思考連鎖指示における大規模言語モデルの否定表現理解

葉夢宇 [1] 栗林樹生 [1,2] 舟山弘晃 [1,3] 鈴木潤 [1,3]
[1] 東北大学 [2]Langsmith 株式会社 [3] 理化学研究所
{ye.mengyu.s1, h.funa}@dc.tohoku.ac.jp
{kuribayashi, jun.suzuki}@tohoku.ac.jp

概要

近年，推論過程の出力を指示することで大規模言語モデルの性能が向上することが示された．しかし既存研究では，推論過程を経て生成された結論のみが評価されており，モデルがどのような推論過程を生成し，また過程から導かれる妥当な結論を下しているのかといった推論内容の大規模な分析は行われていない．本研究では，言語モデルについて，推論過程を踏まえて結論を生成する能力を評価し，特に言語モデルが苦手としてきた否定表現の扱いに焦点を当てる．実験を通して，最先端の 175B 言語モデルですら，推論過程に *not* が存在する場合，結論として *no* を生成するといった浅い理解に基づいた処理が行われている可能性を示す．

1 はじめに

近年，大規模言語モデルに少数事例を提示することで様々な言語タスクを解く，**少数事例指示（few-shot prompting）** [1] が注目を集めている．特に最近では，推論過程を生成するようモデルに指示する**思考連鎖指示 （chain of thought prompting）** [2] の有効性が示された．例えば，「大谷翔平選手が硬式ボールを投げた可能性はあるか」といった質問に対して，思考連鎖指示に基づく推論では，モデルはまず「大谷選手は野球選手であり，硬式ボールは野球で投げるので...」といった推論過程を生成し，続いて「はい」といった結論を導く．既存研究では，思考連鎖指示のもと最終的に得られる結論の質が向上することは示されたものの，モデルが妥当な過推論程を生成しているのか，モデルの結論が過程と一貫しているのかといった推論過程の妥当性については十分に分析されていない．

本研究では，一連の思考指示を (i) 推論仮定を生成する段階と (ii) 与えられた過程から結論を出す段階に分解し，評価の容易さから，まずは後者の与えられた過程から結論を導く能力について統制的に調査を行う．特に，言語モデルが苦手としてきた否定表現の処理に焦点を当て，*only* や *not* といった否定表現が含まれる推論過程から論理的に適切な結論を導くことができるかを調査する（図 1）．

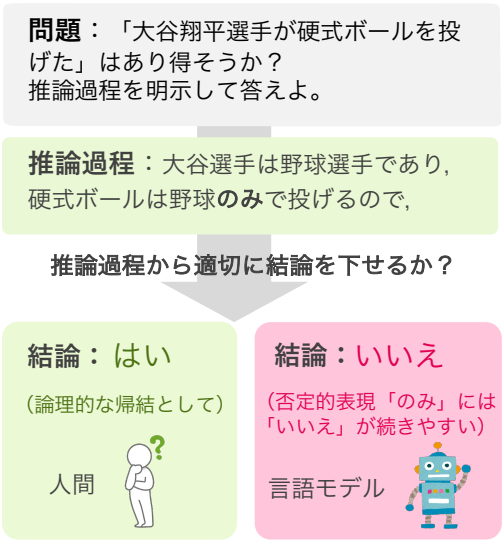

図 1 思考連鎖指示を元に，大規模言語モデルを用いて，否定表現が含まれる推論過程から論理的に適切な結論を導くことができるかを調査する

GPT-3（text-davinci-002，パラメータサイズ 175B）[3, 1] と OPT-66B（Open Pre-trained Transformer Language Models） [4] を含む大規模言語モデルを対象とした実験の結果，*only*，*not*，*implausible* といった表現に**誤誘導 （mispriming）**され，**否定的な表現が推論過程に含まれるだけで，論理的な帰結に反してモデルの下す結果は大きく *no* に偏る**ことが観察された．

この結果から，依然として深い言語理解に基づく論理的な推論が行われていないことが示唆された．

2 実験設定

評価したい能力: 思考連鎖指示 では，少なくとも (i) 推論過程を生成し (ii) 推論過程を踏まえて結論を

Determine whether an artificially constructed sentence relating to sports is plausible or not.

