Distributional Semantics from Text and Images

Elia Bruni, Giang Binh Tran, and Marco Baroni, 2011

Distributional Semantics

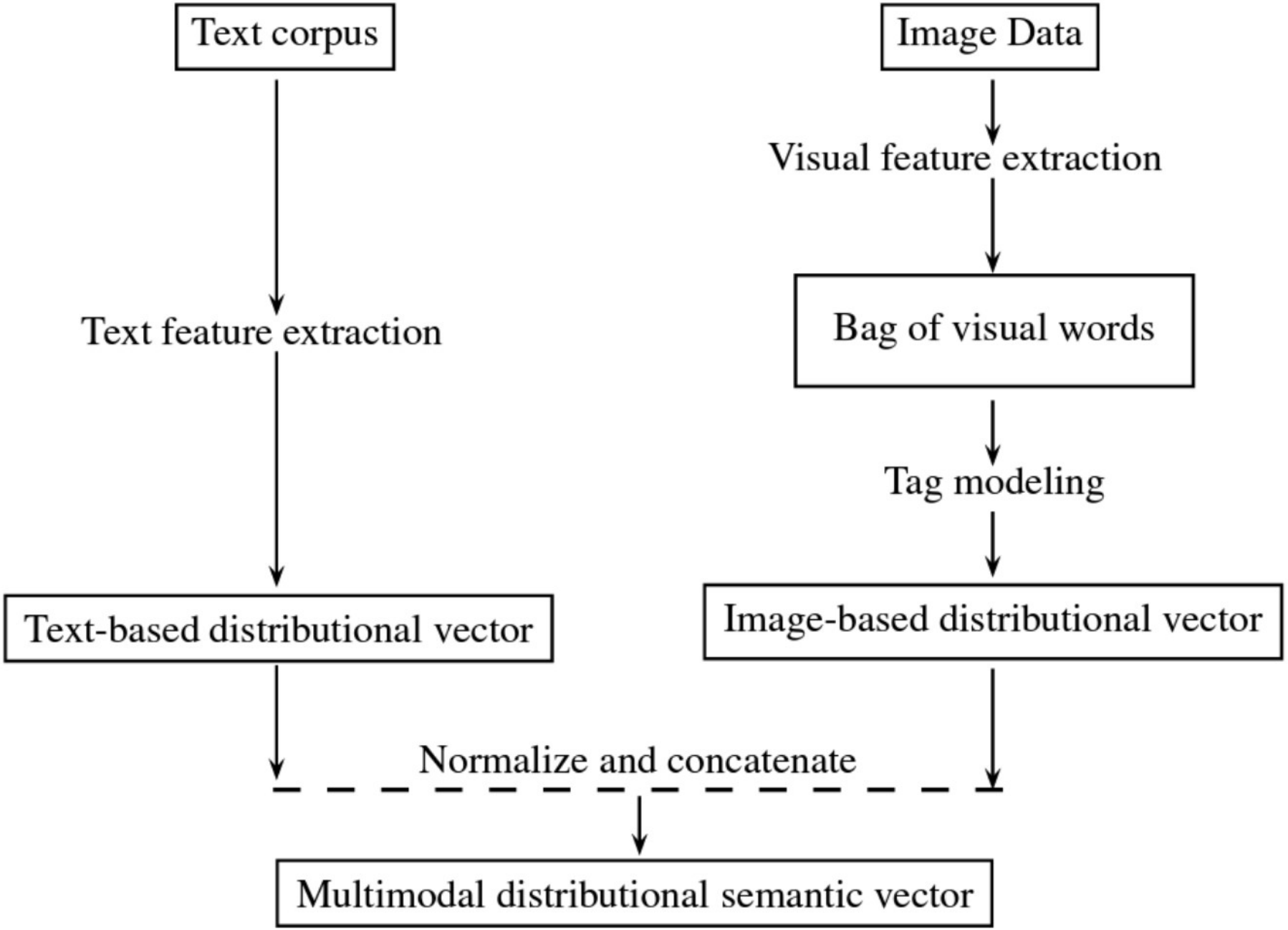
- Most approaches purely linguistic
  - Not cognitively valid
  - Symbols not grounded
- Need non-linguistic information

Aim of Paper

"We try to make these two mutually exclusive accounts communicate, to construct a richer and more human-like notion of meaning"

Previous Work

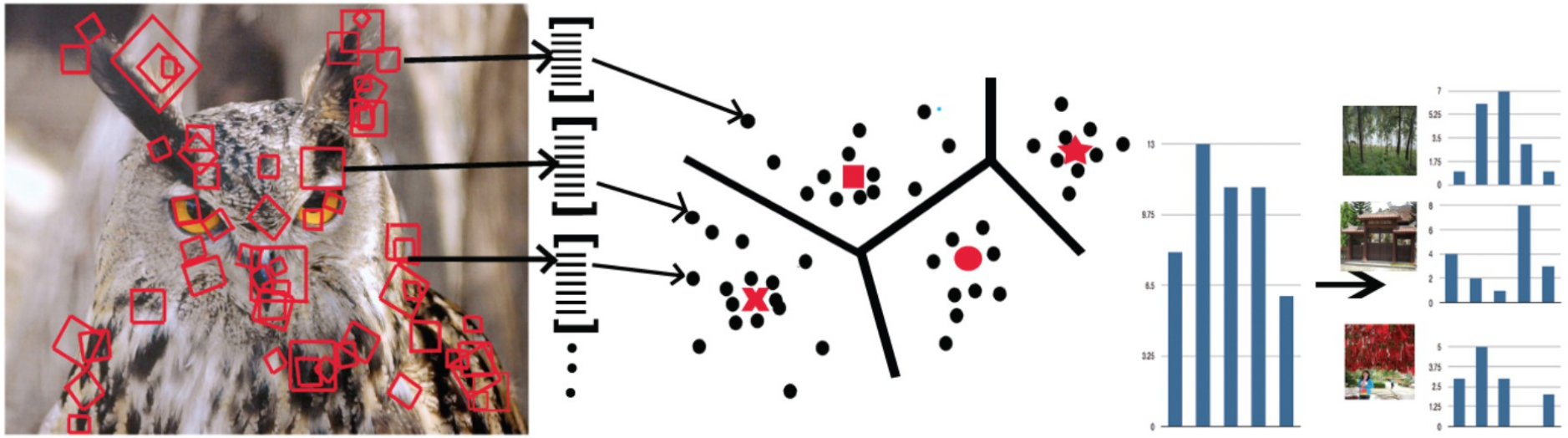
- Feature norms (Andrews et al., 2009)
  - Poor proxy of perceptual experience
- Joint topic model (Feng & Lapata, 2010)
  - Must be trained together

