Distributional semantics for linguists

Lecture 1b

Aurelie Herbelot

Universität Potsdam
Department Linguistik

ESSLLI 2012

Outline

Overview

- Models: which choices must be made when designing a distributional semantics system?
- Building the system: step-by-step example.
- Looking at real distributions.
- Issues: corpus choice, polysemy, fixed expressions.

Outline

The general intuition

- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude (length) and a direction.
- The **semantic space** has dimensions which correspond to possible contexts.
- For our purposes, a distribution can be seen as a point in that space (the vector being defined with respect to the origin of that space).
- *cat* [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2, zebra 0.1...]

The notion of context

- **Context:** if the meaning of a word is given by its context, what does 'context' mean?
  - Word windows (unfiltered): *n* words on either side of the lexical item under consideration (unparsed text).
    Example: n=2 (5 words window):

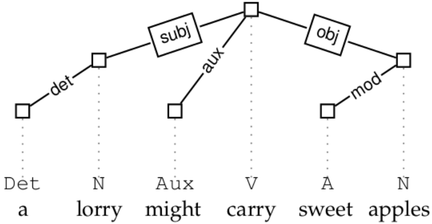
    *... the prime **minister** acknowledged that ...*

  - Word windows (filtered): *n* words on either side of the lexical item under consideration (unparsed text). Some words are not considered part of the context (e.g. function words, some very frequent content words). The stop list for function words is either constructed manually, or the corpus is POS-tagged.
    Example: n=2 (5 words window):

    *... the prime **minister** acknowledged that ...*

The notion of context

- Dependencies: syntactic or semantic. The corpus is converted into a list of directed links between heads and dependents. Context for a lexical item is the dependency structure it belongs to. The length of the dependency path can vary according to the implementation (Padó and Lapata, 2007).



a	[Det,det,N]	lorry
lorry	[N,subj,V]	carry
might	[Aux,aux,V]	carry
apples	[N,obj,V]	carry
sweet	[A,mod,N]	apples

Parsed vs unparsed data: examples

word (unparsed)	**word (parsed)**
meaning_n	or_c+phrase_n
derive_v	and_c+phrase_n
dictionary_n	syllable_n+of_p
pronounce_v	play_n+on_p
phrase_n	etymology_n+of_p
latin_j	portmanteau_n+of_p
ipa_n	and_c+deed_n
verb_n	meaning_n+of_p
mean_v	from_p+language_n
hebrew_n	pron_rel_+utter_v
usage_n	for_p+word_n
literally_r	in_p+sentence_n

Context weighting

- Binary model: if context *c* co-occurs with word *w*, value of vector $\vec{w}$ for dimension *c* is 1, 0 otherwise.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 1} {dog 0} {long 1} {sell 0} {semantics 1}...
- Basic frequency model: the value of vector $\vec{w}$ for dimension *c* is the number of times that *c* co-occurs with *w*.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 2} {dog 0} {long 3} {sell 0} {semantics 1}...

Context weighting

- Characteric model: the weights given to the vector components express how *characteristic* a given context is for *w*. Functions used include:
  - Pointwise Mutual Information (PMI), with or without discounting factor.

$$pmi_{wc} = log(\frac{f_{wc} * f_{total}}{f_w * f_c}) \tag{1}$$

  - Derivatives such as Mitchell and Lapata's (2010) weighting function (PMI without the log).

What semantic space?

- Entire vocabulary.
  - + All information included – even rare, but important contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*)
- Top *n* words with highest frequencies.
  - + More efficient (5000-10000 dimensions). Only ‘real’ words included.
  - - May miss out on infrequent but relevant contexts.

What semantic space?

- Singular Value Decomposition (LSA – Landauer and Dumais, 1997): the number of dimensions is reduced by exploiting redundancies in the data. A new dimension might correspond to a generalisation over several of the original dimensions (e.g. the dimensions for *car* and *vehicle* are collapsed into one).
  - + Very efficient (200-500 dimensions). Captures generalisations in the data.
  - - SVD matrices are not interpretable.
- Other, more esoteric variants...

