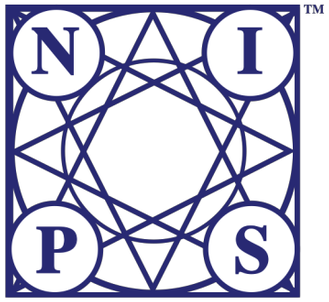


Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

Yu-An Chung Wei-Hung Weng Schrasing Tong James Glass

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Cambridge, Massachusetts, USA

NeurIPS

Montréal, Québec, Canada

December 2018

Machine Translation (MT)

Training data

“the cat is black” → MT system → “le chat est noir”

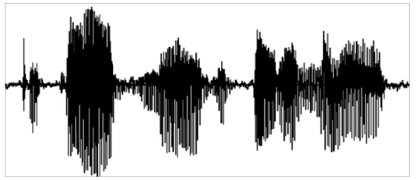
(English text , French translation)

Automatic Speech Recognition (ASR)

 → ASR system → “dogs are cute”

(English audio , English transcription)

Text-to-Speech Synthesis (TTS)

“cats are adorable” → TTS system → 

(English text , English audio)

Machine Translation (MT)

Training data

"the cat is black" → MT system → "le chat est noir"

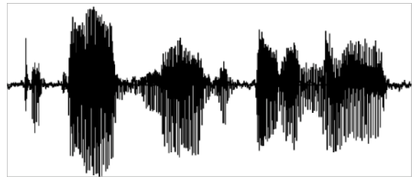
(English text , French translation)

Automatic Speech Recognition (ASR)

 → ASR system → "dogs are cute"

(English audio , English transcription)

Text-to-Speech Synthesis (TTS)

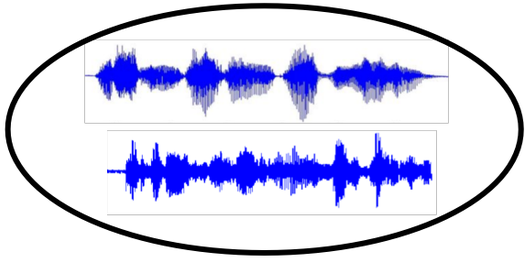
"cats are adorable" → TTS system → 

(English text , English audio)

Parallel corpora for training → Expensive to collect!

Framework

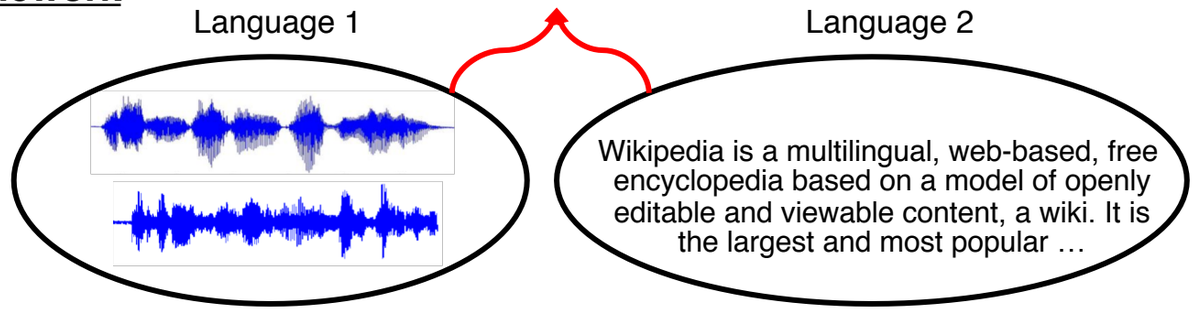
Language 1



Language 2

Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …

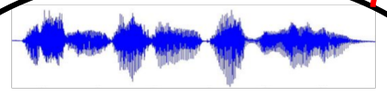
Framework

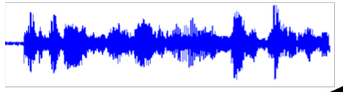

Do not need to be parallel!
Language 1
Language 2
Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …

Framework

Do not need to be parallel!

