NLP WEEK6 논문 요약 : 원아현

제목: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022, NeurIPS)

---

1. 서론 (Introduction)

최근 대규모 언어모델(LLMs)의 발전은 기계번역, 질의응답, 문서 생성 등 다양한 NLP 과제를 크게 개선하였다. 그러나 **산술적 문제 해결, 상식 추론, 상징 조작**등 **복잡한 추론 능력**을 요구하는 과제에서는 여전히 한계가 드러났다.

기존 접근 방식은 크게 두 가지다:

1. **추론 과정을 포함한 학습(Finetuning + Rationale)**: 문제 해결의 중간 reasoning 단계를 담은 데이터셋으로 모델을 학습시킴. 하지만 **데이터 제작 비용이 매우 크고**, 모든 새로운 과제마다 다시 학습이 필요하다.
2. **Few-shot Prompting**: 입력–출력 예시만 주어 모델이 새로운 문제를 해결하도록 유도. 하지만 **추론이 필요한 문제에서는 성능 향상이 미비**하다.

본 논문은 이를 해결하기 위해 **Chain-of-Thought (CoT) Prompting**이라는 간단하면서도 강력한 방법을 제안한다. CoT는 입력–출력 사이에 **사람이 문제를 푸는 과정처럼 단계적 추론을 언어로 기록한 예시**를 추가해 모델이 "생각의 흐름"을 모방하게 하는 방식이다.

---

2. 방법론 (Methodology)

(1) Chain-of-Thought Prompting

- 전통적 few-shot prompting: 〈질문, 정답〉 쌍만 제공.
- CoT prompting: 〈질문, **추론 과정**, 정답〉 형태로 제공.
- 모델이 답을 내리기 전, 자연어로 "중간 계산 및 논리 단계"를 생성하도록 유도한다.

(2) 특징 및 장점

1. **복잡 문제 분해**: 여러 단계로 나누어 처리 가능.
2. **해석 가능성(Interpretability)**: 모델이 왜 그렇게 답했는지 과정을 통해 확인할 수 있음.

3. **범용성**: 수학, 상식, 상징 조작 등 다양한 도메인에 적용 가능.
4. **효율성**: 별도 학습 없이 기존 대형 모델을 그대로 활용할 수 있음.

---

3. 실험 설계 (Experiments)

저자들은 세 가지 영역에서 CoT의 효과를 검증했다.

(1) 산술적 추론 (Arithmetic Reasoning)

- **데이터셋:** GSM8K, SVAMP, ASDiv, AQuA, MAWPS.
- **모델:** GPT-3 (다양한 규모), LaMDA, PaLM, UL2, Codex.
- **비교 방법:**
  - Standard Prompting (질문–답만 제공)
  - CoT Prompting (질문–추론–답 제공)

(2) 상식 추론 (Commonsense Reasoning)

- **데이터셋:**
  - CSQA (일반 상식 질의응답)
  - StrategyQA (다단계 전략적 추론)
  - BIG-bench 과제 (날짜 이해, 스포츠 문장 타당성 판단)
  - SayCan (로봇 행동 계획 생성).

(3) 상징적 추론 (Symbolic Reasoning)

- **과제:**
  - 마지막 글자 연결 (Last Letter Concatenation)
  - 동전 뒤집기(Coin Flip).
- **검증:** 훈련 예시보다 더 긴 단계(out-of-domain) 문제에서도 일반화 가능한지 확인.

---

4. 주요 결과 (Results)

(1) 산술적 추론

- 작은 모델(100B 미만): CoT 효과 거의 없음 → 비논리적이고 헛도는 reasoning 생성.
- 대형 모델(100B 이상): CoT 효과 **폭발적 향상**.

- PaLM 540B + CoT → GSM8K, SVAMP, MAWPS에서 **최신 SOTA 성능 달성**.
- 특히 GSM8K에서는 기존 finetuned GPT-3 + Verifier보다 높은 정확도.

(2) 상식 추론

- PaLM 540B + CoT:
  - **StrategyQA** → 정확도 75.6% (이전 SOTA 69.4% 능가).
  - **Sports Understanding** → 인간 평균(84%)보다 높은 95.4%.
- CSQA에서는 성능 향상이 크지 않았지만, 전반적으로 CoT가 상식 추론 성능을 확실히 높임.

(3) 상징적 추론

- Standard prompting → 실패.
- CoT prompting → **보지 못한 길이 문제(OOD)에서도 성공적으로 일반화.**
- 예: 두 단어 이름 예시만 학습 → 세 단어·네 단어에도 올바른 마지막 글자 연결 수행.

---

5. 추가 분석 (Analysis)

Ablation Study

- 단순 수식만 출력 (Equation only) → 효과 적음.
- 답 먼저 제시 후 reasoning → baseline과 비슷.
- **자연어 기반의 단계적 reasoning 자체가 핵심**임을 보여줌.

Robustness

- 예시 작성자가 달라도, 스타일이 달라도 CoT 성능은 일관되게 baseline보다 우수.
- 예시 순서나 개수를 달리해도 성능 차이는 있지만 CoT의 우위는 유지됨.

---

6. 논의와 한계 (Discussion & Limitations)

- **Emergent Ability**: CoT 효과는 **100B+ 규모의 모델에서만 나타나는 출현적 능력**임.
- **한계점**:
  1. 모델이 실제로 '추론'하는지, 단순히 언어 패턴을 흉내내는지는 불확실.

2. CoT 과정이 항상 올바른 것은 아님 → 틀린 reasoning으로도 정답이 나올 수 있음.
3. 대형 모델에 의존 → 학습/추론 비용이 매우 큼.
4. 예시(CoT annotation)를 사람이 작성해야 하는 부담 존재.

---

7. 결론 (Conclusion)

- **Chain-of-Thought Prompting은 LLM의 추론 능력을 이끌어내는 단순하면서도 강력한 기법**임.
- 산술, 상식, 상징 문제에서 성능을 크게 개선하며, 특히 **대규모 모델에서만 나타나는 새로운 능력**임을 보여줌.
- 향후 연구 과제:
  - 작은 모델에서도 추론 능력을 유도하는 방법.
  - 올바른 reasoning 경로를 보장하는 메커니즘.
  - 자동화된 CoT 데이터 생성.