多言語ゼロショット学習における推論言語に関する分析

大平颯人 [1] 金輝燦 [2] 小町守 [3]
[1] 東北大学 [2] 東京都立大学 [3] 一橋大学
souto@komachi.live kim-hwichan@ed.tmu.ac.jp mamoru.komachi@r.hit-u.ac.jp

概要

現在， Large Language Model (LLM) における in-context learning に関する研究が活発に行われている．in-context learning に用いるプロンプトが適切に設計されていれば， LLM は様々なタスクを実行することができる．本研究では，そのフレームワークを拡張することによって定義される Multilingual Large Language Model (MLLM) の ゼロショット学習に注目する．現在，使用する言語ごとに，どの言語をプロンプトのテンプレートとバーバライザーの言語に用いるかでタスクの性能が大きく異なることが分かっている．そこで本研究では，事前学習時の各言語のデータ量や英語との言語的近さの観点から，この違いの原因を分析する．

1 はじめに

GPT-3 のような LLM において in-context learning は幅広いタスクに適用することができる．これらの利点としては， LLM のパラメータの更新を必要としないため，パラメータ全体を微調整する従来の方法と比較して，遥かに計算機の訓練コストが低いことである [1, 2]．しかし， GPT-3 は非英語のテキストはほとんど含まれていない英語中心の訓練データで学習されており，英語に特化したモデルである．近年では，言語バランスを考慮し，より多言語のデータを用いて学習した Multilingual LLM (MLLM) が開発されている [3, 4]．MLLM は英語中心のデータで学習された LLM と比較して，非英語における in-context learning を高精度に行うことができる．

この事前学習時の言語バランスによってプロンプトに様々な言語の組み合わせを用いることで， MLLM の推論能力を引き出すことが考えられる．実際に Lin ら [3] によって， MLLM の in-context learning において，テンプレート言語およびバーバライザー言語の組み合わせが精度を大きく左右することが示されている．また本研究ではテンプレート言語とバーバライザー言語が同じ場合のみを扱うため，以下ではこれらに使われる言語 2 つをまとめて**推論言語**と呼ぶことにする．**多言語ゼロショット学習**とは，推論言語とターゲット言語に異なる言語を用いたゼロショット学習のことである．

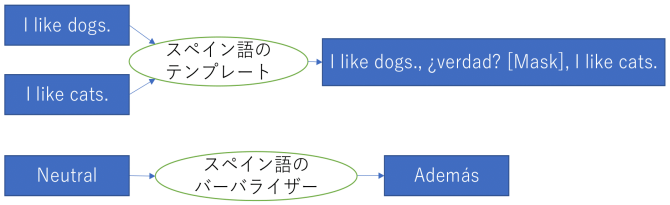

図 1 多言語テンプレートとバーバライザーの例

直感的には，上記に示したように推論言語にターゲット言語と同じ言語を使うことが最適であると考えられる．しかし，全ての設定でそれが最適であるとは限らないことが示されている．Lin ら [3] は XGLM を用いて，多言語ゼロショット学習を行った．Lin らは推論言語にターゲット言語と同じ言語を用いた設定と，ターゲット言語によらず英語を推論言語として用いた設定で実験を行った．その結果， 特定のタスクにおいて推論言語にターゲット言語とは異なる英語を用いた方が大幅に精度が高くなるような言語が存在することが明らかになった．具体的には， XNLI タスクにおける中国語の評価では，中国語を推論言語とするよりも，英語を推論言語として用いた方が，10 ポイントほど精度が高い．また, これは事前学習時のデータ量や言語間の近さが関係している可能性があると指摘されている．

本研究では， MLLM を用いたゼロショット学習の推論において，各ターゲット言語に対し，様々な推論言語を用いた場合について分析し，指摘されていた指標と精度に相関があることを確かめることを目的とする．

