Metaphor recognition on short phrases:
Bulat et al. (2017)

Ines Reinig

HS Figurative Language Resolution
Institut für Computerlinguistik
Universität Heidelberg

05.06.2019
Contents

Bulat et al. (2017)
Overview
Linguistic representations
Property-norm semantic space
Cross-modal mapping
Metaphor classification

Bruni et al. (2012)
Distributional semantic models
Experiments
Goal: Metaphor identification using property-based representations

How is metaphor defined in this work?

→ Conceptual Metaphor Theory (CMT):
  ▶ Introduced in Lakoff and Johnson (1980)
  ▶ Many works on metaphor identification rely on this theory
  ▶ Core idea: metaphor is a cognitive phenomenon and not exclusively linguistic
  ▶ We perceive and conceive things in terms of concepts
  ▶ This conceptual system, which shapes the way we think and express ourselves, is metaphorical
  ▶ Definition of metaphor: understanding of one concept (target domain, e.g. “argument”) in terms of another (source domain, e.g. “war”)
The approach in a nutshell:

- Traditional embeddings (word2vec and count-based) are mapped to attribute vectors, using a supervised system trained on McRae norms
- The resulting vectors are then used as the input to an SVM classifier, which is trained to distinguish literal from metaphorical language
- Experiments show that using the attribute vectors yields a higher F score over using the original vector space.
Approach in Bulat et al. (2017)

Learn linguistic representations: EMBED, SVD

Learn cross-modal maps: ATTR-EMBED, ATTR-SVD

Create property-norm semantic space using MCRAE dataset

Compare models performances on metaphor classification task (SVM): EMBED, SVD, ATTR-EMBED, ATTR-SVD

Figure: overview of the approach in Bulat et al. (2017)
Approach in Bulat et al. (2017): linguistic representations

Learn linguistic representations: EMBED, SVD

Learn cross-modal maps: ATTR-EMBED, ATTR-SVD

Create property-norm semantic space using MCRAE dataset

Compare models performances on metaphor classification task (SVM): EMBED, SVD, ATTR-EMBED, ATTR-SVD

Figure: overview of the approach in Bulat et al. (2017)
Approach in Bulat et al. (2017): linguistic representations

- **EMBED**: context-predicting log-linear skip-gram model (Mikolov et al. 2013):
  1. Vocabulary: lemmas appearing 100+ times
  2. Train skip-gram model to predict context words for each target word in the vocabulary (as opposed to predicting the target word from context words / CBOW) by minimizing a loss-function (negative sampling in this case)

- **SVD**: context-counting model:
  1. Vocabulary: 10K most freq. lemmas
  2. Count word frequencies in sentences (context window: 1 sentence)
  3. Re-weight counts using PPMI
  4. Reduce vector dimensions from 10K to 100 using SVD (so vectors are smaller and denser)
Approach in Bulat et al. (2017)

Learn linguistic representations: EMBED, SVD → Learn cross-modal maps: ATTR-EMBED, ATTR-SVD → Compare models performances on metaphor classification task (SVM): EMBED, SVD, ATTR-EMBED, ATTR-SVD

Create property-norm semantic space using MCRAE dataset

Figure: overview of the approach in Bulat et al. (2017)
MCRAE: property norm dataset collected by McRae et al. (2005):

- one of the largest and most widely used in cognitive science
- 541 concepts annotated with properties (2526 properties total) and production frequencies
- can be extended by predicting properties for new concepts as suggested in this paper
Approach in Bulat et al. (2017): Property norms

Using MCRAE, a property-norm semantic space is created:

<table>
<thead>
<tr>
<th>ACCORDION</th>
<th>CLARINET</th>
<th>CROCODILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>is_loud, 6</td>
<td>has_keys, 9</td>
<td>is_long, 16</td>
</tr>
<tr>
<td>has_keys, 17</td>
<td>is_long 8</td>
<td></td>
</tr>
<tr>
<td>requires_air, 11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: Examples of properties from MCRAE

<table>
<thead>
<tr>
<th>ACCORDION</th>
<th>is_loud</th>
<th>has_keys</th>
<th>requires_air</th>
<th>is_long</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>17</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CLARINET</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>CROCODILE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

Table: Subspace of the property-norm semantic space created in this paper
Approach in Bulat et al. (2017)

- Learn linguistic representations: EMBED, SVD
- Create property-norm semantic space using MCRAE dataset
- Learn cross-modal maps: ATTR-EMBED, ATTR-SVD

Compare models performances on metaphor classification task (SVM): EMBED, SVD, ATTR-EMBED, ATTR-SVD

Figure: overview of the approach in Bulat et al. (2017)
Approach in Bulat et al. (2017): cross-modal maps

MCRAE contains “only” 541 annotated concepts.

As shown in previous publications (Fagarasan et al. 2015; Bulat et al. 2016), cross-modal maps can allow us to get property-based representations for new/unseen concepts.

This work follows the approach in Fagarasan et al. (2015) to get the property-based representations.

