Introduction to Geometry of Cross-Lingual Embeddings

Yoshinari Fujinuma

Agenda

- Understand the importance of orthogonal constraints used to train cross-lingual embeddings

Basics of Cross-Lingual Embeddings

- Shared embedding space across multiple languages
- Assumption: **Geometric relationship of the word vectors are similar across languages**
- Popular methods learn a **linear projection matrix** to map whole embedding space into another
- Pros: Leverage training data from another language

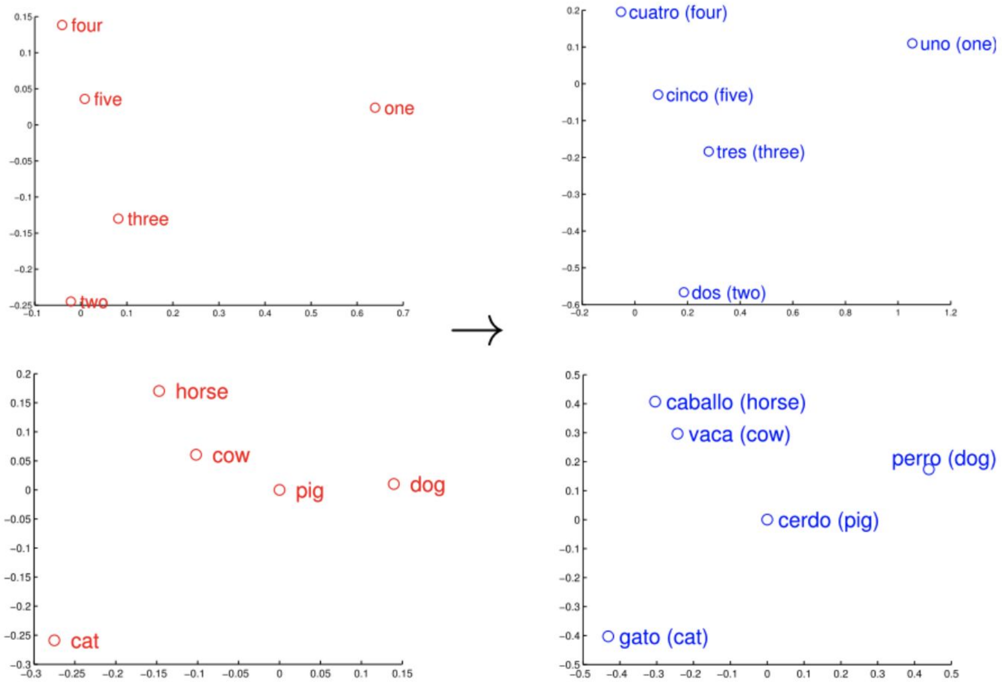


Image from [Mikolov+ 2013]

Survey of Cross-lingual embeddings [Ruder+ 2017]

1. **Mapping-based approach [Mikolov+ 2013, etc.]**
   a. **Understand why orthogonal constraints are important [Xing+ 2015]**
   b. **Unsupervised cross-lingual embeddings [Conneau+ 2017, etc.]**
   c. **On the Limitations of Unsupervised Bilingual Dictionary Induction [Søgaard+ 2018]**
2. Psuedo-parallel corpus approach (i.e., Code-switching approach)
   a. Replace words in a monolingual corpus and make a psuedo-code switched corpus
3. Joint training approach

Understand why orthogonal constraint is important for cross-lingual embeddings

- Geometric interpretation of
    a. dot product
    b. skip-gram with negative sampling models [Mimno+ 2017]
- Length normalization of word vectors
- Orthogonal constraints for mapping two monolingual embeddings [Xing+ 2015]
- Cross-lingual embeddings using mean squared error [Mikolov+ 2013] and orthogonal constraints

Geometric interpretation of Dot Product

- When dot product ($u \cdot v = ||u||\ ||v|| \cos \theta$) is
  - **Negative**: Vectors point **opposite** direction
  - **Positive**: Vectors point the **same** direction

