Capturing Anomalies in the Choice of Content Words in Compositional Distributional Semantic Space

Ekaterina Kochmar, Ted Briscoe

University of Cambridge, UK

RANLP 2013

Capturing Anomalies in the Choice of Content Words

↑

The task

Errors in Content Word Combinations

Adjective–noun (AN) combinations from non-native English texts:

- Now I felt a big anger. → great anger [confused via meaning]
- It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places. → ancient sites [confused via form]
- Deep regards, John Smith → kind regards [(seemingly) unrelated]
- The company had great turnover, which was noticable in this market. → high turnover [context-dependent interpretation]

People rarely intend to generate nonsensical phrases.
Yet, many word confusions result in **semantically anomalous** word combinations

Previous Approaches to Learner Error Detection/Correction

In function words:

- A limited set of possible confusions: *a* $\rightarrow$ $\varnothing$ | *an* | *the*
- Can be learned from the seen examples
- Most often only one suitable correction:
  - I am *_ student $\rightarrow$ I am a student
  - I came *in Tokyo $\rightarrow$ I came to Tokyo
- Machine-learning classifiers with relevant features

Not suitable for content words:

- a much larger set of confusion patterns to be learned
- relevant features – less clear
- errors have more to do with meaning, rather than grammar

Previous Approaches to Learner Error Detection/Correction

In content words:

- Perform error correction for already detected errors (Liu et al., 2009; Dahlmeier and Ng, 2011)
- Writing improvement (Chang et al., 2008; Futagi et al., 2008):
  - for each combination X, check for more fluent/native-like alternatives Y
  - compare alternatives Y to X using some frequency-based measure
  - if $\exists\ Y_i$ more fluent than $X \Rightarrow X$ is an error, Y_i a correction

These approaches:

- do not deal with error detection *per se*
- are unable to deal with previously unseen combinations
- do not make any semantically-motivated decisions

Error Detection in Content Word Combinations

- Many confusions result in **semantically** anomalous combinations
- Learners are creative: many of the combinations are **corpus-unattested**
- **Goal**: detect errors in the choice of content words without punishing learners for creative use of language (falsely identified errors are more harmful for language learning than missed errors)

$\downarrow$

Compositional Distributional Semantics

Distributional Semantic Models (DSMs)

Main points

- Key assumption: word meaning can be approximated by a word's *distribution*
- Method: represent words with distributional vectors, dimensions = co-occurrence with context words
- Hypothesis: semantically similar words occur in similar contexts

Example: rose

- Collect contexts from a corpus:

> ...
> This rose grows up to six feet tall
> The desert rose blooms in the garden
> I bought some roses and lilies the other week for just £2.50
> ...

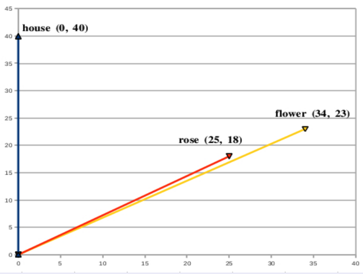
- Construct distributional vectors

Distributional Semantic Models (DSMs)

Distributional Vectors

	bloom	buy	garden	grow	tall	...
rose	25	18	20	33	8	...
flower	34	23	30	38	10	...
house	0	40	24	5	21	...

Graphical Representation



DSMs: From words to phrases

Distributional Vectors

	bloom	buy	garden	grow	tall
rose	25	18	20	33	8
red rose	14	7	5	17	0
old rose	15	3	0	10	0
blue rose	0	0	0	0	0
ignorant rose	0	0	0	0	0

DSMs: Issues

- Data sparsity: less or no occurrences for longer linguistic units
- The longer the phrase, the sparser the vector
- Cannot distinguish between unseen combinations:
  - semantically plausible (but rare, describing false facts, etc)
  - semantically implausible/anomalous

DSMs: From words to phrases

Distributional Vectors

	bloom	buy	garden	grow	tall
rose	25	18	20	33	8
red rose	14	7	5	17	0
old rose	15	3	0	10	0
blue rose	0	0	0	0	0
ignorant rose	0	0	0	0	0

DSMs: Issues

- Less or no occurrences for longer linguistic units – data sparsity
- The longer the phrase, the sparser the vector
- Cannot distinguish between unseen combinations:
  - semantically plausible (but rare, describing false facts, etc)
  - semantically implausible/anomalous

Compositional Models

Key points

- Distributional counts are reliable for words, not for phrases
- $\Rightarrow$ Model phrase vectors from distributional vectors of their constituents
- To combine word representations a and b use:
  - Direct vector combination: $a \circledast b$ (Kintsch, 2001; Mitchell and Lapata, 2008; Erk and Padó, 2008)
  - Linear transformations on vectors: $\mathcal{A}(b)$ (Baroni and Zamparelli, 2010).

