Electronic Supplementary Information

Melting point prediction of small organic molecules by deciphering the chemical structure into a natural language

Discussion

Successful T_m prediction achieved by the NLP approach solely using SMILES

The prediction of T_m from the structure has been extensively studied, and various advanced regression models have been developed to improve the prediction accuracy. However, the efforts were made mainly to optimize the regression methods to translate the structural input to the output (T_m), while the structural input was confined to descriptors or fragments. In this study, we proposed SMILES, a line notation to represent chemical structures, as the only input to the prediction model. SMILES contains basic information, including atoms bonds, and their connections and relationships. The SMILES-based structural input does not require additional computational efforts to calculate a number of descriptors, and most importantly, the similarity between SMILES and natural language entitles the application of NLP approaches to explore the relationship between the structure and T_m.

Another linguistic method based on the fragmentation of SMILES strings to generate the "LINGO" hologram was reported for the similarity search and prediction of logP and the aqueous solubility.[1] Our method has shared in common with the "LINGO" approach regarding the line notation input for structures. However, distinguished from the "LINGO" approach, which relied on the fragmentation of strings into substrings of a defined size (named "LINGO") and the occurrence frequency of the "LINGO" fragment in SMILES strings, we concentrated on exploring the relationship between different components (atoms and bonds) written in a line notation.

Specially, knowing that the typical range of T_m for most drug-like compounds is from 0-300°C,[2] we built a subgroup with experimental T_m values in this range from the original dataset for training and testing. Compared with testing data of the whole dataset (A1), this drug-like subgroup has improved accuracy in terms of the RMSE (39.04°C vs. 37.6°C, whole dataset vs. subgroup) and MAE (30.00°C vs. 28.8°C, whole dataset vs. subgroup). (Fig. S1a).

Furthermore, we separately trained and tested the model for molecules with MW < 600. The RMSE, R^2, and MAE of the test data were 38.04°C, 0.8202, and 29.05°C, respectively. Compared with the model trained and tested using all data (A1), its prediction performance hardly improved. For all data and subset data, the plot of predicted T_m against experimental T_m is shown in Fig. S1b, and the plot of the residual T_m against the molecular weight is shown in Fig. S1c.

Unique and surprising features of the NLP-based T_m prediction model

Unique features of the SMILES-based T_m prediction model were observed, which were distinguished from previous descriptor-based models. The size, polarity, partial atom charge, and rigidity of compounds were found to increase the T_m, whereas the structural flexibility and nonpolar descriptors were reported to lower the T_m.[3] It has been reported that it is more difficult to predict the T_m of molecules with larger sizes, complex and flexible structures, and information on intra- or intermolecular

interactions is generally difficult to capture in previous models.[3,4] However, in our current model, the prediction accuracy did not decrease with increasing size, complexity, and structural flexibility of the molecules. Instead, surprisingly, accuracy tends to be improved for molecules with more complex and flexible structures.

The size of a molecule was reported to be a crucial feature in predicting T_m.[3] Generally, a molecule with a larger size tends to possess a higher MW and a longer length SMILES string. In terms of T_m residuals, an increase in the MW did not increase the variation in the residuals, suggesting that the prediction accuracy remains relatively consistent for compounds with distinct MWs within the range in the dataset. In contrast, the heteroscedasticity of residuals was observed when plotting against the length of the SMILES string, suggesting that the model appeared to perform better for a molecule with a longer SMILES string length.

It was also reported that it is easier to predict the T_m of molecules with more rigid chemical structures, while for molecules with many rotatable bonds and thus a flexible configuration, the T_m is much more difficult to predict.[4] However, in our models, within a certain range, the variance of the prediction residuals decreased with an increasing number of rotatable bonds, i.e., increased flexibility, indicating an enhanced prediction accuracy for compounds with high flexibility, which is quite different from previously reported models.

T_m is known to be dominantly dependent on intermolecular interactions, while for organic molecules, hydrogen bonding plays a major role.[3,4] Generally, the formation of hydrogen bonding within molecules resulted in a higher T_m. The prior knowledge of hydrogen bonding or even the crystal configuration in the molecule certainly may improve the prediction performance of T_m. It was unexpected that a heteroscedasticity pattern was also found in the residual plot against the number of hydrogen bonds, suggesting that improved prediction accuracy was found for compounds with more potential intra- or intermolecular hydrogen bonds.

