Data-driven Chemical Reaction Classification with Attention-Based Neural Networks

Philippe Schwaller
IBM Research – Zurich, Switzerland
DCB, University of Bern, Switzerland
phs@zurich.ibm.com

Alain C. Vaucher
IBM Research – Zurich, Switzerland
ava@zurich.ibm.com

Vishnu H. Nair
IBM Research – Zurich, Switzerland
har@zurich.ibm.com

Teodoro Laino
IBM Research – Zurich, Switzerland
teo@zurich.ibm.com

Abstract

Organic reactions are usually clustered in classes that collect entities undergoing similar structural rearrangement. The classification process is a tedious task, requiring first an accurate mapping of the rearrangement (atom mapping) followed by the identification of the corresponding reaction class template. In this work, we present two transformer-based models that infer reaction classes from the SMILES representation of chemical reactions. The first model, a sequence-2-sequence model, reaches an accuracy of 93.8 % for a multi-class classification task involving several hundred different classes. Initial results show that the second model, a BERT classifier, is also able to achieve a high accuracy (95.3%) on this classification task. The attention weights provided by the sequence-2-sequence model gives an insight into what parts of the SMILES strings are taken into account for classification, based solely on data. We study the incorrect predictions of this model and show that it uncovers different biases and mistakes in the underlying data set.

1 Introduction

Name reactions[1] play a crucial role in the language of organic chemists. They represent an efficient way to communicate what a chemical reaction does, how it is performed in the laboratory, or how it works in terms of atomic rearrangements. It is for this reason that name reactions are currently used to navigate large databases of reactions, to retrieve similar members of the same reaction class with the purpose of helping chemists to analyze and infer optimal reaction conditions. Today, several hundreds of name reactions exist in the RXNO ontology [1]. Quite often their name honors who discovered that chemical reaction or who refined earlier known transformations raising their popularity. An example is the Friedel-Crafts reaction, named after Charles Friedel and James Mason Crafts, who discovered the catalytic effect of aluminum chloride in electrophilic substitutions. Name reactions can also be named after the reaction type, using the initials or referring to structural features.

In the last decade, computer-based systems [2, 3, 4, 5, 6] became an important asset available to chemists for reaction prediction tasks. Nevertheless, the knowledge of the name reaction of a predicted outcome has a great value for expert chemists to assess the quality of the prediction. For this reason, the demand for robust algorithms to categorize chemical reactions is high. The current

[1] For convenience, in this work "name reaction" refers to reaction classes that have an established name in chemistry, and not only to reactions that carry the name of the discoverer(s).

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

state-of-the-art in reaction classification is represented by commercially available tools [7, 8], which classify reactions based on a library of expert-written rules. These tools typically make use of SMIRKS [9], a language to describe transformations in the simplified molecular-input line-entry system (SMILES) format [10]. Classifiers based on machine learning have the potential to increase the robustness to noise in the reaction equations and to avoid the explicit formulation of rules. Among the few attempts made to infer name reaction using AI-based technologies, we report the work of Schneider et al.[11], in which authors developed a reaction classifier based on reaction fingerprints. Unfortunately, the limited set of reaction classes used (only 50 most important ones) makes it difficult to judge how this algorithm would perform on a more comprehensive set.

Here, we use a labelled set of chemical reactions as ground truth to train a FastText classifier [12] and two transformer-based models as architecture [13, 14]. The ground truth data is composed of chemical transformation represented as SMILES, and its labelling (classification) was performed using NameRXN [7]. Instead of relying on the formulation of specific rules and on the need to have every reaction properly atom-mapped, our model learns the atomic motifs that differentiate reactions belonging to different classes. We show that the transformer-based sequence-2-sequence model was able to match the ground-truth classification with an accuracy of 93.8%. The mismatches are mainly related to unrecognized reactions, some of which are correctly classified by our model. Moreover, the transformer-based architecture shows a very high level of robustness towards SMILES errors representation. We report cases where, despite an error in the converted molecules, our model was still able to understand the reaction that was originally described by chemists in the patent procedure text. We also observe that the encoder-decoder attention provides insights into the essence of the chemical transformation of the corresponding class, by providing larger values for the weights of the atoms involved in the reaction center. In an additional experiment we show that BERT classifier [14] is able to predict the ground truth class with an accuracy of 95.3%.