Q: Is the following sentence plausible? "Bam Adebayo scored a reverse layup in the Western Conference Finals."
A: Let's think step by step.
Bam Adebayo is an American basketball player. Scoring a reverse layup only happens in basketball. So the answer is yes.

Q: Is the following sentence plausible? "Santi Cazorla scored a touchdown."
A: Let's think step by step.
Santi Cazorla is a soccer player. Touchdowns are only happen in football. So the answer is no.

Q: Is the following sentence plausible? "DeMar DeRozan was called for the goaltend."
A: Let's think step by step.
DeMar DeRozan is an American basketball player. Goaltending only happens in basketball. So the answer is yes.

図 2　少数事例指示文の例，緑色は推論過程の部分を表す

出力するという 2 つの能力が求められる．本研究では，後者に焦点を当てて評価を行う．後者に焦点を当てる理由としては，*yes/no* の答えのみを出力する段階は分類問題であるため評価が比較的簡単であることと，この能力は思考連鎖指示 が達成されるための必要条件であるため，最低限の要請として達成できることを期待したいという点があげられる．

問題設定: 例題として，「P が A をした」という文の尤もらしさをモデルに自然言語で問う設定を採用した[1)]．ここで P は人名（*player*，スポーツ選手名），A は行動（*action*，ボールを蹴るなど）を表し，推論過程では P が何のスポーツ S の選手であるか，また A がそのスポーツ S で行われているかを段階的に推論するシナリオを採用した．図 2 に示す通り，実際の問題は英語で書かれている．

実験では，問題と推論過程（*So the answer is* まで）をモデルに入力し，モデルが *yes/no* のどちらを回答するか分析することで，推論過程を踏まえて適切な結論を出力する能力を評価する．結論を導くステップでは「P は S の選手である，S では A しない．よって P は A しない」といった簡単な論理推論が求められている．この段階において，例えば推論過程に *not* のような否定表現が出現したら結論を *no* とするといった表層的な処理が行われているかを調査する．

なお，推論過程はあらかじめ準備した適当なものを入力しており，ある人名があるスポーツの選手でありそうか，ある行動がそのスポーツで行われるかといった常識的知識をモデルに問わない．この点は 3.1 節でも更に対処する．

少数事例指示: 先行研究 [6] に倣い，タスクに関する説明と，問題文・模範的な推論過程・結論を合計 3 問をモデルに入力したのち，実際に解かせたい問題の問題文と推論過程を入力する．以降導入する否定語を挿入した設定などにおいても，指示文は共通のものを用いている．

モデル: 実験では，GPT-3（text-davinci-002，パラメータサイズ 175B） [3, 1] と OPT(Open Pre-trained Transformer Language Models) [4] を用いた．OPT については Huggingface[2)] 上に公開されているパラメータサイズの異なる 7 モデル（350M, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B）を評価対象とした．

3　実験

まず初めに，推論過程を踏まえて正しく回答できるかを精緻に問うため，常識知識ではなく推論過程を見なければ解けない問題設定を導入する (3.1 節)．次に，否定表現の有無とモデルの振る舞いについて調査を行う (3.2 節)．

3.1　事前実験: Fiction 設定の導入

設定　BIG-Bench の Sports understanding 問題 [5] から収集したデータセット (Real) に加え，架空の選手名・スポーツ名から構成したデータセット（Fiction）を用意した (付録 A)．既存の Real セットは実世界の常識に即した問題になっているが，本研究の関心のもとでは，モデルが正解した時に推論過程の内容を理解して結論を導いたのか，モデルが持つ常識知識に基づいて回答できたのかの切り分けが難しい．従って，推論過程を踏まえないと答えることのできない Fiction 設定を導入した．なお，正解ラベルの分布は 1:1 であることが期待されるが，乱数シードの都合上，Real 設定では正解ラベルが yes の問題が 496 問，正解ラベルが no の問題が 504 問であり，Fiction 設定では正解ラベルが yes の問題が 495 問，