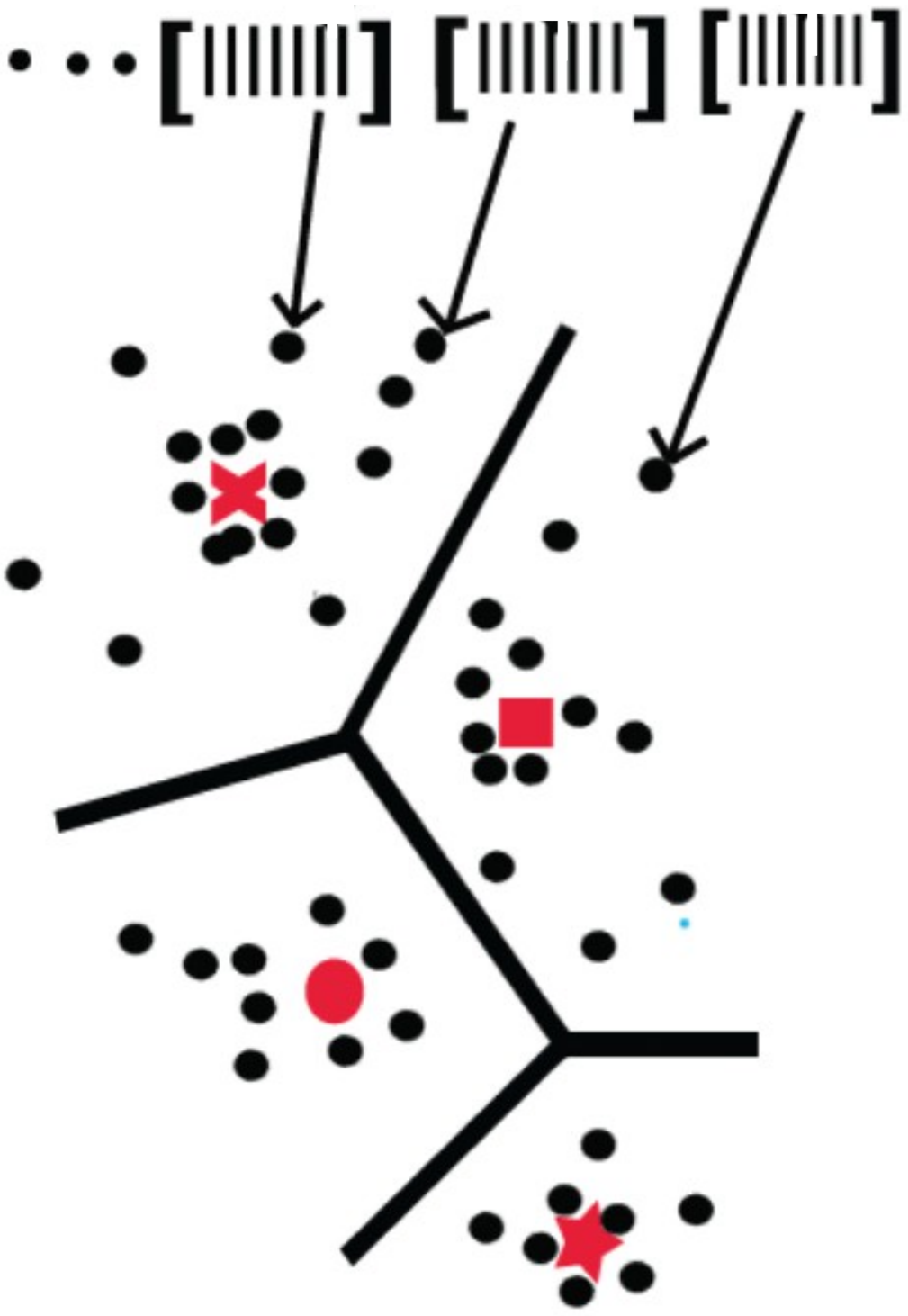

Text corpus
Image Data
Visual feature extraction
Text feature extraction
Bag of visual words
Tag modeling
Text-based distributional vector
Image-based distributional vector
Normalize and concatenate
Multimodal distributional semantic vector

Text-Based Model

- TypeDM (Baroni and Lenci, 2010)
  - Semantic links
  - Count different realisations
  - Local mutual information
- 1.8 billion tokens
  - Web documents, Wikipedia, BNC