2 Data & Models

The data consisted of 2M reactions extracted from the Pistachio database [15] split into train, validation and test sets (90% / 5% / 5%). The reaction data is classified with NameRXN [7], a rule-based software that classifies more than 900 different reactions. The classification is organized in superclasses [16], reaction categories and name reactions according to the RXNO ontology [1]. For more detail on name reactions and their categories, we refer the reader to the work of Schneider et al. [17]. As commonly done, we represent the chemical reactions with SMILES (Simplified Molecular-Input Line-Entry System). We tokenize the reaction SMILES as in Schwaller et al. [5] without enforcing any distinction between reactants and reagents. Therefore, our method is universally applicable, including those reactions where the reactant-reagent distinction is subtle [18].

For the target of the transformer model, we split the class prediction into superclass, category and name reaction prediction. This means, for example, that the target string for the name reaction "[1.2.3]" would be "[1] [1.2] [1.2.3]". To have a baseline, we trained a supervised classifier using FastText [19, 12]. The second model is an autoregressive encoder-decoder transformer model [13]. We constructed a model with 4 encoder layers and 1 decoder layer. As the source and target are dissimilar, we did not share encoder and decoder embeddings. For the remaining hyperparameters, we used the same as were used for the training of the Molecular Transformer [5, 20], which is state-of-the-art in chemical reaction prediction.

One of the major recent advancement in natural language processing is BERT [14], which compared to the seq-2-seq architecture only consists of a transformer encoder with specific heads that can be fine-tuned for different tasks such as multi-class prediction. In an additional experiment, we pretrained a BERT using masked language modeling loss on the chemical reactions and fine-tuned on the name reaction classes. Compared with the hyperparameters of the BERT-Base model in [21], we decreased the hidden size to 256, the intermediate size to 512, and the number of attention heads to 4. For the pretraining we set 500k steps with a learning rate of 1e-4 and a maximum sequence length of 256, the rest of the parameters were kept as suggested in [21]. For the classification training, we only changed the learning rate to 2e-5 and kept the maximum sequence length of 256.

3 Results and Discussion

A summary of the results can be found in Table 1. We observe that a simple N-gram sentence classification model cannot capture the details of the reactions and is only able to correctly match the ground truth 68.4 % of the time. The transformer enc4-dec1 model matched the ground truth classification with 93.8%. The Reaction BERT classifier predicted the correct name reaction with an accuracy of 95.3%, therefore, achieving slightly better results than with the seq-2-seq approach. However, losing the ability of observing the hierarchy in the reaction classes. An elaborate comparison of the two models exceeds the scope of this paper and will be presented in a further study. Here we analyze the transformer enc4-dec1 model and its predictions in more detail.

Table 1: Classication results

Model	Validation Accuracy [%]	Test Accuracy [%]
fasttext default settings	41.4	41.6
fasttext 10-gram, dim 100	68.2	68.4
transformer enc4-dec1	93.7	93.8
BERT classifier	95.1	95.3

We identified different types of incorrect predictions by the transformer enc4-dec1 model. They are summarized in Table 2. Most errors are related to the "Unrecognized" class of the RXNO ontology. The most frequent error type is the prediction of a reaction class for a reaction classified as "Unrecognized" (64.1% of all incorrect predictions), and the second most frequent error type is predicting "Unrecognized" when a class should be predicted (21.7%). Out of the remaining errors, roughly a third predict an incorrect superclass (first number of the class string, 4.8%), a third predict an incorrect category (second number of the class string, 4.6%), and a third predict the incorrect name reaction (third number of the class string, 4.8%).