1)　BIG-Bench [5] 中の Sports understanding タスクを参考にした．

2)　https://huggingface.co/docs/transformers/model_doc/opt

表 1 結果（yes: yes の割合， no: no の割合）

モデル	Real 設定			Fiction 設定		
	正解率	yes	no	正解率	yes	no
GPT-3(175B)	100.0	49.6	50.4	**79.4**	28.9	71.1
OPT-66B	94.7	54.9	45.1	**91.9**	57.6	42.4
OPT-30B	75.1	24.7	75.3	**55.7**	5.2	94.8
OPT-13B	68.2	34.0	66.0	91.0	41.5	58.5
OPT-6.7B	86.5	52.9	47.1	**77.6**	71.1	28.9
OPT-2.7B	77.8	55.6	44.4	**67.0**	81.5	18.5
OPT-1.3B	68.7	18.7	81.3	81.1	43.6	56.4
OPT-350M	51.2	1.6	98.4	**51.1**	2.0	98.0

正解ラベルが no の問題が 505 問である．

結果 表 1 に各設定での正解率と，モデルが *yes/no* と回答した割合を示す．また前述の通り，データセット中の *yes/no* が回答である問題の比はおよそ 1:1 である．事前学習の知識が使えない Fiction 設定では全体的に正解率が下がり，特に OPT-30B および OPT-2.7B においては，出力が yes か no のどちらかに大きく偏ることになった．このことから思考連鎖指示中には生成した推論過程のみでなく，モデルが有する知識も活用されていることが示唆され，推論過程の分析をする上で，事前知識が使えない設定を導入することの意義が確認された．以降の実験では，Fiction 設定に編集を加えていく．

3.2 実験: 否定語の影響

設定： 否定表現に関する 4 つの設定を導入する．Only 設定において，元の問題設定 [5] では，「P は S_1 の選手である．A は S_2 で行われる．よって no」というシナリオで問題が設計されているが，「A が S_2 で行われる」ことは「A が S_1 で行われない」ことを含意しないため，厳密には曖昧性のある問題であった．そこで「A が S_2 **のみ**で行われる」といった制限を導入することで，問題を解きやすくする．Not 設定は，否定表現の意味合いを強まることの影響を調べるために導入する．Implausible+（Not）は Fiction 設定と同じ問題構成で，質問文や推論過程内の機能語を変えた場合でも，モデルがその変化に応じて正しい出力をするかを調査する．各設定の概略を以下に示す．モデルには質問と，回答における { はい/いいえ } の直前までが入力される．また図 3 の例のように，実際はこれらの問題は英語で記述されている．

Only

質問:「P が A した.」という文はあり得そうか？
回答: P は S_1 選手であり， A は {S_1/S_2} **のみ**で行われるので，{ はい/いいえ }

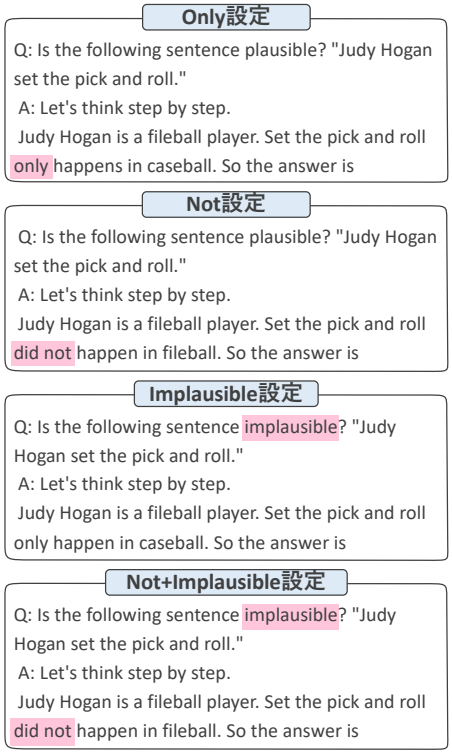

図 3 Only， Not， Implausible および Not+Implausible の設定における入力文の例．

Not

質問:「P が A した.」という文はあり得そうか？
回答: P は S_1 選手であり， A は {S_1 **のみ**で行われる/S_1 で行われ**ない**} ので，{ はい/いいえ }

Implausible

質問:「P が A した.」という文は**あり得ない**か？
回答: P は S_1 選手であり， A は {S_1/S_2} **のみ**で行われるので，{ いいえ/はい }[3)]

