University of Copenhagen
Faculty of Science

Limitations of cross-lingual learning from image search

Mareike Hartmann and Anders Søgaard
Department of Computer Science

Lexicon Induction From Web Search Images

- Bergsma and van Durme (2011) and Kiela et al. (2015) induce lexica for nouns from web search images:

  Get images associated with a word from a web search engine by using the word as a query

  Find a translation for the word by ranking candidates based on similarities between the images associated with the words

- Can the approach be generalized to **verbs** and **adjectives**?

Data

- 3 wordlists of English words (Simlex, MEN, Bergsma) translated into 5 languages:
  German, French, Russian, Italian, Spanish
- 1406 nouns, 206 verbs, 159 adjectives
- For each query word, 50 images returned by Google Search engine

Similarity Computation

- Image feature representations (4069-dimensional) extracted from a convolutional neural network pretrained on ImageNet
- Word translations are ranked according to:
  - Cosine similarity between pairs of individual images
  - Cosine similarity between aggregated representations of image sets
  - Highest number of nearest neighbors

Analysis

- **Image dispersion** is the average cosine similarity between all image pairs in an image set. Adjectives and verbs have higher dispersion values than nouns: nouns: d = 0.60, adjectives: d = 0.66, verbs: d = 0.68

animal (NN)
d = 0.78

differ (VB)
d = 0.76

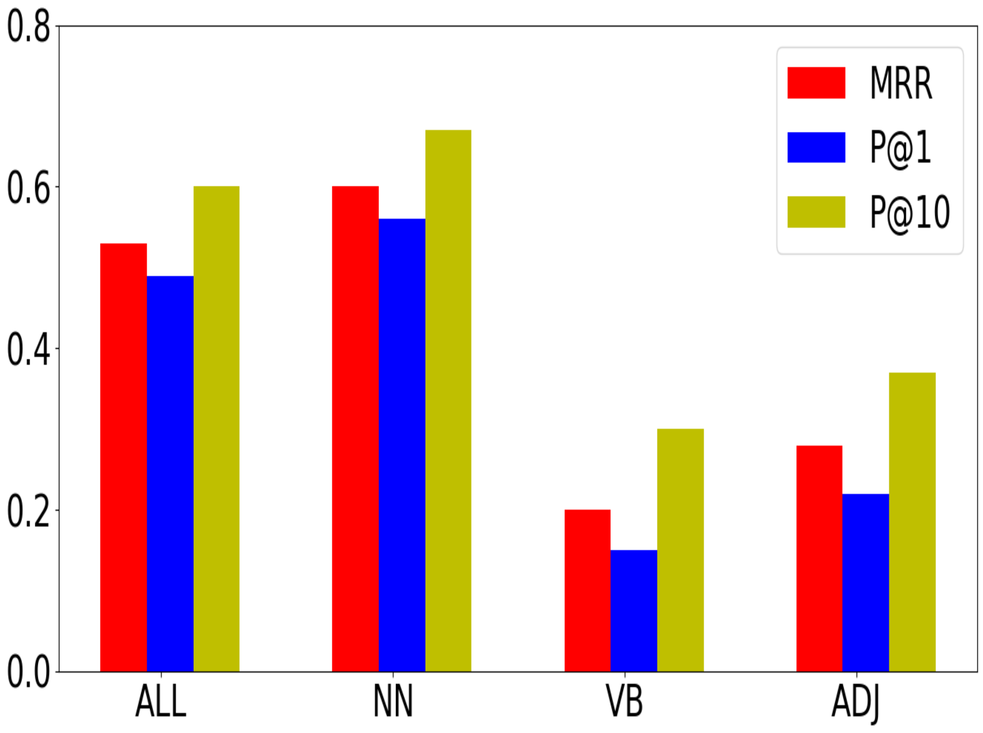
huge(ADJ)
d = 0.79

- Adjectives and verbs have a **higher number of wordsenses** according to WordNet than nouns: nouns: n = 5.08, adjectives: n = 6.88, verbs: n = 9.18
- In many cases, the search engine does not capture the **intended POS** of the query word:

vanish (VB)
d = 0.43

Rank the German candidate translations of the English word **mug** based on similarities between images:

Results



- Results for the best performing ranking method averaged over all language pairs

Conclusion

- The approach does not scale to adjectives and verbs
- One explanation is that adjectives and verbs are more difficult to visualize in an iconic way (higher dispersion values).
- Ideas for improvements:

  Use words in context as queries or collect images based on natural language captions rather than isolated semantic tags

  Train the feature extractor on a different resource than ImageNet categories with its concrete object categories