Going Beyond Word Matching: Syntax Improves In-context Example Selection for Machine Translation

Chenming Tang **Zhixiang Wang** **Yunfang Wu***
National Key Laboratory for Multimedia Information Processing, Peking University
MOE Key Laboratory of Computational Linguistics, Peking University
School of Computer Science, Peking University
{tangchenming@stu, ekko@stu, wuyf@}pku.edu.cn

Abstract

In-context learning (ICL) is the trending prompting strategy in the era of large language models (LLMs), where a few examples are demonstrated to evoke LLMs' power for a given task. How to select informative examples remains an open issue. Previous works on in-context example selection for machine translation (MT) focus on superficial word-level features while ignoring deep syntax-level knowledge. In this paper, we propose a syntax-based in-context example selection method for MT, by computing the syntactic similarity between dependency trees using Polynomial Distance. In addition, we propose an ensemble strategy combining examples selected by both word-level and syntax-level criteria. Experimental results between English and 6 common languages indicate that syntax can effectively enhancing ICL for MT, obtaining the highest COMET scores on 11 out of 12 translation directions.

1 Introduction

In the era of LLMs, ICL has become a popular prompting strategy to elicit the power of LLMs on a wide range of tasks. In ICL, a few demonstrations are given in the input context during inference while not involving parameter tuning (Dong et al., 2022; Min et al., 2022).

As a major natural language processing (NLP) task, there have been several works exploring in-context example selection strategy for MT. Agrawal et al. (2023) propose R-BM25 on the basis of BM25 (Robertson et al., 1994) to enhance word overlap. M et al. (2023) propose CTQScorer combining multiple features. Zhang et al. (2023) conduct a comprehensive study on prompting LLMs for MT and claim that templates, number of examples, features and the quality of example databases all matter for ICL on MT.

In previous studies, for both statistical MT and neural MT, syntax plays an important role to improve model performance (Williams and Koehn, 2014; Wu et al., 2017). However, in case of in-context example selection for MT, previous works focus on superficial word-level matching (like BM25 and R-BM25) or combining multiple straightforward features (like CTQScorer). To the best of our knowledge, no research dig out syntax-level features for ICL demonstrations. To this end, we propose a syntax-based in-context example selection strategy for MT, which selects examples most similar to the test source in syntax based on similarity of dependency trees. In addition, word-level and syntax-level features complement each other and combining both would further elicit LLMs' power on MT. Therefore, we propose an ensemble strategy, concatenating examples selected by BM25 and our syntax-based strategy.

Experimental results between English and 6 common languages indicate that syntax helps find better in-context examples and thus improves LLMs' ability in MT. Comparing with various baselines, our proposed methods obtain the highest COMET scores on 11 out of 12 translation directions.

Our contributions can be summarized as follows:

- For the first time, we propose a novel syntax-based in-context example selection strategy for MT.
- We present a simple but effective ensemble strategy to combine in-context examples selected from different criteria, taking advantage of both superficial word overlapping and deep syntactic similarity.
- We prove that syntax is effective in finding informative in-context examples for MT. We call on the NLP community not to ignore the significance of syntax when embracing LLMs.

* Corresponding author.

2 Preliminary: Syntactic Similarity

Generally, the syntactic structures of sentences are represented by syntax trees, and thus it is natural to measure the syntactic similarity based on similarity or distance between syntax trees. Tree similarity can be measured by Edit Distance (de Castro Reis et al., 2004), Polynomial Distance (Liu et al., 2022), Tree Kernel (Collins and Duffy, 2002; Vishwanathan et al., 2004; Moschitti, 2006), etc.

Compared to other algorithms mentioned above, Polynomial Distance (Liu et al., 2022) is convenient to implement and relatively time-saving to run. So, in this work, we adopt Polynomial Distance between dependency trees to measure the syntactic similarity.