Table 2: Types of incorrect predictions of the transformer enc4-dec1 model on the test set

	Count	Percentage
Correctly predicted	104077	93.78%
Model predicts name reaction instead of "Unrecognized"	4419	3.98%
Model predicts "Unrecognized" instead of name reaction	1500	1.35%
Incorrect name reaction	333	0.30%
Incorrect superclass	330	0.30%
Incorrect category	317	0.29%

In Table 3, we show the reaction classes for which our model makes incorrect predictions most frequently. Due to the little information that can be gained for reactions classes with very few examples, we restricted this analysis to reactions with at least 20 occurrences in the test set. For half of these reaction classes (5 out of 10), the most common error source is the model failing to recognize the reaction class and predicting "Unrecognized". Four other cases involve reaction classes of the superclass 11, “Resolution reactions”. The classification of reactions in the superclass is understandably more complicated since, in general, they do not involve any atomic rearrangements as such. Other than that, the most frequent mistake is the prediction of the "Williamson ether synthesis" class instead of "Ether synthesis". Both reaction classes are related and result in the formation of an ether. The difference is that for the "Williamson ether synthesis" the ether is generated from an alcohol group of the reactant, while it is derived from a carbonyl group for the "Ether synthesis".

Although the large number of "Unrecognized" reactions in Pistachio makes an extensive (human) analysis difficult, a few dozen cases provide interesting insights. Part of the "Unrecognized" reactions, should actually belong to a name reaction. The rules-based systems fails in this assignment, which is instead is successfully completed by our model. This shows that a data-driven approach can be more robust than rule-based models. Another part belongs to examples occurring in multiple reactions. In this case, the reaction cannot be classified into a single name reaction, and our model predicts one of the corresponding reactions. An additional part corresponds to molecules that are incorrectly parsed in Pistachio. If the SMILES string of a molecule involved in the reaction was incorrectly

Table 3: Worst-predicted reaction classes with more than 20 occurrences in the test set for the transformer enc4-dec1 model.

Reaction class		Accuracy [%]	Most frequent incorrectly predicted class	
11.8.3	Chloride salt formation	32.0	11.9	Separation
1.7.13	Ether synthesis	45.5	1.7.9	Williamson ether synthesis
9.7.140	Defluorination	47.8	0.0	Unrecognized
11.7	Racemization	53.2	11.5	Isomerization
11.5	Isomerization	61.5	11.1	Chiral separation
2.5.2	Imidic ester + amine reaction	61.9	0.0	Unrecognized
9.7.148	Imine hydrolysis	64.5	0.0	Unrecognized
9.7.147	Deamination	69.4	0.0	Unrecognized
11.6	Purification	72.5	11.9	Separation
10.4.2	Methylation	72.7	0.0	Unrecognized

derived from the name, rule-based approaches fail to recognize the atomic rearrangements and thus to classify the reaction. For minor parsing errors, our model shows its potential, recognizing the correct transformation in several instances.

Figure 1 shows the attention vectors for four different chemical transformations. Similar to forward chemical reaction prediction [22, 5], we note that the larger weights are associated with the atoms that are part of the reaction center. Just like a human expects to see a certain group of atoms based on the name reaction, the decoder learned to focus on the atoms involved in the rearrangement to classify reactions.

4 Conclusion

In this work, we focused on the data-driven classification of chemical reactions using neural machine translation architectures.

Our transformer-based seq-2-seq model could learn the classification schemes using a broad set of chemical reactions as ground-truth labelled with the use of commercially available reaction classification tool. We match the rule-based classification with an accuracy of 93.8%. Out of the 6.2 % of incorrect predictions, 5.3 % are linked to the "Unrecognized" class of the underlying database. The model is able to learn the atomic environment characteristic of each class and provides a rationale easily interpretable by expert chemists. The possibility to understand the reasoning behind each classification may help the end-user chemists along the adoption process of these technologies.