Image-Based Model





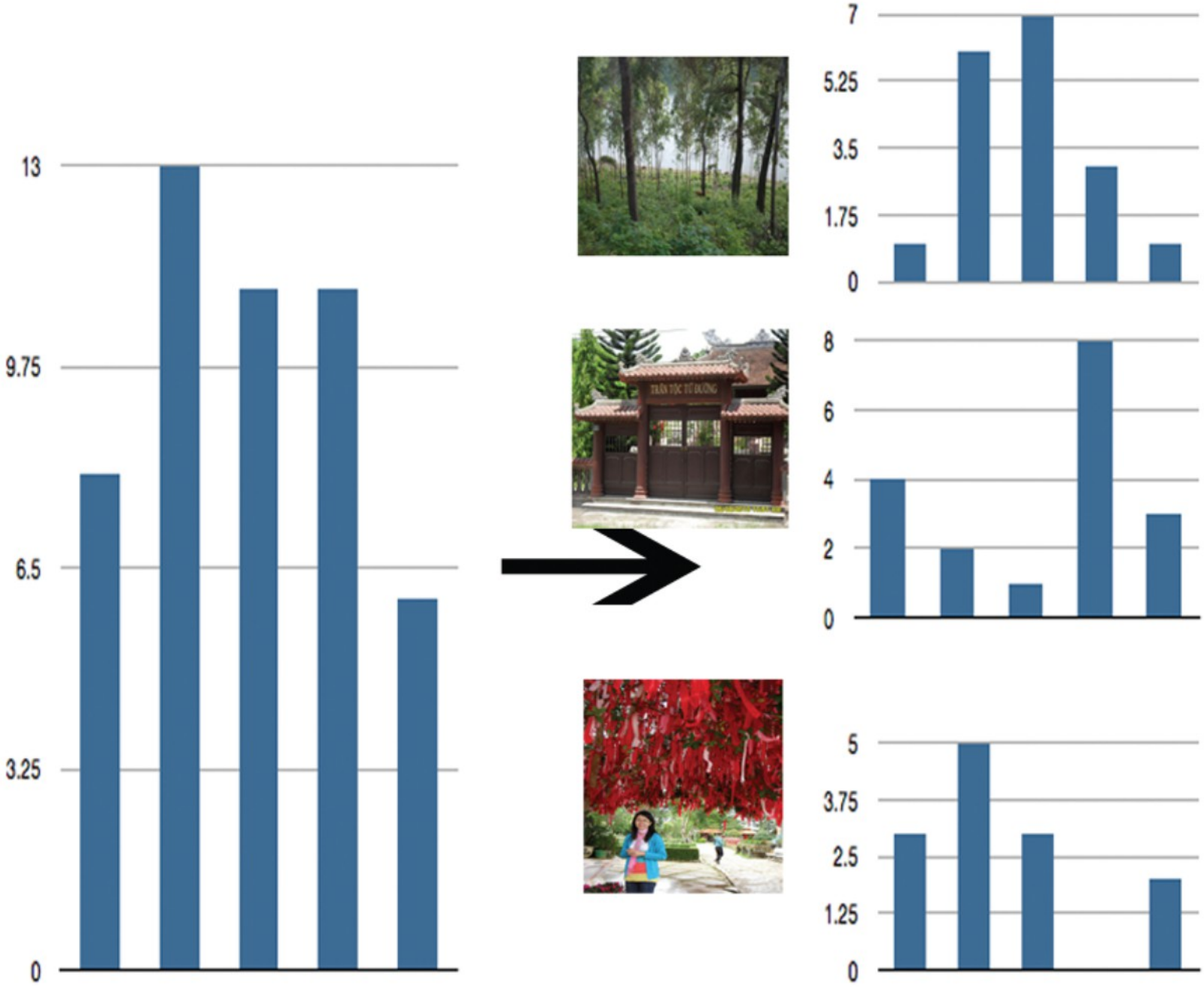

13
9.75
6.5
3.25
0
7
5.25
3.5
1.75
0
8
6
4
2
0
5
3.75
2.5
1.25
0

Image-Based Model

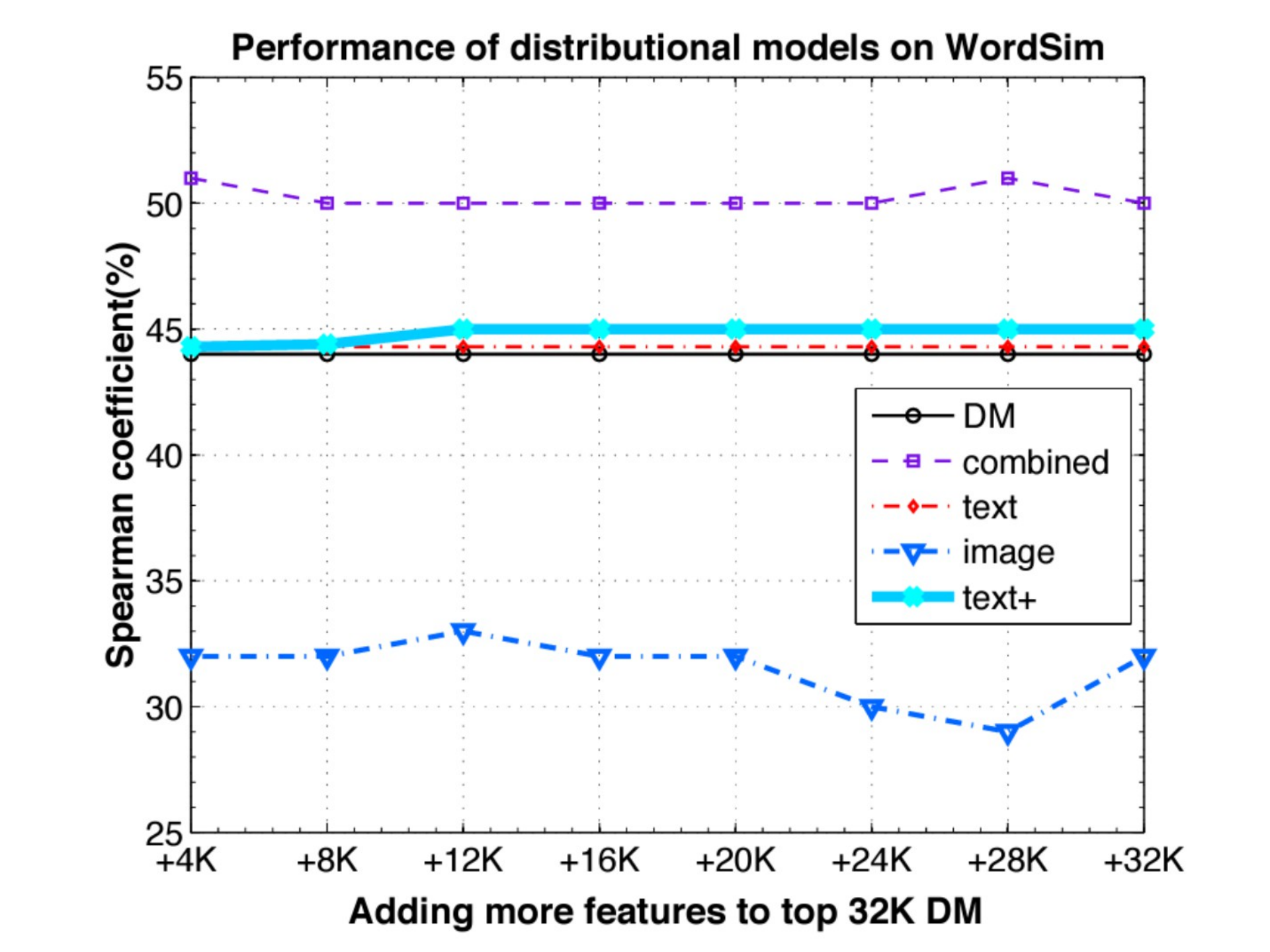
- ESP-Game (von Ahn and Dabbish, 2004)
  - 50K images
  - 6.7 tags/image
- Extract bags of visual words

Evaluation Methods

- Word Similarity (Finkelstein et al., 2002)
- Concept Clustering
  - (Almuhareb, 2006)
  - (Baroni et al, 2010)
- BLESS: Baroni-Lenci Evaluation of Semantic Similarity (GEMS 2011)

BLESS

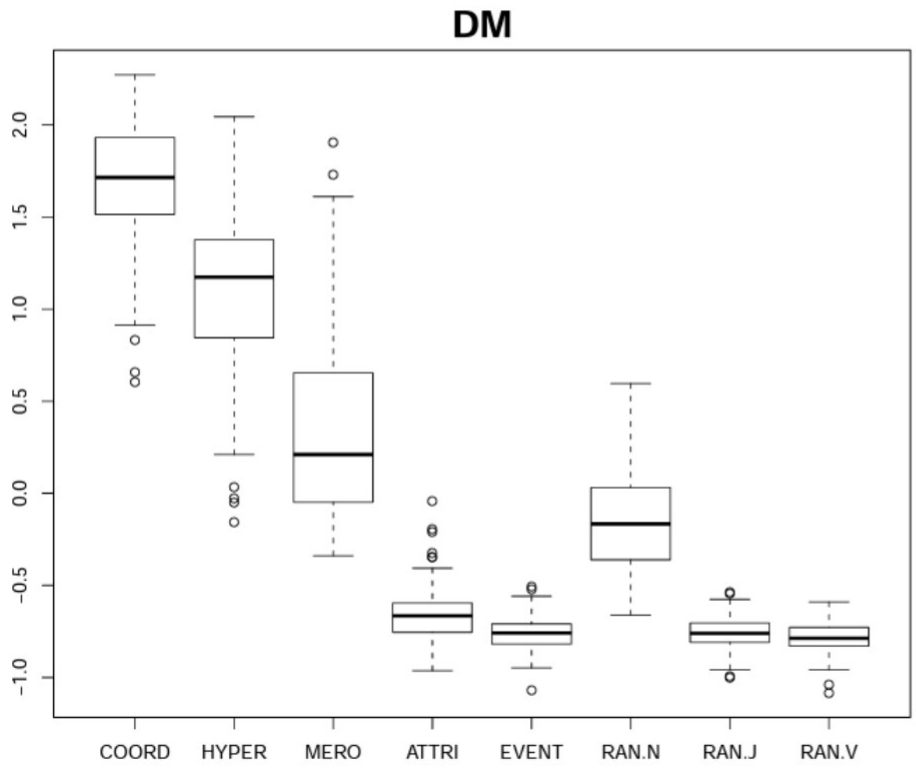
- Hypernymy (spear - weapon)
- Coordination (tiger - coyote)
- Meronymy (castle - hall)
- Attribute (grapefruit - tart)
- Event (cat - hiss)