Concretely, given d as the number of dependency labels, which indicate the grammatical relations between dependents and their heads, we transform dependency trees into polynomials recursively based on two variable sets: $X = \{x_1, x_2 ... x_d\}$ and $Y = \{y_1, y_2, ... y_d\}$. Considering a node with Label l as n^l, its corresponding polynomial is represented as:

$$P(n^l) = \begin{cases} x_l, & n^l \text{ is leaf,} \\ y_l + \prod_{i=1}^{k} P(n_i), & n^l \text{ is non-leaf,} \end{cases} \quad (1)$$

where $n_1, ..., n_k$ are all child nodes of n^l if it is non-leaf. Then, given a root node r, we obtain the polynomial representing the whole tree by computing $P(r)$. A detailed demonstration from Liu et al. (2022) can be found in Appendix A.

Next, we compute the distance between polynomials. Note that the polynomial representing a dependency tree can be described term by term. For each term $cx_1^{e_{x_1}} x_2^{e_{x_2}} ... x_d^{e_{x_d}} y_1^{e_{y_1}} y_2^{e_{y_2}} ... y_d^{e_{y_d}}$ in the polynomial, where e_{x_i}, e_{y_i} and c are the exponent of variable x_i, y_i and the coefficient of the term respectively, we denote it as a term vector:

$$t = [e_{x_1}, e_{x_2}, ..., e_{x_d}, e_{y_1}, e_{y_2}, ..., e_{y_d}, c].$$

In this way, a polynomial P can be written as a set of term vectors $\mathcal{V}_P$. Then, we compute the distance between two polynomials (P and Q) as:

$$d(P,Q) = \frac{\sum_{s \in \mathcal{V}_P} \min_{t \in \mathcal{V}_Q} \| s - t \|_1 + \sum_{t \in \mathcal{V}_Q} \min_{s \in \mathcal{V}_P} \| s - t \|_1}{| \mathcal{V}_P | + | \mathcal{V}_Q |}, \quad (2)$$

where $\| s - t \|_1$ is the Manhattan distance (Craw, 2017) between term vector s and t.

3 Method

3.1 ICL for MT

In recent years, ICL has been proved effective for improving LLMs' performance on various tasks without training or finetuning (Brown et al., 2020; Von Oswald et al., 2023).

To design instructions for MT, we first inform the language pair and the template of presenting the result. Then, we provide several in-context examples selected using our proposed strategy. Last, the source sentence of the test sample is concatenated at the end. An example of our prompt is shown in Figure 1, and we draw inspiration from the work of Agrawal et al. (2023).

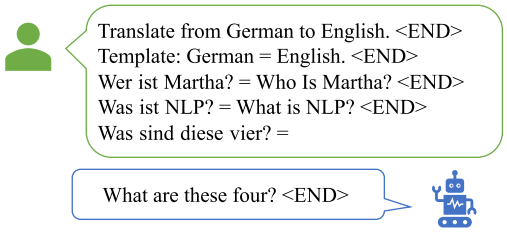

Figure 1: A 2-shot example of our prompt template.

3.2 Syntax-based Example Selection

We parse all our datasets with spaCy (Honnibal et al., 2020) to get dependency trees. Appendix B lists spaCy models we use for different languages.

For each test sample, we compute its Polynomial Distance from each instance of the example database on the source side, and then select the top-k most similar examples. The selected examples serve as demonstrations for ICL, as shown in Figure 1.

3.3 Ensemble of Example Selection

We hypothesize that a combination of word-level closeness and syntax-level similarity would make the LLM generate results good on both counts. So we propose an ensemble strategy, where examples selected by different selection strategies are concatenated. In this work, we explore the ensemble of BM25 and Polynomial Distance, where half of the final examples are from BM25 and the other half are from Polynomial Distance.

4 Experimental Settings

4.1 Large Language Model

Following Agrawal et al. (2023), we adopt $\text{XGLM}_{7.5\text{B}}$ (Lin et al., 2022) for all our experi-

ments, which is a decoder-only multilingual generative language model supporting 30 languages. $XGLM_{7.5B}$ has 32 layers, a hidden dimension of 4096 and 7.5B parameters in total.

4.2 Datasets and Evaluation Metrics

Test Set We perform our evaluation on the *devtest* set of FLORES+ [1] (Costa-jussà et al., 2022), which has 1012 sentences for around 200 languages. We experiment between English and 6 common languages including German, Spanish, French, Japanese, Russian and Chinese.

Example Database We adopt WikiMatrix v1 (Schwenk et al., 2021) as our example database, which has 135M parallel sentences for 1620 language pairs. Detailed statistics of WikiMatrix for each language can be found in Appendix C.