To the best of our knowledge, this is the first work that applied a BERT-like pretraining [14] to chemical reactions. In terms of classification accuracy, with 95.3%, the BERT classifier performed slightly better than the transformer-based seq-2-seq model. The differences will be analyzed in further work. We expect that the Reaction BERT can also be fine-tuned for other tasks than reaction classification.

References

[1] RSC's RXNO Ontology, http://www.rsc.org/ontologies/rxno/index.asp. (Accessed Sep 13, 2019).

[2] Bartosz A Grzybowski, Kyle J M Bishop, Bartlomiej Kowalczyk, and Christopher E Wilmer. The 'wired' universe of organic chemistry. *Nature Chemistry*, 1(1):31–36, April 2009.

[3] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604, 2018.

[4] Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

[5] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, in press, 2019.

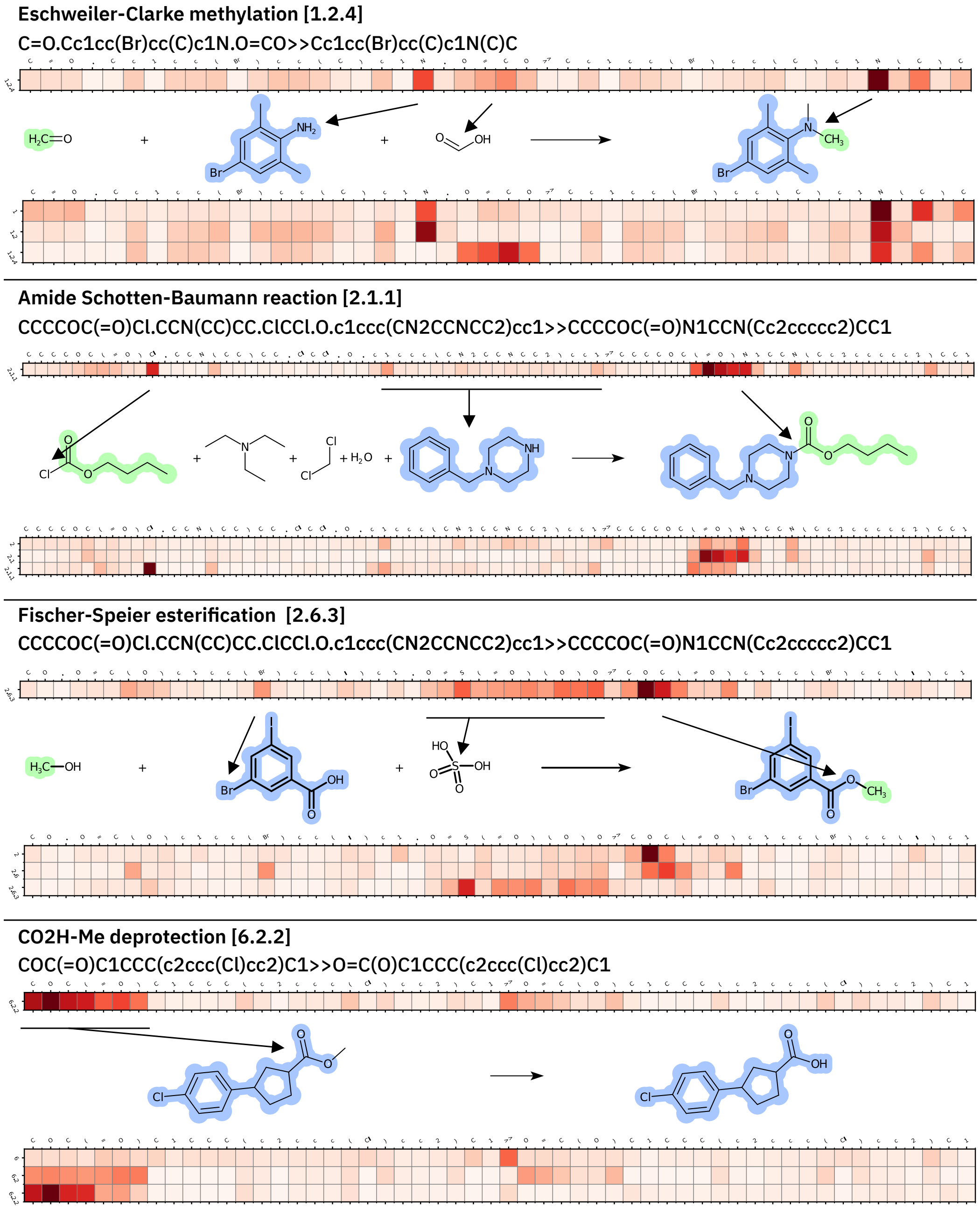
Eschweiler-Clarke methylation [1.2.4]
C=O.Cc1cc(Br)cc(C)c1N.O=CO>>Cc1cc(Br)cc(C)c1N(C)C
1.2.4
H2C=O
NH2
Br
OH
O
N
CH3
1
1.2
1.2.4
Amide Schotten-Baumann reaction [2.1.1]
CCCCOC(=O)Cl.CCN(CC)CC.ClCCl.O.c1ccc(CN2CCNCC2)cc1>>CCCCOC(=O)N1CCN(Cc2ccccc2)CC1
2.1.1
O
Cl
N
+ H2O
NH
2
2.1
2.1.1
Fischer-Speier esterification [2.6.3]
CCCCOC(=O)Cl.CCN(CC)CC.ClCCl.O.c1ccc(CN2CCNCC2)cc1>>CCCCOC(=O)N1CCN(Cc2ccccc2)CC1
2.6.3
H3C—OH
Br
OH
HO
S
O
CH3
2
2.6
2.6.3
CO2H-Me deprotection [6.2.2]
COC(=O)C1CCC(c2ccc(Cl)cc2)C1>>O=C(O)C1CCC(c2ccc(Cl)cc2)C1
6.2.2
Cl
O
OH
6
6.2
6.2.2

