多言語評価極性判定における文法・語彙知識と生成モデルの統合

金山 博　　趙 陽　　大湖 卓也　　岩本 蘭
日本アイ・ビー・エム株式会社 東京基礎研究所
{hkana@jp., yangzhao@, ohkot@jp., ran.iwamoto1@}ibm.com

概要

本研究では、生成モデルを用いた多言語の評価極性判定の際に、構文や語彙の知識に基づく評価表現抽出器と組み合わせることにより、性能の改善を図る。極性を持つ語句の情報をプロンプトに含めることによって、大規模言語モデルにとって分類が難しい言語において特に正解率の向上が見られた。

1　はじめに

大規模言語モデル（LLM）による生成の手法は、トピック分類や評価極性判定などの分類タスクにおいても高い性能を示している [1, 14]。特に、LLM が持つ言語の汎化能力により、正解データに基づくタスク毎の訓練を行わなくても適切なプロンプトを与えることによって問題の解決ができる点で注目を集めている。しかし、対応できる言語は基盤となるモデルによって異なり、学習データの量が少ない言語においては性能が低くなる場合がある。特定の言語をカバーするために基盤モデルを再学習するには大きなコストがかかる上、他の言語の性能を下げてしまうこともある。最近はモデルの知識を編集する試みも盛んであるが、課題も多い [6]。

一方で、LLM に依存しない従来の手法は、タスク毎の学習や知識の構築が必要となるが、出力結果が安定することや、知識の追加などの制御が容易であるといった利点がある。機械学習に依存しない規則に基づく手法も、実応用の局面では価値がある [2]。

本稿では、図 1 のように、多言語の評価極性判定において、構文や語彙の知識に基づいた抽出器の出力をプロンプトに組み込むことにより、LLM による生成の改善を試みる。これによって、対応する言語の拡充や、分野適応などの制御が可能な仕組みを目指すとともに、複数の LLM の特性や、構文や語彙を考慮した知識の補完について検証・考察する。

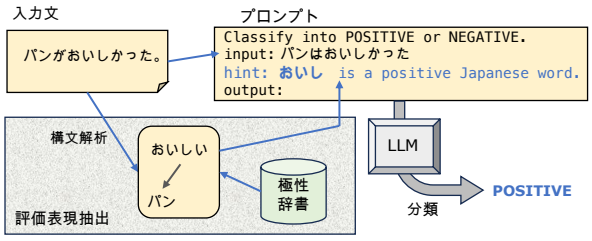

図 1　評価表現抽出器を援用したプロンプトに基づく評価極性判定の概念図。

2　関連研究

2.1　多言語評価表現の抽出

本研究では、著者らが過去に開発した、構文解析と語彙に基づく多言語の評価極性表現の抽出器 UDSA[4] を用いる。これは、Universal Dependencies (UD) [13] の構文構造の共通性を用いて多言語の処理の共通化を図り [15]、各言語の語彙知識の獲得や評価を効率化させたものである [16]。

UDSA は最初に、UD に準拠した構文解析器によって、単語区切り・品詞タグ付け・単語の標準形の取得を経て、入力文を係り受け構造に変換する。その構造を主辞側から辿り、言語毎の辞書項目と照合して、極性を持つ表現を抽出する [5]。これらの操作は、UD の構造を活かして言語共通化を図った構文規則 [15] に基づいている。文中に現れる語句を単純に辞書引きするだけではないため、「おいしいものが欲しい」は検出の対象としない、「おいしいとは思わない」は「おいしい」の極性を反転させるなどの操作ができる。極性表現の辞書として、英語のリソースをもとに翻訳や単語埋め込みの近さを使って整備した多言語の辞書を用いる [16]。

UDSA では、入力文中に含まれる極性表現の構造を、適合率を重視して抽出する。一文中から複数の極性が異なる表現を抽出する場合もあれば、明らか

な極性を持つものでなければ検出をしない。従って 3 節以降で述べる「文単位の極性判定」を直接解決するものではない点に注意されたい。

2.2 推論を用いた極性判定の改良

LLM で評価極性を判定する際に、手掛かりとなる語やそれらを用いた推論を段階的に考えさせるプロンプト「CARP」が提案されている [9]。CARP では、LLM に対するプロンプトを通じて、まず極性を持つ語句を抽出し、それらが文全体の極性に影響する理由を述べさせることにより、多くのデータセットにおいて極性判定の正解率を向上させた。