Evaluation Metrics We report COMET (Rei et al., 2020) from *wmt20-comet-da*, which is considered a superior metric for MT nowadays (Kocmi et al., 2021). As a complement, we report BLEU from sacreBLEU (Post, 2018) in Appendix D.

4.3 Pre-processing

Tokenization We tokenize Chinese with Jieba [2], Japanese with Mecab [3] and other languages with Sacremoses [4].

Cleaning We remove sentences longer than 100 or shorter than 4 tokens and those cannot be identified as their corresponding languages by Langid.py [5]. Sentence pairs with a source/target length ratio exceeding 1.5 are also removed. We have around 85% sentences remaining after preprocessing for all languages except Japanese, of which the percentage is 67%. See Appendix C for detailed statistics.

4.4 Baselines and Comparisons

Random: We report the average result of 3 different random seeds. **BM25:** The BM25 scores are computed using Rank-BM25 (Brown, 2020). **R-BM25:** We evaluate R-BM25 [6] (Agrawal et al., 2023) on our datasets.

Polynomial: Our syntax-based selection method with Polynomial Distance. **BM25 + Polynomial:** The first 4 examples are the top-4 from BM25 and the rest 4 are the top-4 from Polynomial. **Polynomial + BM25:** The first 4 are from Polynomial and the rest 4 are from BM25.

[1] https://github.com/openlanguagedata/flores
[2] https://github.com/fxsjy/jieba
[3] https://github.com/SamuraiT/mecab-python3
[4] https://github.com/hplt-project/sacremoses
[5] https://github.com/saffsd/langid.py
[6] https://github.com/sweta20/inContextMT

5 Results and Analysis

5.1 Main Results

Experimental results on all 6 languages (into and out of English) are shown in Table 1, where the number of in-context examples is set to 8.

Without the help of word-level closeness, Polynomial itself cannot reach perfection in many cases. However, it performs competitively on DE-EN translation, outperforming Random by 1.58 points and BM25 by 3.53 points.

With a combination of word-level closeness and syntax-level similarity, our ensemble strategies take the lead in most cases, outperforming Random by around 1 point on average.

To sum up, the highest scores of 11 out of 12 translation directions are achieved by our proposed methods, which indicates the effectiveness of syntax in in-context example selection for MT.

Surprisingly, the performance of R-BM25 is poor under our experimental settings. This might be due to differences in example database, data preprocessing and the design of prompt templates. We leave exploration of the cause to future work.

We also compare with CTQScorer (M et al., 2023) in Table 2, where the number of examples is set to 4 and we focus on only 3 language pairs , to be in line with their work. Our proposed methods secure all the highest scores on all 6 translation directions. Note that M et al. (2023) use Europarl (Koehn, 2005) and ParaCrawl (Bañón et al., 2020) as example database, and experiment on FLORES-101, which is an earlier version of FLORES+. These factors may lead to an unequal comparison.

5.2 Different Numbers of Examples

We carry out experiments under different numbers of in-context examples. The average COMET scores of our ensemble strategy and other baselines on all 12 translation directions under different numbers of examples are shown in Figure 2. Please refer to Appendix E for our full results. Our proposed ensemble strategy constantly outperforms baselines under different numbers of examples.

Direction	Selection	DE	ES	FR	JA	RU	ZH	Avg.
Into EN	Random	63.57	63.84	71.96	38.03	54.39	45.74	56.26
	BM25	61.62	63.96	70.75	40.83	53.45	48.15	56.46
	R-BM25	56.92	62.39	71.40	39.15	50.81	44.02	54.12
	Polynomial	**65.15**	64.38	72.64	39.51	**54.75**	42.39	56.47
	BM25 + Polynomial	63.24	**65.27**	**73.00**	**41.51**	54.20	47.07	**57.38**
	Polynomial + BM25	62.78	64.12	71.65	41.03	53.93	**48.20**	56.95
Out of EN	Random	44.42	51.92	55.10	**22.92**	48.75	12.46	39.26
	BM25	44.60	53.41	56.10	19.97	51.96	12.17	39.70
	R-BM25	42.09	52.14	51.39	12.24	49.46	3.84	35.19
	Polynomial	44.18	52.70	55.52	18.94	47.51	9.09	37.99
	BM25 + Polynomial	44.45	54.13	55.35	20.46	**52.87**	**12.68**	39.99
	Polynomial + BM25	**44.86**	**54.79**	**56.43**	22.19	51.60	12.10	**40.33**

Table 1: COMET scores for translation into (the top half) and out of (the bottom half) English with 8 examples. The highest scores are in **bold** text.