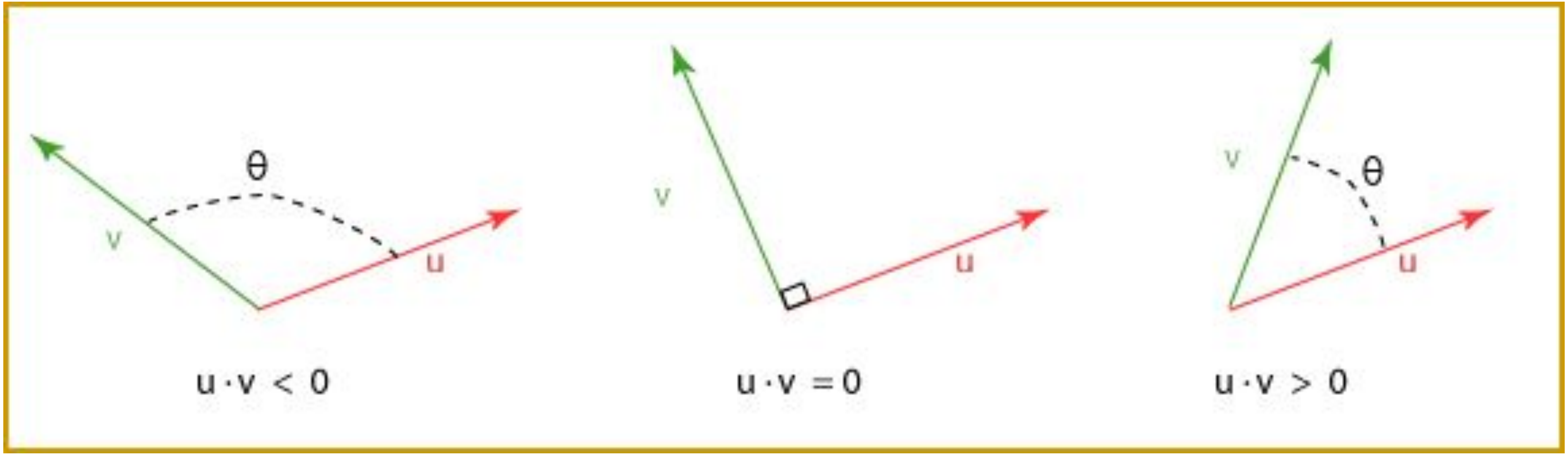


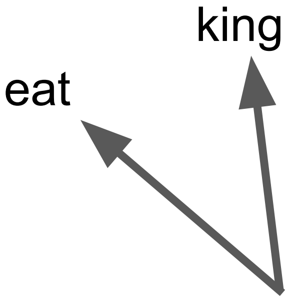
Image from https://www.quora.com/Can-a-scalar-product-be-negative

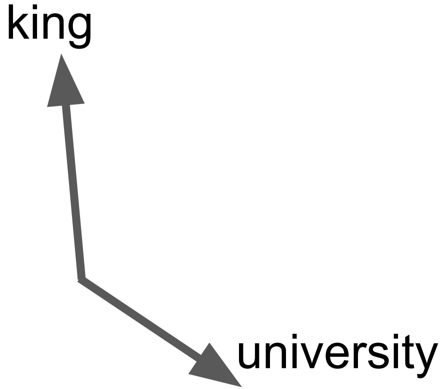
Geometric interpretation of Skip-gram with negative sampling models