本研究では、LLM 自体に手掛かりを生成させるのではなく、外部のツールを用いて極性判定のヒントをプロンプトに与える。これによって、LLM が充分な知識を持たない言語に対応させたり、辞書等のリソースを自由に拡充できるようにして、LLM と既存ツールの統合を試みる。

3 生成モデルを用いた文極性判定

本節では、LLM での生成により文の極性を判定する方法と、そこに UDSA によって極性表現を抽出した結果を反映させる手法について示す。

3.1 プロンプトの作成

入力文を `POSITIVE` または `NEGATIVE` の極性に分類するタスクを解くために、言語モデルに対して以下の英語のプロンプトにより指示を与える。正解事例は用いずに 0-shot で判定をさせる。

```
Classify the next [言語] sentence into POSITIVE
or NEGATIVE. Please just answer the label.
input: [入力文]
output:
```

これにより、`output:` の後に `POSITIVE` または `NEGATIVE` が生成されることを期待する。[**言語**] は “Japanese” “Arabic” などの言語名で埋められる。なお、言語モデルによってプロンプトを一部改変することがあり、それは 4.2 節にて説明する。

3.2 UDSA との統合

UDSA が極性を持つ表現を検出した場合、それをプロンプトにヒントとして加える。`input:` の前に以下の行を加えて、モデルに語彙の知識を与える。

```
hint: [語句] is a [言語] word which have a
[極性] meaning.
```

[**語句**] は極性が検出された部分の表層形、[**極性**] は “positive” または “negative” である。例として、中国語の「我們很失望，...」という文からは、UDSA の解析により “`hint: 失望 is a Chinese word which have a negative meaning.`” というヒントが加えられる。複数の極性表現が抽出された場合には複数行のヒントを加える。

4 実験の設定

4.1 データセット

多言語の評価極性判定のデータセットとして、ここでは Parallel Sentiment [1] を用いる。これは 19 言語のパラレル UD コーパス (PUD) の各 1000 文のうち、極性を持つ 106 文に言語共通のアノテーション（P または N）を付与したものである。今回はそのうち UDSA が対応している 15 言語で実験を行う。

4.2 言語モデル

パラメータの規模、エンコーダの有無、学習時の言語などの多様性を鑑み、4 つの LLM を用いる。

- flan-t5-xxl [3]: エンコーダ・デコーダの T5 モデルを 1800 種のタスク (60 言語をカバー) でインストラクションチューニングしたもの（パラメータ数：11b）。
- flan-ul2 [12]: 上記同様、UL2 モデルをインストラクションチューニングしたもの（20b）。
- mt0-xxl [7]: 多言語 T5 モデルを、45 言語＋プログラムコードでインストラクションチューニングしたもの（13b）。
- llama-2-70b-chat [10]: デコーダのモデル Llama を対話用にチューニングしたもの（70b）。事前学習データの 9 割は英語が占める。

mt0-xxl のみ、極端に `POSITIVE` を出力する傾向が見られたので、プロンプトに以下の記述を追加することにより、ベースの正解率を上げた。

```
In this case all input is either of them,
and do not hesitate to select NEGATIVE when
you think the input is relatively
negative things in some way.
```

1) https://lrec2020.lrec-conf.org/en/shared-lrs/ にて公開されている。

表 1　極性判定の正解率 (%) の 15 言語の平均値。太字は各行での最大値、下線は全体での最大値を示す。

言語モデル	パラメータ数	LLM のみ	プロンプト		UDSA 優先	
			Stanford	Trankit	Stanford	Trankit
LLM 無		54.7			74.5	**75.2**
flan-t5-xxl	11b	75.3	80.4	80.5	**81.9**	81.8
flan-ul2	20b	78.9	79.7	79.4	**83.1**	82.7
mt0-xxl	13b	69.6	76.5	76.5	**80.8**	80.4
llama-2-70b-chat	70b	92.7	**93.6**	93.3	93.5	93.2