How? Learn a mapping function $f : \textbf{LS} \rightarrow \textbf{PS}$ between a linguistic representation $\textbf{LS}$ and the property-norm semantic space $\textbf{PS}$ using PLSR and the 541 concepts in MCRAE as training data.
Cross-modal mapping (Fagarasan et al. 2015)

- We want to predict properties (as in MCRAE) for unseen concepts.
- *Regression analysis* methods allow us to estimate the relationship between variables; this relationship model makes it possible to predict.
- Fagarasan et al. (2015) learn the mapping as a linear relationship between the distr. representation of a word and its featural representation.
- In other words: we want to learn to model the linear relationship between **LS** and **PS**.
- Linear regression is a linear approach to *regression analysis*.
Cross-modal mapping (Fagarasan et al. 2015)

- Linear regression: allows us to understand the relationship between two matrices $X$ and $Y$ for example (or to predict $Y$ from $X$).
- It is a relationship between a “dependent variable” (or “response variable”, e.g. $Y$) and one or more “independent variables” (or “predictors”, e.g. $X$)
- More than one predictor: multiple linear regression
- Application example: predict the taste ($Y$) of jam based on its characteristics $X$ (acidity, amount of sugar, etc.) → non-destructive way of estimating how the product will taste based on its components
- Here: we want to model the relationship between $LS\ (X)$ and $PS\ (Y)$ to get properties for unseen concepts
Cross-modal mapping (Fagarasan et al. 2015)

Given matrices $X$ and $Y$ we have the linear model:

$$ Y = X \beta + \epsilon $$

- $X$ is the known data or “predictors” (here the training data from MCRAE)
- $Y$ is the response variable
- $\beta$ is a vector of regression coefficients; this parameter vector (unknown) is what we are trying to estimate
- $\epsilon$ is an error term (or “noise”); this variable captures all other factors which influence $Y$ other than $X$
Cross-modal mapping (Fagarasan et al. 2015)

\[ Y = X\beta + \epsilon \]

How to estimate the unknown parameter vector \( \beta \)? There are several methods such as (not exhaustive):

- Ordinary Least Squares (OLS)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR), an extension of PCR
Partial Least Squares Regression (PLSR)

- PLSR first decomposes X and Y into their Principal Components (PCs) (using SVD for example) before doing the regression.
- The components have a particularity: they are relevant to X and to Y.
- The set of components (latent vectors) that we keep for regression explain the covariance between X and Y the best possible.
Steps in PLSR

1. Decompose both $X$ and $Y$:

$$X = TP^T$$

$$Y = UQ^T$$

2. Then, to get a regression model relating $Y$ to $X$, fit $\beta$ for $U = T\beta$

3. So we get $Y = UQ^T = T\beta Q^T = XP\beta Q^T$

Given any variable $x$ from $X$ (here: the training data from MCRAE we can use $P$, $Q$ and the fitted $\beta$ to compute the corresponding $y$ value.
Two different maps are learned:

- **ATTR-EMBED**: from EMBED (skip-gram model) to the property-norm semantic space
- **ATTR-SVD**: from SVD (count-based model) to the property-norm semantic space

The result is a representation including linguistic and cognitive information. Since metaphor is a cognitive phenomenon (according to CMT) expressed by means of language, such representations could be useful in the metaphor identification task.
Approach in Bulat et al. (2017)

Learn linguistic representations: EMBED, SVD

Learn cross-modal maps: ATTR-EMBED, ATTR-SVD

Create property-norm semantic space using MCRAE dataset

Compare models performances on metaphor classification task (SVM): EMBED, SVD, ATTR-EMBED, ATTR-SVD

Figure: overview of the approach in Bulat et al. (2017)
Metaphor classification using different semantic representations

- Experiments are conducted using only linguistic representations (EMBED, SVD) and using attribute-based representations (ATTR-EMBED, ATTR-SVD) on one dataset, which makes results comparable.

- Dataset (balanced): TSV-TRAIN (1768 AN-pairs) and TSV-TEST (200 AN-pairs) from web/news domain; no ambiguous instances.

- Classification is performed using SVM (supervised learning); adjective and noun vectors are normalised, then concatenated.
Both attribute-based representations outperform linguistic representations in terms of F1 score:

<table>
<thead>
<tr>
<th>Vectors</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBED</td>
<td>0.84</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>ATTR-EMBED</td>
<td>0.85</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>SVD</td>
<td>0.86</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td>ATTR-SVD</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table: System performance on Tsvetkov et al. test set (TSV-TEST) in terms of precision (P), recall (R) and F-score (F1)

Are differences in performance statistically significant?
Bulat et al. (2017), p. 526:

The best performance is achieved when using the attribute-based representation learned from the embeddings space (ATTR-EMBED), with an improvement of 4% in F1 score over EMBED.

