English-to-Chinese Transliteration with a Phonetic Auxiliary Task

October 2020

Yuan He

Department of Computer Science

University of Oxford

Shay Cohen

School of Informatics

University of Edinburgh

What is the Task?

- Transliteration: translation of named entities based on pronunciation.

Source Name	Target Name	IPA	Pinyin
Amy	艾米	/ˈeɪ.mi/	ài mǐ

- General Motivations:
  - Information Retrieval – Proper and Common nouns that carry important information are often transliterated (Saravanan et al., 2010).
  - Address Verification – Names of streets, cities, provinces, etc.
  - Machine Translation – Cooperate with the Named Entity Recognition system (Marton and Zitouni, 2014).
- Why study English-to-Chinese Transliteration?
  - Challenging because:
    - English is alphabetical, characters are not individually meaningful.
    - Chinese is logographic – i.e. each character represents a whole word or phrase.

Our Contributions

- A new English-to-Chinese named entities dataset (“DICT”) particular to names of people, based on the dictionary:

  A Comprehensive Dictionary of Names in Roman-Chinese (Xinhua News Agency, 2007)

  Our data is available at: https://github.com/Lawhy/Multi-task-NMTransliteration/mnmt/datasets

- Substitution-based metric called Accuracy with Alternating Character Table (ACC-ACT):
  - It gives a better estimation of the system’s quality than the traditional word accuracy (ACC).

- Multi-task learning transliteration model with a phonetic auxiliary task:
  - It attains better scores than single-main-task or single-auxiliary task models.
  - It achieves similar performance as the previous state of the art with a model of a much smaller size (22M parameters vs. 133M parameters).

Related Work

- Sound representation as an intermediate:
  - Knight and Graehl, 1997
- Exploiting both graphemes and phonemes:
  - Oh et al., 2006 (Correspondence-based)
  - Le and Sadat, 2018 (G2P)
- Incorporating phonetic information:
  - Jiang et al., 2009 (Pinyin)
  - Salam et al., 2011 (IPA)
- Using Multi-task Learning on models through joint learning of various NLP tasks:
  - Luong et al., 2016
  - Dong et al., 2015

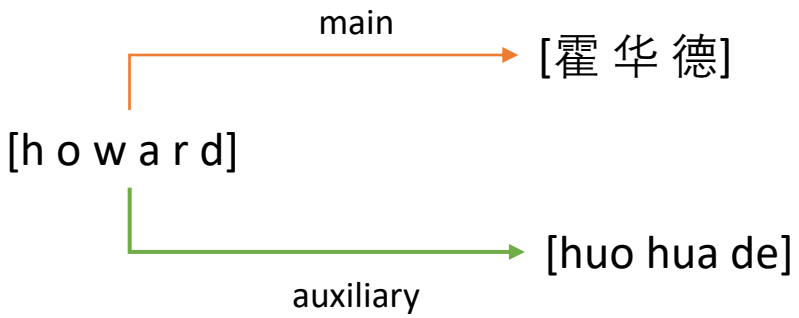
Core Idea: Incorporating Sound Information

- Why using sound information?

  Answer: Most of the transliterations are based on the pronunciation instead of meaning.

- How to include it in the model design?

  Answer: We tried multi-task learning because in theory, training closely related tasks can improve the performance on both. We define the En2Ch transliteration as the main task and En2Pinyin as the auxiliary one (see an example below).

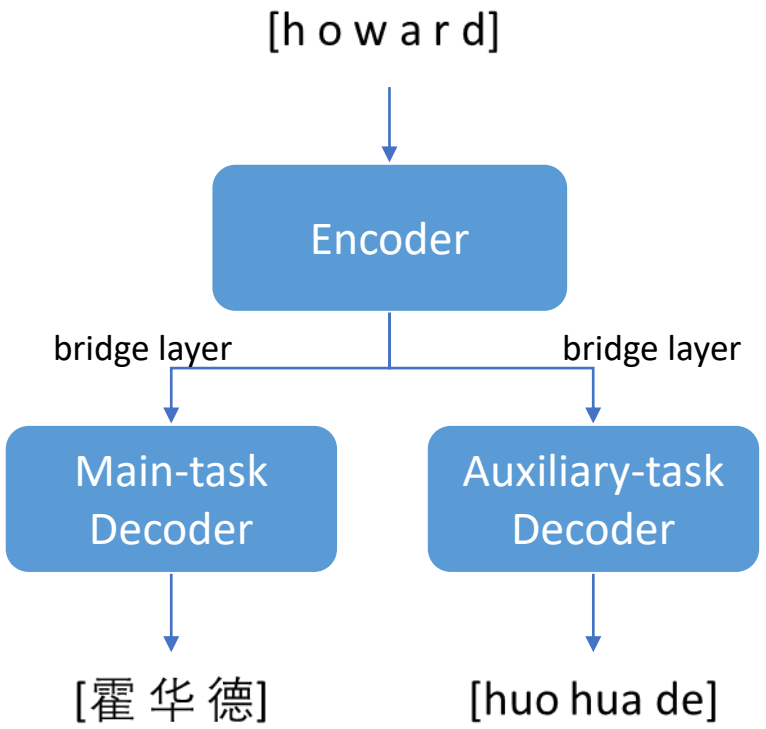


Our Model

- Shared encoder and dual decoders

- Customized loss function controlled by lambda:

$$L_{mtl} = \lambda \cdot L_{main} + (1 - \lambda) \cdot L_{aux}$$



Our code is available at:
https://github.com/Lawhy/Multi-task-NMTransliteration

- Shared encoder and dual decoders
- Customized loss function controlled by lambda:

$L_{mtl} = \lambda \cdot L_{main} + (1 - \lambda) \cdot L_{aux}$

Adaptive Evaluation Metric

- Why the traditional one is criticized?