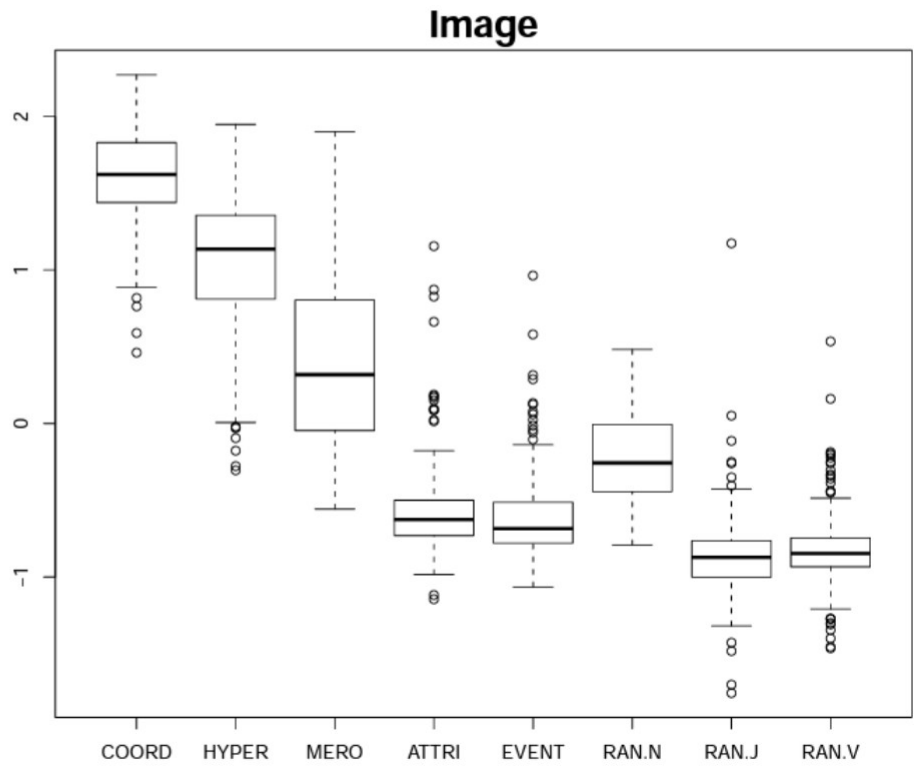

Performance of distributional models on WordSim
Spearman coefficient(%)
55
50
45
40
35
30
25
+4K
+8K
+12K
+16K
+20K
+24K
+28K
+32K
Adding more features to top 32K DM
DM
combined
text
image
text+

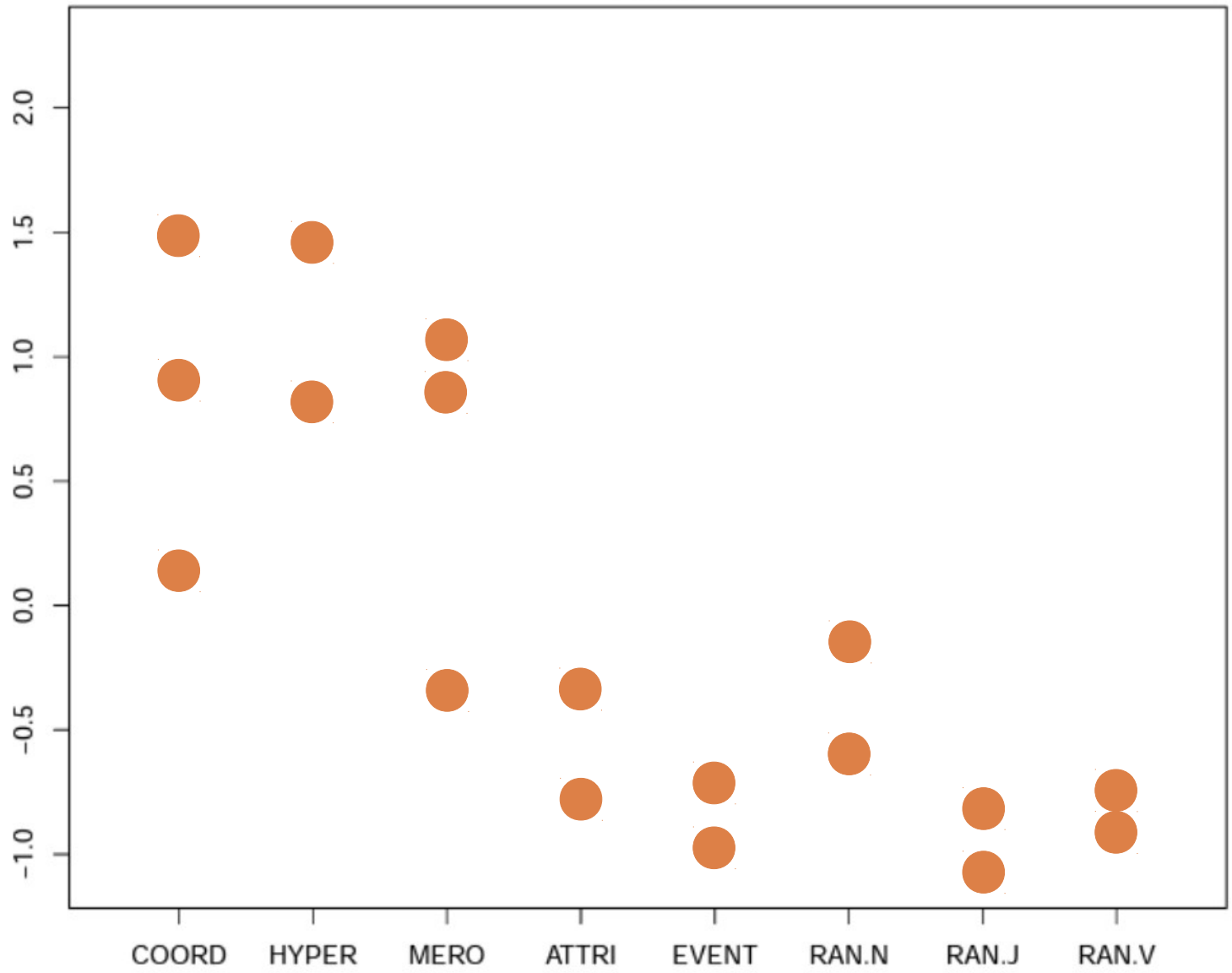
Results (clustering)