具体的には以下の 3 つの問いを立て分析を行う．

1. 各ターゲット言語に対して，推論言語によって精度が大きく異なるという現象は，複数の

MLLM で同様に起こるものなのか
2. 事前学習データに含まれる言語比率と性能に相関があるか
3. ターゲット言語が，事前学習時に含まれる最もデータ量の多い英語と言語的に近いほど，英語を推論言語にすることで性能に高くなる傾向があるか

実験では， XGML [3]， mGPT [4] という 2 つの MLLM に対し XNLI [5] をタスクとして性能評価を行った．全ての MLLM で各ターゲット言語に対して，最適な推論言語が必ずしもターゲット言語ではないことを確認した．次に，モデルの事前学習時の各言語のデータ量とターゲット言語・推論言語ごとの性能の相関を見たが，明確な相関は得られなかった．最後に，推論言語をターゲット言語と英語の 2 つに絞り，2 つの観点（構造的，音韻論的）での英語との言語的な近さと性能の差の相関を測った．前者ではトルコ語を除けば，構造的に近いほど性能の差が縮まるといった結果が得られた．後者についても音韻論的に近いほど性能の差が縮まるといった結果が得られた．

2 研究課題

Brown ら [1] はゼロショット学習の一つであるフレームワークを提案し， LLM がタスク固有の微調整をしなくともある程度のタスク実行能力を獲得していることを示した．彼らが提案したフレームワークでは，ターゲット言語 t と同じ言語のテンプレート $\mathcal{T}_t$， バーバライザー v_t を用いる．ここで，テンプレート $\mathcal{T}_t$ はテスト事例 x_t を [MASK] トークンを含む穴埋め形式の文に変換する関数で [1]， バーバーライザー $v_t : \mathcal{Y} \to \mathcal{V}_t$ は各候補ラベル $y \in \mathcal{Y}$ を自然言語の単語，またはトークン $\mathcal{V}_t$ に対応づける関数である．このフレームワークでは， $\mathcal{T}_t(x_t)$ の [MASK] を $v_t(y)$ に置換する関数 $\mathcal{P}$ を用いて，以下のように定義される値を最大化する予測ラベル $\hat{y}$ を求める．

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}} \sigma(\mathcal{M}, \mathcal{P}(\mathcal{T}_t(x_t), v_t(y))). \quad (1)$$

ただし σ は $\mathcal{P}(\mathcal{T}_t(x_t), v_t(y))$ を入力した時のモデル $\mathcal{M}$ から得られる尤度を示す．

Lin ら [3] は $\mathcal{M}$ を XGLM として，このフレームワークを拡張し，ゼロショット学習を行った．Brown らとの相違点として， Lin らはターゲット言語 t と同じ言語のテンプレート $\mathcal{T}_t$，バーバライザー v_t を用いた設定に加えて，ターゲット言語によらず英語のテンプレート $\mathcal{T}_{\text{En}}$， バーバライザー v_{En} を用いた設定で実験を行った．

そこで本研究ではこのフレームワークを用いて，以下の 3 つの問いを立て分析を行う．

RQ1: 推論精度の言語依存の一般性 Lin ら [3] は XGLM を用いてテンプレート言語とバーバライザー言語によってターゲット言語における推論精度が大幅に変わることを示した．しかし， XGLM 以外の MLLM に関しては上記の現象が起こるかは示されていない．

本研究では，ターゲット言語と最適な推論言語が必ずしも一致しないという問題が， XGLM 特有の現象ではないことを示すために，複数の MLLM を用いて実験を行う．Lin らに倣い，各ターゲット言語 t のテンプレート $\mathcal{T}_t$， バーバライザー v_t を用いる設定と，英語をテンプレート言語とした $\mathcal{T}_{\text{En}}$ と英語をバーバライザー言語とした v_{En} を用いる設定で実験を行い性能を比較する．

RQ2: 事前学習データの言語比率との関係 Lin ら [3] によって，事前学習時のデータ内に多く含まれる言語が推論言語として有効に働く可能性があると指摘されている．そのため，ターゲット言語 t のデータが事前学習時のデータに十分に含まれる場合は，ターゲット言語のテンプレートとバーバライザーを用いて有効に働くことが考えられる．一方で，ターゲット言語のデータが乏しい場合，英語のような事前学習時のデータ内に多く含まれる言語のテンプレートとバーバライザーを使った方が良いと考えられる．

