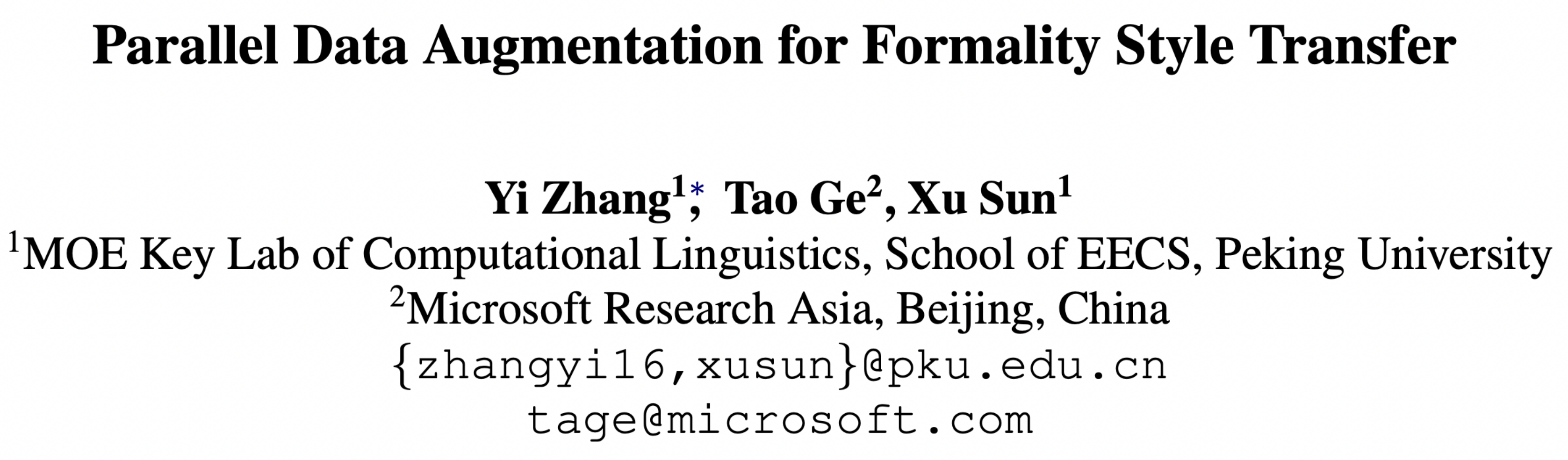
RELATED WORK

ACL 2020



Parallel Data Augmentation for Formality Style Transfer

Yi Zhang[1]*, Tao Ge[2], Xu Sun[1]
[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Microsoft Research Asia, Beijing, China
{zhangyi16,xusun}@pku.edu.cn
tage@microsoft.com

SUMMARY

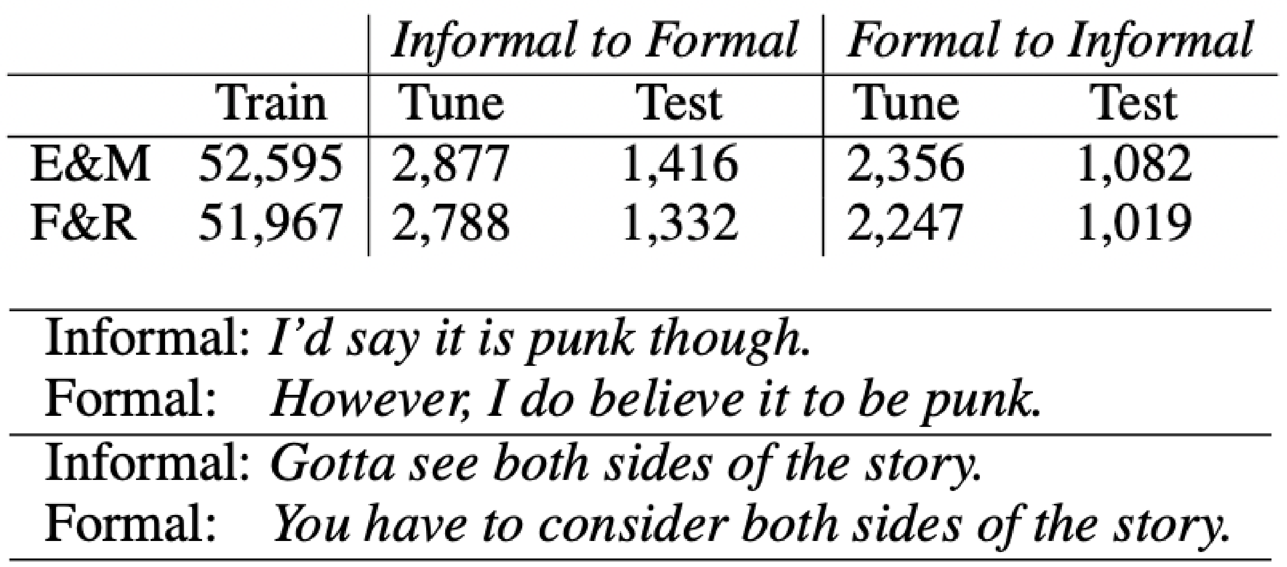
Zhang et.al present a formality test transfer model that experiments on three different types of domain. To help improve the model's generalization ability and reduce the overfitting risk, they focused on methods on augmenting the data into large amount of parallel datasets.
Three types of data augmentation methods used were back translation, formality discrimination, and multi-task transfer. Similar to machine translation, data augmentation were explored using seq2seq model.