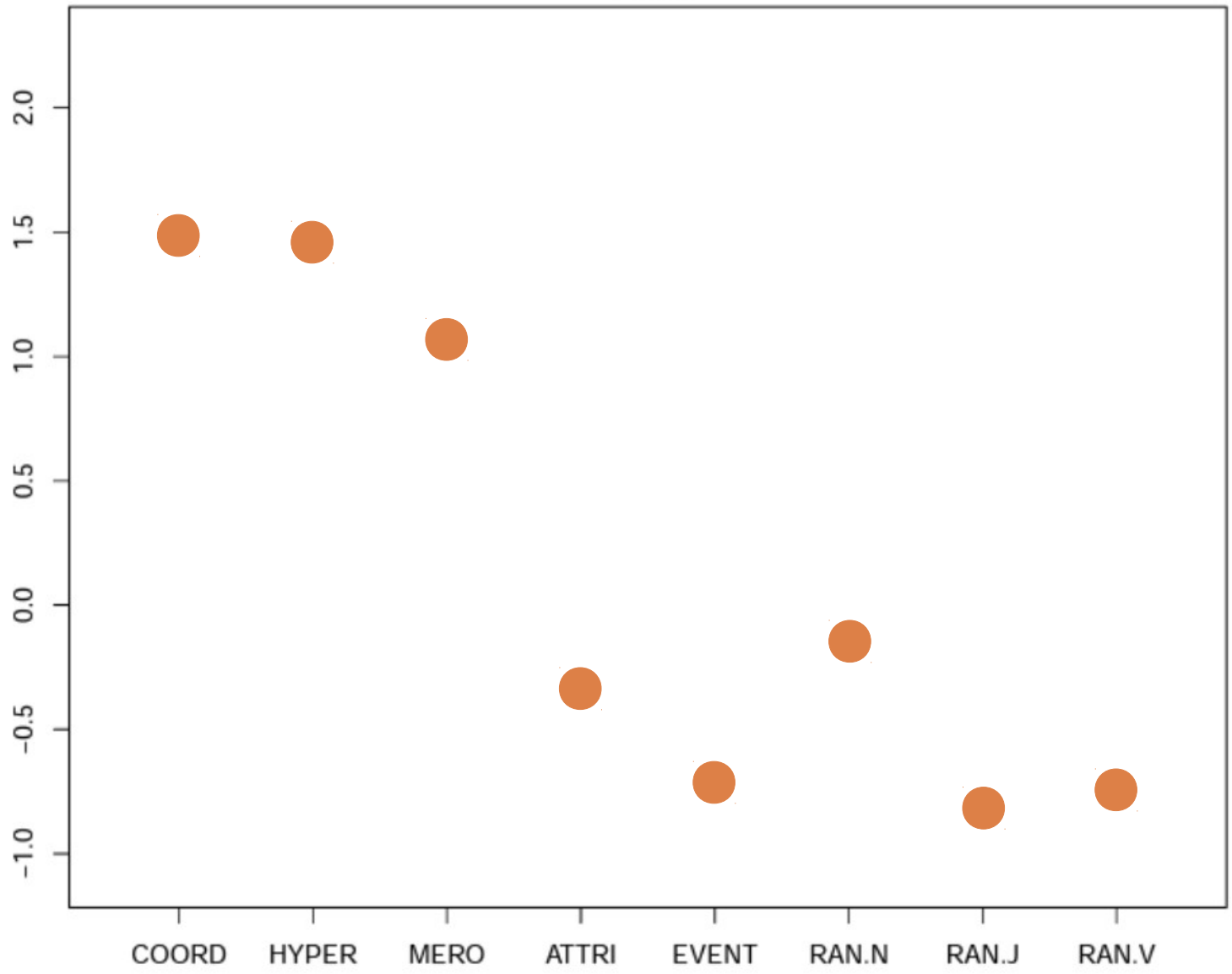
Model	AP	Battig
DM	**81**	**96**
text	79	83
text+	80	86
image	25	36
combined	78	**96**

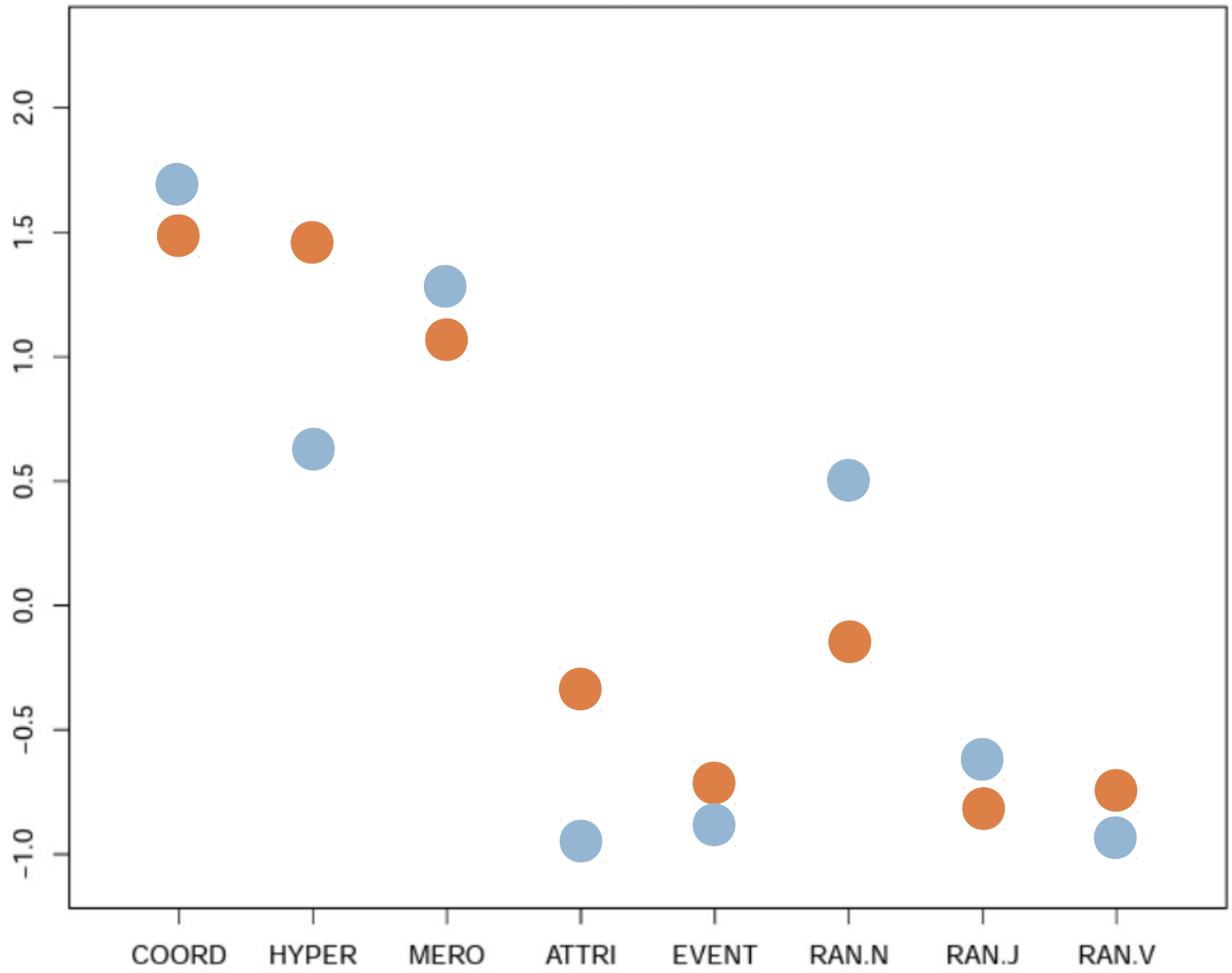
Results (BLESS)