To assess

Test in relevant NLP tasks:

- similarity detection, paraphrase ranking, adjective–noun (AN) vector prediction
- semantic anomaly detection in AN combinations (Vecchi et al., 2011):
  - Ability of the models for account for linguistic creativity
  - Unseen semantically acceptable vs unseen semantically anomalous ANs
- novel task: error detection in content word combinations in real learner data

Error Detection Using Compositional Distributional Semantics

Error detection in content word combinations $\sim$ semantic anomaly detection (Vecchi et al., 2011):

- **Semantically anomalous** combinations can be detected ✓
- Can deal with **corpus-unattested** examples ✓
- **Goal**: detect errors in the choice of content words without punishing learners for creative use of language (falsely identified errors are more harmful for language learning than missed errors)

- Use 3 models of semantic composition: *additive (add)*, *multiplicative (mult)* and *adjective-specific linear maps (alm)*;
- Detect a difference between model-generated vectors for correct and incorrect combinations

Test Data

- AN examples from the Cambridge Learner Corpus FCE dataset (Yannakoudakis et al., 2011)
- Error coding used to detect ANs with the incorrect adjective and/or noun used
- Test set: skewed towards correct combinations
  - 4681 correct ANs
  - 530 incorrect ANs
- Wide range of constituent adjectives and nouns
- Many test combinations attested in the BNC
- $\rightarrow$ Different from Vecchi et al.'s setting, but a natural setting to test the semantic models

Semantic Space Construction

Source Corpus

- British National Corpus (`http://www.natcorp.ox.ac.uk/`)
- Lemmatised, tagged and parsed with the RASP system (Briscoe et al., 2006)
- Statistics extracted at the lemma level, no inflectional information

Semantic Space: a Collection of Distributional Vectors

- Target words and combinations:
  - 8,364 nouns including 8K most frequent in the corpus + test ones
  - 4,353 adjectives including 4K most frequent in the corpus + test ones
  - 63,336 ANs generated, >100 in the corpus + test ones
- Context words:
  - 10K most frequent nouns, adjectives and verbs
  - Co-occurrence counts converted into Local Mutual Information scores (Evert, 2005)

Additive and multiplicative models (Mitchell and Lapata)

- Use component-wise vector addition and multiplication:

$$c_i = a_i + b_i \qquad\qquad c_i = a_i \times b_i$$

- Advantages :
  - Simple to implement and interpret
  - Require no training or tuning
  - Promising results in other NLP tasks, including anomaly detection
- Weak points:
  - Commutative $\rightarrow$ do not distinguish between heads and modifiers, grammatical functions
  - Examples: same vectors generated for *vector component* and *component vector*, *man chase dog* and *dog chase man*

Adjective–specific linear maps (Baroni and Zamparelli)

- Words in the combination have different grammatical functions
- Nouns represented by their distributional vectors in a usual way
- Adjectives: e.g., *new* in *new friend* $\neq$ *new* in *new shoes*
  $\rightarrow$ Distribution does not capture the meaning
- Adjectives **not vectors**, but **matrices** encoding distributional functions
- AN vector as matrix-by-vector multiplication:
  $\mathcal{ADJ}(noun) = \mathbf{F}_{adj} \times \overrightarrow{noun} = \overrightarrow{AN}$
- A separate matrix learned for each adjective – *adjective-specific*
- Mapping from one nominal meaning (noun) to another (AN) – *linear maps*

Alm model

- For each *adj*, use all seen [*noun* :: *adj*–*noun* (AN)] pairs to derive the *adj* matrix
- Apply partial least squares regression algorithm
- Learn the correspondences between nouns and correspondent ANs in the seen pairs
- The *ij*-th cell in the matrix defines how much the components corresponding to the *j*-th **input** (=noun) context element contributes to the value of the *i*-th context element in the **output** (=AN) vector:

