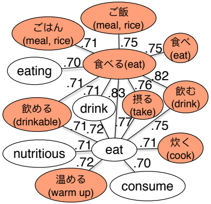
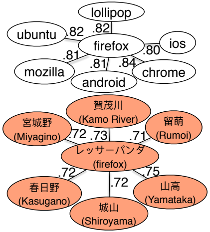
A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings based on Graph Modularity

Yoshinari Fujinuma[1] Jordan Boyd-Graber[2] Michael J. Paul[1]

[1]University of Colorado [2]University of Maryland





Motivation

- You encounter a disaster in Ethiopia
- Want a document classifier, but no labeled data in Amharic
- One solution: Exploit labeled data in English

Cross-Lingual Word Embedding

slow: -0.21, 0.35, ...
ቀርፋፋ (slow): -0.32, 0.45, ...
...

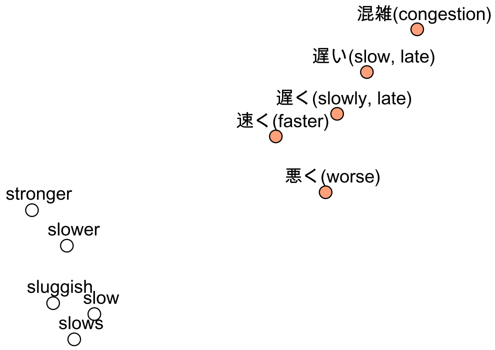
- How do you evaluate when labeled data is not available in Amharic?

Outline

- Motivation
- Limitations: Clustering by language
- Graph modularity
- Correlations of graph modularity and downstream tasks
- Comparing to other metrics
- Conclusion

Limitations of Cross-Lingual Word Embeddings

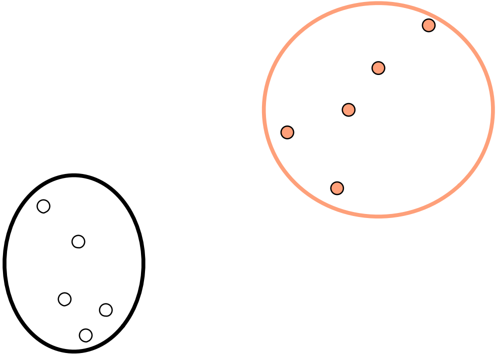
- Words within a language tend to be closer than words from another language
- We call this "Clustering by language"
- Discourages the transfer of knowledge from one language to another



t-SNE projection of an EN-JA embedding

Limitations of Cross-Lingual Word Embeddings

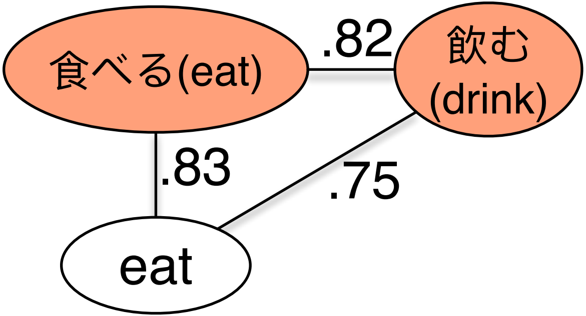
- Words within a language tend to be closer than words from another language
- We call this "Clustering by language"
- Discourages the transfer of knowledge from one language to another



t-SNE projection of an EN-JA embedding

Quantifying Clustering by Language Using a Lexical Graph

- Main Idea: Use “Clustering by Language” to evaluate embeddings
- Convert cross-lingual word embeddings into cross-lingual lexical graphs
- k-nearest neighbor graph
  - Nodes: Words
  - Edges: Cosine similarity between words