Outline

Our reference text

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Example:** Produce distributions using a word window, frequency-based model

The semantic space

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- We assume that we only keep content words in the semantic space.
- **Dimensions:**

difference	impossible	thing
get	major	turns
go	possibly	usually
goes	repair	wrong

Frequency counts...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Counts:**

difference 1	impossible 1	thing 3
get 1	major 1	turns 1
go 3	possibly 2	usually 1
goes 1	repair 1	wrong 4

Conversion into 5-word windows...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- ∅ ∅ **the** major difference
- ∅ the **major** difference between
- the major **difference** between a
- major difference **between** a thing
- ...

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (frequencies):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 1	possibly 1	usually 1
goes 2	repair 0	wrong 2

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (PMIs):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 0.22184875	possibly 0.397940009	usually 0.698970004
goes 1	repair 0	wrong 0.397940009

Outline

Corpus description

- Obtained from the entire English Wikipedia.
- Corpus parsed with the English Resource Grammar (Flickinger, 2000) and converted into DMRS form (Copestake, 2009).
- Dependencies considered include:
  - For nouns: head verbs (+ any other argument of the verb), modifying adjectives, head prepositions (+ any other argument of the preposition).
    e.g. cat: chase_v+mouse_n, black_a, of_p+neighbour_n
  - For verbs: arguments (NPs and PPs), adverbial modifiers.
    e.g. eat: cat_n+mouse_n, in_p+kitchen_n, fast_a
  - For adjectives: modified nouns; rest as for nouns (assuming intersective composition).
    e.g. black: cat_n, chase_v+mouse_n

System description

- Semantic space: top 100,000 contexts.
- Weighting: normalised PMI (Bouma 2007).

$$pmi_{wc} = \frac{log(\frac{f_{wc} * f_{total}}{f_w * f_c})}{-log(\frac{f_{wc}}{f_{total}})} \tag{2}$$

An example noun

- *language*:

0.541816::other+than_p()+English_n
0.525895::English_n+as_p()
0.523398::English_n+be_v
0.48977::english_a
0.481964::and_c+literature_n
0.476664::people_n+speak_v
0.468399::French_n+be_v
0.463604::Spanish_n+be_v
0.463591::and_c+dialects_n
0.452107::grammar_n+of_p()
0.445994::foreign_a
0.445071::germanic_a
0.439558::German_n+be_v
0.436135::of_p()+instruction_n

0.435633::speaker_n+of_p()
0.423595::generic_entity_rel_+speak_v
0.42313::pron_rel_+speak_v
0.42294::colon_v+English_n
0.419646::be_v+English_n
0.418535::language_n+be_v
0.4159::and_c+culture_n
0.410987::arabic_a
0.408387::dialects_n+of_p()
0.399266::part_of_rel_+speak_v
0.397::percent_n+speak_v
0.39328::spanish_a
0.39273::welsh_a
0.391575::tonal_a

An example adjective

- *academic*:

0.517031::Decathlon_n
0.512661::excellence_n
0.449711::dishonesty_n
0.445393::rigor_n
0.426142::achievement_n
0.421246::discipline_n
0.397311::vice_president_n+for_p()
0.391978::institution_n
0.38937::credentials_n
0.378062::journal_n
0.373727::journal_n+be_v
0.372052::vocational_a
0.371873::student_n+achieve_v
0.361359::athletic_a

0.356562::reputation_n+for_p()
0.354674::regalia_n
0.353712::program_n
0.351601::freedom_n
0.347751::student_n+with_p()
0.34621::curriculum_n
0.342008::standard_n
0.34151::at_p()+institution_n
0.340271::career_n
0.337857::Career_n
0.329923::dress_n
0.329358::scholarship_n
0.329281::prepare_v+student_n
0.328009::qualification_n

Outline

Corpus choice

- As much data as possible?
  - British National Corpus (BNC): 100 m words
  - Wikipedia: 897 m words
  - UKWac: 2 bn words
  - ...
- In general preferable, *but*:
  - More data is not necessarily the data you want.
  - More data is not necessarily realistic from a psycholinguistic point of view. We perhaps encounter 50,000 words a day. BNC = 5 years' text exposure.