Alm model

OLD	*bloom*	*buy*
bloom	10	0
buy	6	15

$\times$

	tree
bloom	34
buy	10

$=$

	OLD(tree)
bloom	$(10 \times 34) + (0 \times 10) = 340$
buy	$(6 \times 34) + (15 \times 10) = 354$

Alm model

OLD	*bloom*	*buy*
bloom	10	0
buy	6	15

$\times$

	tree
bloom	34
buy	10

$=$

	OLD(tree)
bloom	$(10 \times 34) + (0 \times 10) = 340$
buy	$(6 \times 34) + (15 \times 10) = 354$

Alm model

OLD	*bloom*	*buy*
bloom	10	0
buy	6	15

$\times$

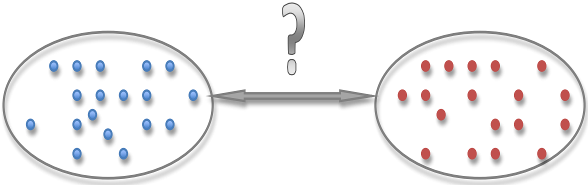
	tree
bloom	34
buy	10

$=$

	OLD(tree)
bloom	$(10 \times 34) + (0 \times 10) = 340$
buy	$(6 \times 34) + (15 \times 10) = 354$

Measures of Semantic Anomaly

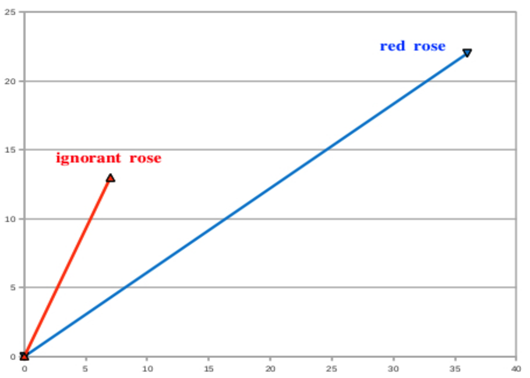
We have modelled vectors representing correct and incorrect AN combinations. How do we distinguish between them?



Measures of Semantic Anomaly

1. Vector Length (Vecchi et al.)

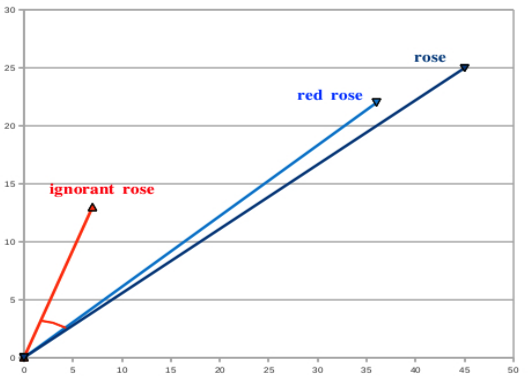
In anomalous ANs, the counts in the input vectors are distributed differently → some “incompatible dimensions” would receive low counts → anomalous AN vectors are expected to be shorter:



Measures of Semantic Anomaly

2. Cosine to the Component Noun (Vecchi et al.)

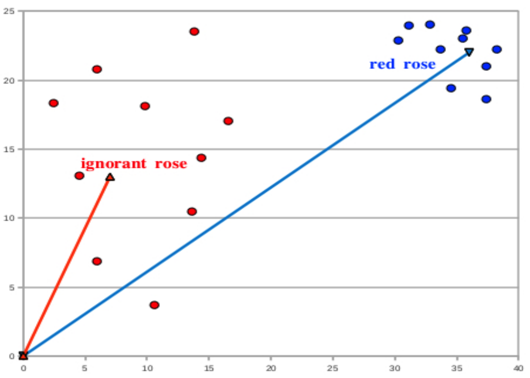
Anomalous ANs are less similar to the input nouns → their vectors are expected to have lower cosine to the input noun vector:



Measures of Semantic Anomaly

3. Neighbourhood Density (Vecchi et al.)

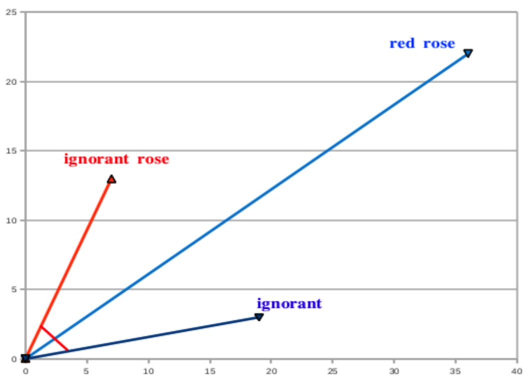
Anomalous AN vectors are expected to have sparser neighbourhoods (measured as an average cosine/distance to the 10 nearest neighbours):