Just as a sentence is composed of words, phrases, and clauses, a SMILES more like a chemical language string is a “structural sentence” composed of atoms, bonds, fragments, backbone, and branches. In NLP, many failures result from the complex syntactic structures of sentences, which confuse the model and render it harder to interpret semantics and pragmatics. Since the line notation of a chemical structure has a similar hierarchical structure as a language, the complexity of SMILES strings, such as the appearance of many branches in a SMILES string depicted by brackets, might complicate the information extraction. Therefore, a residual analysis against the number of branches was performed. The symmetric distribution of residuals on both sides of the x-axis coupled with slight heteroscedasticity suggests that prediction accuracy might increase when increasing the number of branches, which was similar to the correlation between the residual distribution and the length of the SMILES string. Although the growing complexity of SMILES strings does not necessarily correspond to the elongation of SMILES strings and the increasing number of branches, the results did suggest an improved prediction accuracy for molecules with complex structures.

These unique phenomena were largely distinguished from previous descriptor-based models, suggesting that the classification of molecular similarity in our current model might be different from that in fragment-based or descriptor-based models. Even though almost all QSPR (Quantitative Structure Property Analysis) studies were based on the molecular similarity principle, it remains unclear and controversial to define the molecular structure similarity in regard to T_m prediction, since the value of T_m profoundly relies upon not only the 3D structures of the molecules but also the lattice packing of these molecules. Graph-based fingerprints[5] were also established to compare the molecular structure and used for similarity searches in cheminformatics. Nevertheless, the compromised prediction accuracy of T_m has brought into question how to appropriately define molecular similarity when predicting properties related to the 3D structure or spatial and topological properties. The application of the NLP approach might offer another way to define the molecular similarity by coupling the relationship between various components in the structure.

Do we need anything else to improve NLP-based T_m prediction?

In the field of NLP,[6] the introduction of recurrent neural networks (RNNs) has brought significant performance improvements. Long-short-term memory (LSTM)[7] used in our model is a special type of RNN that avoids long-term dependency problems by introducing different "gate" structures that control the flow of information. Through the improvement and popularization of many scholars,[8,9] this method has been successful in language models[10] and speech recognition.[11] In our model, LSTM proved its ability to extract information from line notations of chemical structures. In addition to LSTM, we also tested other structures, including bidirectional LSTM,[12] gated recurrent units (GRUs), convolutional neural networks (CNNs)+LSTM, and transformer[13]. The RMSEs obtained under the setting of experiment A1 are listed (Table S3). LSTM offered the best performances, indicating that it is the most suitable algorithm for this application.

We can see that the effect of using the Transformer is not good, which also confirms that it does not apply to all linguistic problems[14,15]. We believe that this is since the remote dependency capture capability of the Transformer is worse than the RNN-like structure, and it cannot model location information well. In machine translation, the computer uses punctuation to split (such as commas and periods) long sentences, so there is less long-range information loss, which means the Transformer is very suitable for machine translation tasks. However, in processing SMILES strings, the longest SMILES string can reach a length of 288, it is too long for the Transformer. Furthermore, the melting point prediction has a huge dependence on each element position (The melting point will change dramatically after swapping two-element positions of a SMILES string. When the combination of two different canonicalized SMILES forms is used to train the model, the prediction accuracy improves.). These result in the Transformer being unable to capture the most accurate information from SMILES strings for structure-related property prediction. This point can also be confirmed numerically because the experimental results of the Transformer are similar to the experimental results of A3 and A4 in the main text. A3 and A4 are the

experiments that different canonicalized SMILES forms were used for training and validation.

Generally, adding prior knowledge could improve the prediction accuracy of a deep learning model. To test this hypothesis, we introduced the atomic mass or molecular weight (Table S4) into the model. The results, however, demonstrated that the addition of the atomic mass as prior knowledge did not improve prediction accuracy. The RMSE, R^2, and MAE of the test data were 38.5°C, 0.82, and 29.1°C, respectively, which provided little improvement compared with the original predictions. There are two possible explanations. First, the model could have already extracted the knowledge of the atomic mass or molecular weight from the SMILES strings during the independent learning process. Second, the atomic mass or molecular weight did not play a crucial role in the prediction of T_m in this model. This exercise also confirms one of the key findings discussed above; i.e., the prediction accuracy remains relatively consistent for compounds with distinct molecular weights within the range in the dataset.