Language 1

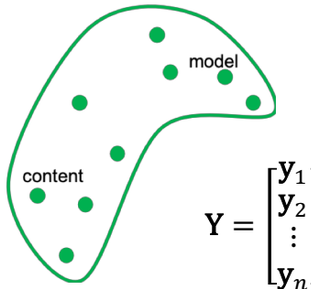




Language 2

Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …

Word2vec
[Mikolov et al., 2013]

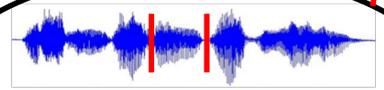


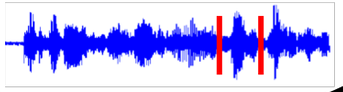
$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}^T \in \mathbb{R}^{d \times n}$$

Framework

Do not need to be parallel!

Language 1



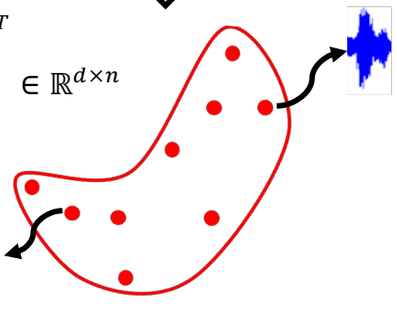


Language 2

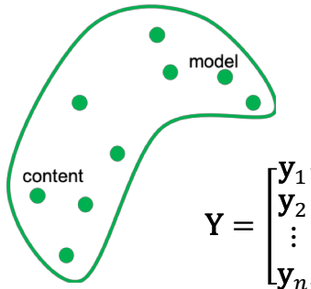
Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …

Speech2vec
[Chung & Glass, 2018]

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}^T \in \mathbb{R}^{d \times n}$$