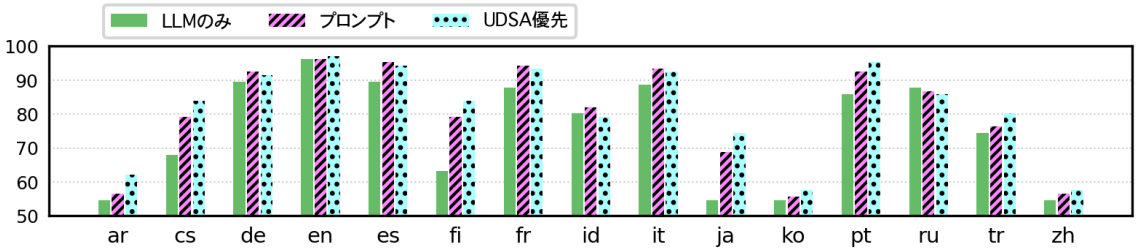

図 2　flan-t5-xxl の各言語の極性判定の正解率。

4.3　構文解析と評価表現抽出

UDSA の前段の構文解析器として、StanfordNLP [8] と Trankit [11] を用いた。構文解析の精度は Trankit の方が高いが、極性表現の抽出の規則や辞書リソースとの整合性から、StanfordNLP の方が後段の処理に適する場合もあるため、実験では両方を用いた。

5　実験結果

5.1　全言語平均の評価

表 1 に 15 言語平均の正解率を示す。ベースラインとして、常に NEGATIVE とする（LLM・UDSA ともに用いない）場合の正解率は、全言語共通で 54.7% である。また、LLM を用いずに、UDSA が極性表現を検出した時にそれを文の極性とした場合[2)]は、75% 前後の正解率であった。すなわち、評価表現抽出により半数弱の `POSITIVE` の文を正しく判定できたことになる。

次に、「LLM のみ」の列で、各 LLM による極性判定の正解率を見る。flan-t5-xxl, flan-ul2, mt0-xxl は 70~80% の正解率であり、llama-2-70b-chat は 92.7% とさらに高い正解率を示している。

3.2 節で述べた手法により、UDSA の抽出結果をヒントとしてプロンプトに埋め込んだ場合、flan-t5-xxl と mt0-xxl では 5~7 ポイントの正解率向上が見られた。flan-ul2 と llama-2-70b-chat の上がり幅は小さいが、llama-2-70b-chat では LLM のみで高い正解率が得られているところから、さらに改善されている。

「UDSA 優先」は、UDSA が評価表現を抽出した場合には LLM の結果によらずその極性（複数ある場合は主辞に近い側）を答えた場合の正解率である。仮に UDSA の抽出が常に正しいとしたら、最も理想的なヒントを付与した時の上限値とも捉えられる。

5.2　各言語の評価

図 2 から図 5 に、15 言語それぞれに対し、「LLM のみ」「UDSA のプロンプトを統合」「UDSA の結果を優先」の 3 つを比較したグラフを示す。ここでは StanfordNLP を用いた結果を報告する。

図 2 の flan-t5-xxl では、英語など欧米言語は LLM での正解率が高いのに対し、アラビア語・日本語・韓国語・中国語ではベースラインに近く、ほぼ判定ができていない。UDSA の知識で補完することにより、特にチェコ語、フィンランド語、日本語で大きな性能改善が見られた。

図 3 の flan-ul2 も同様の傾向だが、プロンプトの正解率が UDSA 優先に比べて低い傾向がある。すなわち 3.2 節の手法によりプロンプトに埋め込まれたヒントが解釈されづらいモデルであるといえる。

