https://go.umd.edu/lchange21_ancient_trans

Time-Aware Ancient Chinese Text Translation and Inference

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, Vera Demberg

Department of Language Science and Technology, Saarland University

Department of Computer Science, University of Maryland, College Park

Outline

- Background
- Task Formulation
- Proposed Framework:

  Multi-label Prediction of Translation & Chronological Context
- Experimental Results & Discussion
- Conclusion

Background

Why to people need a better understanding of ancient Chinese, including its translation to modern text?

- The Chinese language is spoken by more than 1.3 billion people
  - Chinese inherits a lot of phrases from ancient time.
  - However, its ancient variant (or *ancient Chinese*) is mastered only by a few.
  - This proves to be a bottleneck in understanding the essence of the Chinese culture.

Background

What important purposes may an ancient Chinese translation & chronology inference system serve?

1. **Education:** Ancient Chinese is an essential part of the curriculum in all Chinese-speaking regions.
    - A translation system can bolster the understanding of ancient texts.
2. **Historical Linguistics:** Chronology inference may help to settle the linguistic debate with regard to the era of origin of text.
    - Especially useful where carbon dating cannot pinpoint the exact era
    - But where their linguistic features can formulate a clear-cut time period

Background

What are the challenges for building such a translation system?

Need more efficient ways of utilizing available data

- Extensive timeline of ancient texts
  - E.g. Pre-Qin (先秦) era and Song dynasty (宋朝) are roughly 700 years apart.
  - This gap witnessed a **drastic evolution of linguistic properties**, especially the meaning and usage of words and phrases.
- Different eras often consist of various amounts of available data
  - Data imbalance complicates the design of the translation systems and limits their generalizability.
  - Past attempts at building such translation systems yield limited performance mainly due to data scarcity in some eras.

Background

Why at a fine-grained view, "ancient Chinese" **may not be considered a single language** with a static word-meaning mapping?

Language change across time suggests a time-aware modeling approach

- The dynamic word-meaning mapping over time can be seen in, e.g., the usage of polysemous Chinese words, many of which are highly ambiguous.
  - Example 看 ('kàn')
  - **Meaning Lost:** Back in early ancient time, it has many meanings such as "to visit" and "to listen", in addition to the major modern meaning, "to look"
  - **Subtle Lexical Semantic Shift:**
    - Earliest known meaning of "看" is "to look into the distance"
    - Meaning of "to look at something closely" emerged during the Han (汉) period and eventually became the prominent meaning

Task Formulation

- Our approach is inspired by recent advances in machine translation and text style transfer/generation
  - Utilize semi-supervised techniques to tackle similar data scarceness challenges
  - Align latent representations of different languages/styles

- Data for translation model
  - Nonparallel datasets A and M: sentences in Ancient Chinese (zh-a) and Modern Chinese (zh-m) respectively
  - Parallel dataset P: pairs of sentences in both variants of text
  - Data size $|A|, |M| \gg |P|$

Task Formulation

Main Task

Convert input ancient Chinese text a to its modern variant m

a：王曰：先王不同俗，何古之法？

m：王说：先王习俗不同，哪种古法可以仿效？

(The king said, "there are various conventions used by previous kings. Which one can we follow?")

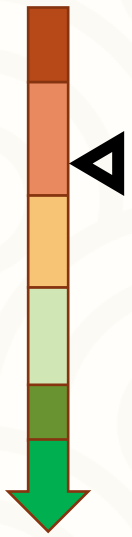
Which king?

Of which dynasty?