Direction	Selection	DE	FR	RU	Avg.
Into EN	CTQ (M et al., 2023)	64.77	71.28	50.85	62.30
	Random	63.08	71.84	54.09	63.00
	BM25	61.12	72.64	54.19	62.65
	R-BM25	58.24	70.96	50.79	60.00
	Polynomial	**65.23**	72.42	**54.36**	**64.00**
	BM25 + Polynomial	64.30	**73.16**	53.76	63.74
	Polynomial + BM25	63.92	72.34	53.32	63.19
Out of EN	CTQ (M et al., 2023)	38.05	41.41	44.26	41.24
	Random	41.92	54.01	47.08	47.67
	BM25	40.11	52.12	49.75	47.33
	R-BM25	34.30	43.27	43.59	40.39
	Polynomial	42.16	52.44	47.07	47.22
	BM25 + Polynomial	**42.73**	**54.74**	47.93	**48.47**
	Polynomial + BM25	42.59	52.73	**49.82**	48.38

Table 2: COMET scores for translation between DE, FR, RU and EN with 4 examples. The highest scores are in **bold** text.

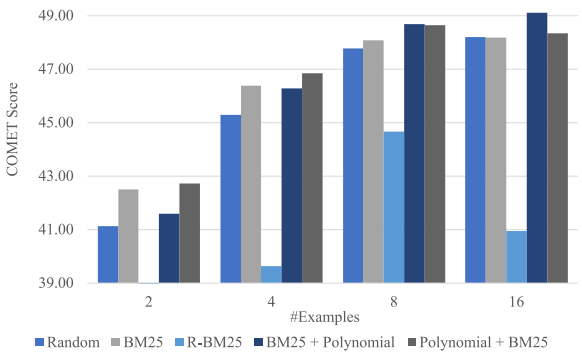

Figure 2: Average COMET scores on all 12 translation directions under different numbers of in-context examples (2, 4, 8, and 16). The score of R-BM25 under 2 examples is 34.56, which we do not show in the chart to save space. Note that when there are 16 in-context examples, the context sometimes exceeds the LLM's maximum length, which may hurt the precision of the result.

Input	International sanctions have meant that new aircraft cannot be purchased.
BM25	They cannot be purchased on board buses.
Polynomial	The CJLS has stated that this particular ceremony should not be performed.

Table 3: Top-1 examples of EN-DE translation selected by BM25 and Polynomial.

5.3 Case Analysis

We give an instance to compare the examples selected by different strategies, as shown in Table 3. BM25 selects sentences with more word overlapping ("cannot be purchased") while Polynomial selects sentences with the similar syntactic structure (perfect tense, "that" clause, negative sentence, passive voice). For closely related language pairs like EN-DE, when the source sentences share similar syntax or patterns, it is likely that the target sentences also share these features. See Appendix F for an instance of model outputs.

6 Conclusion

In this work, we investigate whether syntactic information can help find better in-context examples for MT. We propose selecting examples based on similarity of dependency trees, and present a simple but effective ensemble method by selecting examples from both word-overlap-based and syntax-similarity-based selection. Our proposed methods obtain the highest COMET scores on 11 out of 12 translation directions, indicating that injecting syntax information during in-context example selection is helpful for MT. We call on the NLP community to pay more attention to syntactic knowledge for syntax-rich tasks like MT.

Limitations

First, due to limited time and computational resources, we have not evaluated our methods on low-resource languages and other LLMs. Second, in our ensemble strategy, we have only explored combining word-level selection and syntax-level selection. Other types of selection strategies (e.g., selection based on semantics) are not explored in the ensemble. Third, we have not explored methods based on constituent trees and other algorithms of tree similarity besides Polynomial distance. Last, our selection is based on the similarity on the source side. However, similarity on the target side or between two sides is also worth trying.

