Compositionality prediction using semantic spaces

ACL 2011 shared task

Author: Lubomir Krcmar

Date: 24.11. 2011

What is the compositionality?

Non-compositional

- reinvent wheel
- blue chip
- catch eye
- right wing
- beg question

Compositional

- short distance
- student learn
- answer questions
- olive oil
- lemon juice

"The meaning of the whole is less related to the meaning of the parts"

ACL 2011 shared task

- People judgments → Data sets

Type	compound	Coarse	Score
EN_V_OBJ	beg question	low	18
EN_V_OBJ	pull plug	low	21
EN_V_SUBJ	company take	medium	50
EN_ADJ_NN	hard work	medium	51
EN_ADJ_NN	short distance	high	97
EN_V_SUBJ	student learn	high	98

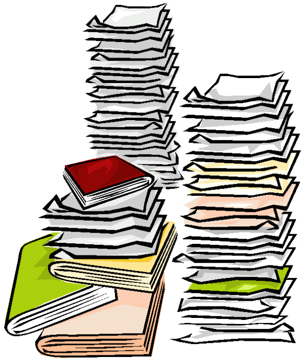
- How to do it automatically?

What is the semantic space?

- Output of algorithms
  - Words associated with vectors
  - How to build a semantic space?
    - Differs – LSA, HAL, COALS

The principle:

Corpora



Input matrix

```
2 0 0 0 0 0 1 2
4 0 0 0 3 1 0 0
0 1 3 0 1 0 1 1
1 0 0 0 0 0 0 0
0 0 1 0 0 0 1 0
0 0 2 0 1 0 1 1
```

Employed algorithm

COALS

Semantic space

```
Word1: 0.7 0.7 1.3 3.1 0.5 1.8 0.2 0.1
Word2: 0.3 0.7 0.7 3.2 1.0 1.0 8.2 2.0
Word3: 7.6 0.7 0.2 0.1 0.2 1.0 7.6 0.6
Word4: 2.9 1.7 0.3 0.8 0.0 1.3 2.2 0.0
Word5: 1.1 0.5 0.3 0.1 1.0 1.0 0.9 0.9
Word6: 0.2 0.0 0.3 0.1 4.2 1.0 0.1 0.3
```

Hyperspace-Analogue to Language

- Harris distributional hypothesis
  - Terms are similar to the extent to which they share similar linguistic contexts
    - “red car”, “blue car” “fast car”, “slow car”
    - “red paper”, “blue paper”
    - Etc..

- COALS ~ HAL with special settings..
- LSA – based on “bag of words principle”

Sample COALS results:

“red”

Neighbor	Similarity
yellow	0,71
white	0,68
blue	0,66
pink	0,64
black	0,61
purple	0,61
grey	0,6
green	0,6
coloured	0,58
brown	0,58

“wheel”

Neighbor	Similarity
brake	0,55
tyre	0,52
lever	0,48
rim	0,47
cylinder	0,46
roller	0,46
shaft	0,45
chassis	0,45
plate	0,45
screw	0,45

Part of car?

Round shape?

How the COALS space was built?

- UKWAC corpora tagged corpora
  - Lemmas and tags used
- Sspace package
  - COALS alg., default settings, 1/3 of UKWAC used so far
- Morphological restriction
  - Similar words – same morphological category
- Low occurring words treated as stopwords
- Metacentrum facilities

How to use SS for compositionality prediction?

- Consider occurrences of alternatives!
  - Non-compositional
    - “Reinvent wheel” X “Reinvent brake” X “Reinvent tyre”
    - “Blue chip” X “Yellow chip” X “White chip”
  - Compositional
    - “Short distance” X “Long distance” X “Short length”

- However, how to transform occurrences to the compositionality measure?

“Blue chip”

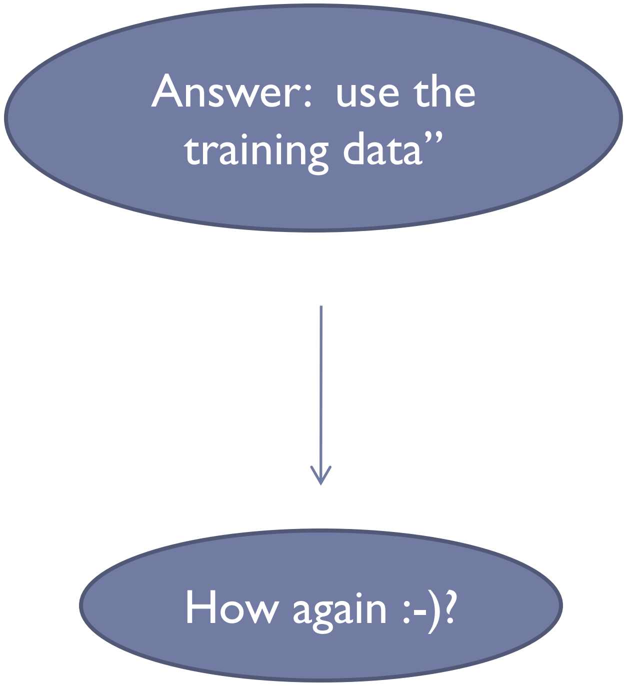
Compound	#	Compound	#
blue chip	1299		
blue poker	5	yellow chip	1
blue holdem	0	pink chip	0
blue stud	3	white chip	7
blue casino	0	red chip	11
blue texas	0	purple chip	1
blue strip	15	coloured chip	3
blue clay	13	black chip	1
blue tournament	0	green chip	5
blue card	52	grey chip	1
blue dice	0	pale chip	0

“Short distance”

Alternative	#	Alternative	#
short distance	3125		
short length	453	long distance	3725
short mile	6	brief distance	2
short radius	14	lengthy distance	7
short height	7	extended distance	18
short km	0	straight distance	4
short angle	4	quick distance	1
short walk	2950	introductory distance	0
short kilometre	0	quiet distance	0
short velocity	0	narrow distance	0
short speed	10	slow distance	6

What is the right model?

- How many neighbors to use?
  - 2, 10, 20?
  - Depends on compound type?
- How to weight neighbors?
  - Closer neighbors – higher significance?
- How to weight counts?
  - Use log?
- Use both words in compounds?
  - Or use just the “head” one?



Answer: use the training data”

How again :-)?

Alternatives to my approach

- Use Wordnet
  - Could be used very similarly
  - But:
    - Manually constructed
    - No “part of” relations and other ones?
    - Synsets – often phrases

- Comparison of distributions (part X whole)
  - Distribution of “red tape” X “tape”
  - Distribution of “student learn” X “student”
  - But:
    - Seems to be useful for “EN_ADJ_NN “ type only

After “student” verbs are expected

Conclusion

- The approach is transparent and might work well!
- Work done:
  - Create semantic space
  - Find similar words and built alternative compounds
  - Count occurrences of compounds
- Work to be done:
  - Create the right scoring model ? How to use training data
  - Build semantic space from the whole UKWAC corpora
  - Test and evaluate

References

- Johannsen, A., Martinez, H., Rishøj, C., & Søgaard, A. (2011). Shared task system description : Frustratingly hard compositionality prediction. *Computational Linguistics*, (June), 29–32.
- Harris, Z. (1954). Distributional structure. (J. Katz, Ed.) Word Journal Of The International Linguistic Association, 10(23), 146-162. Oxford University Press.
- Jurgens and Stevens, (2010). The S-Space Package: An Open Source Package for Word Space Models. In System Papers of the Association of Computational Linguistics.

- The access to the MetaCentrum computing facilities provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is appreciated.