Methods

Dataset

Two datasets are used for model establishment and validation. One is 20110803ONSMP030, available on the Open Notebook Science (ONS) wiki community, containing 19933 small molecules; the other is a collection reported by Tetko et al.[2] with 275133 molecules. Molecules in both datasets are all free forms. In each data set, a molecule is excluded if it is already in the other data set, is not an organic molecule, or its T_m range is larger than 5°C. This pre-processing excluded 864 (~4.3%) and 18604 (~6.8%) molecules in the two datasets. The average T_m was used while training the model. The average T_m distribution of the remaining molecules in the two datasets is shown in Fig. S2.

Training and test data for the experiments

The canonicalization of the SMILES form on our datasets is performed only by Open Babel, if not specified. The training-test split ratio is 9:1. In section "*Impact of the SMILES form on the prediction accuracy"*, both experiments A1 and A2 use the same training and test molecules from the 20110803ONSMP030 dataset, but A2 applied RDKit to canonicalize the SMILES strings. A3 and A4 use the same training data as A1 and A2 with the test data exchanged. The training data in experiments B1 and B2 are the combination of the training data in A1 and A2, and the test data in B1 and B2 are the same as the test data in A1 and A2, respectively. In experiment C, compared with A1, its training dataset has 20,000 more molecules. These added molecules are selected randomly from Tetko's dataset [2] and are all different from the molecules in the 20110803ONSMP030 dataset. In section "*'Language learning' is important for NLP-based Tm prediction"*, experiments 20K and 40K are the same as A1 and C mentioned above, respectively. The datasets used by experiments 2K, 4K, and 8K are 2,000, 4,000, and 8,000 molecules randomly selected from the 20110803ONSMP030

dataset. The 270K experiment uses Tetko's dataset. Except for these two sections, all the other experiments are just A1 or the analyses and visualizations of A1's results.

Model structure

Each character in the canonicalized SMILES strings, including elements, bonds, and symbols, was encoded by a defined number to construct the sequence vector (Table S5). Elements *Cl* and *Br*, even with two characters, were encoded by only one number. The length of sequence vectors was fixed at 288, which is the maximum length of the SMILES strings that appeared in the dataset. Vectors with fewer dimensions were padded with zeros to the full length. The 288-dimensional vectors were then mapped into a high-dimensional space through the embedding layer. Next, 48 features were extracted from these vectors through two layers of LSTM, where the first LSTM layer returned the full sequence. Finally, the T_m of a molecule was output through two fully connected layers. This network configuration yields 99569 trainable parameters. The process of the model is presented through an example in Fig. S3. In addition to LSTM, the same exercise was also performed when bidirectional LSTM, GRU, CNN+ LSTM, and Transformer were applied. Based on the comparison listed in Table S3, LSTM was selected for the majority of the work.

Network training

We select RMSprop as the optimizer, and the loss function is MSE. The initial learning rate is set as 0.001, and it will be reduced by 0.1 when the test loss does not decrease for ten epochs. Early stopping is applied to reduce the risk of overfitting. The loss function curves rapidly decrease within the first tenth epoch and level off afterward (Fig. S4). The patterns of the loss curves are almost consistent for both the training and test data, suggesting proper use for the hyperparameter selection and timely stopping during the training process. However, there is an obvious difference between the RMSE of the training data and that of the test data, suggesting a certain extent of overfitting, which is common in nonlinear deep learning models.

Evaluation of the prediction performance

The prediction accuracy of the model was evaluated by the root mean square error (RMSE), correlation coefficient (R^2), and absolute mean error (MAE) calculated by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{\exp}^{i} - y_{pred}^{i}\right)^2}{N}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{\exp}^{i} - y_{pred}^{i}\right)^2}{\sum_{i=1}^{N}\left(y_{\exp}^{i} - \bar{y}\right)^2}$$

$$MAE = \frac{\sum_{i=1}^{N}\left|y_{\exp}^{i} - y_{pred}^{i}\right|}{N}$$

where N is the total number of molecules to evaluate; $y^i_{\exp}$ is the experimental T_m of the i_{th} molecule; y^i_{pred} is the predicted T_m of the i_{th} molecule; and $\overline{y}$ is the mean experimental T_m of all molecules.