Ethics Statement

Task	Average Time
Pre-processing	$\sim$ 2 hrs
Dependency Parsing	$\sim$ 1.5 hrs
BM25 Selection	$\sim$ 5 hrs
Polynomial	$\sim$ 8 hrs
LLM Inference	$\sim$ 1hr

Table 4: Average computation time on all languages.

Computational Budget We run pre-processing and in-context example selection on Intel® Xeon® Gold 5218 CPU and the LLM's inference on NVIDIA A40. Table 4 shows the average computation time.

Reproducibility All the experiments are completely reproducible since our selection methods are deterministic and sampling is disabled during LLM generation.

Artifact	License
spaCy	MIT
Jieba	MIT
Mecab	GPL, LGPL, BSD
Sacremoses	MIT
Rank_BM25	Apache-2.0
XGLM	MIT
COMET	Apache-2.0
sacreBLEU	Apache-2.0
FLORES+	CC-BY-SA-4.0
WikiMatrix	BSD

Table 5: Licenses of scientific artifacts we use.

Scientific Artifacts We cite all the creators of scientific artifacts we use in this paper. Licenses of these scientific artifacts are shown in Table 5.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 263–270, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Susan Craw. 2017. Manhattan distance. *Encyclopedia of Machine Learning and Data Mining*, pages 790–791.

Davi de Castro Reis, Paulo Braz Golgher, Altigran Soares da Silva, and Alberto H. F. Laender. 2004. Automatic web news extraction using tree edit distance. In *The Web Conference*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengyu Liu, Tinghao Feng, and Rui Liu. 2022. Quantifying syntax similarity with a polynomial representation of dependency trees. *arXiv preprint arXiv:2211.07005*.

Aswanth M, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining multiple features for in-context example selection for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning: ECML 2006*, pages 318–329, Berlin, Heidelberg. Springer Berlin Heidelberg.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

SVN Vishwanathan, Alexander Johannes Smola, et al. 2004. Fast kernels for string and tree matching. *Kernel methods in computational biology*, 15(113-130):1.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Philip Williams and Philipp Koehn. 2014. Syntax-based statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Doha, Qatar. Association for Computational Linguistics.

Shuangzhi Wu, M. Zhou, and Dongdong Zhang. 2017. Improved neural machine translation with source syntax. In *International Joint Conference on Artificial Intelligence*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

A Demonstration of Polynomial Distance

The original demonstration of converting dependency trees to polynomials taken from Liu et al. (2022) is shown in Figure 3.

B The spaCy Models We Use for Parsing

The spaCy models we use for dependency parsing are listed in Table 6.

C Data Statistics

Statistics on the size of WikiMatrix for each language are shown in Table 7.

D BLEU Results

BLEU scores of translation with 8 in-context examples are shown in Table 8.

