# What makes NLP difficult?

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

# Ambiguity

| 원문 | 차를 마시러 공원에 가던 차 안에서 나는 그녀에게 차였다. |
|---|---|
| G* | I was kicking her in the car that went to the park for tea. |
| M* | I was a car to her, in the car I had a car and went to the park. |
| N* | I got dumped by her on the way to the park for tea. |
| K* | I was in the car going to the park for tea and I was in her car. |
| S* | I got dumped by her in the car that was going to the park for a cup of tea. |

# Ambiguity

## 중의성 해소(word sense disambiguation)

| 원문 | 차를 마시러 공원에 가던 차 안에서 나는 그녀에게 차였다. |
|---|---|
| G* | I was kicking her in the car that went to the park for tea. |
| M* | I was a car to her, in the car I had a car and went to the park. |
| N* | I got dumped by her on the way to the park for tea. |
| K* | I was in the car going to the park for tea and I was in her car. |
| S* | I got dumped by her in the car that was going to the park for a cup of tea. |

# Ambiguity

| 원문 | 나는 철수를 안 때렸다. |
| --- | --- |
| 1 | 철수는 맞았지만, 때린 사람이 나는 아니다. |
| 2 | 나는 누군가를 때렸지만, 그게 철수는 아니다. |
| 3 | 나는 누군가를 때린 적도 없고, 철수도 맞은 적이 없다. |

# Ambiguity

**문장 내 정보의 부족으로 인한 모호성이 발생**

| 원문 | 나는 철수를 안 때렸다. |
|---|---|
| 1 | 철수는 맞았지만, 때린 사람이 나는 아니다. |
| 2 | 나는 누군가를 때렸지만, 그게 철수는 아니다. |
| 3 | 나는 누군가를 때린 적도 없고, 철수도 맞은 적이 없다. |

# Ambiguity

| 원문 | 선생님은 울면서 돌아오는 우리를 위로 했다. |
|---|---|
| 1 | (선생님은 울면서) 돌아오는 우리를 위로 했다. |
| 2 | 선생님은 (울면서 돌아오는 우리를) 위로 했다. |

# Ambiguity

**문장 내 정보의 부족이 야기한 구조 해석의 문제**

| 원문 | 선생님은 울면서 돌아오는 우리를 위로 했다. |
|---|---|
| 1 | (선생님은 울면서) 돌아오는 우리를 위로 했다. |
| 2 | 선생님은 (울면서 돌아오는 우리를) 위로 했다. |

# Why Language has Ambiguity?

- 언어는 마치 생명체와 같이 진화하며, 특히 효율성을 극대화 하는 방향으로 진화

- 따라서 최대한 짧은 문장 내에 많은 정보를 담고자 한다.
  - 정보량이 낮은 내용(context)은 생략
  - 여기에서 모호함(ambiguity)이 발생

- 생략된 context를 인간은 효율적으로 채울 수 있지만,
  기계는 이러한 task에 매우 취약함.

# Paraphrase

# Paraphrase

| 번호 | 문장 표현 |
|---|---|
| 1 | 여자가 김치를 어떤 남자에게 집어 던지고 있다. |
| 2 | 여자가 어떤 남자에게 김치로 때리고 있다. |
| 3 | 여자가 김치로 싸대기를 날리고 있다. |
| 4 | 여자가 배추 김치 한 포기로 남자를 때리고 있다. |
| 5 | 여자가 김치를 사용해 남자를 때리고 있다. |
| 6 | 남자가 여자에게 김치로 싸대기를 맞고 있다. |
| 7 | 남자가 여자로부터 김치로 맞고 있다. |

# Paraphrase

**문장의 표현 형식은 다양하고,
비슷한 의미의 단어들이 존재하기 때문에
paraphrase의 문제가 존재**

| 번호 | 문장 표현 |
|---|---|
| 1 | 여자가 김치를 어떤 남자에게 집어 던지고 있다. |
| 2 | 여자가 어떤 남자에게 김치로 때리고 있다. |
| 3 | 여자가 김치로 싸대기를 날리고 있다. |
| 4 | 여자가 배추 김치 한 포기로 남자를 때리고 있다. |
| 5 | 여자가 김치를 사용해 남자를 때리고 있다. |
| 6 | 남자가 여자에게 김치로 싸대기를 맞고 있다. |
| 7 | 남자가 여자로부터 김치로 맞고 있다. |

# Discrete, not Continuous

- 이산 값을 갖는 자연어는 사람의 입장에서 인지가 쉬울 수 있으나,
  기계의 입장에서는 매우 어려운 값.

- One-hot 인코딩으로 표현된 값은 유사도나 모호성을 표현할 수 없다.
  - 서로 다른 One-hot 벡터끼리의 유사도나 거리는 모두 동일하다.

- 따라서, 아래의 질문에 대답할 수 없다.
  - <파랑>과 <핑크> 중에서 <빨강>에 가까운 단어는 무엇인가?
  - 하지만 사람의 어휘 체계는 계층적 구조를 띄고 있다.

- 또한 높은 차원으로 표현되어 매우 sparse하게 된다.

# Discrete, not Continuous

- 이산 값을 갖는 자연어는 사람의 입장에서 인지가 쉬울 수 있으나,
  기계의 입장에서는 매우 어려운 값.

- One-hot 인코딩으로 표현된 값은 유사도나 모호성을 표현할 수 없다.
  - 서로 다른 One-hot 벡터끼리의 유사도나 거리는 모두 동일하다.

- 따라서, 아래의 질문에 대답할 수 없다.
  - <파랑>과 <핑크> 중에서 <빨강>에 가까운 단어는 무엇인가?
  - 하지만 사람의 어휘 체계는 계층적 구조를 띄고 있다.

- 또한 높은 차원으로 표현되어 매우 sparse하게 된다.

**딥러닝에서는**
**Word Embedding을 통해 해결**

# Summary

- Ambiguity
- Paraphrase
- Discrete, not Continuous