# Distributional Semantics and Compositionality 2011:
## Shared Task Description and Results


Chris Biemann
TU Darmstadt
Germany

Eugenie Giesbrecht
FZI Karlsruhe
Germany

DiSCo 2011 Workshop @ ACL-HLT 2011, June 24, 2011, Portland, Oregon, USA

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Overview of the Shared Task

- Motivation
- Preparation
  - Corpora
  - Semi-automatic candidate extraction
  - Mturk for collecting judgments
- Data
- Evaluation scoring
- Results

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Why a shared task on graded compositionality?

- Distributional models assume compositionality
- Non-compositional phrases should be treated as multi-word units
- Multi-word definition is application-dependent
- some phrases are more compositional than others
- for some phrases, compositionality depends on the context
- First data set for graded compositionality

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Why call for corpus-based models?

- DMs have been successfully applied to a number of semantic tasks
- Compositionality in DMs still a research topic
- Corpus–based acquisition of MWUs is language-independent
- Corpus-based models for graded compositionality would enable MWU lists tailored to applications by
  - computing them on the application domain
  - thresholding on compositionality score based on performance

# Preparation: Corpora

- WaCky:
  - large (1-2B tokens) enough for corpus-based methods
  - freely available in
  - English, German, Italian, French
  - POS-tagged
  - lemma information
  - uniform format
  - web-based: realistic distribution
  - cleaned

# Target Constructions

- To restrict the focus, we only look at word pairs in three highly frequent constructions

- ADJ_NN: adjectives modifying nouns, as in “red herring”, “blue skies”
- V_SUBJ: verbs and nouns in subject position, e.g. “flies fly”, “people transfer (sth.)”
- V_OBJ: verbs and nouns in object position, e.g. “lose keys” , “kick bucket”

# From WaCky to Phrases

- Extract candidates, overgenerate
  - POS patterns
  - window-based approach
- Sort in descending order of frequency
- Filter manually for plausible candidates: typical pairs in syntactic positions
- Select “balanced” set based on subjective compositionality of phrases

➔Must bias selection since non-compositional phrases are rare

# From Phrases to Contexts

- Extract 7 sentences per phrase from corpus
- Exclude very long, very short or spurious sentences
- Exclude phrases that appear in very fixed contexts
- Use 5 sentences per phrase for collection of judgments

# Example contexts for "bucking the trend"

- I would like to **buck** the **trend** of complaint !
- One company that is **bucking** the **trend** is Flowcrete Group plc located in Sandbach , Cheshire . "
- We are now moving into a new phase where we are hoping to **buck** the **trend** .
- With a claimed 11,000 customers and what look like aggressive growth plans , including recent acquisitions of Infinium Software , Interbiz and earlier also Max international , the firm does seem to be **bucking** the **trend** of difficult times .
- Every time we get a new PocketPC in to Pocket-Lint tower , it seems to offer more features for less money and the HP iPaq 4150 is n't about to **buck** the **trend** .

# Mturk Human Intelligence Task

## How literal is this phrase?

Can you infer the meaning of a given phrase by only considering their parts literally, or does the phrase carry a 'special' meaning?
In the context below, how literal is the meaning of the phrase in bold?
Enter a number between 0 and 10.

- 0 means: this phrase is not to be understood literally at all.
- 10 means: this phrase is to be understood very literally.
- Use values in between to grade your decision. Please, however, try to take a stand as often as possible.

In case the context is unclear or nonsensical, please enter "66" and use the comment field to explain. However, please try to make sense of it even if the sentences are incomplete.

Example 1 :
There was a red truck parked curbside. It looked like someone was living in it.
YOUR ANSWER: 10
reason: the color of the truck is red, this can be inferred from the parts "red" and "truck" only - without any special knowledge.

Example 2 :
What a tour! We were on cloud nine when we got back to headquarters but we kept our mouths shut.
YOUR ANSWER: 0
reason: "cloud nine" means to be blissfully happy. It does NOT refer to a cloud with the number nine.