図 4 の mt0-xxl は、多言語をカバーするモデルであるが、flan の 2 モデルと同様に日本語や中国語の性能が低い。プロンプトにヒントを加えた時の改善

2)　評価表現が抽出されなかった場合は NEGATIVE とする。

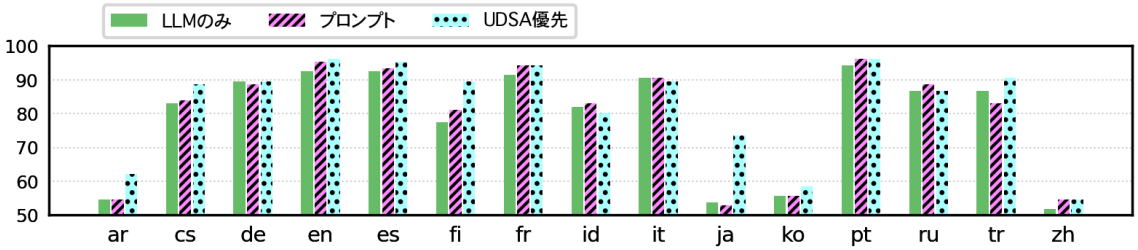

図 3　flan-ul2 の各言語の極性判定の正解率。

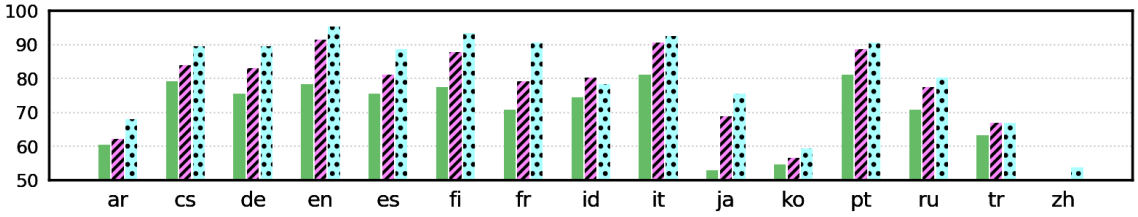

図 4　mt0-xxl の各言語の極性判定の正解率。

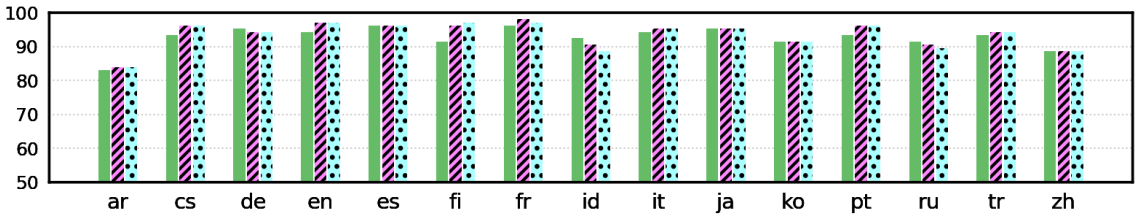

図 5　llama-2-70b-chat の各言語の極性判定の正解率。

の幅は全モデルの中で最も大きく、特に英語・フィンランド語・日本語で大きく改善している。

図 5 の llama-2-70-chat は全言語において性能が非常に高く、アラビア語・中国語以外の正解率は 91% を超えている。この高い水準でも、UDSA 経由のヒントの追加により 8 言語で正解率が改善した。他のモデルと異なり、UDSA 優先の正解率が最大となったのはフィンランド語のみである。このことから、言語の情報を充分に持つモデルにおいては、外部ツールによる判定結果よりも詳細な語彙知識をプロンプトに含めるほうが効果が高いことがわかる。インドネシア語では UDSA により正解率が下がっているのは、StanfordNLP の係り受け誤りに起因しており、Trankit を用いた場合には正解率が向上した。

5.3　誤りの例

flan-t5-xxl 及び flan-ul2 では、ASCII 外の文字が多い言語（アラビア語・日本語・韓国語・中国語）に対して、`NEGATIVE` と判定する場合が非常に多く、UDSA のプロンプトに加えられたヒントによる出力の変化は、それにより `POSITIVE` と判定できるようになったという向きがほぼ全てだった。

llama-2-70-chat が UDSA のヒントを使っても正解しなかった例として、フランス語の “Ils n’explosent généralement pas...” (通常は爆発しない...) という否定の表現があった。この場合のヒントは “hint: explosent is a French word which have a *positive* meaning.” という、極性を反転させた形で与えていたが、これがモデルを混乱させたようである。より詳細に “explosent is a French word which have a negative meaning, but the word is used in a negated form so it should have an opposite polarity.” というヒントを与えれば、正しく `POSITIVE` と判定できるようになった。