Quantifying Clustering by Language Using a Lexical Graph

- Distinguish good vs. bad embeddings by looking at the structure

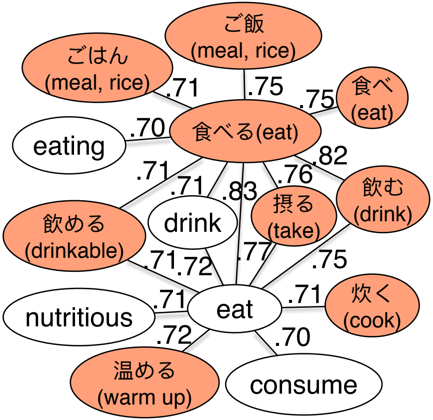


Figure 1: Good

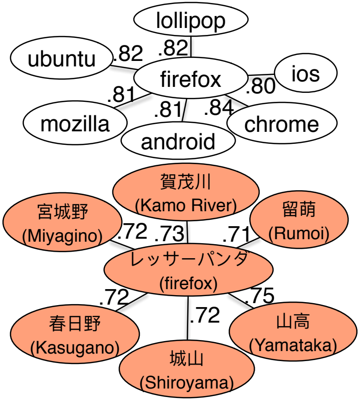


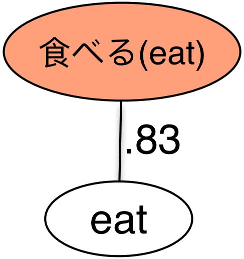
Figure 2: Bad

Outline

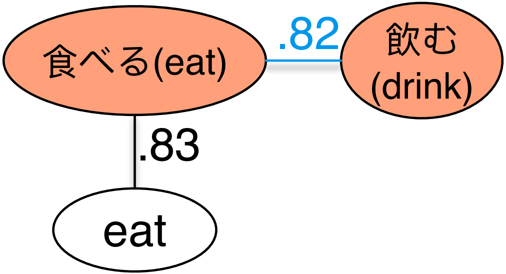
- Motivation
- Limitations: Clustering by language
- Graph modularity
- Correlations of graph modularity and downstream tasks
- Comparing to other metrics
- Conclusion

Defining Graph Modularity (Newman, 2003)

- Focus on edges that are connected to the **same language**
- Modularity = "actual intra-lingual edges" - "expected intral-lingual edges"



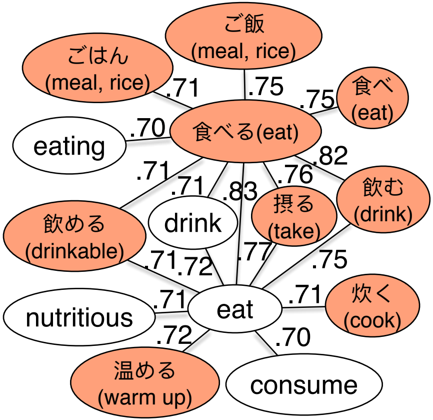
$<$



$$-\left(\frac{0.83}{2}\right)^2 \times 2 \approx -0.34$$

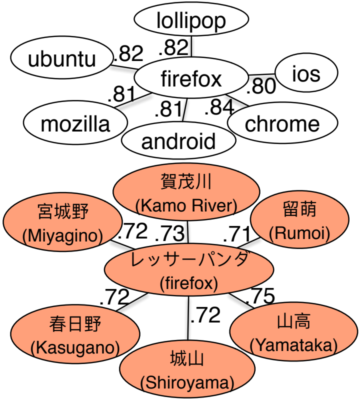
$$-\left(\frac{0.83}{4}\right)^2 + \frac{0.82}{4} - \left(\frac{2.47}{4}\right)^2 \approx -0.14$$

Defining Graph Modularity



Modularity: 0.143

<



Modularity: 0.672

Outline

- Motivation
- Limitations: Clustering by language
- Graph modularity
- Correlations of graph modularity and downstream tasks
- Comparing to other metrics
- Conclusion

Experiment Setup

- Cross-Lingual Embedding Methods
  - Supervised
    - Mean Squared Error (Mikolov et al., 2013, MSE)
    - MSE+Orthogonal Constraint (Xing et al., 2015)
    - Canonical Correlation (Faruqui and Dyer, 2014, CCA)
  - Unsupervised
    - Vecmap (Artetxe et al., 2018)
    - MUSE (Conneau et al., 2018)

Experiment Setup

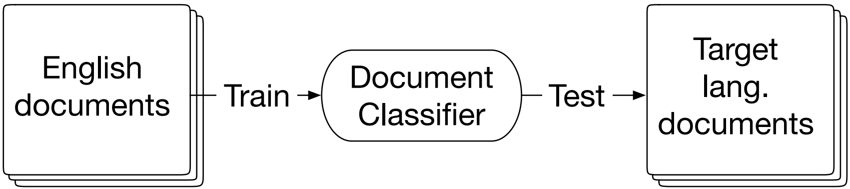
- Cross-Lingual Embedding Methods
  - Supervised
    - Mean Squared Error (Mikolov et al., 2013, MSE)
    - MSE+Orthogonal Constraint (Xing et al., 2015)
    - Canonical Correlation (Faruqui and Dyer, 2014, CCA)
  - Unsupervised
    - Vecmap (Artetxe et al., 2018)
    - MUSE (Conneau et al., 2018)