- Word vector w_i
- Context vector c_j
- Negative context vector c_s
- “The king likes to eat cakes” -> (w_i, c_j) = (“king”, “eat”)
- E.g., (w_i, c_s) = (“king”, “university”)

$$l = \log(\sigma(w_i^T c_j)) + \sum_s^S (\log(\sigma(-w_i^T c_s)))$$

Geometric interpretation of Skip-gram with negative sampling models





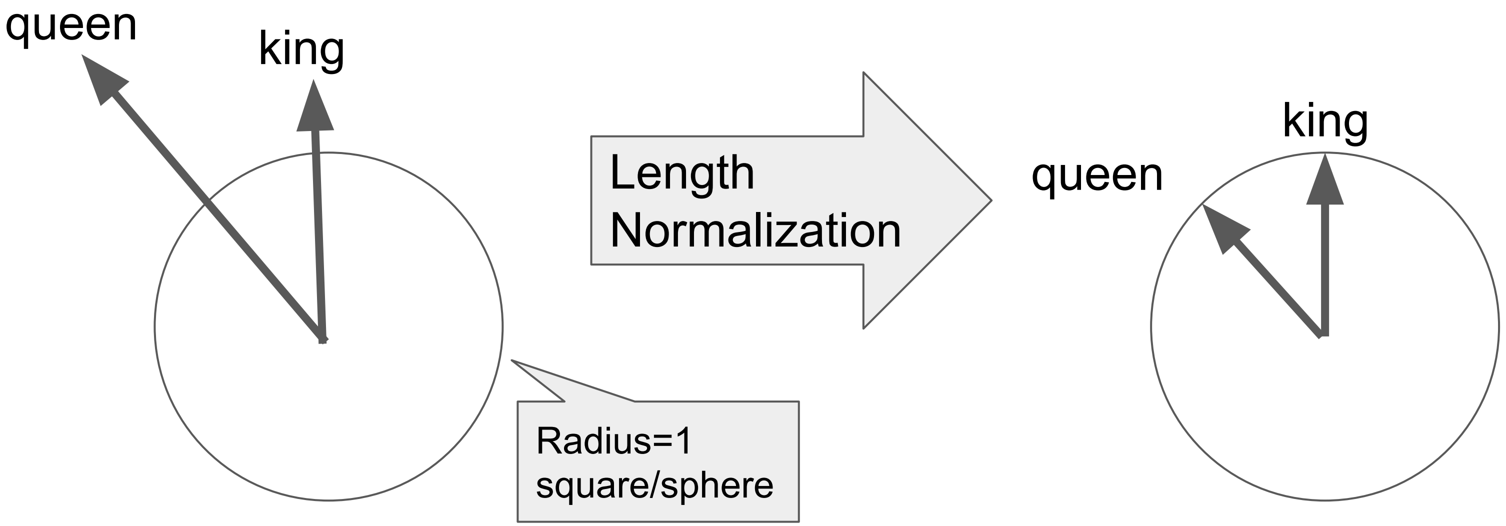
Making w_i and c_j point the **same** direction

Making w_i and negative context vector c_s point the **opposite** direction

$$l = \boxed{\log(\sigma(w_i^T c_j))} + \boxed{\sum_s^S (\log(\sigma(-w_i^T c_s))}\,)$$

Length normalization of vectors

- Make the length of the vector being ||u|| = 1
- Dot product becomes equivalent to cosine similarity
  - $u \cdot v = ||u||\ ||v|| \cos\theta = \cos\theta$