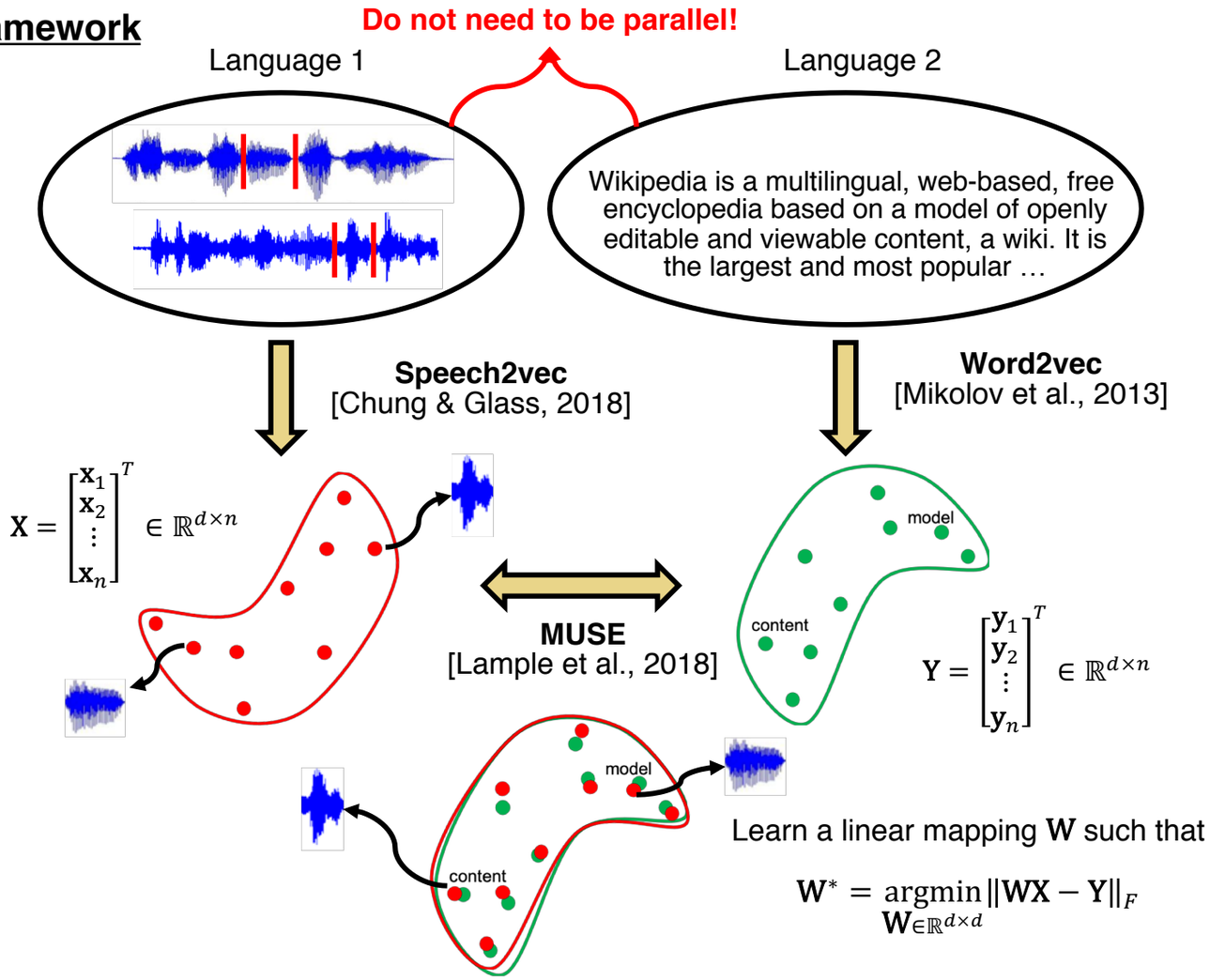


Word2vec
[Mikolov et al., 2013]