NOVEL CONTRIBUTION

❶ Address the issue of lack of parallel data in NLP task, especially text style transfer
❷ Presents a solution regarding the issue
❸ Demonstrates different methods to augment the data for models to improve performance
❹ Evaluates own procedures to better exploit the augmented data
❺ Further inspire others in the field of text style transfer

DATA

Our data is **GYFAC** (Grammarly's Yahoo Answers Formality Corpus) introduced in *'Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Be nchmarks and Metrics for Formality Style Transfer'* (ACL, 2018) which provides a parallel corpus between informal and formal sentences.
It consists of 2 different domains:
1. Entertainment & Music (E&M)
2. Family Relations (F&R)
each with following number of sentences:



		Informal to Formal		*Formal to Informal*	
	Train	Tune	Test	Tune	Test
E&M	52,595	2,877	1,416	2,356	1,082
F&R	51,967	2,788	1,332	2,247	1,019

Informal: *I'd say it is punk though.*
Formal: *However, I do believe it to be punk.*

Informal: *Gotta see both sides of the story.*
Formal: *You have to consider both sides of the story.*

Baseline

Only consider the original GYFAC corpus.
❶ train dataset size : 104,562
❷ validation dataset size : 4,603
❸ test dataset size : 2,101

Augmented data

Augment the original corpus through different methods:
(synonym replacement, round-trip translation)
❶ train dataset size : 209,124

자연어처리 Proposal [prahdazi]
통계학과 2019150419 기다연
컴퓨터학과 2019320062 이지수

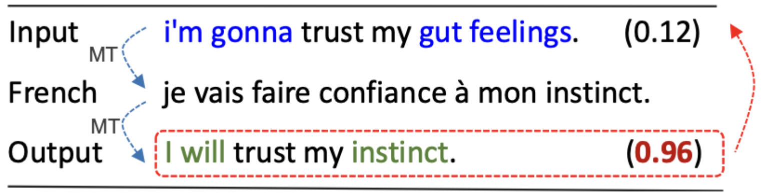
METHOD

01 Back Translation (BT)

: generate source language sentences from target monolingual sentences
❶ Feed formal sentences to the pre-trained model with original parallel data
❷ Generate informal counterparts of sentences
❸ Pair up {fed formal sentence - generated informal sentece} as augmented data