そこで，本研究では，各ターゲット言語 $t_1, ..., t_n$ の事前学習データに含まれる割合 $r = \{r_{t_1}, ..., r_{t_n}\}$ と各言語のテンプレート，バーバライザーを用いた場合の精度 $acc^t = \{acc^{t_1}_{t_1}, ..., acc^{t_n}_{t_n}\}$ を計測し，相関を測る．正の相関がある場合，事前学習データにターゲット言語が多く含まれるほど，ターゲット言語のテンプレート，バーバライザーを用いることで MLLM が高精度に推論を行えることが示される．

RQ3: 英語との言語的な類似性との関係 Lauscher ら [6] は，言語間転移においてソース言語とターゲット言語の関係性について分析を行い，両者が言語的に類似している場合や事前学習時のデータ量が多い言語ほど，多言語の転移学習を高

1) 例えば， XNLI タスクの場合，英語テンプレート $\mathcal{T}_{\text{EN}}$ は前提，仮説の二文 $x^{pre}_{\text{EN}}, x^{hyp}_{\text{EN}}$ を入力としてとり， “x^{pre}_{EN}, right? [Mask], x^{hyp}_{EN}” のような形式に変換する．

精度に行えることを示した．

そこで，本研究では，ターゲット言語の推論において，異なる言語のバーバライザーを用いることはある種の言語間転移と考え，英語と言語的に近いターゲット言語の推論を行う場合，英語のテンプレート，バーバライザーが有効に機能するという仮説を立て検証を行う．具体的には，英語のテンプレートとバーバライザーを用いた場合の精度 $acc_{t_i}^{\mathrm{En}}$ と各ターゲット言語 t_i 自身を用いた場合の精度 $acc_{t_i}^{t_i}$ を計算し，差分スコア $acc_{t_i}^{\mathrm{En}} - acc_{t_i}^{t_i}$ を求める．次に各ターゲット言語について，英語との言語間の近さを構造的，音韻論的観点で測り，最後に差分スコアとの相関を測る．

各ターゲット言語について，距離と差分に正の相関がある場合，英語と近いターゲット言語の推論を行うほど，英語のテンプレート，バーバライザーを使うことで高精度な推論を行うことができることを意味する．

3 実験

3.1 実験設定

本研究では，XGLM-1.7B [3][2]，mGPT-1.3B [4][3] の 2 つの MLLM を用いる．タスクとしては XNLI [5] を用いて，評価指標としては accuracy を採用する．本研究では英語 (En)，フランス語 (Fr)，スペイン語 (Es)，中国語 (Zh)，トルコ語 (Tr) の 5 つの言語に関して実験を行う．表 2 に実験で用いた各言語のテンプレートとバーバライザーを示す．これらは Lin ら [3] と同様のものである．

英語とターゲット言語の言語的近さを測るために Lang2Vec [7][4] を用いて，構造的，音韻論的な類似度の両方を測定し，accuracy との相関を観察した．Lang2Vec では WordNet などの様々なデータセットを用いて，特徴を機械学習で抽出し，その特徴の有無を 0 と 1 から成るベクトルで表現している．

3.2 実験結果

RQ1 XGLM (1.7B)，mGPT (1.3B) 2 つの MLLM に対して ゼロショット学習を用いて XNLI によって性能評価を行い，表 1 に結果を示した．2 つのモデルで似たような結果が得られたことから，おおよそ推論言語の性能の傾向はモデルに依らないことが分かった．このことは RQ2, RQ3 の結果を見るとより明確に分かる．