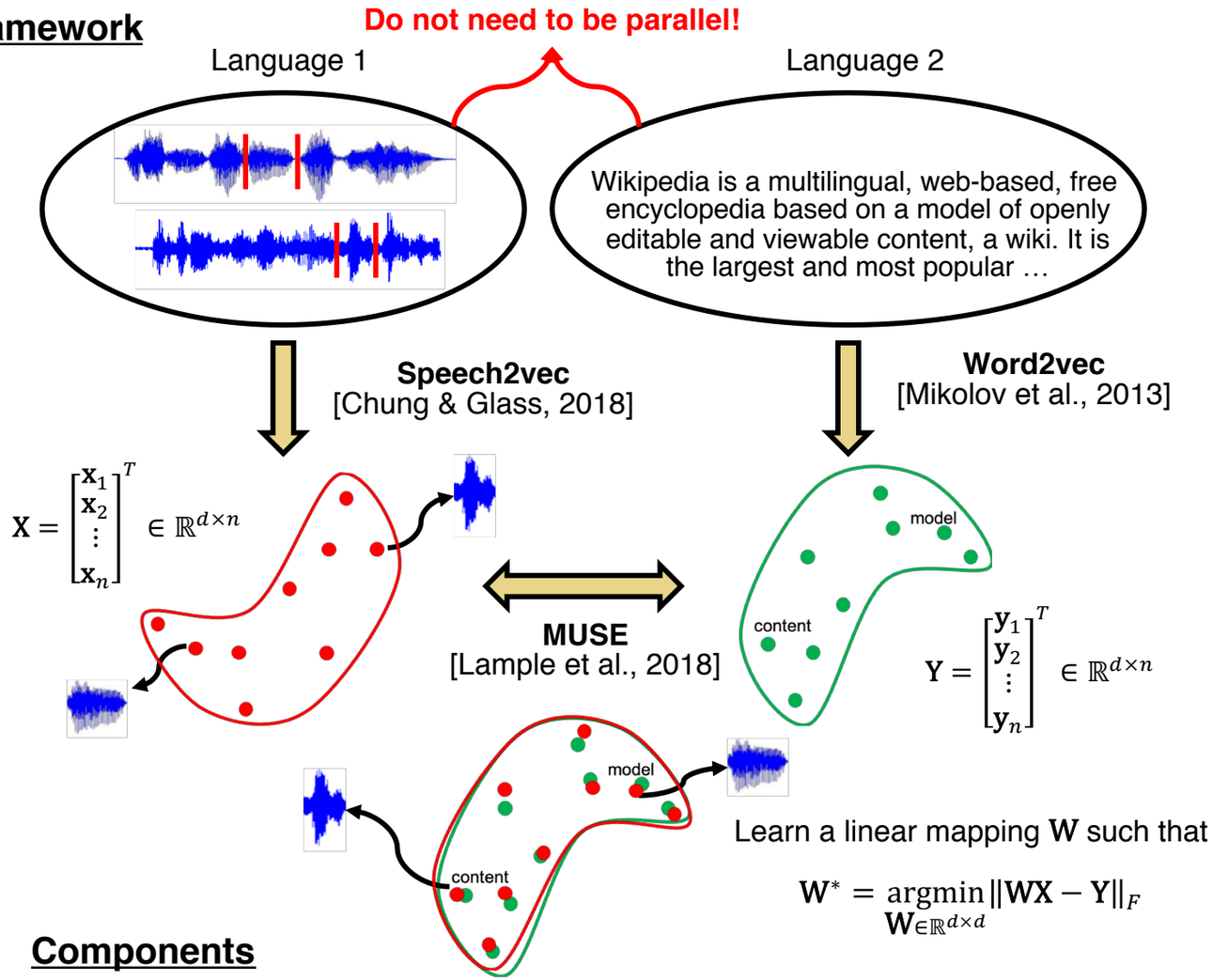


$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}^T \in \mathbb{R}^{d \times n}$$

Framework


Do not need to be parallel!
Language 1
Language 2
Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …
Speech2vec
[Chung & Glass, 2018]
Word2vec
[Mikolov et al., 2013]
$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}^T \in \mathbb{R}^{d \times n}$
MUSE
[Lample et al., 2018]
model
content
$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}^T \in \mathbb{R}^{d \times n}$
model
content
Learn a linear mapping $\mathbf{W}$ such that
$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|\mathbf{WX} - \mathbf{Y}\|_F$