Implausible+Not

質問:「P が A した.」という文は**あり得ない**か？
回答: P は S_1 選手であり， A は {S_1 **のみ**で行われる/S_1 で行われ**ない**} ので，{ いいえ/はい }

3.3 結果・考察

表 2 に結果を示す．全てのタスクにおいて，no が出力された問題の割合の大幅な上昇を観測した．この結果から，出力が否定的な意味合いを持つ単語に誤誘導されていることが示唆される．

Only 設定: この設定では，問題が明確化されたにも関わらず，Fiction 設定と比較すると *no* の生成率が高まることから，言語モデルは *only* によって誤誘

3) 機能語が「あり得る」から「あり得ない」に変化したため，結論は反転する．下の Implausible+Not も同様．

表 2 結果（yes: yes の割合， no: no の割合）

モデル	Only 設定			Not 設定			Implausible 設定			Implausible+Not 設定		
	正解率	yes	no	正解率	yes	no	正解率	yes	no	正解率	yes	no
GPT-3(175B)	63.4	12.9	**87.1**	63.0	12.5	**87.5**	49.2	0.3	**99.7**	49.2	0.3	**99.7**
OPT-66B	77.2	28.7	71.3	78.2	27.7	72.3	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-30B	50.5	0.0	**100.0**	50.5	0.0	**100.0**	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-13B	51.1	0.6	**99.4**	51.1	0.6	**99.4**	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-6.7B	71.3	20.8	79.2	71.3	20.8	79.2	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-2.7B	66.5	17.2	**82.8**	67.1	16.6	**83.4**	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-1.3B	50.5	0.0	**100.0**	50.5	0.0	**100.0**	49.5	0.0	**100.0**	49.5	0.0	**100.0**
OPT-350M	50.7	0.2	**99.8**	50.7	0.2	**99.8**	49.5	0.0	**100.0**	49.5	0.0	**100.0**

導されたことが確認された．この点からも，言語モデルが論理的な推論といった深い言語理解に基づいて思考連鎖指示 をできているとは考えづらい．

Not 設定：この設定では，*yes* の問題は *only* を用いた Only 設定と同様のものを，*no* の問題は *not* を用いた推論過程に書き換えたものである．この設定でも Fiction 設定と比較すると *no* の生成率が高くなり，依然として否定語の扱いに苦戦していることが観察された．

Implausible（+Not）設定：Implausible 設定と Implausible+Not 設定の共通点は，少数事例指示文の部分に変更を加えずに，*implausible* かという質問に変更したことである．この 2 つの設定でも，Fiction 設定より多くの *no* が生成され，否定表現に誤誘導され，GPT-3，そしてどんなパラメータ数の OPT に関しても，*implausible* という単語に誤誘導され，ほぼ *no* しか答えない現象が観察された．

以上の観察より，単なる否定的な単語の出現により，論理的に妥当な帰結に反して，言語モデルの回答は *no* に誤誘導されていくことが確認された．従って，思考連鎖指示 では生成された推論過程に含まれる浅い特徴によって，結論を導いている可能性が示唆された．

4 関連研究

思考連鎖指示による性能向上 これまでの思考連鎖指示に関する研究では，主に最終的な結論の正当性が評価され，推論過程の妥当性に関しての大規模な評価は行われていなかった．例えば，BIG-Bench [5] を代表とした横断的評価において，思考連鎖指示を用いることによって，多数のタスクにおいて人間の正解率を上回ったことが報告された [6]．

本研究は，結論の正当性と共に，推論の過程の妥当性や，モデルの推論におけるある種の癖を分析したものであり，本研究の結果から，思考連鎖指示の評価の解像度を上げ，近年の言語モデルの能力について知見を深めることも期待される．

言語モデルにおける誤誘導効果と否定表現 事前学習済み言語モデルでは，例えば *Talk? Birds can [MASK]* と入力すると *Birds can talk* と生成してしまうように，質問に直接関係のない単語に出力が誘導される誤誘導効果が報告されている [7]．

