Searching for the X-Factor: Exploring Corpus Subjectivity for Word Embeddings

Maksim Tkachenko and Chong Cher Chia and Hady W. Lauw

Singapore Management University
Word Embeddings

• Dense vectors of words
• Unsupervised training: GloVe, Word2Vec
• Words in similar context tend to have similar meaning

\[
good \rightarrow (\ldots 0.0335, -0.1018, 0.2300, \ldots) \in \mathbb{R}^{300}
\]

• Words with similar meanings tend to be close in embedding space
Training Word Embeddings

Wikipedia

The Free Encyclopedia

Target Word

This camera is good for high quality ...

Context Words

Counting Contexts

good → (... 8321, 235, 63444, ...) ∈ \( R^{\text{Vocabulary Size} (\approx 300k)} \)

Reducing Dimensionality

good → (... 0.0335, −0.1018, 0.2300, ...) ∈ \( R^{300} \)
Different Input Corpora

WIKIPEDIA
The Free Encyclopedia

Counting Contexts

good $\rightarrow (..., ?, ?, ?, ...) \in \mathbb{R}^\text{Vocabulary Size ($\approx 300k$)}$

Reducing Dimensionality

good $\rightarrow (..., ?, ?, ?, ...) \in \mathbb{R}^{300}$
An article must be written from a neutral point of view, which among other things means “representing fairly, proportionately, and, as far as possible, without editorial bias, all of the significant views that have been published by reliable sources on a topic.”
“Amazon values diverse opinions” and that “content [customer reviews] you submit should be relevant and based on your own honest opinions and experience.”
Subjectivity Scale

More Objective

Objective Embeddings (OE)

More Subjective

Subjective Embeddings (SE)

WIKIPEDIA
The Free Encyclopedia

amazon
Binary Classification Tasks

• Sentiment Classification (**positive** vs. **negative**):
  • Amazon Reviews (24 categories) + Rotten Tomatoes Reviews
  
  “A very funny movie” vs. “One lousy movie”

• Subjectivity Classification (**subjective** vs. **objective**)
  • Rotten Tomatoes Reviews

  “The story needs more dramatic meat” vs. “She's an artist”

• Topic Classification (**in-topic** vs. **out-of-topic**)
  • Newsgroups Dataset (6 categories)
Methodology

• Cross-validation on balanced samples

• Binary logistic regression classifier

• Sentence embedding = average of word embeddings

• The same number of sentences and the same vocabulary when training embeddings
Empirical Findings

SE and OE are very similar on “objective” tasks.

SE understand sentiment words better than OE?
Top Words Similar to “good”

### Objective Embeddings

| Word | Similarity |
|------|------------|
| bad  | 0.68       |
| decent | 0.67      |
| nice  | 0.62       |
| poor  | 0.61       |
| ...   | ...        |

### Subjective Embeddings

| Word    | Similarity |
|---------|------------|
| decent  | 0.78       |
| great   | 0.76       |
| nice    | 0.69       |
| terrific| 0.64       |
| ...     | ...        |
Sentiment Words Still Cause Troubles!

![Subjective Embeddings](amazon)

| Word A | Word B | Their Similarity |
|--------|--------|------------------|
| ...    | ...    | ...              |
| waste  | Save   | 0.51             |
| love   | hate   | 0.60             |
| loves  | hates  | 0.68             |
| easy   | difficult | 0.56          |
| ...    | ...    | ...              |
SentiVec Embeddings

| Similar to “good” | Similarity |
|------------------|------------|
| bad              | 0.68       |
| decent           | 0.67       |
| nice             | 0.62       |
| poor             | 0.61       |
| ...              | ...        |

Objective Word2Vec Embeddings

| Similar to “good” | Similarity |
|------------------|------------|
| decent           | 0.79       |
| nice             | 0.76       |
| perfect          | 0.75       |
| excellent        | 0.73       |
| ...              | ...        |

Objective SentiVec Embeddings
SentiVec: Infusing Sentiment

\[ \text{SentiVec} = \text{Word2Vec} + \text{Lexical Resource} \]

- Predicts context words as in Word2Vec Skip-gram
- Predicts word category

**Lexical Resource**

- **Negative**: waste, junk, horrible, defective, ...
- **Positive**: love, great, recommend, easy, ...
Logistic SentiVec

This camera is good for high quality ...

Word2Vec Skip-gram objective

(good, camera)
(good, is)
(good, for)
(good, high)

vs.

Random Noise
(good, frog)
(good, duck)
...

Lexical objective of SentiVec (two classes)

\[
P(\text{good is POSITIVE}) = \sigma \left( \overline{\text{good}} \cdot \phi \right)
\]

\[
P(\text{good is NEGATIVE}) = 1 - P(\text{good is POSITIVE})
\]

\[
P(\text{good is POSITIVE}) \rightarrow \text{MAXIMIZE}
\]
Spherical SentiVec

Positive Words $\phi_{POSITIVE}$

Neutral Words $\phi_{NEUTRAL}$

Negative Words $\phi_{NEGATIVE}$
Empirical Findings

- **Objective Embeddings**
  - Amazon Sentiment (average over 24 categories): 0.8%
  - Rotten Tomatoes Sentiment: 0.3%

- **Subjective Embeddings**
  - Amazon Sentiment (average over 24 categories): 0.2%
  - Rotten Tomatoes Sentiment: 0%

SentiVec does not affect “objective” classification tasks.
Changes in Similarity

Target Word: Good

Target Word: Bad
Conclusion

• Explored effects of corpus subjectivity for word embeddings

• SentiVec, a method for infusing lexical information into word embeddings

• Sentiment-infused SentiVec embeddings space facilitate better sentiment-related similarity

Pre-trained Word Embeddings & Code: https://sentivec.preferred.ai/