# NewsEdits: A News Article Revision Dataset and a Document-Level Reasoning Challenge

Alexander Spangher, Xiang Ren, Jonathan May, Nanyun Peng

NAACL 2022

발제자: HUMANE Lab Research Intern 최종현

2024.12.27 랩 세미나

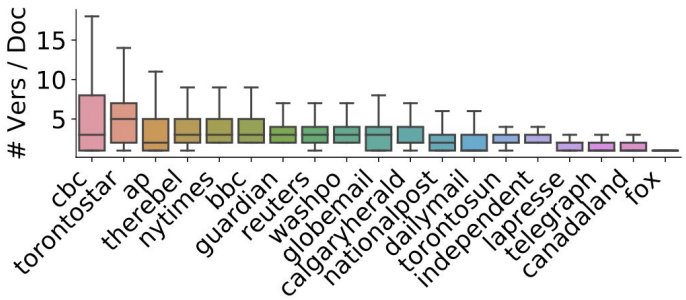
CC BY SA

# Introduction

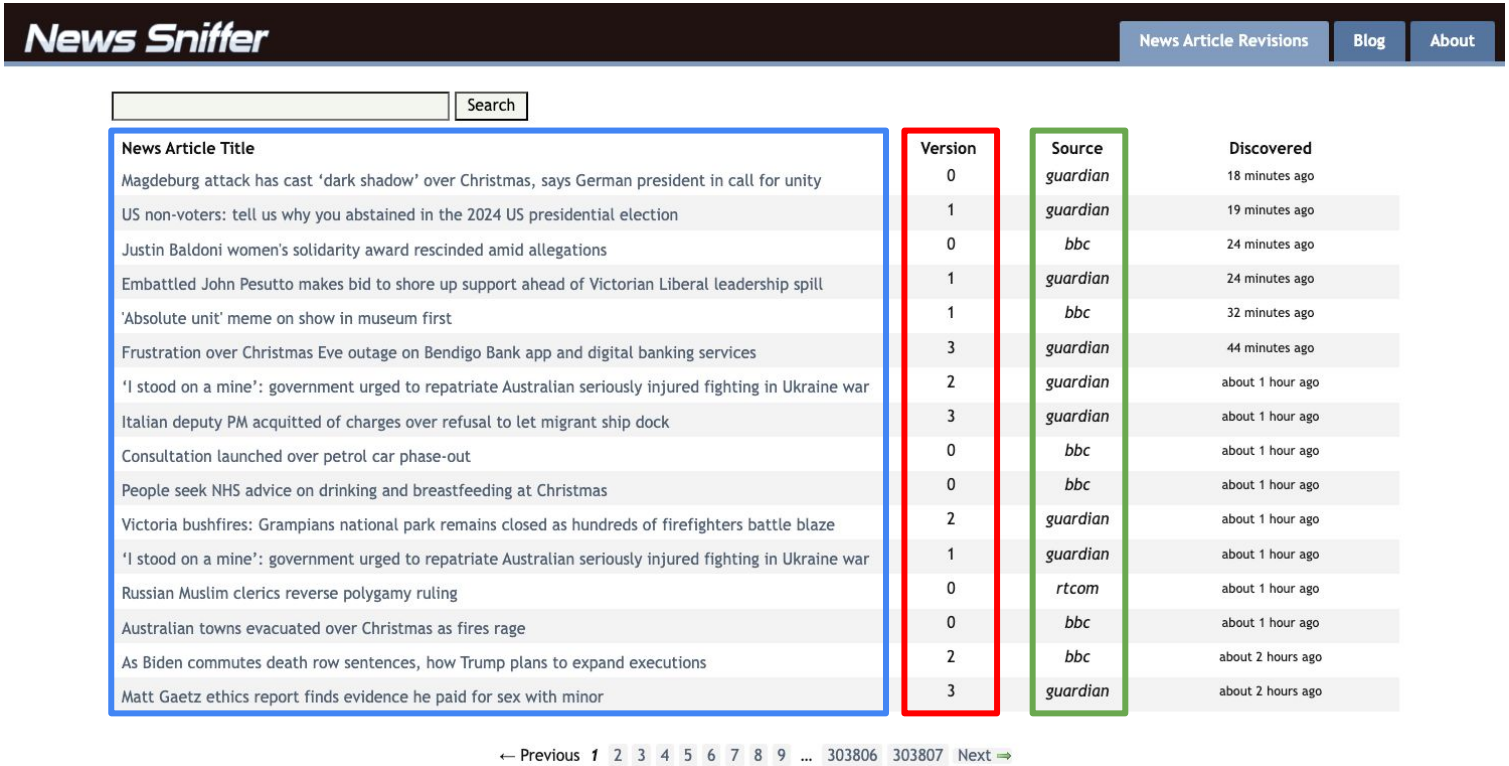
- 뉴스 기사는 작성 후에도 오타 수정, 새로운 정보 추가 등의 수정을 거칠 수 있음
- 뉴스 기사의 수정 기록은 새로운 정보를 담거나, 사건을 업데이트 하거나, 관점을 넓히는 등의 특징을 가지고 있음
- 이러한 뉴스 기사의 수정 기록을 담은 데이터셋 - ***NewsEdits***
- 저자는 다음 질문들에 대한 답을 하고자 함
    - 기사의 수정사항들은 어떠한 형태로 존재하는지?
    - 모델이 기사의 수정사항을 예측할 수 있는지?