Before and after the length normalization

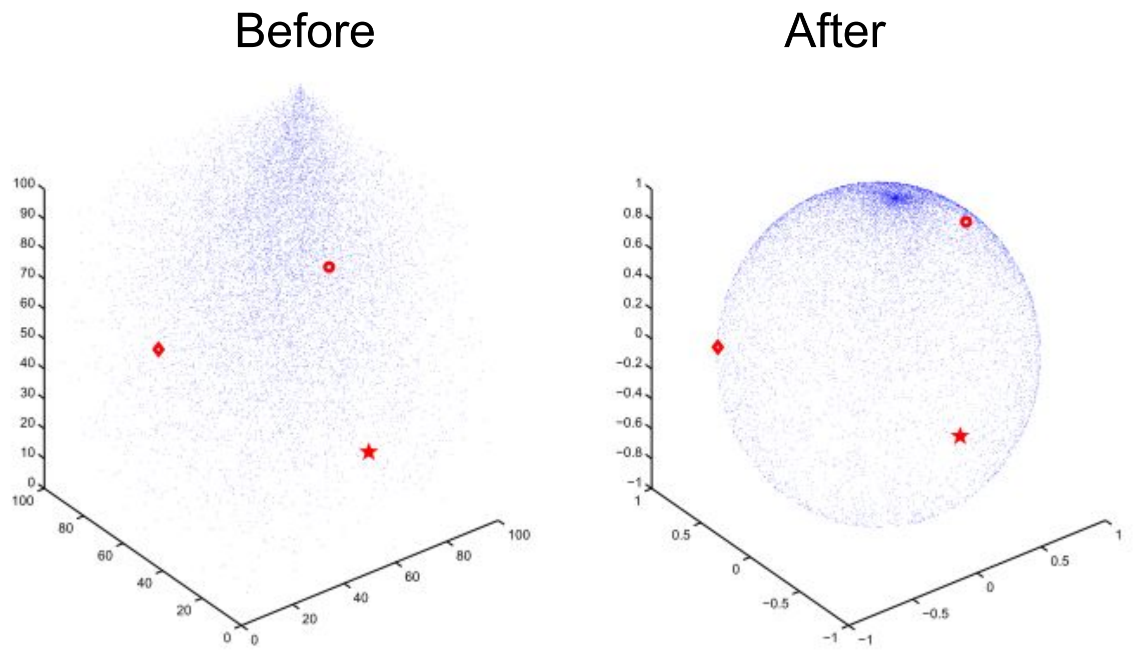
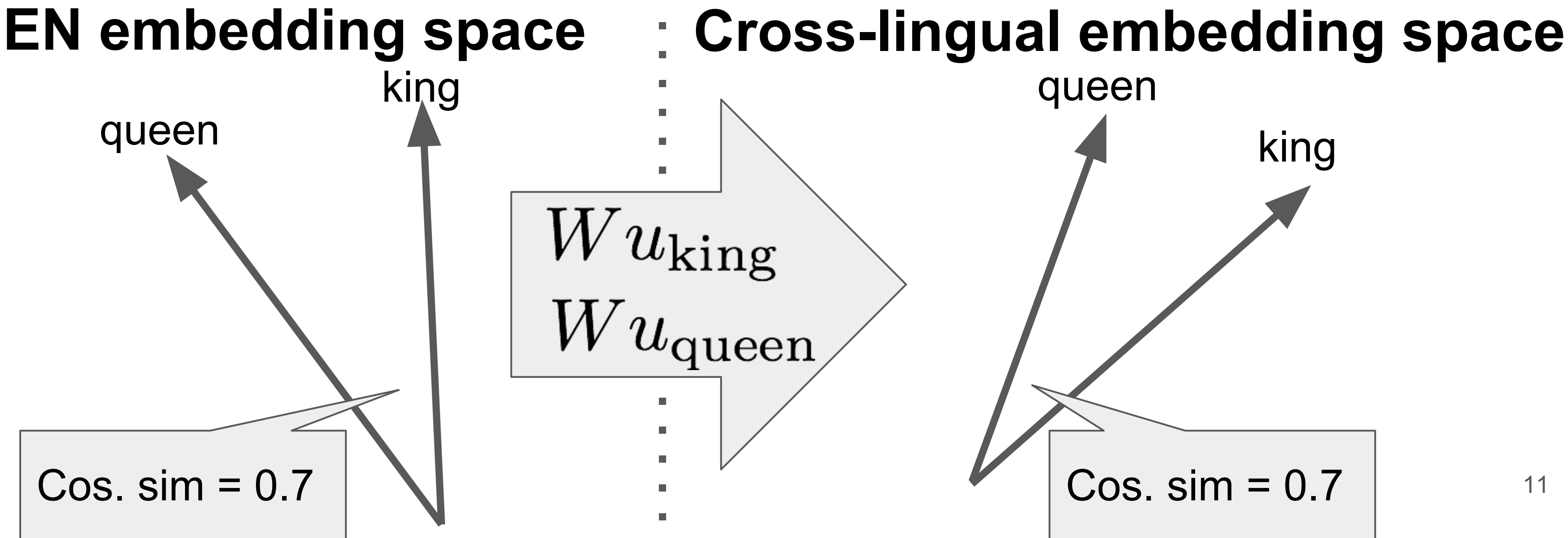