Experiment Setup

- Source Language
  - English
- Target Languages
  - Spanish
  - Italian
  - Danish
  - Japanese
  - Hungarian
  - Amharic

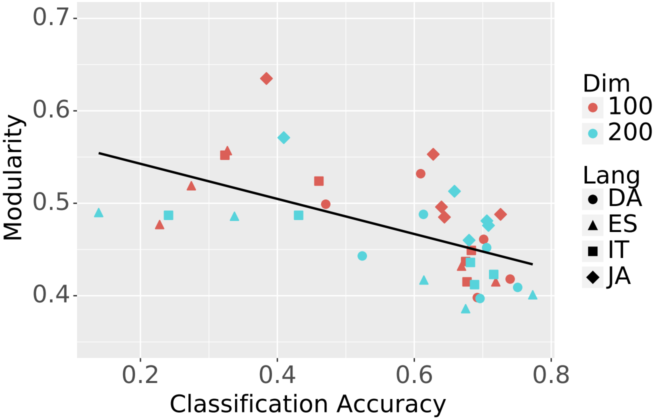
See results in the paper

Task 1: Cross-Lingual Document Classification



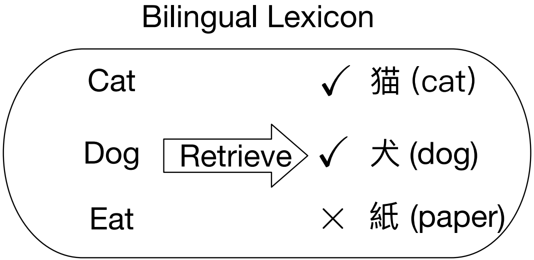
- Classification task of four topics
- Dataset: Reuters RCV1, RCV2 corpora (Lewis et al., 2004)

Task 1: Cross-Lingual Document Classification



- Spearman's Correlation $= -0.665$

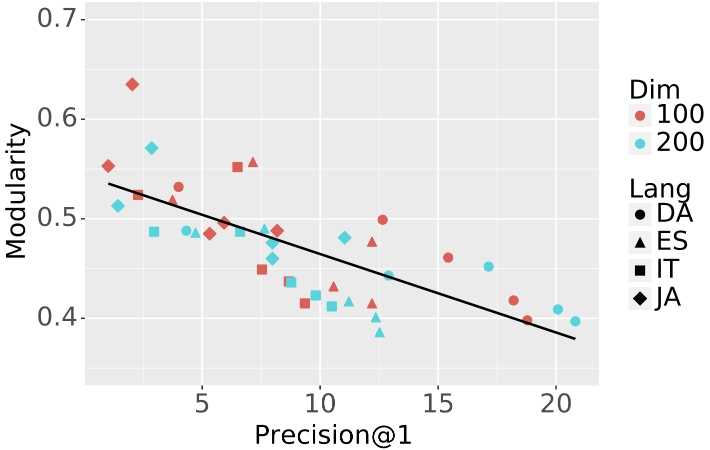
Task 2: Bilingual Lexical Induction



Precision@1 $= 0.67$

- Translate words from a source language to a target language
- Dataset: MUSE test set

Task 2: Bilingual Lexical Induction



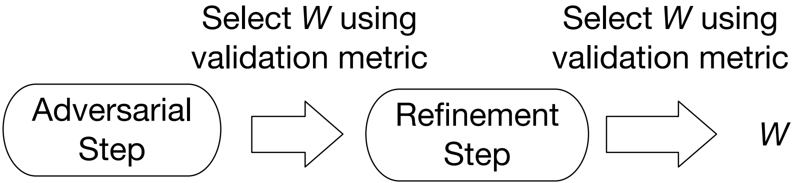
- Spearman's correlation to graph modularity $= -0.789$