Example 3 :
Yellow fever is found only in parts of South America and Africa.
YOUR ANSWER: 7
reason: "yellow fever" refers to a disease causing high body temperature. However, the fever itself is not yellow. Overall, this phrase is fairly literal, but not totally, hence answering with a value
between 5 and 8 is appropriate.

We take rejection seriously and will not reject a HIT unless done carelessly. Entering anything else but numbers between 0 and 10 or 66 in the judgment field will automatically trigger rejection.

YOUR CONTEXT with **big day**
Special Offers : Please call FREEPHONE 0800 0762205 to receive your free copy of ' Groom ' the full colour magazine dedicated to dressing up for the **big day** and details of Moss Bros Hire rates .
How literal is the bolded phrase in the context above between 0 and 10?
[ ]
OPTIONAL: leave a comment, tell us about what is broken, help us to improve this type of HIT:
[ ]

# Quality worker selection

1. Open task: $0.02
   - anyone can submit answers.
   - Clear-cut test examples.
   - high volume, high quality people get invited for the closed task
2. Closed task: $0.03
   - 4 workers per HIT
   - eyeballing for quality check

# Sample Answers and Score Calculation

| | Responses |
|---|---|
| • I look towards the **big picture** , what 's really happening behind the illusions of the separate ego . | 0; 3; 1; 0 |
| • " I think the things which have longevity will be the things that have a bit of depth to them , that are part of a **bigger picture** . | 5; 5; 0; 0 |
| • The ' close look at **the big picture** ' series of conferences kicked off in Manchester in November . | 0; 0; 3; 4 |
| • Click here for a **bigger picture** | 10; 10; 10; 10 |
| You see a picture, but when you click, you can view a larger picture. The size increases. | |
| • In order to see the **bigger picture** you have to be personally and interpersonally aware . | 0; 4; 1; 5 |

TECHNISCHE UNIVERSITÄT DARMSTADT

Sum = 71, Avg = Sum/#judgments = 3.55,
**Score =** round(10*Avg) = **36**

FZI

# Data Sets in Numbers

| EN | ADJ_NN | V_SUBJ | V_OBJ | Sum |
|---|---|---|---|---|
| Train | 58 (43) | 30 (23) | 52 (41) | 140 (107) |
| Vali. | 10 (7) | 9 (6) | 16 (13) | 35 (26) |
| Test | 77 (52) | 35 (26) | 62 (40) | 174 (118) |
| All | 145 (102) | 74 (55) | 130 (94) | 349 (251) |

| DE | ADJ_NN | V_SUBJ | V_OBJ | Sum |
|---|---|---|---|---|
| Train | 49 (42) | 26 (23) | 44 (33) | 119 (98) |
| Vali. | 11 (8) | 9 (8) | 9 (7) | 29 (23) |
| Test | 63 (48) | 29 (28) | 57 (44) | 149 (120) |
| All | 123 (98) | 64 (59) | 110 (84) | 297 (241) |

- coarse scoring (numbers in parentheses)
  - low: 0..25
  - medium: 38..62
  - high: 75..100

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Evaluation Scoring

$$NUMSCORE(S,G) = \frac{1}{N}\sum_{i=1..N}|g_i - s_i|$$

$$COARSE(S,G) = \frac{1}{N}\sum_{i=1..N}\begin{cases} s_i == g_i : 1 \\ otherwise : 0 \end{cases}$$

- S=($s_1$,$s_2$, … $s_n$) system responses
- G=($g_1$,$g_2$, … $g_n$) gold standard
- missing system responses are filled with 50 / medium