We repeated experiments A1, A2, B1, and B2 five times by using different training and test data, which are split from the original dataset by using different random number seeds. For each of A1, A2, B1, and B2, the average values of five-times experiments are reported in Table 1 in parentheses.

Properties and descriptors of the molecules

The physicochemical properties and descriptors of the molecules in the dataset, including the MW, length of the SMILES string, number of branches, number of rotatable bonds, number of hydrogen bond acceptors, and number of hydrogen bond donors, were all obtained by RDKit. The number of branches was defined by the occurrence of brackets.

Supplementary Figures

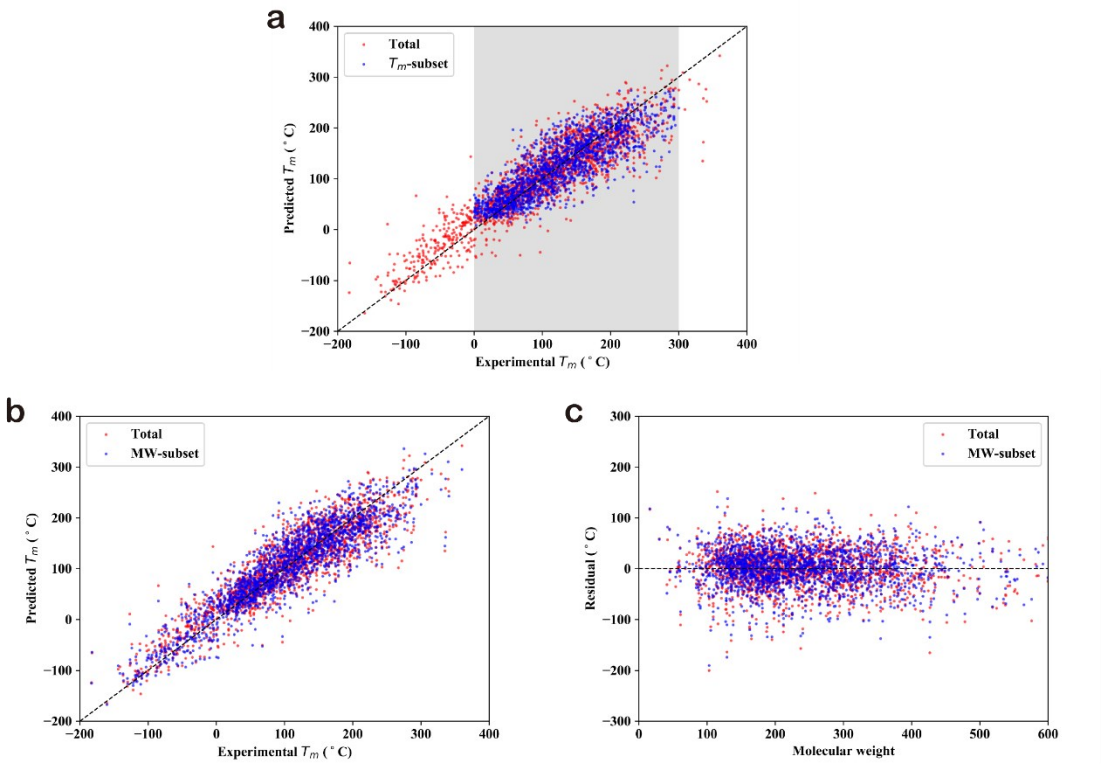

Fig. S1 Some supplementary experiment results. (a) The distribution of the predicted T_m and experimental T_m on both the original test data and the extracted test data. The extracted data are molecules with the experimental T_m ranging from 0-300°C. (b) The joint distribution of the predicted T_m and experimental T_m on both the total data and the subset data. The extracted data are the molecules with MW<600. (c) The joint distribution of the residual and molecular weight for the total data and the subset data. The extracted data are the molecules with MW<600.