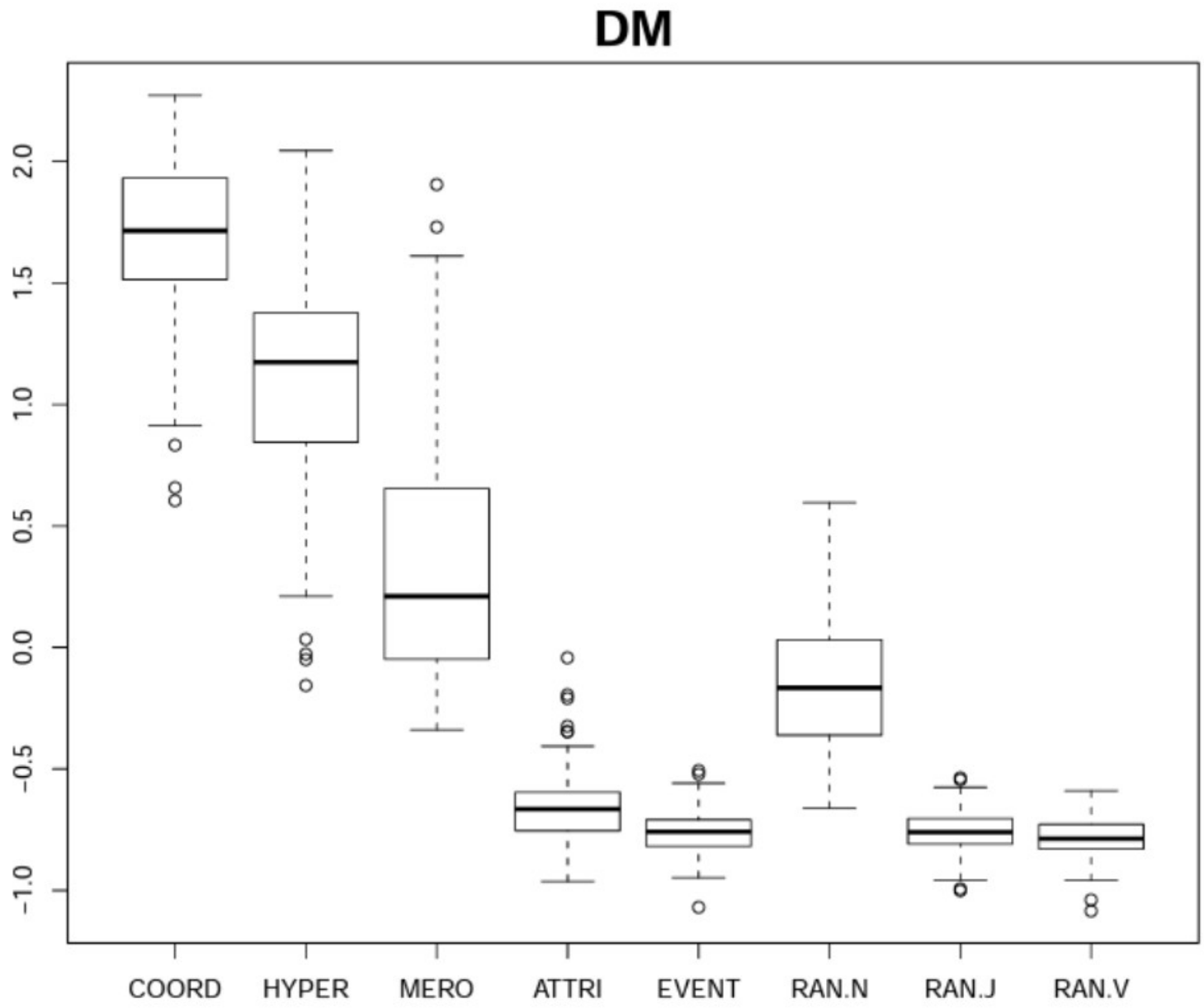





2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V


DM
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V

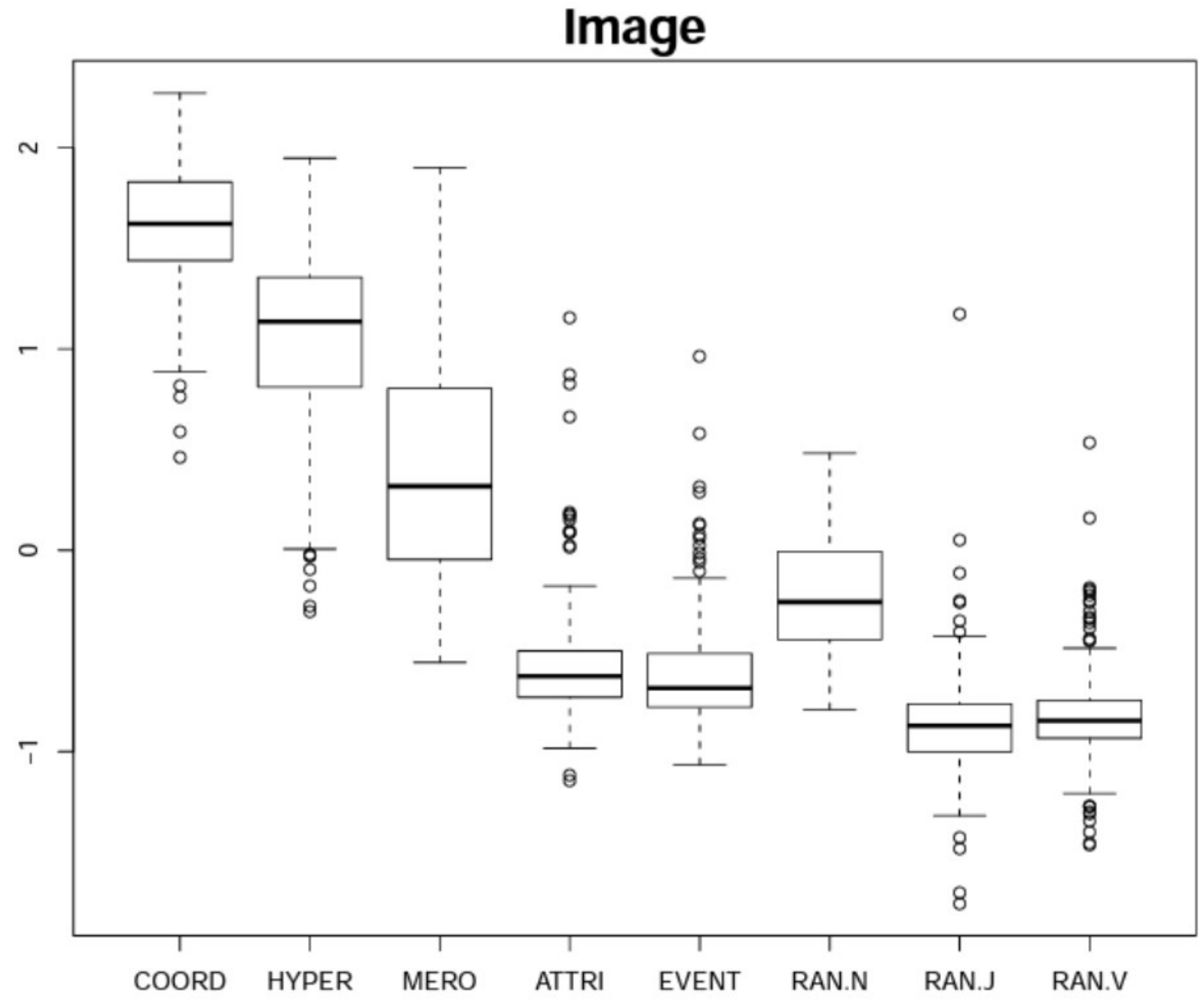


Image
2
1
0
−1
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V

Summary of Results

- WordSim
  - Promising but not significant
- Concept clustering
  - Mixed results
- BLESS
  - Can distinguish related attributes and events from random words