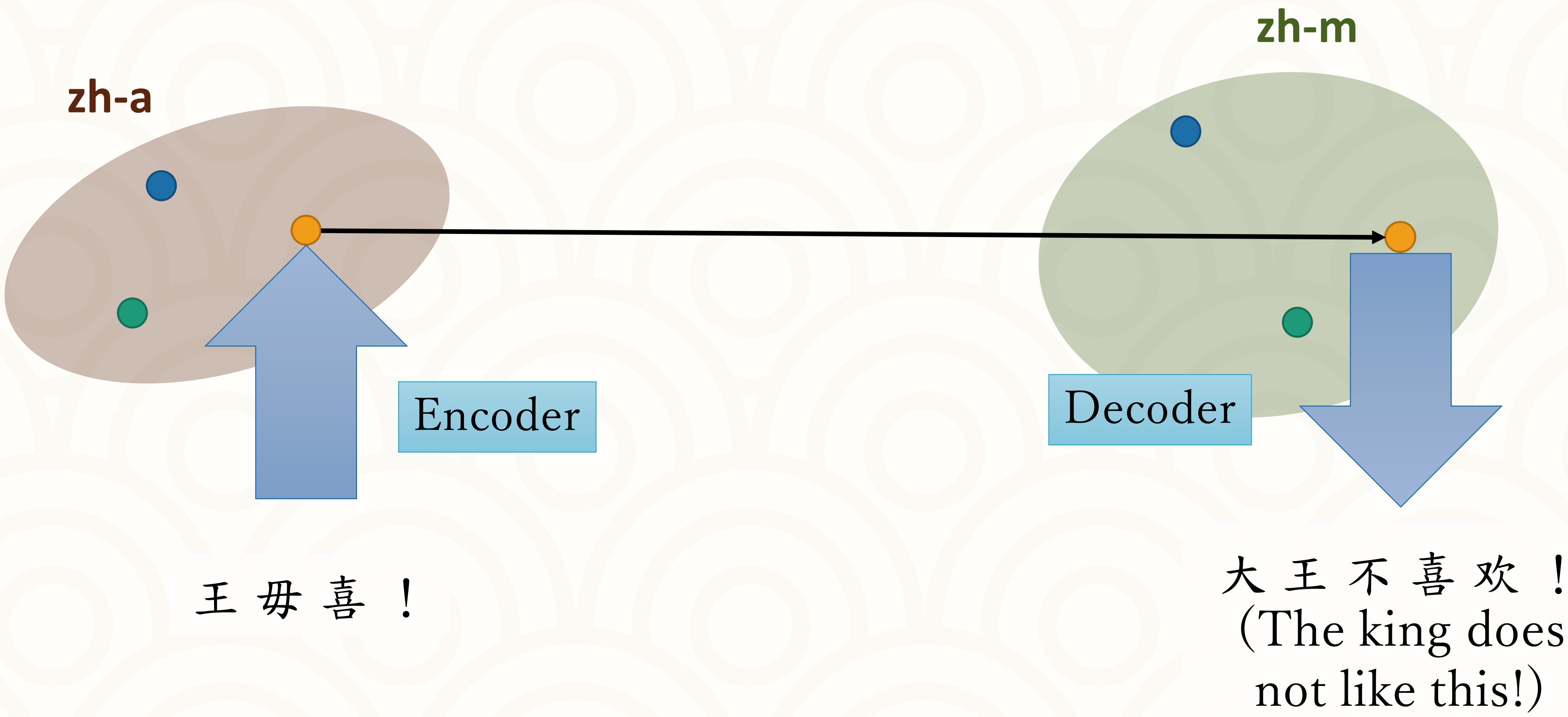
What are the previous conventions?

Auxiliary Task

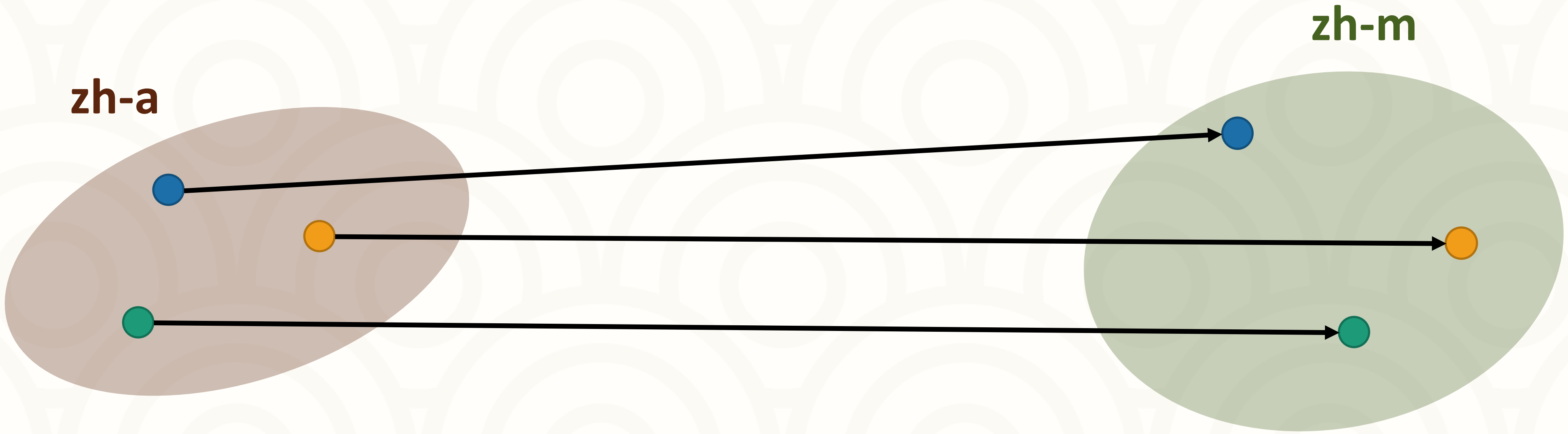
Predict the chronological period c of the ancient text a