02 Formality Discrimination (F-Dis)

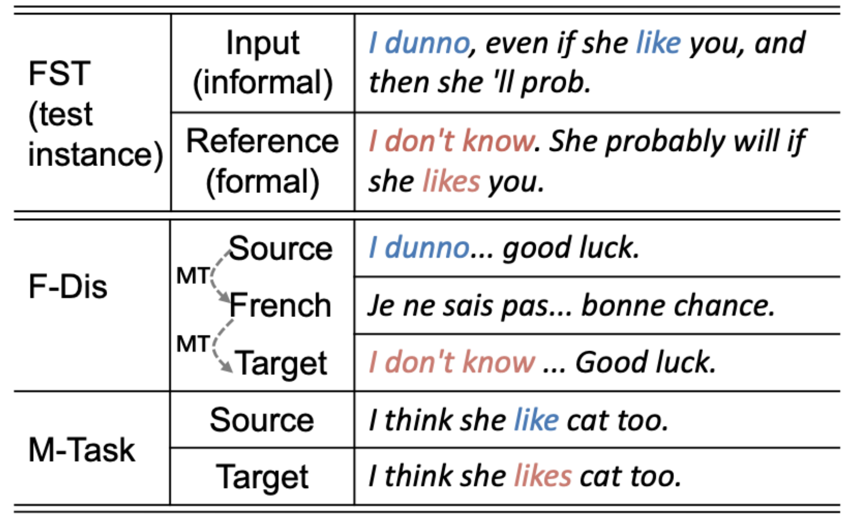
: round-trip translate informal sentences by MT models trained with formal text to generate formal rewrites



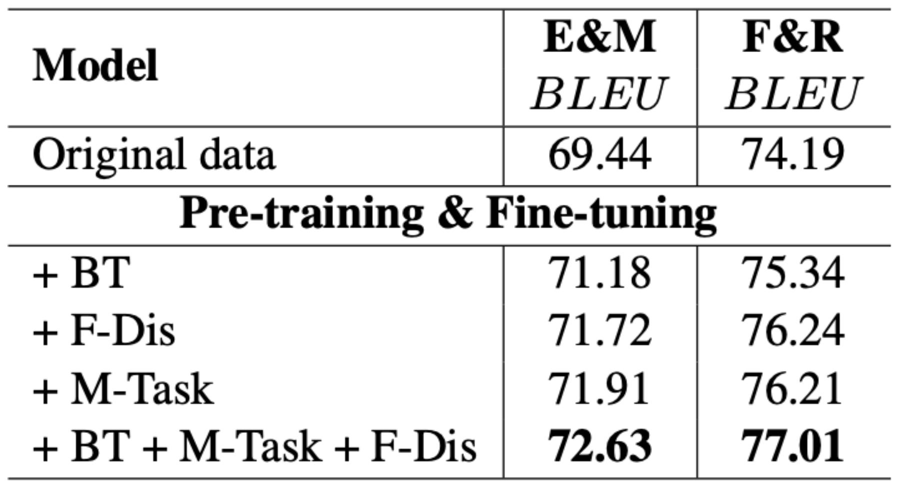
Input i'm gonna trust my gut feelings. (0.12)
French je vais faire confiance à mon instinct.
Output I will trust my instinct. (0.96)

03 Multi-task transfer (M-Task)

: transfer knowledge from **GEC** (Grammatical Error Correction) task to **FST** (formality style transfer) to improve formality of informal sentences
(mostly, informal sentences are deemed to be mor grammatically incorrect)



FST (test instance)	Input (informal)	*I dunno, even if she like you, and then she 'll prob.*
	Reference (formal)	*I don't know. She probably will if she likes you.*
F-Dis	Source	*I dunno... good luck.*
	MT French	*Je ne sais pas... bonne chance.*
	MT Target	*I don't know ... Good luck.*
M-Task	Source	*I think she like cat too.*
	Target	*I think she likes cat too.*



