

Neural Machine Translation for Mathematical Formulae

ACL 2023

Felix Petersen, Moritz Schubotz, André Greiner-Petter, Bela Gipp

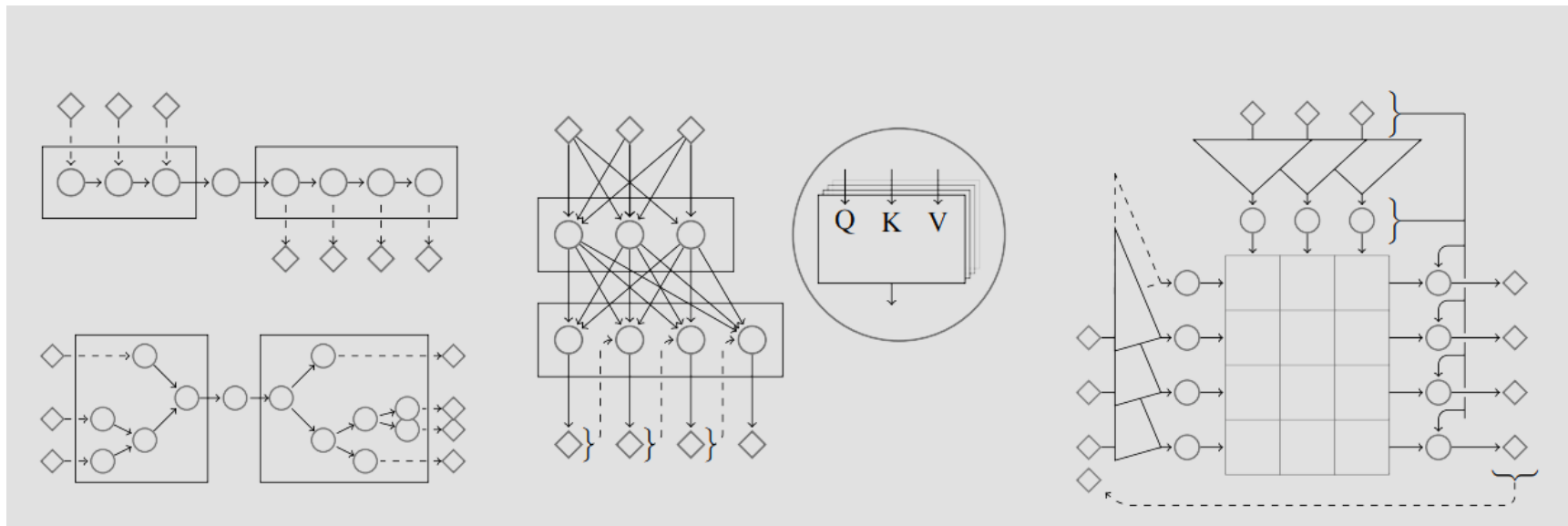
- Introduction
- Related Work
- Training Data Sets & Preprocessing
- Methods
- Evaluation of the Convolutional Network

- Introduction

$$(x)_n := x^{\overline{n}} = \prod_{k=0}^{n-1} (x + k).$$

$$(x)_n := x^{\underline{n}} = \prod_{k=0}^{n-1} (x - k).$$

- Related Work



- Training Data Sets & Preprocessing

- Mathematical Functions Site Data set
- Semantic LATEX Data Set
- Preprocessing

Data Set	Formulae	Input (\LaTeX)	Output (Mat. / <i>sem.</i> L.)
Mathematica	307 672	345.5 ± 534.4 (195)	320.7 ± 585.7 (168)
<i>semantic</i> \LaTeX	11 639	163.8 ± 246.2 (116)	145.6 ± 230.1 (103)

- Methods

- RNN
- Recursive NN
- Transformer
- Convolutional Seq-to-Seq Networks

• Evaluation of the Convolutional Network

- Exact Match (EM) Accuracy
- Levenshtein Distance (LD)
- BLEU
- Perplexity

$$\text{ppl}(p) = 2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$$

Perplexity의 경우 수학기공식 번역에서 실 번역 성능 간 불일치 발생. 일반적으로 perplexity가 낮을 수록 성능이 잘 나오는 것으로 해석. 그러나 Validation set에 대한 perplexity가 학습 진행에 따라 오히려 증가함. 이는 perplexity가 모델 번역 정확도 보다는, 예측의 확신도를 반영하기 때문.

- Evaluation Techniques

- Evaluation Techniques
- Round trip experiments
- DLMF (Digital Library of Mathematical Functions)

• Evaluation Teechniques

- Mathematica \leftrightarrow LaTeX 변환:
95.1% EM, 99.68 BLEU (논문 모델)
vs. 2.7% EM (Mathematica)
- LaTeX \rightarrow sLaTeX 변환:
90.7% EM, 96.79 BLEU (논문 모델)
- im2latex-100k round-trip 평가:
0.698% EM (논문 모델)
vs. 0.153% EM (Mathematica)

Method	EM	Import	LD _{≤ 5}	LD
Mathematica	0.153%	15.3%	2.30%	18.3
Conv. Seq2Seq	0.698%	16.3%	2.56%	12.9

Table 4

Table 2: Main results for the back-translation.

Metric	LaTeX \rightarrow Mathematica	LaTeX \rightarrow <i>semantic</i> LaTeX
EM	95.1%	90.7%
BLEU	99.68	96.79

Table 3: Comparison between Mathematica and our model on back-translating the formulae of the Mathematical Functions Site data set. Import denotes the fraction of formulae that can be imported by Mathematica, i.e., whether Mathematica can import the LaTeX format or whether our model produces valid Mathematica syntax, respectively.

Method	EM	Import	LD _{≤ 5}	LD
Mathematica	2.7%	88.5%	16.4%	88.7
Conv. Seq2Seq	95.1%	98.3%	96.7%	0.615

- Qualitative Analysis

- Equation B.1
 $\sigma_k(n)$ 번역 성공
- Equation B.2
 \wp (\wp)
 $\wp-1$ 를 제대로 해석
- Equation B.7
적분기호 정확히 변환

- 브라켓(bracket) 불일치 사례
- \wedge (Wland)등의 기호 변환오류
- DLMF에서 일부 기호가 Mathematica-export 어휘에 없는 토큰이 포함되어 있으므로 모델은 이를 해석할 수 없는 사례
- 하지만 vocab제한 상황에서도 일부의 사례에서 감마 함수와 오일러 e를 정확히 예측하였음

- Limitations

- 학습데이터 편향
- 도메인 특화 문제
- 데이터셋 확장 문제
- 문맥 정보 활용 문제