). Attribute-based and value-based clustering: an evaluation. *Proceedings of EMNLP 2004*.

M. Hearst (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING 1992*.

A. Martin (2007). The representation of object concepts in the brain. *Annual Review of Psychology 58*.

J. Pustejovsky (1995). *The generative lexicon*. MIT Press.

R. Rapp (2003). Word sense discovery based on sense descriptor dissimilarity. *Proceedings of the Ninth Machine Translation Summit*.

R. Rapp (2004). A freely available automatically generated thesaurus of related words. *Proceedings of LREC 2004*.

M. Sahlgren (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm University.

L. Wu and L. Barsalou (Submitted). *Grounding concepts in perceptual simulation: I. Evidence from property generation.*