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Participants

| Systems | Institution | Team | Approach |
|---|---|---|---|
| Duluth-1<br>Duluth-2<br>Duluth-3 | Dept. of Computer Science,<br>University of Minnesota | Ted Pedersen | statistical<br>association measures:<br>t-score and pmi |
| JUCSE-1<br>JUCSE-2<br>JUCSE-3 | Jadavpur University | Tanmoy Chakraborty, Santanu Pal<br>Tapabrata Mondal, Tanik Saikh,<br>Sivaju Bandyopadhyay | mix of statistical<br>association measures |
| SCSS-TCD:conf1<br>SCSS-TCD:conf2<br>SCSS-TCD:conf3 | SCSS,<br>Trinity College Dublin | Alfredo Maldonado-Guerra,<br>Martin Emms | unsupervised WSM,<br>cosine similarity |
| submission-ws<br>submission-pmi | Gavagai | Hillevi Hägglöf,<br>Lisa Tengstrand | random indexing<br>association measures (pmi |
| UCPH-simple.en | University of Copenhagen | Anders Johannsen, Hector Martinez,<br>Christian Rishøj, Anders Søgaard | support vector regression<br>with COALS-based<br>endocentricity features |
| UoY: Exm<br>UoY: Exm-Best<br>UoY: Pro-Best | University of York, UK;<br>Lexical Computing Ltd., UK | Siva Reddy, Diana McCarthy,<br>Suresh Manandhar,<br>Spandana Gella | exemplar-based WSMs<br><br>prototype-based WSM |
| UNED-1: NN<br>UNED-2: NN<br>UNED-3: NN | NLP and IR Group at UNED | Guillermo Garrido,<br>Anselmo Peas | syntactic VSM,<br>dependency-parsed UKW<br>SVM classifier |

Table 3: Participants of DiSCo'2011 Shared Task

# English Numeric Results

| | responses | Spearman's $\rho$ | Kendall's $\tau$ | EN all | EN_ADJ_NN | EN_V_SUBJ | EN_V_OBJ |
|---|---|---|---|---|---|---|---|
| number of phrases | | | | 174 | 77 | 35 | 62 |
| 0-response baseline | 0 | N/A | N/A | 23.42 | 24.67 | 17.03 | 25.47 |
| random baseline | 174 | (0.02) | (0.02) | 32.82 | 34.57 | 29.83 | 32.34 |
| UCPH-simple.en | 174 | 0.27 | 0.18 | **16.19** | 14.93 | 21.64 | **14.66** |
| UoY: Exm-Best | 169 | **0.35** | **0.24** | 16.51 | 15.19 | **15.72** | 18.6 |
| UoY: Pro-Best | 169 | 0.33 | 0.23 | 16.79 | **14.62** | 18.89 | 18.31 |
| UoY: Exm | 169 | 0.26 | 0.18 | 17.28 | 15.82 | 18.18 | 18.6 |
| SCSS-TCD: conf1 | 174 | 0.27 | 0.19 | 17.95 | 18.56 | 20.8 | 15.58 |
| SCSS-TCD: conf2 | 174 | 0.28 | 0.19 | 18.35 | 19.62 | 20.2 | 15.73 |
| Duluth-1 | 174 | (-0.01) | (-0.01) | 21.22 | 19.35 | 26.71 | 20.45 |
| JUCSE-1 | 174 | 0.33 | 0.23 | 22.67 | 25.32 | 17.71 | 22.16 |
| JUCSE-2 | 174 | 0.32 | 0.22 | 22.94 | 25.69 | 17.51 | 22.6 |
| SCSS-TCD: conf3 | 174 | 0.18 | 0.12 | 25.59 | 24.16 | 32.04 | 23.73 |
| JUCSE-3 | 174 | (-0.04) | (-0.03) | 25.75 | 30.03 | 26.91 | 19.77 |
| Duluth-2 | 174 | (-0.06) | (-0.04) | 27.93 | 37.45 | 17.74 | 21.85 |
| Duluth-3 | 174 | (-0.08) | (-0.05) | 33.04 | 44.04 | 17.6 | 28.09 |
| submission-ws | 173 | 0.24 | 0.16 | 44.27 | 37.24 | 50.06 | 49.72 |
| submission-pmi | 96 | - | - | - | - | 52.13 | 50.46 |
| UNED-1: NN | 77 | - | - | - | 17.02 | - | - |
| UNED-2: NN | 77 | - | - | - | 17.18 | - | - |
| UNED-3: NN | 77 | - | - | - | 17.29 | - | - |