6　まとめ

本論文では、多言語の評価極性判定において、外部からの知識をプロンプトに組み入れることによって正解率を向上させられるという結果が得られ、LLM による判定結果の生成と、構文や語彙の知識に由来する抽出が相互補完的であることが示された。言語と LLM の組み合わせによっては、評価表現抽出の結果でオーバーライドするほうが正解率が高かったことから、外部の知識を伝えるプロンプトには改良の余地があることが示唆される。

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, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[2] Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 827–832, Seattle, Washington, USA, October 2013.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[4] Hiroshi Kanayama and Ran Iwamoto. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 4063–4073, 2020.

[5] Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pp. 494–500, Geneva, Switzerland, 2004.

[6] Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models, 2023.

[7] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

[8] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160–170, Brussels, Belgium, October 2018.

[9] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8990–9005, Singapore, December 2023.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[11] Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 80–90, 2021.

[12] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

[13] Daniel Zeman and et al. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19, Vancouver, Canada, August 2017.

[14] Yang Zhao, Tetsuya Nasukawa, Masayasu Muraoka, and Bishwaranjan Bhattacharjee. A simple yet strong domain-agnostic de-bias method for zero-shot sentiment classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3923–3931, 2023.

[15] 金山博, 岩本蘭. 多言語評価表現抽出を通じた Universal Dependencies の検証. 言語処理学会第 26 回年次大会予稿集, March 2020.

[16] 岩本蘭, 金山博. 多言語極性辞書の構築とその包括的評価. 言語処理学会第 26 回年次大会予稿集, March 2020.

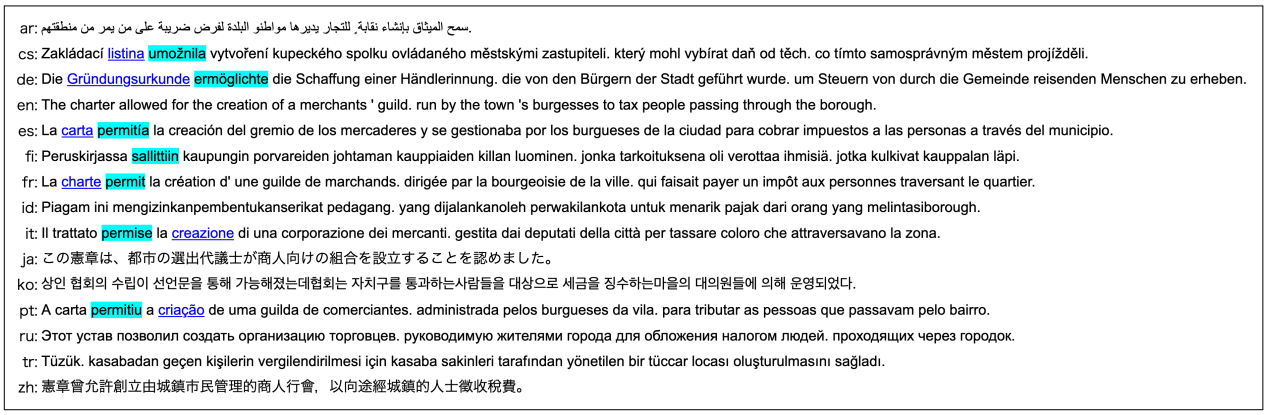

図 6　Parallel Sentiment 中の 15 言語の文の例と、UDSA（StanfordNLP 利用）による極性表現の検出の結果。
青と赤でハイライトされた部分がそれぞれ positive, negative と自動検出された語句で、下線の部分は評価の対象を示す。
正解のアノテーションは POSITIVE。アラビア語で句点の位置が乱れているのは表示上の都合。