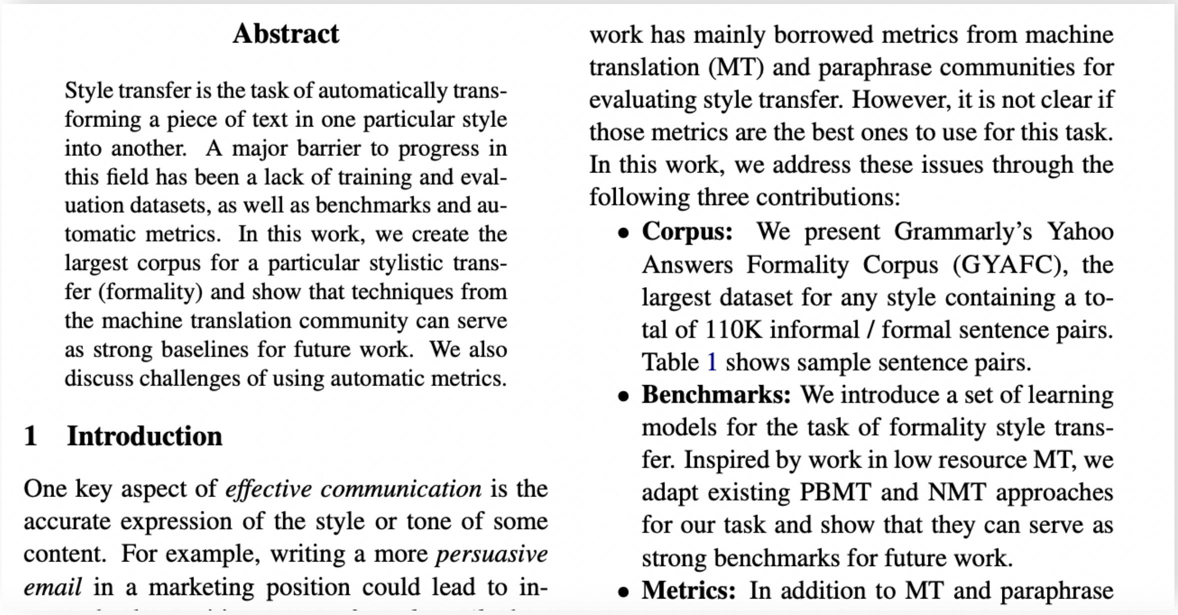
Model	E&M *BLEU*	F&R *BLEU*
Original data	69.44	74.19
Pre-training & Fine-tuning		
+ BT	71.18	75.34
+ F-Dis	71.72	76.24
+ M-Task	71.91	76.21
+ BT + M-Task + F-Dis	**72.63**	**77.01**

IDEA

Our main idea is to **transfer text style from academic text to spoken text**.
It will be mainly used when preparing for speech interviews based on your CV or preparing for a presentation with research papers.

Contributions of our project will be:
❶ Add beam-search decoder and attention to basic sequence-to-sequence model
❷ Analyze the effects of data augmentation (with synonyms)
❸ Compare the effects of attention variants:
dot, general, concatenate scoring function

Academic Text



Abstract

Style transfer is the task of automatically transforming a piece of text in one particular style into another. A major barrier to progress in this field has been a lack of training and evaluation datasets, as well as benchmarks and automatic metrics. In this work, we create the largest corpus for a particular stylistic transfer (formality) and show that techniques from the machine translation community can serve as strong baselines for future work. We also discuss challenges of using automatic metrics.

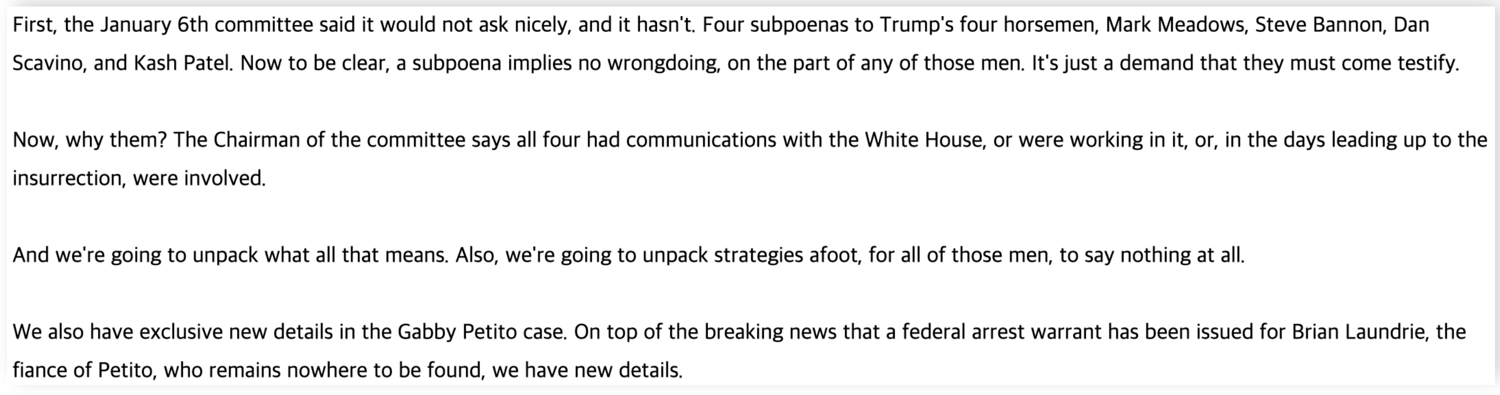
1 Introduction

One key aspect of *effective communication* is the accurate expression of the style or tone of some content. For example, writing a more *persuasive email* in a marketing position could lead to in-