Image from [Xing+ 2015]

Intuition of orthogonal projection $W^T W = I$

- Preserves the dot product of any two vectors after mapped to the shared cross-lingual embedding space

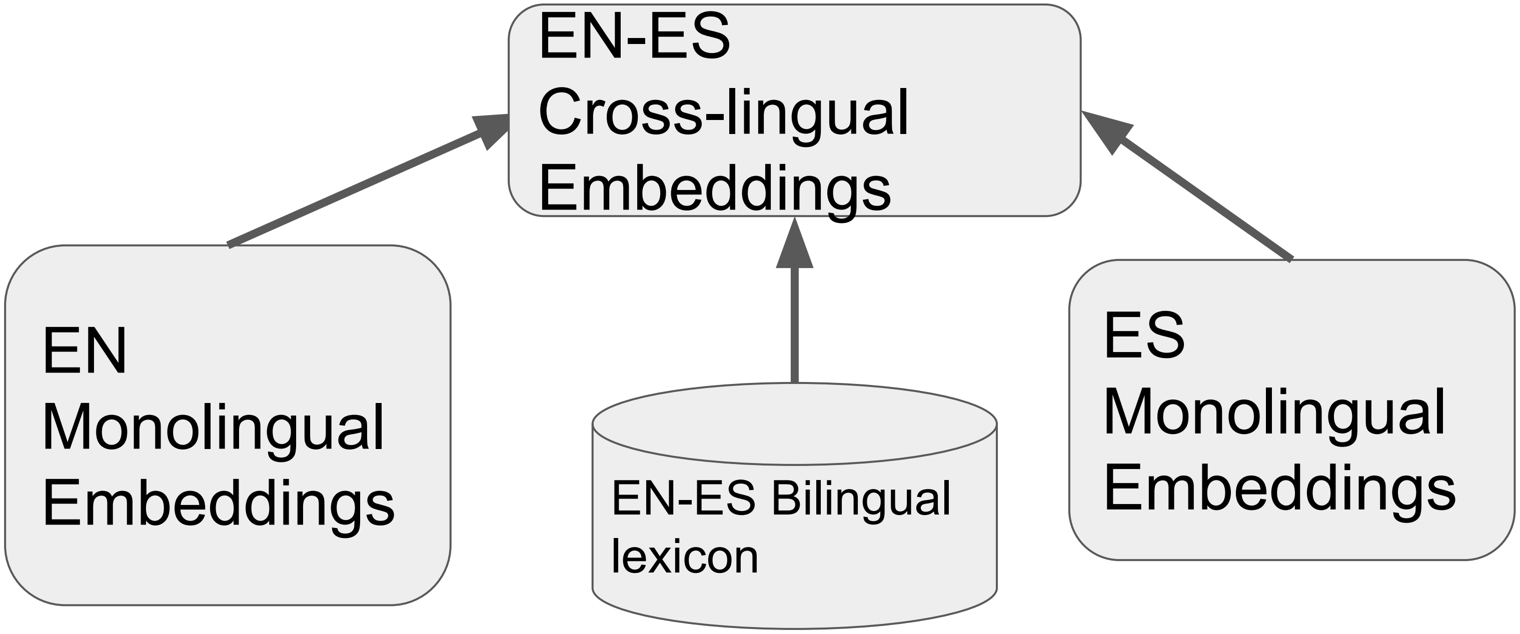


Intuition of orthogonal projection $W^T W = I$

- Preserves the dot product of any two vectors after mapped to the shared cross-lingual embedding space