Proposed Framework:
Multi-label Prediction of Translation & Chronological Context



Proposed Framework:
Multi-label Prediction of Translation & Chronological Context



$$L = L_{supervised}(P)$$

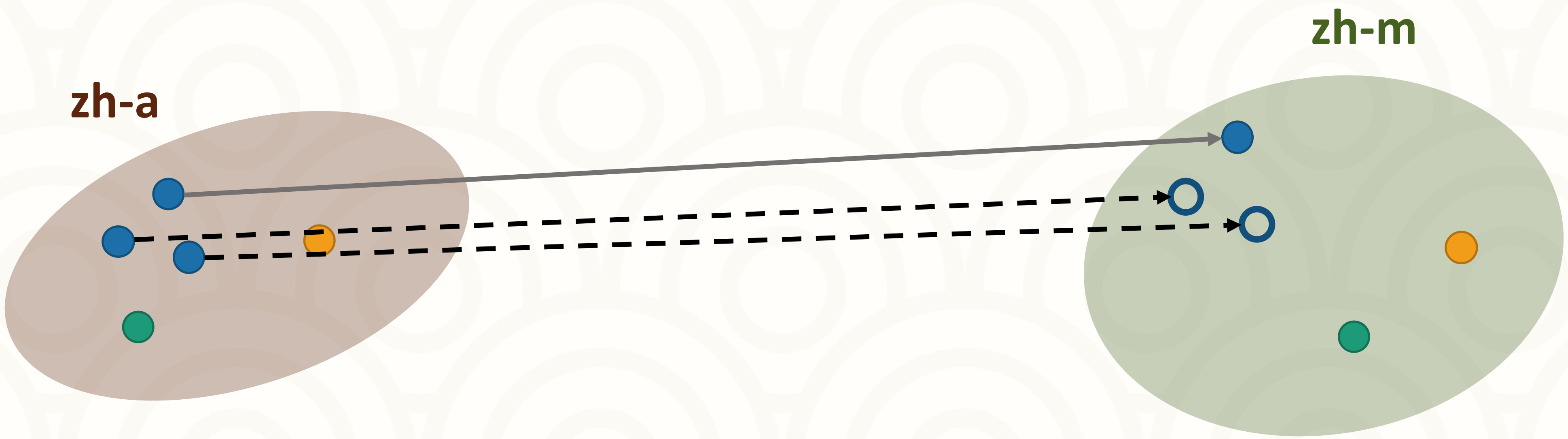
Limited Parallel Data

Proposed Framework:
Multi-label Prediction of Translation & Chronological Context

$$L = L_{supervised}(P) + L_{lm}(A)$$

Augment with ancient text
w/o translation

Proposed Framework:
Multi-label Prediction of Translation & Chronological Context