# 데이터 수집 방법

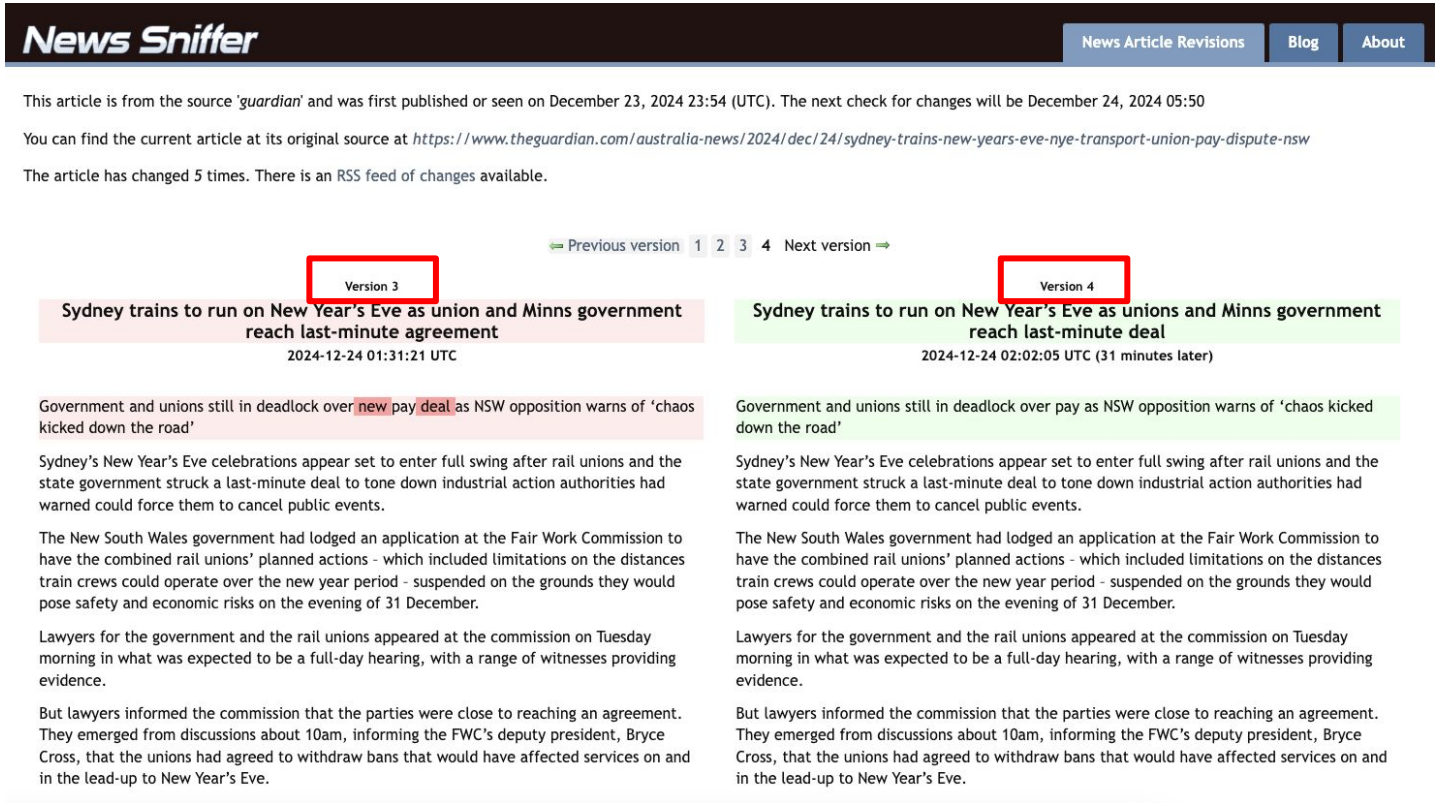
- 기사는 URL로 관리되고, 원본 기사와 수정 버전들은 같은 URL로 관리 됨
- 기사는 2가지 소스에서 가져옴
  - NewsSniffer
  - DiffEngine을 사용하는 트위터 계정
  - 각 소스들은 CBC, TorontoStar, NYTimes, DailyMail 등의 기사를 포함함
  - 2006년~2021년 15년 동안의 기사
  - 120만개의 뉴스 기사와 460만개의 수정 버전