# English Coarse Results

| | responses | EN all | EN_ADJ_NN | EN_V_SUBJ | EN_V_OBJ |
|---|---|---|---|---|---|
| number of phrases | | 118 | 52 | 26 | 40 |
| zero-response baseline | 0 | 0.356 | 0.288 | 0.654 | 0.250 |
| random baseline | 118 | 0.297 | 0.288 | 0.308 | 0.300 |
| Duluth-1 | 118 | **0.585** | 0.654 | 0.385 | 0.625 |
| UoY: Exm-Best | 114 | 0.576 | 0.692 | 0.500 | 0.475 |
| UoY: Pro-Best | 114 | 0.567 | **0.731** | 0.346 | 0.500 |
| UoY: Exm | 114 | 0.542 | 0.692 | 0.346 | 0.475 |
| SCSS-TCD: conf2 | 118 | 0.542 | 0.635 | 0.192 | **0.650** |
| SCSS-TCD: conf1 | 118 | 0.534 | 0.64 | 0.192 | 0.625 |
| JUCSE-3 | 118 | 0.475 | 0.442 | 0.346 | 0.600 |
| JUCSE-2 | 118 | 0.458 | 0.481 | 0.462 | 0.425 |
| SCSS-TCD: conf3 | 118 | 0.449 | 0.404 | 0.423 | 0.525 |
| JUCSE-1 | 118 | 0.441 | 0.442 | 0.462 | 0.425 |
| submission-ws | 117 | 0.373 | 0.346 | 0.269 | 0.475 |
| UCPH-simple.en | 118 | 0.356 | 0.346 | 0.500 | 0.275 |
| Duluth-2 | 118 | 0.322 | 0.173 | 0.346 | 0.500 |
| Duluth-3 | 118 | 0.322 | 0.135 | **0.577** | 0.400 |
| submission-pmi | - | - | - | 0.346 | 0.550 |
| UNED-1-NN | 52 | - | 0.289 | - | - |
| UNED-2-NN | 52 | - | 0.404 | - | - |
| UNED-3-NN | 52 | - | 0.327 | - | - |

TU D

FZI

# German Results

| numerical scores | responses | $\rho$ | $\tau$ | DE all | DE_ADJ_NN | DE_V_SUBJ | DE_V_OBJ |
|---|---|---|---|---|---|---|---|
| number of phrases | | | | 149 | 63 | 29 | 57 |
| 0-response baseline | 0 | - | - | 32.51 | 32.21 | 38.00 | 30.05 |
| random baseline | 149 | (0.005) | (0.004) | 37.79 | 36.27 | 47.45 | 34.54 |
| UCPH-simple.de | 148 | 0.171 | 0.116 | 24.03 | 27.09 | 15.55 | 24.06 |

| coarse values | responses | DE all | DE_ADJ_NN | DE_V_SUBJ | DE_V_OBJ |
|---|---|---|---|---|---|
| number of phrases | | 120 | 48 | 28 | 44 |
| 0-response baseline | 0 | 0.158 | 0.208 | 0.071 | 0.159 |
| random baseline | 120 | 0.283 | 0.313 | 0.214 | 0.295 |
| UCPH-simple.de | 119 | 0.283 | 0.375 | 0.286 | 0.182 |

- we have a clear winner here ☺

TECHNISCHE UNIVERSITÄT DARMSTADT

FZI

# Conclusions

- seven groups, 19 submissions
- two kinds of approaches:
  - lexical association measures
  - word space models of various flavors
- no clear winner for EN dataset, with *UoY: Exm-Best* being the most robust of the systems
- a slight favor for approaches based on word space model, esp. in numerical evaluation.

A pure corpus-based acquisition of graded compositionality is a hard task!

# Thanks!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

FZI