$$(Wu)^T (Wv) = u^T W^T W v = u^T v$$

Cross-lingual Embeddings at High-Level



Objective Function and Orthogonal Constraint

- Mean squared error [Mikolov+ 2013] with orthogonal constraints [Xing+ 2015, etc]
- X: English word vectors in a bilingual lexicon
- Z: Target language (e.g., Spanish) word vectors in a bilingual lexicon
- W: Projection matrix from EN to target lang (or vice versa)

$$\arg\min_{W} \|XW - Z\|_F^2$$

$$W^T W = I$$

Minimize the mean squared error of the vectors we want to align:
E.g.,
- X = (u_“king”, u_“queen”)
- Z = (v_“el rey”, v_“la reina”)

Objective Function and Orthogonal Constraint

- Mean squared error [Mikolov+ 2013] with orthogonal constraints [Xing+ 2015, etc]
- X: English word vectors in a bilingual lexicon
- Z: Target language (e.g., Spanish) word vectors in a bilingual lexicon
- W: Projection matrix from EN to target lang (or vice versa)

$$\arg\min_{W} \|XW - Z\|_F^2$$

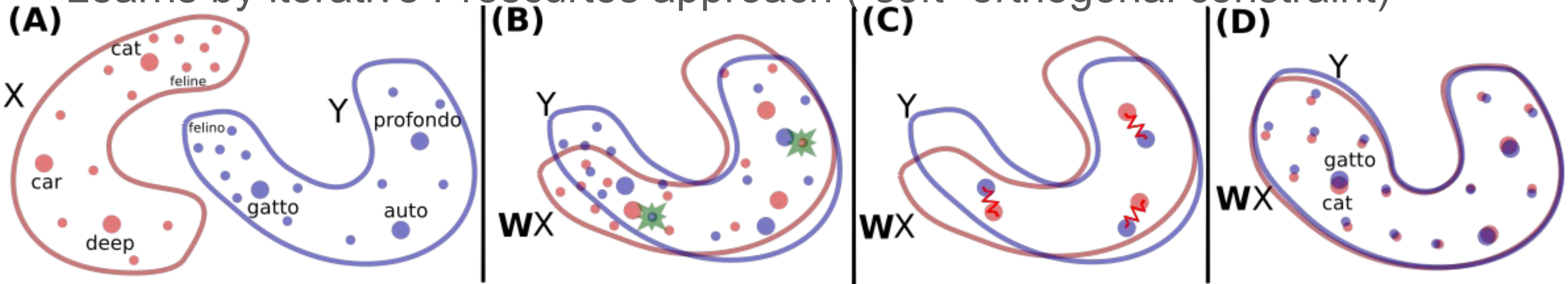
$$W^T W = I$$

Pros of Orthogonal constraint

1. Preserves the dot product in the original embedding space
2. Avoids overfitting *W* to the translation pairs in the bilingual lexicon
3. Has closed form solution using SVD (Proscurtes problem)

Unsupervised Cross-lingual Embedding [Conneau+ 2018, etc.]

- Input: Two monolingual embeddings
  - Does not use any form of bilingual resources (e.g., parallel corpus, bilingual lexicon)
- Used in the following papers:
  - Two "unsupervised machine translation" papers [Lample+ 2018a, Artexte+ 2018a]
  - More recent version of those [Lample+ 2018b, Artexte+ 2018b]
- Learns by iterative Proscurtes approach ("soft" orthogonal constraint)



What is not covered in this talk

1. CCA-based approach [Faraqui+ 2014]
2. Non-linear approach [Lu+, 2015]
3. Unsupervised Machine Translation [Lample+ 2018a, Artexte+ 2018a, etc.]
4. Hubness problem [Dinu+, 2015] and its solution discussed in [Conneau+ 2018]
5. Few recent papers on cross-lingual embedding