Measures of Semantic Anomaly

4. Cosine to the Component Adjective (new metric)

For the *add* and *mult* model, both input vectors contribute equally. Then, why not calculating the distance to the input adjective:

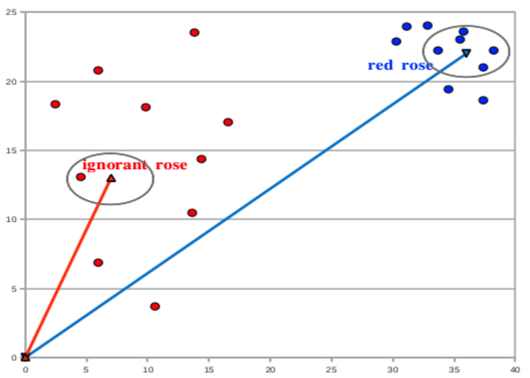


Measures of Semantic Anomaly

5. Ranked Density within Close Proximity (new metric)
6. Number of Close Neighbours (new metric)

Close proximity – a neighbourhood populated by vectors for which the cosine is >0.8.

$RDens = \sum_{i=1}^{N} rank_i distance_i$ and N itself.



Measures of Semantic Anomaly

7. Component Overlap (new metric)

Hypothesis: semantically acceptable ANs would be placed in the neighbourhoods populated by similar words and combinations

red rose	*ignorant rose*
(x) **rose**	people
red (x)	blind people
flower	like-minded
...	...

Method: a proportion of neighbours (among 10 nearest ones) containing the same constituent words as in a tested AN.

Evaluation

- Use the 7 measures
- Compute the difference between the mean values for the two groups of vectors
- Apply t-test, statistical significance level $p < 0.05$
- Evaluate on:
  - the full test set
  - corpus-attested examples only (context-dependent errors)
  - corpus-unattested examples only (similar to Vecchi et al.)

What next

Test **reliability** of the measures

Those that detect the difference between vectors reliably can further be used by an error detection algorithm

Results: *add* model

Measure	*all*	*attest*	*unattest*
VLen	0.1992	0.6226	0.1840
CosN	0.0797	0.1538	**0.00001**
Dens	0.9792	0.3921	0.5589
CosA	0.6867	0.3790	**0.0026**
RDens	0.6915	0.7493	0.1414
Num	0.8756	0.5753	0.1050
COver	0.6028	0.2126	0.1200

Table : p values for the *add* model ($p < 0.05^*$)

Conclusion: performs well only with 2 measures, and only on one subset

Results: *mult* model

Measure	*all*	*attest*	*unattest*
VLen	**0.0033**	0.1549	**0.0004**
CosN	**0.0017**	**0.0182**	**0.0083**
Dens	0.3531	0.6656	0.2703
CosA	**0.00002**	**0.0144**	0.3352
RDens	**0.0002**	**0.0300**	**0.0001**
Num	**0.0001**	**0.0091**	**0.0001**
COver	**0.0041**	**0.0096**	0.7317

Table : p values for the *mult* model ($p < 0.05^*$)

Conclusion: performs well with wide variety of measures and on all subsets

Results: *alm* model

Measure	*all*	*attest*	*unattest*
VLen	0.6537	0.2840	0.5557
CosN	**0.00003**	**0.0003**	0.1555
Dens	0.8160	0.4902	0.1799
CosA	**0.0188**	**0.0070**	0.8440
RDens	0.9106	0.6804	0.8588
Num	0.5959	0.9619	0.1402
COver	**0.00001**	**0.0004**	0.1484

Table : p values for the *alm* model ($p < 0.05^{*}$)

Conclusion: is not helpful for previously unseen examples.

Conclusions

Results

- Semantic models can provide some reliable clues for error detection in content word combinations
- Our new metrics show promising results with all the models
- The *mult* model performs the best, followed by the *alm* model
- The cosine measures are most reliable, and density is less reliable of all
- We have established a link between two NLP areas

Future Work

- Explore the features of the semantic space setting and parameter setting for the models
- Consider how to extend semantic models to consider context information
- Use the output of semantic models to build an error detection classifier

Thank you!

Questions?