Qualitative Inspection

combined	text+
tennis / racket	physics / proton
planet / sun	championship / tournament
closet / clothes	profit / loss
king / rook	registration / arrangement
cell / phone	mile / kilometer

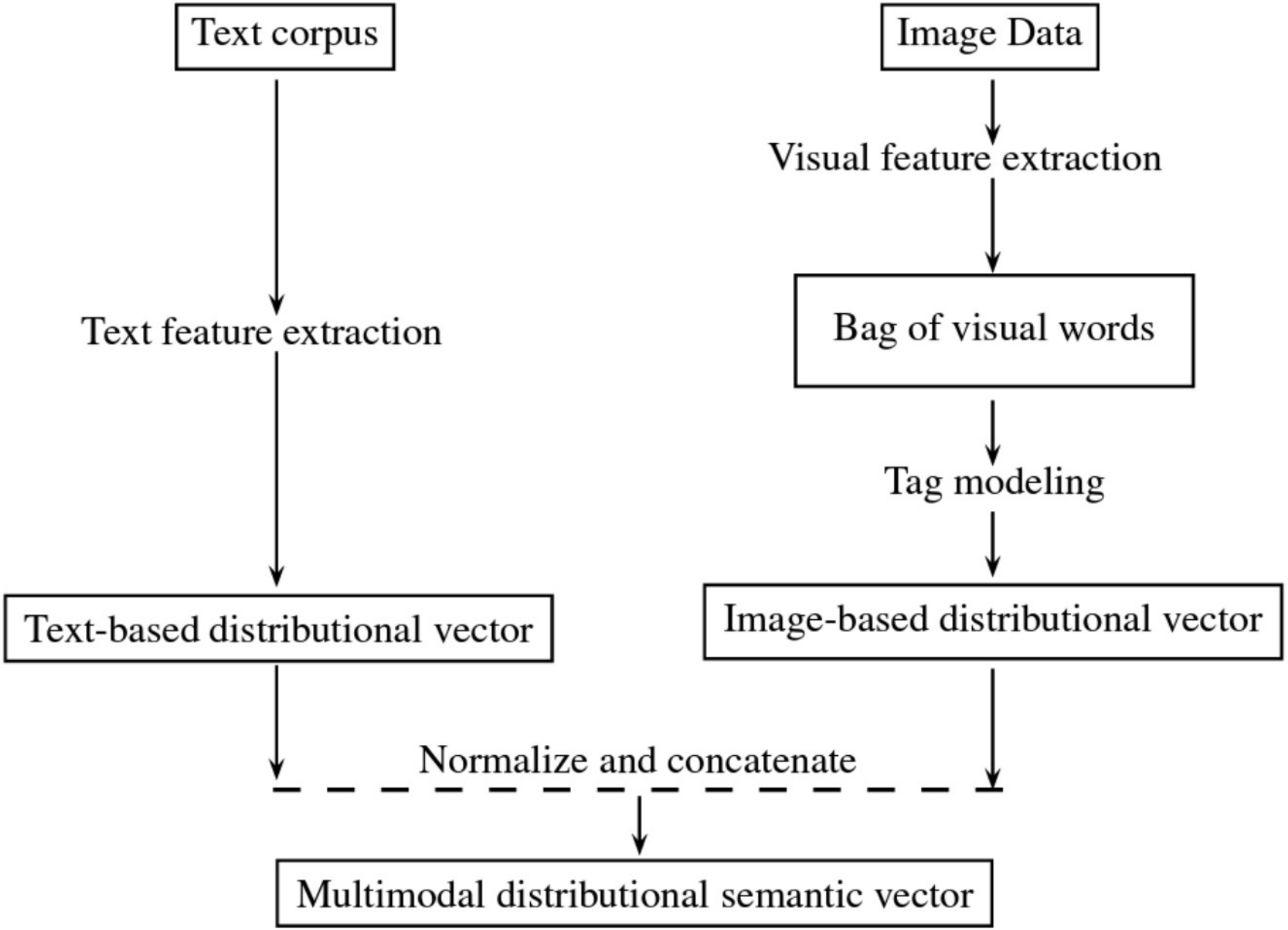
concept	*DM*	*image*	*concept*	*DM*	*image*
ant	small	black	potato	edible	red
axe	powerful	old	rifle	short	black
cathedral	ancient	dark	scooter	cheap	white
cottage	little	old	shirt	fancy	black
dresser	new	square	sparrow	wild	brown
fighter	fast	old	squirrel	fluffy	brown
fork	dangerous	shiny	sweater	elegant	old
goose	white	old	truck	new	heavy
jet	fast	old	villa	new	cosy
pistol	dangerous	black	whale	large	gray

Aim of Paper

"We try to make these two mutually exclusive accounts communicate, to construct a richer and more human-like notion of meaning"

Better semantics?