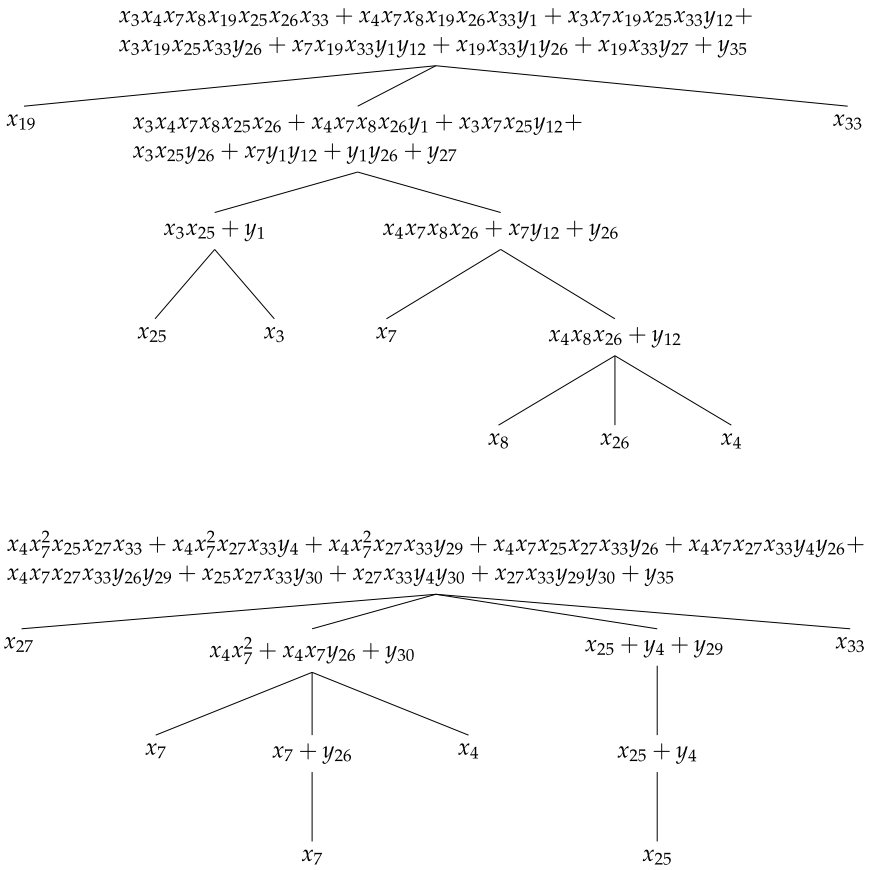

Figure 3: Original demonstration of converting dependency trees to polynomials taken from Liu et al. (2022).

Language	spaCy Model
DE	*de_core_news_sm*
EN	*en_core_web_sm*
ES	*es_core_news_sm*
FR	*fr_core_news_sm*
JA	*ja_core_news_sm*
RU	*ru_core_news_sm*
ZH	*zh_core_web_sm*

Table 6: The spaCy models used for different languages.

E Full Results under Different Numbers of Examples

COMET scores of translation with 2, 4, annd 16 in-context examples are shown in Table 9, 10 and 11, respectively. Note that when there are 16 in-context examples, the context sometimes exceeds the LLM's maximum length, which may hurt the precision of the result.

F Case Analysis on Model Output

In addition to Section 5.3, we display an example output with different selection strategies. In the case shown in Table 12, BM25 omits a constituent ("de plus" in the source sentence) while Polynomial translates not accurately enough. With the help of both word-level and syntax-level demonstrations, our ensemble strategy "BM25 + Polynomial" produces a relatively high-quality translation. This indicates that both word-level closeness and syntax-level similarity matters in in-context example selection for MT.

Language	ISO Code	#Pairs (M) after Pre-processing	#Pairs (M) before Pre-processing	Percentage
German	DE	1.33	1.57	85%
Spanish	ES	2.84	3.38	84%
French	FR	2.44	2.76	88%
Japanese	JA	0.57	0.85	67%
Russian	RU	1.39	1.66	84%
Chinese	ZH	0.68	0.79	86%

Table 7: Size of WikiMatrix before and after pre-processing and the percentage of reserved sentences after pre-processing for 6 languages. Each entry refers to the data between English and the corresponding language.

Direction	Selection	DE	ES	FR	JA	RU	ZH	Avg.
Into EN	Random	38.60	29.60	40.35	18.26	31.21	21.71	29.96
	BM25	38.50	**31.15**	41.20	19.05	32.16	22.32	30.73
	R-BM25	36.93	29.88	40.33	18.61	30.82	21.65	29.70
	Polynomial	39.20	29.62	40.88	18.40	31.34	21.10	30.09
	BM25 + Polynomial	39.37	30.57	**41.70**	**19.21**	**32.34**	**22.68**	**30.98**
	Polynomial + BM25	**39.40**	30.76	41.54	19.11	31.90	22.03	30.79
Out of EN	Random	28.02	24.16	35.87	12.21	27.18	13.21	23.44
	BM25	28.75	24.78	**37.89**	13.23	28.37	13.35	**24.40**
	R-BM25	27.38	24.26	35.88	10.98	27.32	11.33	22.86
	Polynomial	28.15	24.03	36.41	12.12	27.06	12.06	23.31
	BM25 + Polynomial	28.76	24.65	37.40	13.13	**28.49**	**13.64**	24.35
	Polynomial + BM25	**28.86**	**25.13**	37.32	**13.28**	28.25	13.47	24.39

Table 8: BLEU scores for translation into and out of English. The highest scores are in **bold** text.