Figure 1: Encoder-decoder attention for few reaction examples predicted by the transformer enc4-dec1 model. SMILES tokens in dark red indicate parts of the molecules that were relevant for the model prediction. The upper attention vector represents the normalized sum of the three bottom vectors for the superclass, category and name reaction prediction.

[6] IBM RXN for Chemistry, https://rxn.res.ibm.com. (Accessed Sep 13, 2019).

[7] Nextmove Software nameRXN, http://www.nextmovesoftware.com/namerxn.html. (Accessed Jul 29, 2019).

[8] Hans Kraut, Josef Eiblmaier, Guenter Grethe, Peter Löw, Heinz Matuszczyk, and Heinz Saller. Algorithm for reaction classification. *Journal of chemical information and modeling*, 53(11):2884–2895, 2013.

[9] Daylight Theory Manual, Chapter 5. http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed may 25, 2014).

[10] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988.

[11] Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of chemical information and modeling*, 55(1):39–53, 2015.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[15] Nextmove Software Pistachio, http://www.nextmovesoftware.com/pistachio.html. (Accessed Jul 29, 2019).

[16] John S Carey, David Laffan, Colin Thomson, and Mike T Williams. Analysis of the reactions used for the preparation of drug candidate molecules. *Organic & biomolecular chemistry*, 4(12):2337–2347, 2006.

[17] Nadine Schneider, Daniel M Lowe, Roger A Sayle, Michael A Tarselli, and Gregory A Landrum. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry*, 59(9):4385–4402, 2016.

[18] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

[19] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[20] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[21] BERT code, https://github.com/google-research/bert#sentence-and-sentence-pair-classification-tasks. (Accessed Oct 15, 2019).

[22] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.