work has mainly borrowed metrics from machine translation (MT) and paraphrase communities for evaluating style transfer. However, it is not clear if those metrics are the best ones to use for this task. In this work, we address these issues through the following three contributions:

- **Corpus:** We present Grammarly's Yahoo Answers Formality Corpus (GYAFC), the largest dataset for any style containing a total of 110K informal / formal sentence pairs. Table 1 shows sample sentence pairs.
- **Benchmarks:** We introduce a set of learning models for the task of formality style transfer. Inspired by work in low resource MT, we adapt existing PBMT and NMT approaches for our task and show that they can serve as strong benchmarks for future work.
- **Metrics:** In addition to MT and paraphrase

Interview/Presentation



First, the January 6th committee said it would not ask nicely, and it hasn't. Four subpoenas to Trump's four horsemen, Mark Meadows, Steve Bannon, Dan Scavino, and Kash Patel. Now to be clear, a subpoena implies no wrongdoing, on the part of any of those men. It's just a demand that they must come testify.

Now, why them? The Chairman of the committee says all four had communications with the White House, or were working in it, or, in the days leading up to the insurrection, were involved.

And we're going to unpack what all that means. Also, we're going to unpack strategies afoot, for all of those men, to say nothing at all.

We also have exclusive new details in the Gabby Petito case. On top of the breaking news that a federal arrest warrant has been issued for Brian Laundrie, the fiance of Petito, who remains nowhere to be found, we have new details.

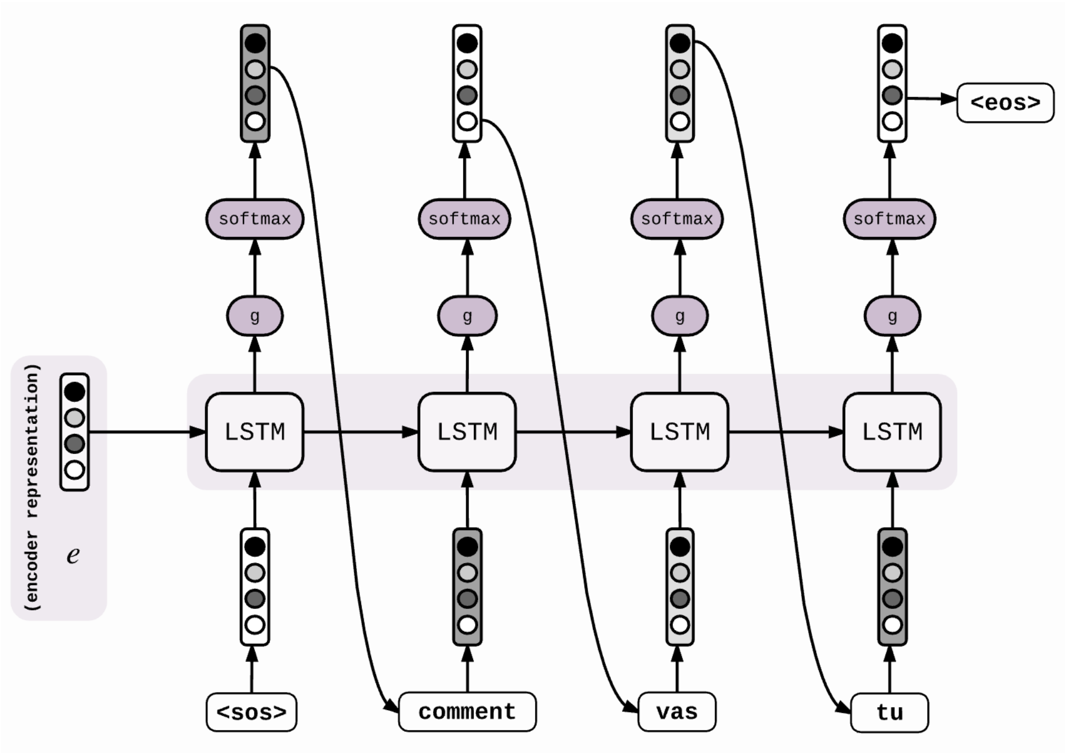
자연어처리 Proposal [prahdazi]
통계학과 2019150419 기다연
컴퓨터학과 2019320062 이지수