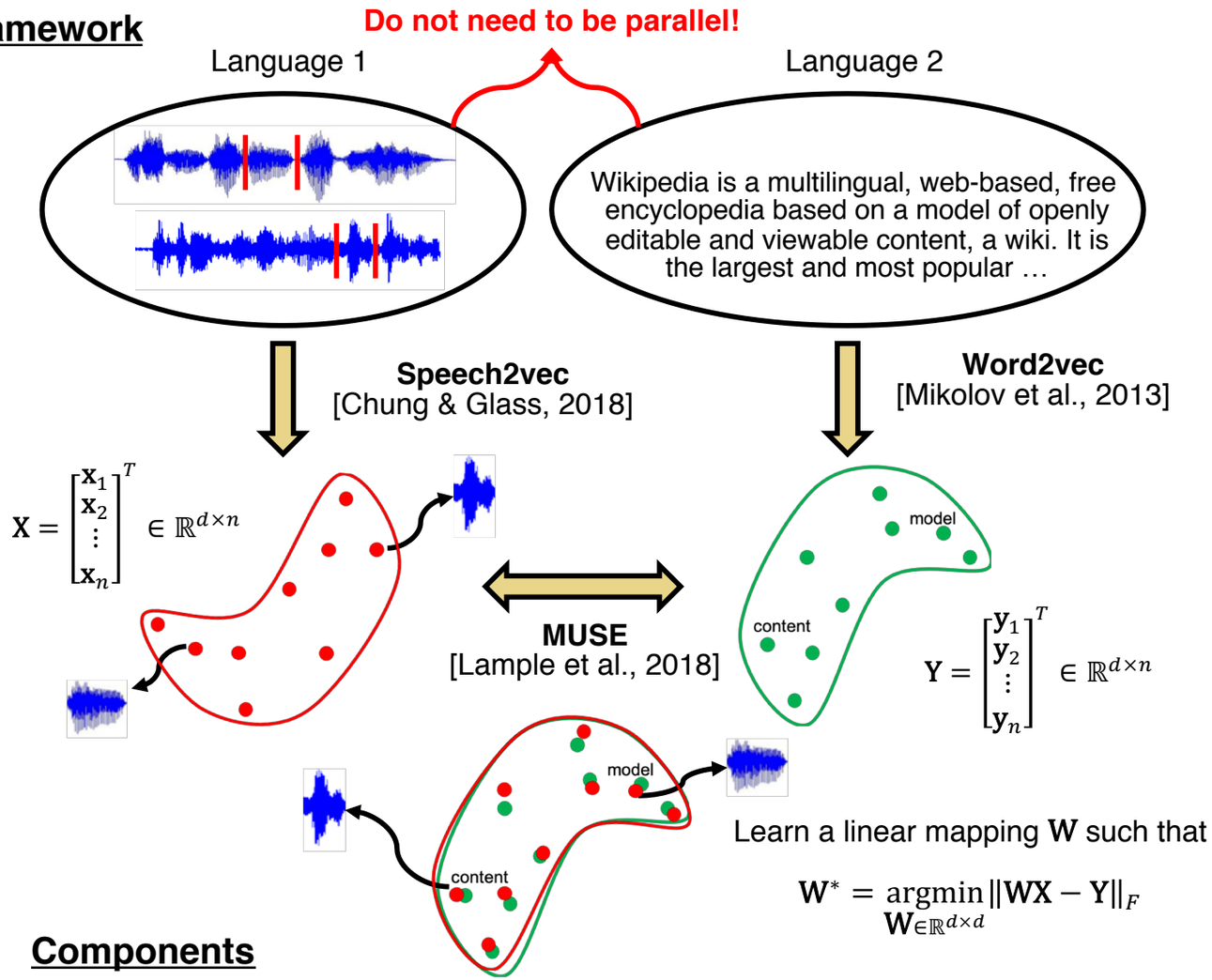
Framework



Components

- **Word2vec**
  - Learns distributed representations of words from a **text** corpus that model word semantics in an unsupervised manner
- **Speech2vec**
  - A speech version of word2vec that learns semantic word representations from a **speech** corpus; also unsupervised
- **MUSE**
  - An unsupervised way to learn $\mathbf{W}$ with the assumption that the two embedding spaces are approximately isomorphic