- Richer
- More human-like
- Grounded

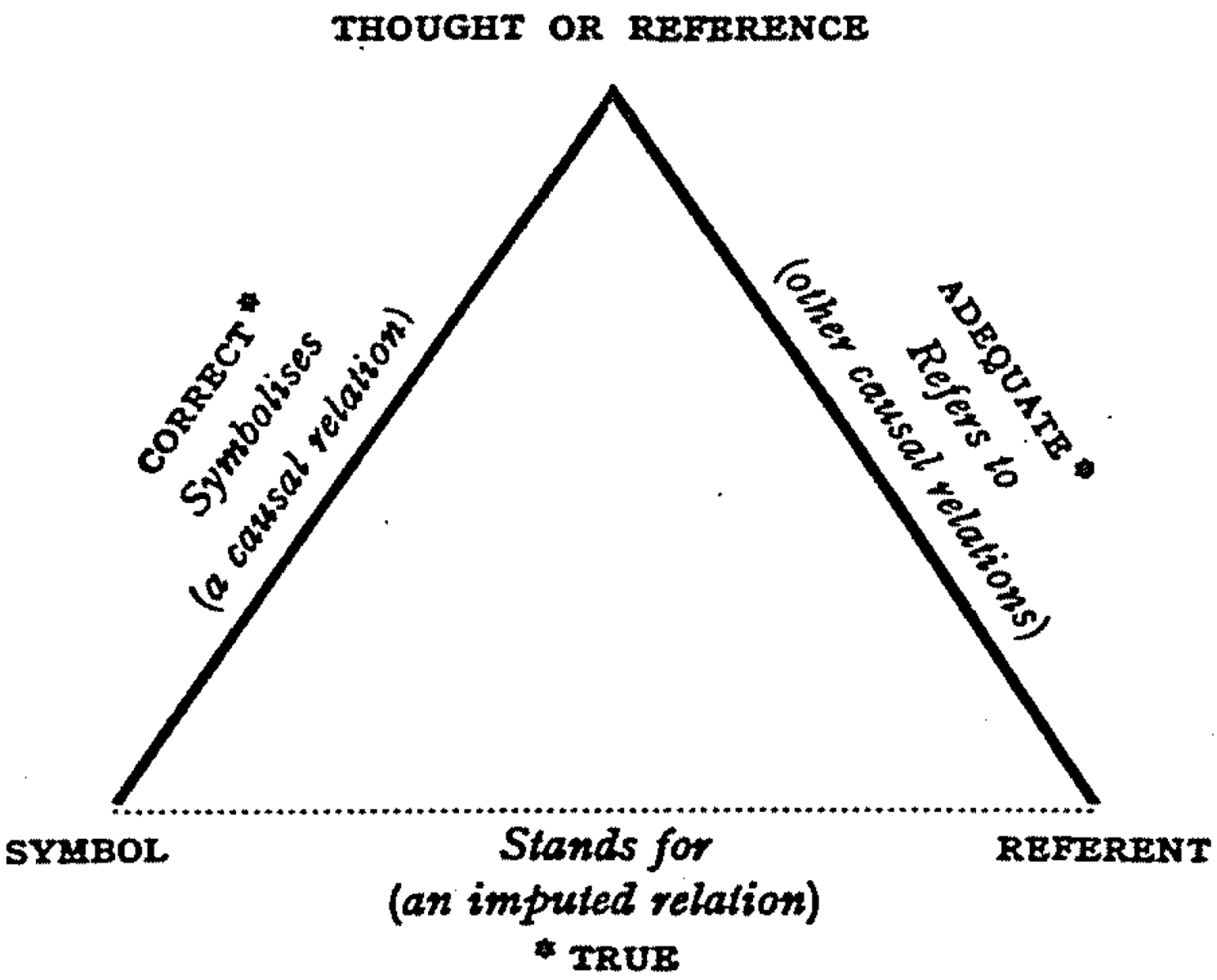

Text corpus
Image Data
Visual feature extraction
Text feature extraction
Bag of visual words
Tag modeling
Text-based distributional vector
Image-based distributional vector
Normalize and concatenate
Multimodal distributional semantic vector

$$\text{Sim}_{\text{combined}} = \frac{\text{Sim}_{\text{text}} + \text{Sim}_{\text{image}}}{2}$$

Symbol Grounding

"How can the meanings of meaningless symbol tokens... be grounded in anything but other meaningless symbols?"

(Harnad, 1990)



(Ogden and Richards, 1923)

Better semantics?

- Multimodal
- Still no “communication”
- Only partly grounded

Complementarity

"We thus see here the first evidence of the complementary nature of visual and textual information"

Comments on Slides

5 The method proposed by the authors - separate models are trained for textual and visual information, which allows them to exploit independent datasets.

6 The text-based model is a state-of-the-art off-the-shelf tool. This builds distributional vectors using semantic links as contexts, counting how many different realisations of each context appears in the corpus. For example, "animal skin" and "skin of an animal" would be two different realisations of the same semantic link.

7-11 The image-based model uses existing image processing software which finds points of interest in an image. Each of point of interest can be expressed as a vector, so an image can then be described by a bag of vectors. After calculating the points of interest of all images in the corpus, a clustering algorithm is run on the corresponding vectors. Each cluster represents a "visual word", and the bag of vectors for an image can then be converted to a "bag of visual words".

13 Word similarity and concept clustering are standard evaluation tasks, but the BLESS dataset requires more explanation. It comprises words paired with sets of related words, organised into five categories. In the given examples, the target word is on the left, and the related word on the right. For every target word, there is at least one word for each category, and typically more. This will be more fully explained in slides 17-21.

14 The black line shows performance for the full textual distributional model, and the dashed red line shows performance for the textual model when only using the top 32K dimensions. (Note that for both of these lines, the data points do not really mean anything, as the y-values are always the same.) The other three lines show performance for a varying number of dimensions, for different models. The thick light blue line represents the textual model, from 36K dimensions on the far left, to 64K dimensions on the far right - we can see that performance increases over the 32K model (red line), but must eventually decrease to that of the full model (black line). The dashed blue line at the bottom represents the visual model, again from 36K dimensions on the far left, to 64K dimensions on the far right - we can see that this performs worse than the textual model. Finally, the dashed purple line at the top represents a combined model, with 32K textual dimensions, and additional visual dimensions (4K on the left, to 32K on the right). The combined model exhibits the highest performance, but the authors note that the improvement over the textual model is not statistically significant.

15 For the Battig dataset, the combined model is competitive with the full textual model, and the visual features improve performance more than an equivalent number of additional textual features. For the AP dataset, the visual features actually decrease performance.

17-19 For each target word in the BLESS dataset, we take all related words, and calculate the similarity with the target. We also do this for a set of random nouns, adjectives, and verbs. We then choose the highest similarity score in each of these eight categories (five BLESS categories, and three types of random word), and normalise them as z-scores, based on the standard deviation. We can do this for every target word in the dataset, which produces a distribution of z-scores for each category. The distributions can be visualised using box-and-whisker plots, as shown in the next two slides.

20-21 The visual model can significantly distinguish related attributes and events from random words, while the textual model cannot. However, whether it really makes sense to say that an attribute of an object is similar to that object is debatable.

23 For each pair of words in the WordSim dataset, we can calculate the ratio between the similarities produced by the combined and by the textual models. On the left, we have those pairs which are judged most similar by the combined model but not the textual model, and vice versa. We can see that the combined model prefers words that can be easily expressed with pictures, while the textual model prefers abstract concepts.

24 Given a target word, we can find the most similar attribute - a random selection of target words are displayed here. As with the previous slide, we can see the combined model prefers visible attributes, such as colours, while the textual model prefers abstract attributes. However, note there is also a preference for common words such as "old", which are presumably nearer the centre of the semantic space.

25-26 At this point, we can ask whether the authors have achieved what they wanted to - have they produced a better semantic model, which is richer, more human-like, and grounded in the real world?

27-28 The proposed method only combines the textual and visual information by concatenating vectors. Since they normalise the vectors before doing this, and all of their evaluation methods rely on cosine similarity, the extent of the "communication" between the two sources of information reduces to averaging the similarity scores.

30 The so-called "semiotic triangle" gives a helpful perspective on the symbol grounding problem. At the bottom left, we have symbols (such as words); at the top, abstract concepts (either psychological or part of a formal semantic system); and at the bottom right, referents in the real world. A symbol can be said to be grounded if it represents a concept that has a referent. Textual distributional semantic models try to represent concepts using a semantic vector space, but since this is built using a contexts of words, there is no link to referents in the real world - the symbols are therefore not grounded. In the visual model, the semantic space is built

using visual information, so the symbols are grounded. In the combined model, we can only say that these concepts are partly grounded, given the lack of real "communication" between its components.

31-32 Despite the inconclusive results in terms of performance, and the simplistic way that visual and textual information is combined, we can still see this paper as demonstrating a complementarity between the two modalities.