$$L = L_{supervised}(P) + L_{lm}(A)$$

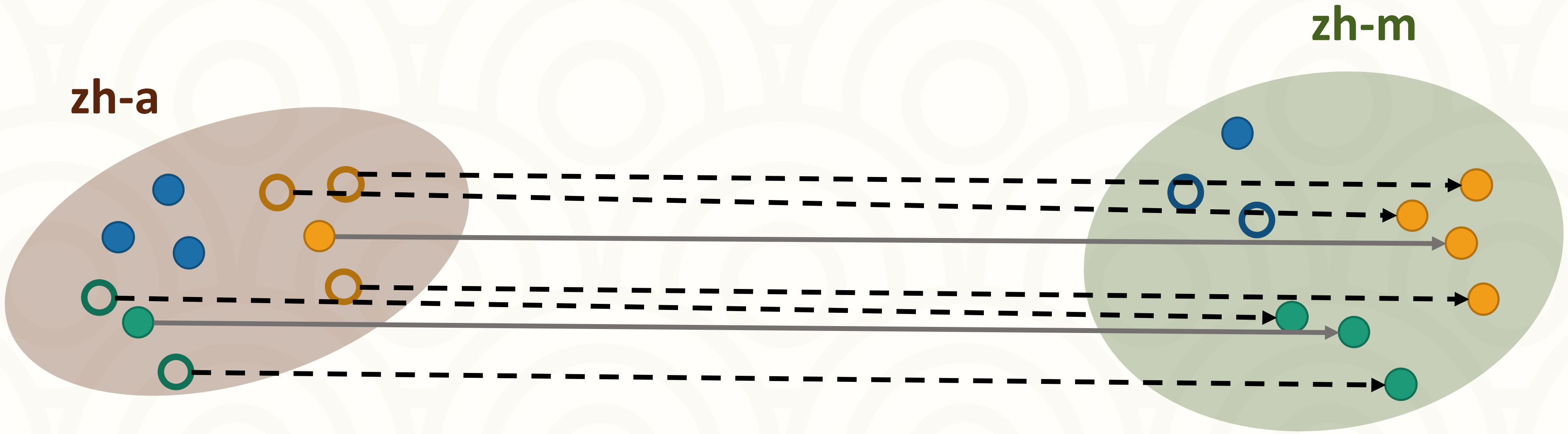
Proposed Framework:
Multi-label Prediction of Translation & Chronological Context

$$L = L_{supervised}(P) + L_{lm}(A) + L_{lm}(M)$$

Augment with modern text
(usually easy to obtain)

Proposed Framework:

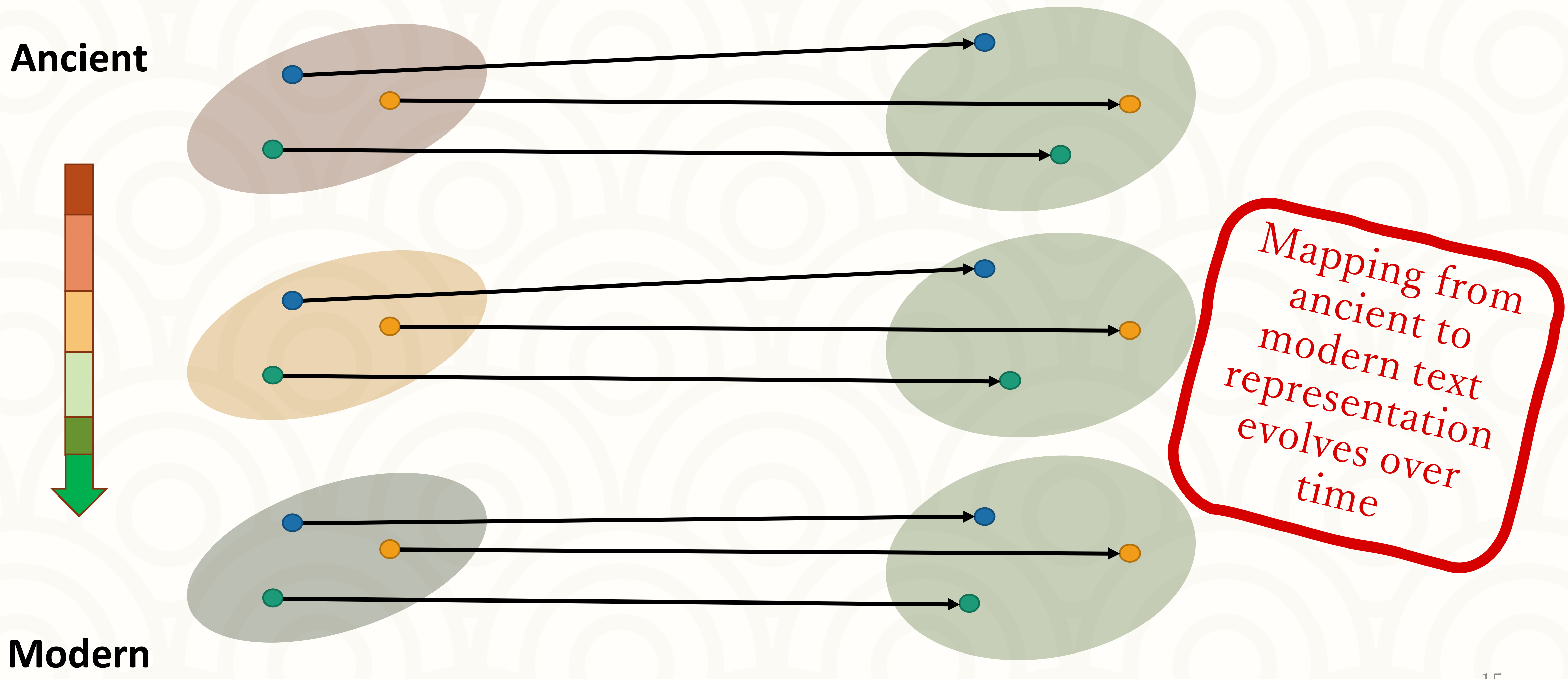
Multi-label Prediction of Translation & Chronological Context