Framework



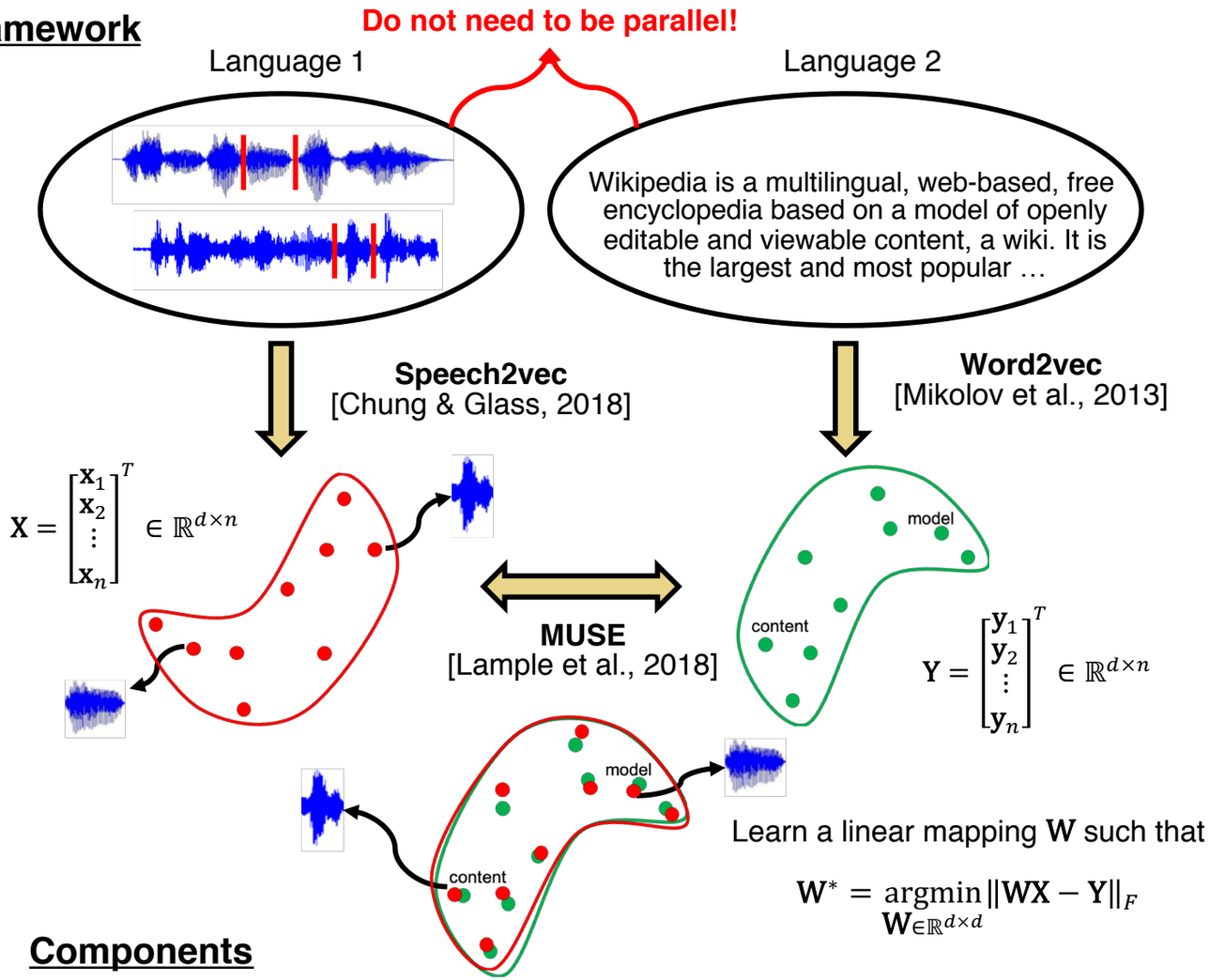
Components

- **Word2vec**
  - Learns distributed representations of words from a **text** corpus that model word semantics in an unsupervised manner
- **Speech2vec**
  - A speech version of word2vec that learns semantic word representations from a **speech** corpus; also unsupervised
- **MUSE**
  - An unsupervised way to learn $\mathbf{W}$ with the assumption that the two embedding spaces are approximately isomorphic

Advantages

- Rely only on monolingual corpora of speech and text that:
  - Do not need to be parallel
  - Can be collected independently, greatly reducing human labeling efforts
- The framework is unsupervised:
  - Each component uses unsupervised learning
  - Applicable to low-resource language pairs that lack bilingual resources

Framework



Components

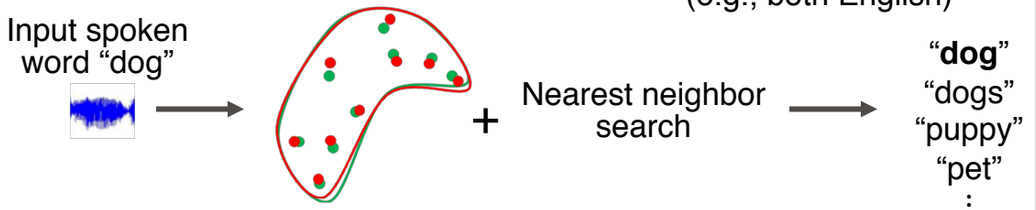
- **Word2vec**
  - Learns distributed representations of words from a **text** corpus that model word semantics in an unsupervised manner
- **Speech2vec**
  - A speech version of word2vec that learns semantic word representations from a **speech** corpus; also unsupervised
- **MUSE**
  - An unsupervised way to learn $\mathbf{W}$ with the assumption that the two embedding spaces are approximately isomorphic

Advantages

- Rely only on monolingual corpora of speech and text that:
  - Do not need to be parallel
  - Can be collected independently, greatly reducing human labeling efforts
- The framework is unsupervised:
  - Each component uses unsupervised learning
  - Applicable to low-resource language pairs that lack bilingual resources

Usage of the learned $\mathbf{W}$

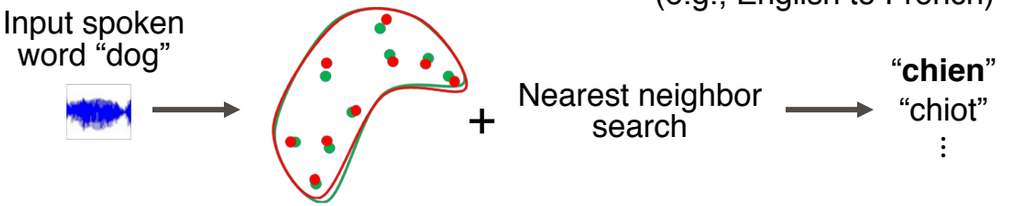
#1: Unsupervised spoken word recognition: when **language 1 = language 2** (e.g., both English)



Foundation for ***Unsupervised Automatic Speech Recognition***

An interesting property of our approach: **synonym retrieval**
→ The list of nearest neighbors actually contain both synonyms and different lexical forms of the input spoken word.

#2: Unsupervised spoken word translation: when **language 1 ≠ language 2** (e.g., English to French)



Foundation for ***Unsupervised Speech-to-Text Translation***

Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

10:45 AM – 12:45 PM

Room 210 & 230 AB #156