Selection	Into EN						Out of EN						Avg.
	DE	ES	FR	JA	RU	ZH	DE	ES	FR	JA	RU	ZH	
Random	62.06	63.43	71.66	22.68	53.29	20.00	33.63	50.86	**50.31**	**13.62**	43.86	**7.98**	41.12
BM25	60.66	**64.68**	71.44	**38.50**	52.93	**36.54**	31.25	50.92	49.20	3.73	45.94	4.17	42.50
R-BM25	53.59	62.89	68.28	31.75	49.61	33.13	18.78	41.72	39.45	-12.95	38.97	-10.52	34.56
Polynomial	**63.97**	63.69	72.04	27.55	53.81	23.53	27.05	49.46	48.84	6.20	42.96	-1.36	39.81
BM25 + Polynomial	63.71	64.51	72.24	32.79	**53.83**	25.46	**34.89**	50.49	48.86	0.60	**46.81**	4.98	41.60
Polynomial + BM25	62.40	64.30	**72.49**	33.24	52.55	33.09	34.53	**51.93**	49.79	7.46	44.30	6.52	**42.72**

Table 9: COMET scores for translation with 2 in-context examples.

Selection	Into EN						Out of EN						Avg.
	DE	ES	FR	JA	RU	ZH	DE	ES	FR	JA	RU	ZH	
Random	63.08	64.29	71.84	30.94	54.09	35.04	41.92	51.06	54.01	**19.85**	47.08	**10.17**	45.28
BM25	61.12	64.67	72.64	**40.49**	54.19	**45.34**	40.11	53.63	52.12	14.33	49.75	8.22	46.38
R-BM25	58.24	62.37	70.96	34.52	50.79	38.05	34.30	47.95	43.27	-2.57	43.59	-5.82	39.64
Polynomial	**65.23**	64.53	72.42	33.88	**54.36**	36.71	42.16	51.53	52.44	15.55	47.07	5.97	45.15
BM25 + Polynomial	64.30	**65.23**	**73.16**	39.93	53.76	37.67	**42.73**	**53.69**	**54.74**	13.90	47.93	8.31	46.28
Polynomial + BM25	63.92	64.96	72.34	39.45	53.32	44.62	42.59	53.04	52.73	18.02	**49.82**	7.28	**46.84**

Table 10: COMET scores for translation with 4 in-context examples.

Selection	Into EN						Out of EN						Avg.
	DE	ES	FR	JA	RU	ZH	DE	ES	FR	JA	RU	ZH	
Random	64.25	63.83	71.47	38.80	54.26	46.84	45.33	51.38	55.65	**24.39**	49.04	13.00	48.19
BM25	60.72	62.70	70.14	40.42	52.05	**47.90**	46.31	53.87	56.15	21.01	52.69	14.21	48.18
R-BM25	55.36	59.05	68.97	34.97	47.21	42.51	37.48	47.53	48.52	0.89	47.93	0.82	40.94
Polynomial	**64.42**	63.76	**72.09**	40.86	**54.35**	45.48	44.85	52.20	55.57	21.09	49.21	7.52	47.62
BM25 + Polynomial	63.26	**64.07**	70.82	**41.58**	53.99	47.43	**46.39**	54.23	57.43	22.61	**52.90**	14.61	**49.11**
Polynomial + BM25	61.20	63.82	69.73	38.49	52.38	46.93	46.09	**54.55**	**58.52**	21.26	51.89	**15.08**	48.33

Table 11: COMET scores for translation with 16 in-context examples.

Source (FR)	De plus, le Centre d'alerte des tsunamis dans le Pacifique a déclaré qu'il n'y avait aucun signe de tsunami.
Reference (EN)	Also the Pacific Tsunami Warning Center said that there was no Tsunami indication.
BM25	The Pacific Tsunami Warning Center said there was no sign of a tsunami.
Polynoimal	In addition, the Pacific Tsunami Warning Center has declared there is no sign of a tsunami.
BM25 + Polynomial	In addition, the Pacific Tsunami Warning Center said there was no sign of a tsunami.

Table 12: Outputs of FR-EN translation under different selection strategies.