  Answer: The traditional word accuracy (ACC) only considers one possible reference.

- How about including more references in the dataset?

  Answer: Unrealistic because no dataset is guaranteed to contain all possible references.

- How to solve the problem?

  Answer: Given a properly constructed Alternating Character Table (ACT), our metric (ACC-ACT) can cover most of the possible alternatives (see next slide).

Accuracy with Alternating Character Table (ACC-ACT)

- Alternating Character Table:
  - A table with each row storing a list of interchangeable characters;
    only suitable for a logographic system where every character is independent
  - Created based on the one-source-multiple-target name pairs in the DICT dataset.
    can be improved further by linguists

- Example of how the algorithm captures the multiple references:

Source Name	Target Name 1	Target Name 2	MED
Mona	莫纳	莫娜	1

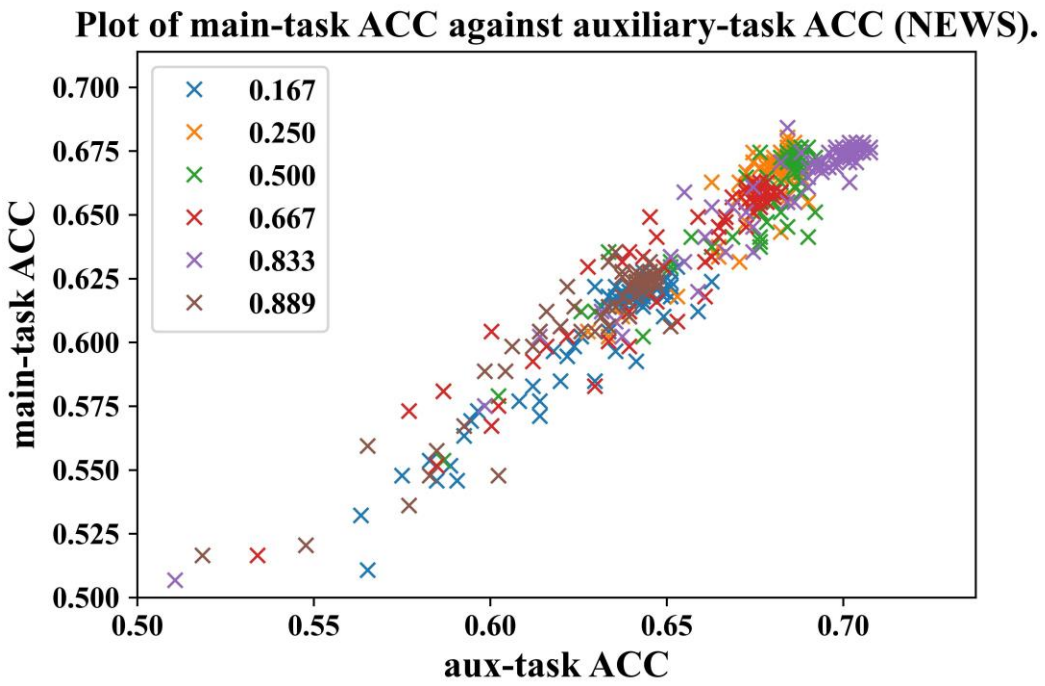
Datasets

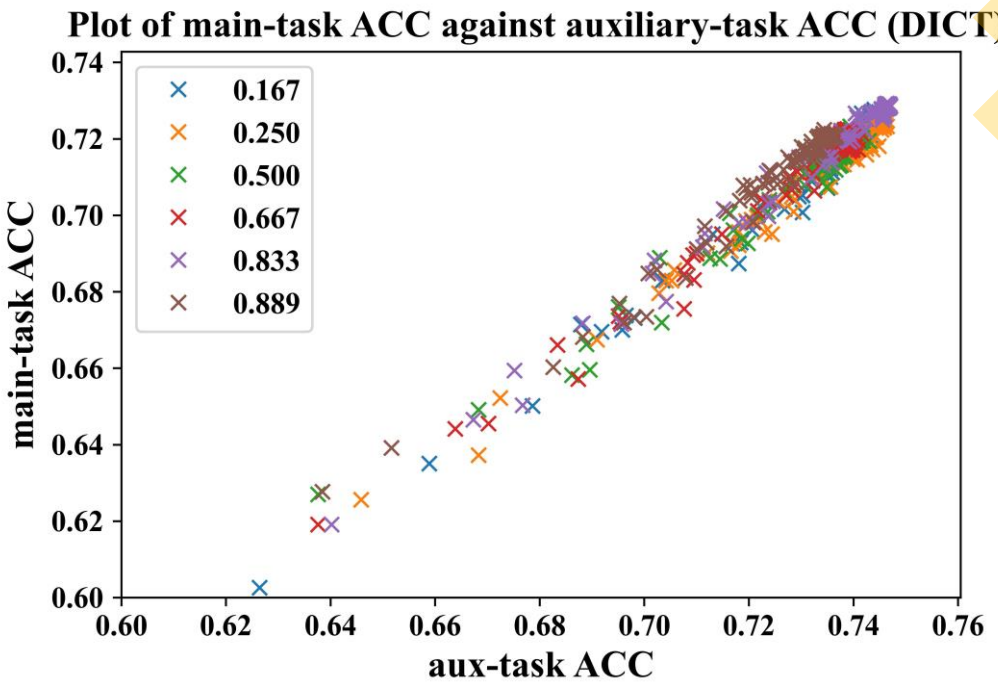
- DICT dataset
  - Extracted from the giant dictionary (see previous slide).