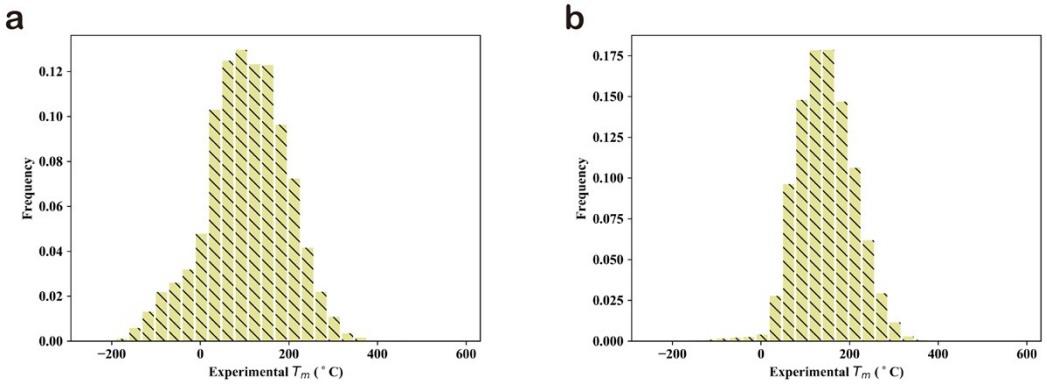

Fig. S2 The frequency distribution of the average T_m in two datasets. (a) 20110803ONSMP030. (b) dataset reported by Tetko et al. (Note: All the bar widths are 30°C.)

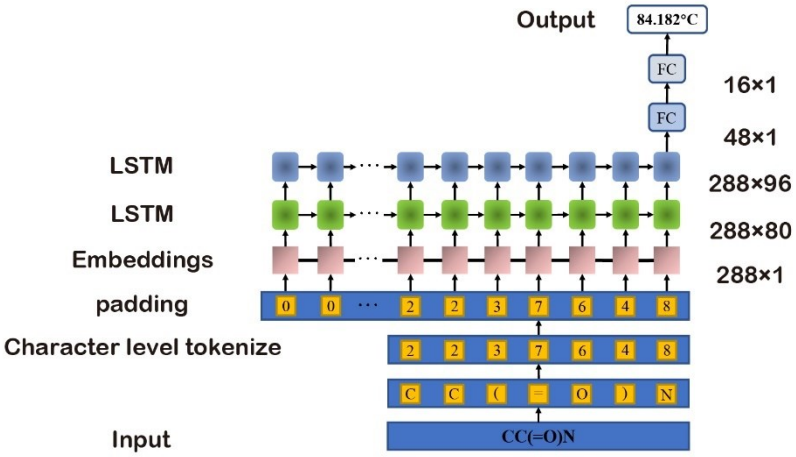

Fig. S3 The model process presented through an example *CC(=O)N*.

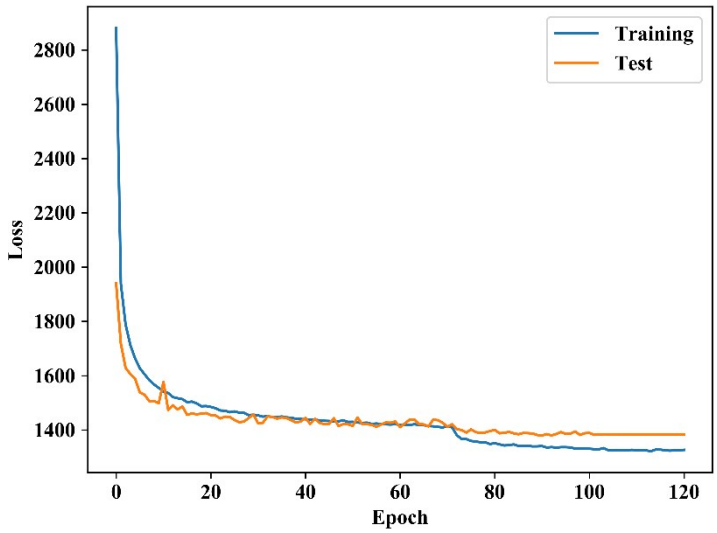

Fig. S4 Loss function curve of the training and the test data.