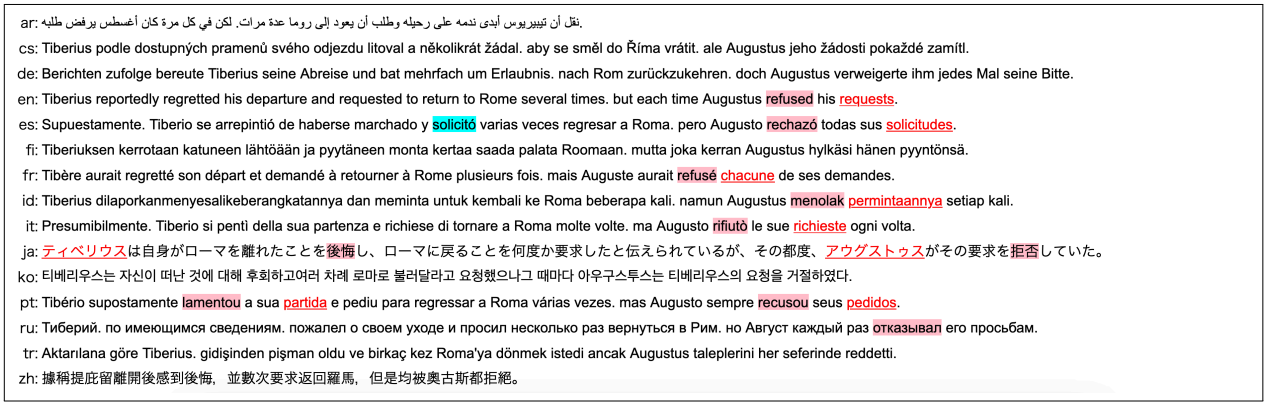

図 7　もう一つの例文。正解のアノテーションは NEGATIVE。
並列構造と語彙の関係でスペイン語の “solicitar”（懇願する）が positive として検出されている。

	flan-t5-xxl		flan-ul2		mt0-xxl		llama-2-70b-chat	
	−	+	−	+	−	+	−	+
ar	N	N	N	N	P	P	N	N
cs	N	P	P	P	P	P	N	P
de	N	P	P	P	P	P	P	P
en	P	P	P	P	P	P	N	N
es	N	P	N	N	N	P	N	P
fi	N	N	N	N	N	P	N	P
fr	N	P	P	P	N	P	N	P
id	N	N	N	N	P	P	N	N
it	N	P	P	P	P	P	N	P
ja	N	N	N	N	P	P	P	P
ko	N	N	N	N	P	P	N	N
pt	N	P	N	P	P	P	N	P
ru	N	N	N	N	P	P	N	N
tr	N	N	P	N	P	P	P	P
zh	N	N	N	N	P	P	N	N

表 2　図 6 の例に対する各モデルの評価極性の判定の結果。− は LLM のみを用いた場合、+は UDSA によるヒントをプロンプトに与えた場合である。P, N はそれぞれ POSITIVE,NEGATIVE の出力で、P が正解である。
llama を含むすべてのモデルで判定が改善された言語がある。flan-ul2 のトルコ語では、ヒントの付与が無い場合にもかかわらず結果が変わっている。

	flan-t5-xxl		flan-ul2		mt0-xxl		llama-2-70b-chat	
	−	+	−	+	−	+	−	+
ar	N	N	N	N	P	P	N	N
cs	N	N	N	N	N	N	N	N
de	N	N	N	P	P	P	N	N
en	N	N	N	N	P	N	N	N
es	N	N	N	N	P	P	N	N
fi	N	N	N	N	N	N	N	N
fr	N	N	N	N	P	N	N	N
id	N	N	N	N	P	N	N	N
it	N	N	N	N	P	N	N	N
ja	N	N	N	N	P	N	N	N
ko	N	N	N	N	P	P	N	N
pt	N	N	N	N	N	N	N	N
ru	N	N	N	N	P	N	N	N
tr	N	N	N	N	P	P	N	N
zh	N	N	N	N	P	P	N	N

表 3　図 7 の例に対する各モデルの評価極性の判定の結果。N が正解である。
mt0-xxl は POSITIVE と判定する傾向にあるが、UDSA の結果を受けて 6 言語で判定が正しくなった。
flan-ul2 のドイツ語では表 2 と同様の挙動が見られる。
スペイン語では positive,negative の 2 つのヒントが加えられたが、それによる悪影響は無かった。