	LSTM+Cosine	LSTM+Tanh	**GRU+Cosine**	GRU+Tanh
F1-score	0.82	0.7788	**0.8396**	0.7463

Table 3: The comparison among different configurations of the proposed Siamese models: LSTM + Cosine, LSTM + Tanh, GRU + Cosine, and GRU + Tanh.

	Accuracy	F1	Precision	Recall
Triplet Siamese (A)	**0.9736**	**0.8396**	**0.9082**	0.7807
Triplet Siamese (B)	0.9527	0.5323	0.44	0.6735
Jaro Winkler	0.904	0.4633	0.3083	0.9318
Levenshtein	0.9333	0.56	0.3962	0.9545
Damerau Levenshtein	0.9323	0.5563	0.3925	0.9545
Q-Gram	0.9202	0.5212	0.3554	0.9773
Cosine	0.9444	0.5985	0.4409	0.9318
Smith Waterman	0.9576	0.6769	0.5116	**1.0**
LCS	0.9556	0.6562	0.5	0.9545
Logistic Regression	0.9313	0.5641	0.3929	**1.0**

Table 4: The performance on the validation set of seven distance-related methods, the ensemble logistic model, and the triplet Siamese model trained by two training sets - (A) is the dataset A, (B) is the dataset B.

	Accuracy	F1	Precision	Recall
Triplet Siamese (A)	**0.983**	**0.6836**	**0.5839**	0.8246
Triplet Siamese (B)	0.9806	0.6426	0.546	0.7807
Jaro Winkler	0.9587	0.2749	0.1619	0.9091
Levenshtein	0.9601	0.2817	0.1667	0.9091
Damerau Levenshtein	0.9595	0.2787	0.1646	0.9091
Q-Gram	0.9758	0.3922	0.25	0.9091
Cosine	0.9671	0.328	0.199	**0.9318**
Smith Waterman	0.9693	0.3431	0.2103	**0.9318**
LCS	0.9787	0.4233	0.2759	0.9091
Logistic Regression	0.9671	0.3226	0.1961	0.9091

Table 5: The performance on the testing set of seven distance-related methods, he ensemble logistic model, and the triplet Siamese model trained by two training sets - (A) is the dataset A, (B) is the dataset B.

University at Ho Chi Minh City. We acknowledge the support of Science Foundation Ireland under grant number SFI/13/RC/2106 and L. Meltzers Høyskolefonds, UiB 2019/2259-NILSO.

REFERENCES

[1] Han K. Cao, Duyen T. Ly, Duy M. Nguyen, and Binh T. Nguyen. 2019. Automatically Generate Hymns Using Variational Attention Models. In *Advances in Neural Networks – ISNN 2019*, Huchuan Lu, Huajin Tang, and Zhanshan Wang (Eds.). Springer International Publishing, Cham, 317–327.

[2] Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 459–474.

[3] J. Gao, Y. He, X. Zhang, and Y. Xia. 2017. Duplicate short text detection based on Word2vec. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. 33–37. https://doi.org/10.1109/ICSESS.2017.8342858

[4] Yukiko Homma. 2017. Detecting Duplicate Questions with Deep Learning.

[5] Doris Hoogeveen, Andrew Bennett, Yitong Li, Karin M. Verspoor, and Timothy Baldwin. 2018. Detecting Misflagged Duplicate Questions in Community Question-Answering Archives. In *ICWSM*.

[6] Bromley Jane, Guyon Isabelle, LeCun Yann, Säckinger Eduard, and Shah Roopak. 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS'93)*. Morgan Kaufmann Publishers Inc., San Francisco CA USA, 737–744. http://dl.acm.org/citation.cfm?id=2987189.2987282

[7] Matthew A. Jaro. 1976. *UNIMATCH: A Record Linkage System: User's Manual*. Technical Report. U.S. Bureau of the Census, Washington, D.C.

[8] Matthew A. Jaro. 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Amer. Statist. Assoc.* 84, 406 (1989), 414–420. https://doi.org/10.1080/01621459.1989.10478785 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.10478785

[9] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition.

[10] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. ACM, New York, NY, USA, 19–34. https://doi.org/10.1145/3183713.3196926

[11] D. M. H. Nguyen, H. T. Vu, H. Q. Ung, and B. T. Nguyen. 2017. 3D-Brain Segmentation Using Deep Neural Network and Gaussian Mixture Model. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 815–824.

[12] Thorsten Papenbrock, Arvid Heise, and Felix Naumann. 2015. Progressive Duplicate Detection. *Knowledge and Data Engineering, IEEE Transactions on* 27 (05 2015), 1316–1329. https://doi.org/10.1109/TKDE.2014.2359666

[13] Esko Ukkonen. 1992. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science* 92, 1 (1992), 191 – 211. https://doi.org/10.1016/0304-3975(92)90143-4

[14] William E. Winkler and Yves Thibaudeau. 1991. *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census*. Technical Report Statistical Research Report Series RR91/09. U.S. Bureau of the Census, Washington, D.C.