Why? Leads to the question: why could such embeddings be more suitable than count-based models for metaphor detection (discussion)?
Bulat et al. (2017): conclusion

- **Initial hypothesis:**
  
  *In this paper we hypothesise that such attribute-based representations provide a suitable means for generalisation over the source and target domains in metaphorical language [...].*

- **Conclusion:**
  
  *Our results demonstrate that [attribute-based semantic representations] provide a suitable level of generalisation for capturing metaphorical mechanisms.*
Hypothesis as to why attribute-based representations perform better:

\[
\text{[...]} \text{attribute-based dimensions are cognitively-motivated and represent cognitively salient properties for concept distinctiveness.}
\]

(Bulat et al. 2017, p. 526).

Question: which additional features could be interesting in the metaphor detection task?
Bulat et al. (2017): criticism

Positive points:

- Using cognitively relevant features is an interesting and novel contribution in the metaphor detection task (metaphor is a cognitive phenomenon according to CMT).
- The hypothesis, goal and conclusions drawn from experiments are expressed clearly in this paper.
- The attribute-based approach is compared to two baselines (an additional random baseline would have been good too).
Bulat et al. (2017): criticism

But:

- No tests for statistical significance are reported to show that models outperforming the baselines are not a mere coincidence.

- The approach followed to create cross-modal maps between linguistic representations and property-norm semantic spaces is explained very briefly even though it is an important aspect of the work (however, references to similar or previous works are mentioned).
Bruni et al. (2012)

Distributional Semantics in Technicolor

- Comparison of models using textual, visual and both types of information on semantic relatedness tasks
- Important result: models combining visual and textual features outperform purely textual models regarding words with visual correlation (such as color terms)
Overview of all distributional semantic models implemented in this work:

- **Textual models:**
  - Two models based on counting co-occurrences with collocates within fixed windows (nearest 2 and nearest 20 content words)
  - One model based on a word-by-document matrix
  - The Distributional Memory model: it exploits lexico-syntactic and dependency relations between words

- **Visual models:** in the data, each image is tagged with one or more words; all the following models extract visual features using **BoVW**:
  - One model extracting features suited for characterizing parts of objects (**SIFT**)
  - Three models extracting color information
Bruni et al. (2012)

- Multimodal models: created by normalizing, then concatenating the two vectors from textual and visual representations (8 different models)
- Hybrid models: they represent patterns of co-occurrence of words as tags of the same images:
  - One model carrying the information about a word’s co-occurrence with other words in the image label (label = set of all words associated to the image)
  - One model carrying the information about a word’s co-occurrence with images (1 image is 1 dimension)
Bruni et al. (2012): Textual models

Same vocabulary for all textual models: 30K lemmas extracted from ukWaC and Wackypedia corpora (approx. 2B tokens together).

- Two models based on counting co-occurrences with collocates within a window of fixed width: one model considers a window of 2 words, the other model considers a window of 20.
- Another model follows a “topic-based” approach: based on a word-by-document matrix recording the distribution of all target words across the 30K documents with the largest cumulative LMI mass.
- Distributional Memory model (Baroni and Lenci, 2010): grammar based model that reaches state-of-the-art performance in numerous semantic tasks. It relies on co-occurrences and encodes morphological, structural and pattern information.
The dataset used (ESP-Game dataset) contains 100K images tagged with one or more words (or “tags”): avg. 4 words per image; 20K distinct tags. One vector with visual features is built for each tag in the dataset. **BoVW**: the bag-of-visual-words approach models images using “visual words”. In each image, relevant areas are identified and a feature vector is built for each area. For a new image, the nearest visual words are identified, such that the image can be represented by a BoVW feature vector.

Which types of features are extracted from the images?

Scale-Invariant Feature Transform (SIFT) vectors, which are suited to characterize parts of objects, and LAB features, which only encode color information.
In multimodal models, textual and visual models are combined by normalizing and concatenating feature vectors. The combination is performed using a linear weighted combination function in which a weighting parameter is tuned on a separated development set. Tuning results in an optimal weight parameter that gives equal importance to visual and textual features.
Two hybrid models that carry patterns of co-occurrence of words as tags of the same images are built.
Two types of experiments are conducted:

- Experiments for general semantic models: all models are evaluated in terms of their Spearman correlation to human ratings on two distinct datasets.
- Experiments for models of the meaning of color terms:
  - one evaluating each model’s ability to associate nouns denoting concrete things (crow, wood, grass) with colors (black, brown, green)
  - one evaluating each model’s ability to distinguish between literal and nonliteral language.
Bruni et al. (2012): Conclusion / discussion

- Main conclusion: for words where vision is relevant, multimodal models often outperform purely textual models. In the particular task of distinguishing literal from nonliteral language, multimodal models also perform significantly better.

- Textual models in this work (from 2012) are count-based (co-occurrences). What other types of semantic representations of words could be used as well?
Final remarks

- We discussed two papers showing that information that is not purely textual can be relevant and even improve models for semantic tasks such as metaphor identification.
  - Bulat et al. (2017) shows that cognitive information (attributes of concepts) are relevant.
  - Bruni et al. (2012) shows that visual information helps to improve textual information.