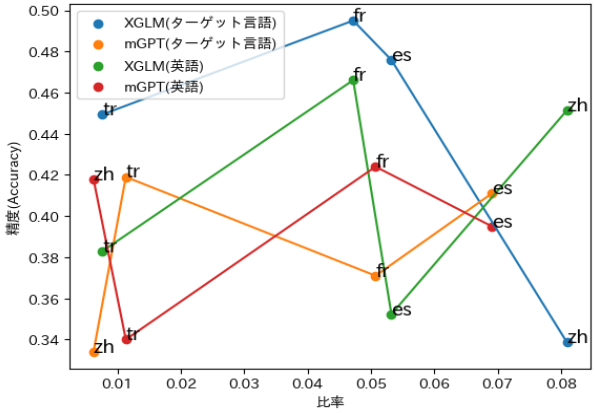

図 2 MLLM の事前学習データ量の比率とターゲット言語自身と英語を推論言語とした場合の精度

RQ2 図 2 に事前学習時のデータ量の比率と，ターゲット言語と同じ推論言語を用いた場合と，英語を推論言語とした場合の性能の関係をそれぞれ描いたが，両者に関して明確な相関は得られず， Lin ら [3] によって指摘されていた，事前学習時のデータ量と性能との相関の関係性が今回の結果では否定された．[5]

RQ3 英語との構造的な近さによって，英語を推論言語とした方が性能が高いのかを確かめた．その結果，図 3 が示しているように，トルコ語を除けば英語と構造的に近ければ近いほど性能が下がるといった，距離と性能の差の間に正の相関が得られた．また次に，音韻論的な近さに注目した実験を行った．その結果，図 4 が示しているように距離と性能の差に正の相関が得られた．

4 考察

RQ1 については表 1 の実験結果より， MLLM における多言語 in-context learning での高い性能を達成するような推論言語の選択方法を，モデルの種類に依らずに言語ごとに決められる可能性が示唆される．図 2 より，今回の場合は事前学習時のデータ量と性能との関係はモデルごとに違っていることが分かったが，これは mGPT の結果が事前学習時の全データ量ではなく，その一部である mC4 のデータ量に基づいた結果であるなどの理由が考えられる．

2） https://huggingface.co/facebook/XGLM-1.7B
3） https://huggingface.co/ai-forever/mGPT
4） https://github.com/antonisa/lang2vec

5） ただし，mGPT に関しては，事前学習時に mC4 と Wikipedia からのデータを用いていたが， Wikipedia からの詳細なデータ量が得られなかったため， 図 2 については， mC4 のデータ比率のみを考慮して描いている．

表 1　RQ1（推論精度の言語依存の一般性）の実験結果

	英語ラベル					ターゲット言語ラベル			
モデル	en	fr	es	zh	tr	fr	es	zh	tr
XGLM	0.478	0.466	0.352	**0.451**	0.383	**0.478**	**0.476**	0.339	**0.449**
mGPT	0.482	**0.424**	0.395	**0.418**	0.340	**0.424**	**0.411**	0.334	**0.419**