Supplementary Tables

Table S1 Prediction performance of the SMILES-based model using different canonicalized SMILES forms

	R2		RMSE (°C)		MAE	
	Training	Test	Training	Test	Training	Test
A1	0.8746	0.8103	32.34	39.04	23.87	30.00
A2	0.8691	0.8154	33.06	38.52	24.27	29.07
A3	0.8242	0.7491	38.32	44.90	28.15	34.73
A4	0.8195	0.7366	38.82	46.00	28.69	35.54
B1	0.8748	0.8264	31.55	37.35	23.20	28.33
B2	0.8796	0.8307	30.94	36.88	22.82	27.88
C	0.8643	0.8216	31.14	37.86	23.11	28.54

Table S2 Prediction performance over different sizes of training datasets.

	RMSE		MAE	
	Training	Test	Training	Test
2k	43.92°C	54.40°C	33.54°C	42.05°C
4k	41.01°C	48.85°C	31.15°C	35.29°C
8k	34.44°C	43.72°C	25.58°C	32.64°C
20k	32.34°C	39.04°C	23.87°C	30.00°C
40k	31.14°C	37.86°C	23.11°C	28.54°C
270k	34.22°C	37.15°C	25.82°C	28.03°C

Table S3 RMSE of A1 using four different model structures.

Method	RMSE (test)
LSTM	39.04°C
bidirectional LSTM	40.34°C
GRU	40.67°C
CNN + LSTM	42.36°C
Transformer	43.43°C

Table S4 Characteristic weight in the section *"Do we need anything else to improve NLP-based Tm prediction?"*.

character	weight	character	weight
C	12	+	1
c	12	S	32
(	1	Br	80
)	1	#	22
1	1	4	23
O	16	/	24
=	1	o	25
2	1	s	32
N	14	I	127
[	1	P	31
]	1	Si	28
F	19	B	11
n	14	5	1
Cl	35.5	\\	1
@	1	6	1
3	1	7	1
H	1	8	1
-	1	p	31

Table S5 Character number defined in the model.

character	number	character	number
c	1	+	19
C	2	S	20
(	3	Y	21
)	4	#	22
1	5	4	23
O	6	/	24
=	7	o	25
2	8	s	26
N	9	I	27
[	10	P	28
]	11	Q	29
F	12	B	30
n	13	5	31
G	14	\\	32
@	15	6	33
3	16	7	34
H	17	8	35
-	18	p	36

References

1 D. Vidal, M. Thormann and M. Pons, *Journal of chemical information modeling*, 2005, **45**, 386-93.

2 I. V. Tetko, D. M. Lowe and A. J. Williams, *Journal of cheminformatics*, 2016, **8**, 2.

3 C. A. Bergström, U. Norinder, K. Luthman and P. Artursson, *Journal of chemical information computer sciences*, 2003, **43**, 1177-85.

4 M. Karthikeyan, R. C. Glen and A. Bender, *Journal of chemical information and modeling*, 2005, **45**, 581-90.

5 C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, *Journal of chemical information and modeling*, 2017, **57**, 1757-72.

6 O. Vinyals, A. Toshev, S. Bengio and D. Erhan, presented in part at Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

7 S. Hochreiter and J. Schmidhuber, *Neural computation*, 1997, **9**, 1735-80.

8 A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, *IEEE transactions on pattern analysis machine intelligence*, 2008, **31**, 855-68.

9 J. Schmidhuber, *Neural networks*, 2015, **61**, 85-117.

10 I. Sutskever, O. Vinyals and Q. V. Le, presented in part at Advances in neural information processing systems, 2014.

11 A. Graves, A.-r. Mohamed and G. Hinton, presented in part at 2013 IEEE international conference on acoustics, speech and signal processing, 2013.

12 M. Schuster and K. K. Paliwal, *IEEE Transactions on Signal Processing*, 1997, **45**, 2673-81.

13 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., presented in part at Neural Information Processing Systems, 2017.

14 T. Domhan, presented in part at Annual Meeting of the Association for Computational Linguistics, 2018.

15 Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le and R. Salakhutdinov, presented in part at Annual Meeting of the Association for Computational Linguistics, 2019.