$$L = L_{supervised}(P) + L_{lm}(A) + L_{lm}(M)$$

Semi-supervised Training Objective

Proposed Framework:
Multi-label Prediction of Translation & Chronological Context



Proposed Framework:
Multi-label Prediction of Translation & Chronological Context

- Solution: **Time-aware reranking** with a large-scale language model
  - Pre-train a GPT-2 model with a large modern Chinese corpus
  - Continue to train it on chronologically-annotated ancient-modern sentence pairs, maximizing $p_{\mathrm{GPT-2}}(a, m, c)$

- **Use #1: Chronology Inference**
  - Predict the era of the text to provide more context for understanding

- **Use #2: Quality Estimation for Reranking**
  - Consider all translation candidates m' and chronological periods c'
  - Select the best candidate according to $p_{\mathrm{GPT-2}}(a, m', c')$

Experiments:
Dataset Construction

Chronology label	# sentence pairs
pre-qin (先秦)	1,244
han (汉)	20,460
song (宋)	7,103

Released for future research

- Chronology Annotation
  - 28,807 ancient Chinese prose sentences (Liu et al., 2019)
  - Consider 3 chronology labels: pre-qin (先秦), han (汉), song (宋)
  - Annotation is based on which ancient book the sentences are taken from
- Parallel Data
  - 10% chronologically-annotated sentences for development and test set respectively
  - Additional 4,760 sentences from ancient Chinese poems w/ modern Chinese translation (Shang et al., 2019)

Experiments:
Dataset Construction

- Nonparallel data
  (Shang et al., 2019)
  - Source side (zh-a): sentences from ancient poems, w/o translations
  - Target side (zh-m): sentences from modern lyrics

		# sentences	# characters (source, target)
Nonparallel	zh-a	269,409	4M
	zh-m	77,687	826K
Parallel	Train	27,807	(524K, 797K)
	Dev	2,880	(59K, 88K)
	Test	2,880	(60K, 90K)

- GPT-2 pre-training data: 1.2GB of Chinese Wikipedia text

Experiments:
Model Configurations

- Translation model based on Fairseq
  - Tokenization: character-based
  - Source vocabulary size: 4,824
  - Target vocabulary size: 4,600
  - Transformers, ~ 54M parameters
- Reranking model based on GPT-2
  - ~ 82M parameters

Results & Discussion:
Ancient-to-modern Translation

Training objectives	BLEU-4
$L_{supervised}$ (Liu et al., 2019)	19.59
$L_{supervised} + L_{lm}(M)$	23.05
$L_{supervised} + L_{lm}(M) + L_{lm}(A)$	23.15
+ share decoder embeddings	24.38
+ time-aware reranking	**24.51**

zh-m nonparallel data enhances the decoder's ability to generate modern Chinese.

zh-a nonparallel data may help the encoder to maintain crucial semantic information.

- Source and target side vocabularies have a large overlap
- May also serve as evidence that there are still ancient components in modern Chinese

Human Evaluation (randomly sampled 100 sentences)	Adequacy (avg. of 3 experts)	Fluency (avg. of 3 experts)
0-5 scale	4.06	3.68