PROPOSAL

01 [MODEL]

Seq2seq with attention

❶ Design the econder, decoder
- Hidden Layer : LSTM
- Beam search decoding



❷ Add attention

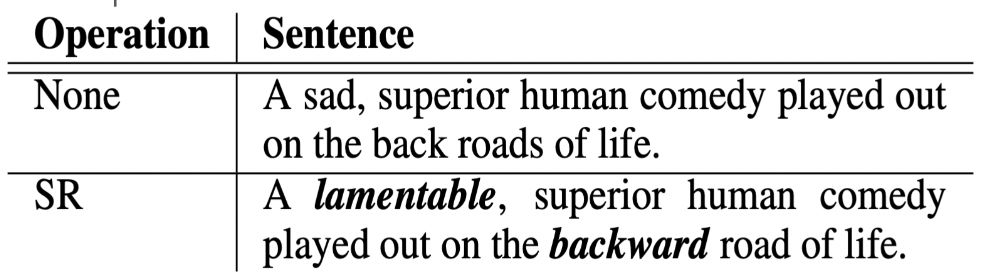
$$\alpha_{t'} = f(h_{t-1}, e_{t'}) \in \mathbb{R} \qquad \text{for all } t'$$
$$\bar{\alpha} = \text{softmax}(\alpha)$$
$$c_t = \sum_{t'=0}^{n} \bar{\alpha}_{t'} e_{t'}$$

02 [DATA AUGMENTATION]

❶ Baseline model
- GYAFC benchmark dataset (110k)
  (1) Entertainment & Music
  (2) Family & Replationships

❷ Synonym replacement
- Substitute similar words based on
  (1) WordNet (155k)
  (2) PPDB (Over 220M)

Example

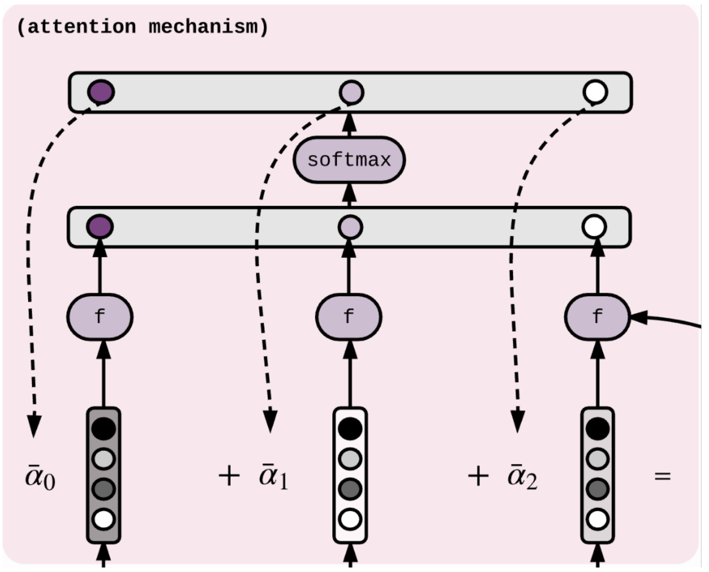


Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.

❸ Round trip translation
- Two parallel corpus
- Train two models for Machine Translation
  (1) English → Korean
  (2) Korean → English

03 [OUTPUT] Scoring Function

❶ Attention variants
- Compute with 3 different functions



$$f(h_{t-1}, e_{t'}) = \begin{cases} h_{t-1}^T e_{t'} & \text{dot} \\ h_{t-1}^T W e_{t'} & \text{general} \\ v^T \tanh(W[h_{t-1}, e_{t'}]) & \text{concat} \end{cases}$$

❷ BLEU (Bilingual Evaluation Understudy)
- N-gram precision

$$BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)$$