- NEWS dataset
  - Raw data came from the NEWS 2018 Shared Task (http://workshop.colips.org/news2018/).
  - Preprocessed data came from the work of Grundkiewicz and Heafield, 2018.

- The NEWS official test set is anonymized.

Analysis of the Main-Auxiliary Relationship





Experiment Results (NEWS)

Experiment results on the NEWS internal test set (official development set)				
	Main			Auxiliary
System	**ACC**	**ACC+**	**ACC-ACT**	**ACC**
Baseline (Ours)	0.724	0.733	0.742	0.736
Multi-task (Ours)	0.739	0.749	0.760	0.757
BiDeep (Grundkiewicz and Heafield, 2018)	0.731	0.739	0.746	0.740
BiDeep+ (Grundkiewicz and Heafield, 2018)	NA	0.765 (reported)	NA	NA

NEWS Leaderboard (third-party)

Table of the NEWS leaderboard		
User	**ACC+**	**F-score**
romang (Grundkiewicz and Heafield, 2018)	0.3040 (1)	0.6791 (2)
Ours	0.2990 (2)	0.6799 (1)
saeednajaf	0.2820 (3)	0.6680 (3)
soumyadeep	0.2610 (4)	0.6603 (4)

Available at https://competitions.codalab.org/ competitions/18905#results, accessed 19 June 2020.

Experiment Results (DICT)

Experiment results on the DICT test set			
	Main		Auxiliary
System	**ACC**	**ACC-ACT**	**ACC**
Baseline	0.726	0.748	0.738
Multi-task	0.729	0.751	0.749
BiDeep (Grundkiewicz and Heafield, 2018)	0.732	0.755	0.760

Discussion & Conclusion

- Incorporating phonetic information benefits transliteration task in a neural setting.
- But still unable to handle the irregular cases in transliteration well.
- ACC-ACT covers all cases of ACC and captures more acceptable transliterations.
- The idea can be generalized to other transliteration tasks.

Thank you!

References I

Yuval Marton and Imed Zitouni. 2014.
Transliteration normalization for information extraction and machine translation.
Journal of King Saud University - Computer and Information Sciences, 26(4):379 – 387. Special Issue on Arabic NLP.

K. Saravanan, Raghavendra Udupa and A Kumaran. 2010.
Crosslingual Information Retrieval System Enhanced with Transliteration Generation and Mining.
The Forum for Information Retrieval Evaluation (FIRE-2010) Workshop, Kolkata, India, Proceedings of the FIRE 2010 Shared Evaluation.

Knight and Graehl, 1997
Machine transliteration.
In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98/EACL '98, page 128–135, USA. Association for Computational Linguistics.

References II

Jong-Hoon Oh, Key-Sun Choi, and Hitoshi Isahara. 2006.
A machine transliteration model based on correspondence between graphemes and phonemes.
ACM Trans. Asian Lang. Inf. Process., 5:185–208.

Ngoc Tan Le and Fatiha Sadat. 2018.
Low-resource machine transliteration using recurrent neural networks of Asian languages.
In Proceedings of the Seventh Named Entities Workshop, pages 95– 100, Melbourne, Australia. Association for Computational Linguistics.

Xue Jiang, Le Sun, and Dakun Zhang. 2009.
A syllable-based name transliteration system.
In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pages 96–99, Suntec, Singapore. Association for Computational Linguistics.

Khan Md. Anwarus Salam, Yamada Setsuo, and Tetsuro Nishino. 2011.
Translating unknown words using wordnet and ipa-based-transliteration.
14th International Conference on Computer and Information Technology (ICCIT 2011), pages 481–486.

References III

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016.
Multitask sequence to sequence learning.
CoRR, abs/1511.06114.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015.
Multi-task learning for multiple language translation.
In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Roman Grundkiewicz and Kenneth Heafield. 2018.
Neural machine translation techniques for named entity transliteration.
In Proceedings of the Seventh Named Entities Workshop, pages 89–94, Melbourne, Australia. Association for Computational Linguistics.