アリストテレスの「命題論」では，全ての宣言文は肯定と否定のいずれかに分類されている [8]，否定表現は自然言語を扱うための重要な概念であるとされている．否定表現に焦点を絞ったデータセットが多数存在している [9, 10, 11, 12]．それらのデータセットを用いた評価結果から，否定表現を理解することは事前学習済みモデルにとって挑戦的なタスクであることが確認されている [13]．本研究の結果では，推論過程や問題文に否定表現が含まれる際，最近の大規模言語モデルでも，*no* と結論を下してしまうような誤誘導効果が生じることを確認した．

5 おわりに

本稿では，大規模言語モデルの推論能力の向上に寄与する思考連鎖指示を対象に，推論過程から結論を導く部分，特に否定表現に注目し，大規模言語モデルは妥当な推論過程を行なっているかどうかを検証した．GPT-3 と OPT を対象とした実験では，論理関係ではなく，否定表現が存在するから *no* を生成するというような浅い推論を行なっている可能性を明らかにした．

今後はより多くの思考連鎖指示が有効とされるタスクに検証対象を拡張しさらに深堀りしていく．また，これまでに BERT [14] などを対象とした内部解析手法を用い，大規模言語モデルの内部挙動を分析する方向性も興味深い．

謝辞

本研究は，JSPS 科研費 JP21H04901， JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research)， JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の助成を受けて実施されたものである．

参考文献

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. **CoRR**, Vol. abs/2201.11903, , 2022.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. **CoRR**, Vol. abs/2203.02155, , 2022.

[4] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. **CoRR**, Vol. abs/2205.01068, , 2022.

[5] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. **CoRR**, Vol. abs/2206.04615, , 2022.

[6] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. **CoRR**, Vol. abs/2210.09261, , 2022.

[7] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020**, pp. 7811–7818. Association for Computational Linguistics, 2020.

[8] John L Ackrill and et al. **Categories and De interpretatione**. Clarendon Press, 1975.

[9] Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupala, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, **Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020**, pp. 163–173. Association for Computational Linguistics, 2020.

[10] Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. **CoRR**, Vol. abs/2211.00295, , 2022.

[11] Roser Morante and Eduardo Blanco. *sem 2012 shared task: Resolving the scope and focus of negation. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, **Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada**, pp. 265–274. Association for Computational Linguistics, 2012.

[12] Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. Negation scope detection for twitter sentiment analysis. In Alexandra Balahur, Erik Van der Goot, Piek Vossen, and Andrés Montoyo, editors, **Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2015, 17 September 2015, Lisbon, Portugal**, pp. 99–108. The Association for Computer Linguistics, 2015.

[13] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 9106–9118. Association for Computational Linguistics, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

A FICTION データセット

FICTION 設定では，架空のスポーツ名を 5 個，そして一種目につき 100 個の架空の選手名，合計 500 個の選手名を作った．ここで架空スポーツ名の全部および架空選手名の一部（50 個）を表 3 に示す．

表 3 FICTION データセットの一部詳細

FICTION スポーツ名	FICTION 選手名の一部	
FANBALL	Tilda Pruitt,	Sansone Brady
	Judy Tate,	Petrina Norman
	Rutherford Lucas,	Way Franklin
	Jannel Stanton,	Ora Law
	Owen McGee,	Kalvin Barr
HOLDERBALL	Phoebe Richardson,	Michel Allen
	Alyssa McIntyre,	Dosi Sykes
	Francisco McCoy,	Lorain Reid
	Neda Rose,	Sonni Burnett
	Agathe Frederick,	Darrick Rogers
HILEBALL	Hussein Whitfield,	Larisa Keller
	Wilburn Anderson,	Ernesto Hall
	Douggie Barbour,	Celia Jain
	Raynard Kemp,	Gregor O'Neill
	Carlton Morris,	Katti Davies
CASEBALL	Rob Hancock,	Chas Morrow
	Sonja Fletcher,	Kaleb Graham
	Garwin Shields,	Gunter Payne
	Elliott Blum,	Hailey Hatcher
	Cornelius McCarthy,	Parker Baxter
HOCKBALL	Vita Elmore,	Jay Fowler
	Stefan Camp,	Malcolm Pearson
	Cassi Cooke,	Linnet Page
	Myrilla Anderson,	Tobye Washington
	Granville White,	Corny Reid