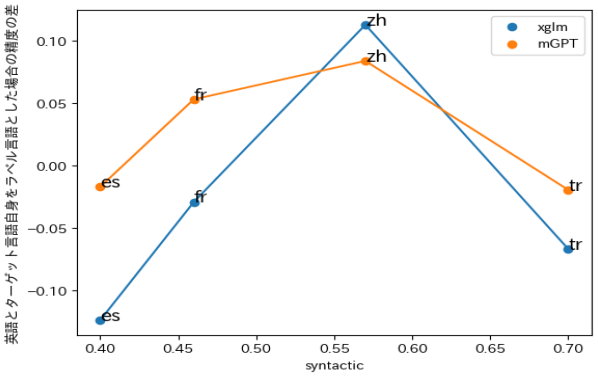

図 3　構造的な近さと英語およびターゲット言語を推論言語とした場合の精度の差

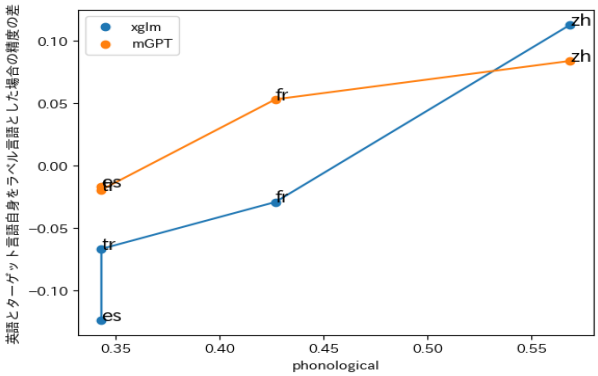

図 4　音韻的な近さと英語およびターゲット言語を推論言語とした場合の精度の差

一方，図 3，4 から分かるように，言語的近さの観点についてはおおよそモデルに依らない結果が得られたため，この観点は推論言語の選択の際に役立ちうると考えられる．このようなモデルに依らない結果が得られたのは，XGLM と mGPT が Brown ら [1] によって提案されたアーキテクチャーと訓練手順にどちらもほぼ従っているなどの事前学習時の共通点によると考えられる [3, 4].

また RQ2 については図 2 が示すように，事前学習時のデータ量と性能との間に明確な相関は見られなかった．Lin ら [3] によってモデルサイズが大きいほど多言語 few-shot in-context learning でのプロンプト内の例を効率的に活用できる可能性があるということが指摘されている．今回の場合はモデルサイズが小かったため，プロンプトによる多言語ゼロショット学習での言語間転移の恩恵が十分に得られなかった可能性が考えられる．

RQ3 の実験では図 3 と図 4 が示すように，構造的，音韻論的どちらの観点の距離に関しても，一部のデータ点（図 3 でのトルコ語）を除けば，英語を推論言語とした場合とターゲット言語自身を推論言語とした場合の性能の差について，正の相関が得られた．これらは小さいモデルではプロンプトによる言語間転移を十分に行えないために，英語と構造的，音韻論的に近い言語ほどターゲット言語との区別がつかず，ノイズとなってしまうなどの可能性が考えられる．

5　まとめ

今回は，MLLM の多言語ゼロショット学習での推論言語と性能との間の関係性が，モデルに依らず，事前学習時のデータ量や言語的近さの観点から説明できるのか確かめるために，様々な実験を行った．

まず得られた結果が，モデルに依らないということを確かめるために 2 つのモデルを用い，おおよそ性能の傾向はモデルに依らないということが確かめられた．

また，モデルの事前学習時のデータ量と英語またはターゲット言語自身をラベルに用いた場合の性能との関係を分析したが，今回の限られた実験設定では，この 2 つには目立った相関が得られず，モデル間で傾向が違っていた．

最後に構造的，音韻論的 2 つの言語的近さに注目して，これらと英語を用いたときにターゲット言語自身を用いた場合と比べた性能の関係を分析した．トルコ語を除けばどちらの場合でも英語に近いほど英語を用いた方が性能が下がるといった結果が得られた．

今後は様々なモデルや言語を扱い，テンプレート言語とバーバライザー言語を異なるものとした実験など様々な状況下で実験を行うことがより深い分析を行う上で重要になってくると考えられる．

参考文献

[1] 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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. **ArXiv**, 2021.

[3] 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. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[4] Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mgpt: Few-shot learners go multilingual, 2022.

[5] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics.

[7] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

表 2 MLLM によるゼロショット XNLI 実験で用いたテンプレート

Lang	Template	Candidate Verbalizer		
		Entailment	Contradiction	Neutral
En	{ Sentence 1}, right? [Mask], {Sentence 2}	Yes	No	Also
Fr	{ Sentence 1}, non? [Mask], {Sentence 2}	Oui	Non	De plus
Es	{ Sentence 1}, ¿verdad? [Mask], {Sentence 2}	Sí	No	Además
Zh	{ Sentence 1}[Mask], {Sentence 2}	由此可知,	所以, 不可能	同 时
Tr	{ Sentence 1}, değil mi? [Mask], {Sentence 2}	Evet	Hayır	Ayrıca