Corpus choice

- Distribution for *unicycle*, as obtained from Wikipedia.

0.448051::motorized_a
0.404372::pron_rel_+ride_v
0.238612::for_p()+entertainment_n
0.235763::half_n+be_v
0.235407::unwieldy_a
0.230275::earn_v+point_n
0.216627::pron_rel_+crash_v
0.190785::man_n+on_p()
0.186325::on_p()+stage_n
0.185063::position_n+on_p()
0.168102::slip_v
0.162611::and_c+1_n
0.159627::autonomous_a
0.155822::balance_v
0.133084::tall_a
0.124242::fast_a
0.106976::red_a
0.0714643::come_v
0.0601987::high_a

Polysemy

- Distribution for *pot*, as obtained from Wikipedia.

0.566454::melt_v
0.442374::pron_rel_+smoke_v
0.434682::of_p()+gold_n
0.40773::porous_a
0.401654::of_p()+tea_n
0.39444::player_n+win_v
0.393812::money_n+in_p()
0.376198::of_p()+coffee_n
0.33117::amount_n+in_p()
0.329211::ceramic_a
0.326387::hot_a
0.323321::boil_v
0.313404::bowl_n+and_c
0.306324::ingredient_n+in_p()
0.301916::plant_n+in_p()

0.298764::simmer_v
0.292397::pot_n+and_c
0.284539::bottom_n+of_p()
0.28338::of_p()+flower_n
0.279412::of_p()+water_n
0.278914::food_n+in_p()
0.262501::pron_rel_+heat_v
0.260375::size_n+of_p()
0.25511::pron_rel_+split_v
0.254363::of_p()+money_n
0.2535::of_p()+culture_n
0.249626::player_n+take_v
0.246479::in_p()+hole_n
0.244051::of_p()+soil_n
0.243797::city_n+become_v

Fixed expressions

- Distribution for *time*, as obtained from Wikipedia.

0.462949::of_p()+death_n
0.448965::same_a
0.446277::1_n+at_p(temp)
0.445338::Nick_n+of_p()
0.423542::spare_a
0.418568::playoffs_n+for_p()
0.416471::of_p()+retirement_n
0.405288::of_p()+release_n
0.397135::pron_rel_+spend_v
0.389886::sand_n+of_p()
0.385954::pron_rel_+waste_v
0.382816::place_n+around_p()
0.37777::of_p()+arrival_n
0.376466::of_p()+completion_n
0.374797::after_p()+time_n
0.374682::of_p()+arrest_n
0.371589::country_n+at_p()
0.370736::age_n+at_p()
0.370626::space_n+and_c
0.370555::in_p()+career_n

0.370464::world_n+at_p()
0.363982::and_c+space_n
0.363241::generic_entity_rel_+mark_v
0.361872::of_p()+introduction_n
0.357929::in_p()+year_n
0.357565::of_p()+appointment_n
0.356229::of_p()+trouble_n
0.355658::of_p()+merger_n
0.354794::on_p()+ice_n
0.353891::practice_n+at_p()
0.351994::of_p()+birth_n
0.351556::full_a
0.348029::of_p()+accident_n
0.34785::state_n+at_p()
0.347753::to_p()+time_n
0.345147::of_p()+election_n
0.345088::area_n+at_p()
0.342571::and_c+money_n
0.342113::time_n+after_p()
0.341877::allotted_a

Outline

Conclusion

- Various models for distributional systems, with various consequences on the output.
- Known issues: corpus-dependence (which notion of concept is at play here?), word senses are collapsed (perhaps not such a bad thing...), fixed expressions create noise in the data.