Results & Discussion:
Ancient-to-modern Translation

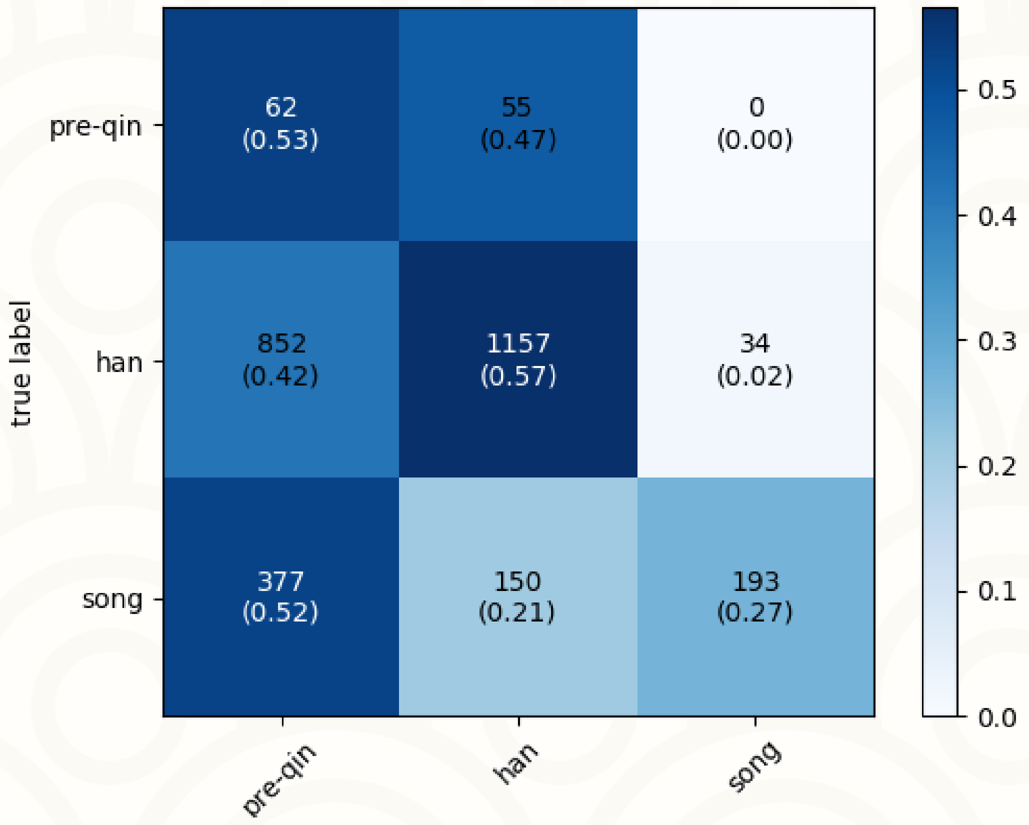
Training objectives	BLEU-4	BLEU-4 on subsets of test data: pre-qin	han	song
$L_{supervised}$ (Liu et al., 2019)	19.59	14.41	**20.02**	19.13
$L_{supervised} + L_{lm}(M)$	23.05	15.97	**23.32**	23.17
$L_{supervised} + L_{lm}(M) + L_{lm}(A)$	23.15	14.15	23.34	**23.72**
+ share decoder embeddings	24.38	15.70	24.52	**24.99**
+ time-aware reranking	24.51	15.50	24.62	**25.24**

- Translations of ancient text chronologically closer to modern Chinese (han and song) tend to yield better performances
  - Semantic gaps are generally smaller
- Leveraging zh-m nonparallel data is most helpful for the song period
  - Much closer to modern Chinese compared to the text from the other two periods

https://go.umd.edu/lchange21_ancient_trans

Results & Discussion:
Chronology Inference

Period (# test)	Precision	Recall	F1
pre-qin (117)	0.05	0.53	0.09
han (2043)	0.85	0.57	0.68
song (720)	0.85	0.27	0.41
Accuracy			0.49
Macro avg.	0.58	0.45	0.39
Weighted avg.	0.82	0.49	0.59



- Performance depends very much on the data **scarcity** and the **closeness** of chronological periods.
  - On the Chinese historical timeline, `han` is very close to `pre-qin`, but `han` and `song` are more separated.

Results & Discussion: Chronology Inference

- Another source of difficulty: ancient Chinese writings tend to quote a considerable amount of text written in previous time periods
  - For example, a history book written in the `song` period may inherit narratives written in `pre-qin` and `han` for the history before `han`
  - So it is challenging to perform chronology inference based solely on the linguistic properties of individual sentences
  - Chronology inference can still provide useful signals for the translation model to better capture semantic differences across time.

Conclusion

- We present a framework that translates ancient Chinese texts into its modern correspondence in low resource scenarios with:
  - Very little parallel data
  - Large number of monolingual sentences without ancient-modern alignment
- We display the usefulness of chronology inference as an auxiliary task that hints at potential diachronic semantic gaps.
- We hope to extend this research to further model additional contextual information about each era.

Thank you!

感 謝 聆 聽

Time-aware Ancient Text Translation and Inference

https://github.com/orina1123/time-aware-ancient-text-translation