Outline

- Motivation
- Limitations: Clustering by language
- Graph modularity
- Correlations of graph modularity and downstream tasks
- Comparing to other metrics
- Conclusion

Metric Used in MUSE (Conneau et al., 2018)

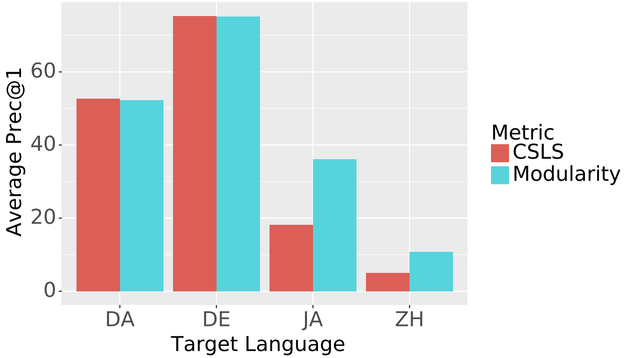
- MUSE trains a cross-lingual mapping matrix W without any bilingual lexicon



- Default validation metric is cross-lingual similarity local scaling (CSLS) (Conneau et al., 2018)

CSLS vs. Modularity for MUSE

- Replace CSLS with modularity and compare them
- Modularity makes MUSE stable on distant language pairs
- MUSE(+CSLS) is unstable on distant language pairs (Søgaard et al., 2018)

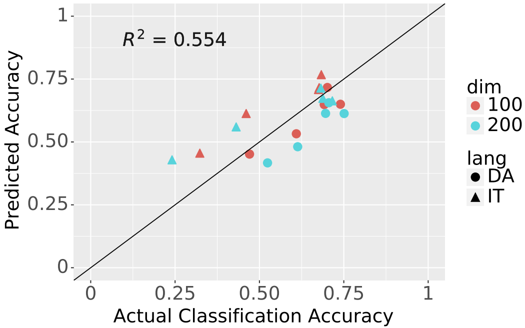


Comparing to Other Metrics

- A good metric captures information not captured by other metrics
- Predict classification accuracy by linear regression using
  - Resource-free metrics
    - Modularity
    - CSLS (Conneau et al., 2018)
  - Resource-dependent metrics
    - QVEC-CCA (Ammar et al., 2016)
    - Average cosine similarity between translations
- Ablation study by omitting each metric

Comparing to Other Metrics

- Using all metrics: $R^2 = 0.814$



~~Modularity~~
QVEC-CCA
Cosine similarity
CSLS
$R^2 :\downarrow 0.260$

Comparing to Other Metrics

- Using all metrics: $R^2 = 0.814$

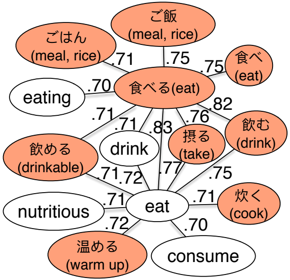
Modularity	Modularity	Modularity	~~Modularity~~
QVEC-CCA	QVEC-CCA	~~QVEC-CCA~~	QVEC-CCA
Cosine similarity	~~Cosine similarity~~	Cosine similarity	Cosine similarity
~~CSLS~~	CSLS	CSLS	CSLS
$R^2 :\downarrow 0.023$	$R^2 :\downarrow 0.044$	$R^2 :\downarrow 0.111$	$R^2 :\downarrow 0.260$

- Modularity is a good metric and captures information not captured by other metrics

Outline

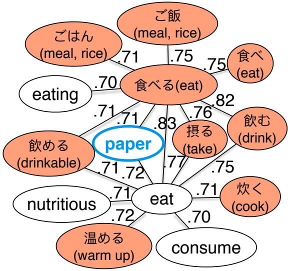
Conclusion & Summary

- Graph modularity is a good & cheap evaluation measure for cross-lingual embeddings
  - Correlated to downstream tasks
  - Successful as a validation metric (for MUSE)
- But combine with other metrics if possible
  - Modularity looks at only the structure, not the meanings.



Modularity: 0.143

=



Modularity: 0.143

Q & A

- Questions?

References

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. *Computing Research Repository*, arXiv:1602.01925. version 2.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *Computing Research Repository*, arXiv:1309.4168. version 1.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67(2).

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the Association for Computational Linguistics*.

Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.