# NewsEdits 데이터셋 개요 (News Sniffer)



# NewsEdits 데이터셋 개요 (News Sniffer)



News Sniffer

News Article Revisions | Blog | About

This article is from the source '*guardian*' and was first published or seen on December 23, 2024 23:54 (UTC). The next check for changes will be December 24, 2024 05:50

You can find the current article at its original source at *https://www.theguardian.com/australia-news/2024/dec/24/sydney-trains-new-years-eve-nye-transport-union-pay-dispute-nsw*

The article has changed 5 times. There is an RSS feed of changes available.

⇦ Previous version 1 2 3 4 Next version ⇨

**Version 3**

**Sydney trains to run on New Year's Eve as union and Minns government reach last-minute agreement**

**2024-12-24 01:31:21 UTC**

Government and unions still in deadlock over new pay deal as NSW opposition warns of 'chaos kicked down the road'

Sydney's New Year's Eve celebrations appear set to enter full swing after rail unions and the state government struck a last-minute deal to tone down industrial action authorities had warned could force them to cancel public events.

The New South Wales government had lodged an application at the Fair Work Commission to have the combined rail unions' planned actions – which included limitations on the distances train crews could operate over the new year period – suspended on the grounds they would pose safety and economic risks on the evening of 31 December.

Lawyers for the government and the rail unions appeared at the commission on Tuesday morning in what was expected to be a full-day hearing, with a range of witnesses providing evidence.

But lawyers informed the commission that the parties were close to reaching an agreement. They emerged from discussions about 10am, informing the FWC's deputy president, Bryce Cross, that the unions had agreed to withdraw bans that would have affected services on and in the lead-up to New Year's Eve.

**Version 4**

**Sydney trains to run on New Year's Eve as unions and Minns government reach last-minute deal**

**2024-12-24 02:02:05 UTC (31 minutes later)**

Government and unions still in deadlock over pay as NSW opposition warns of 'chaos kicked down the road'

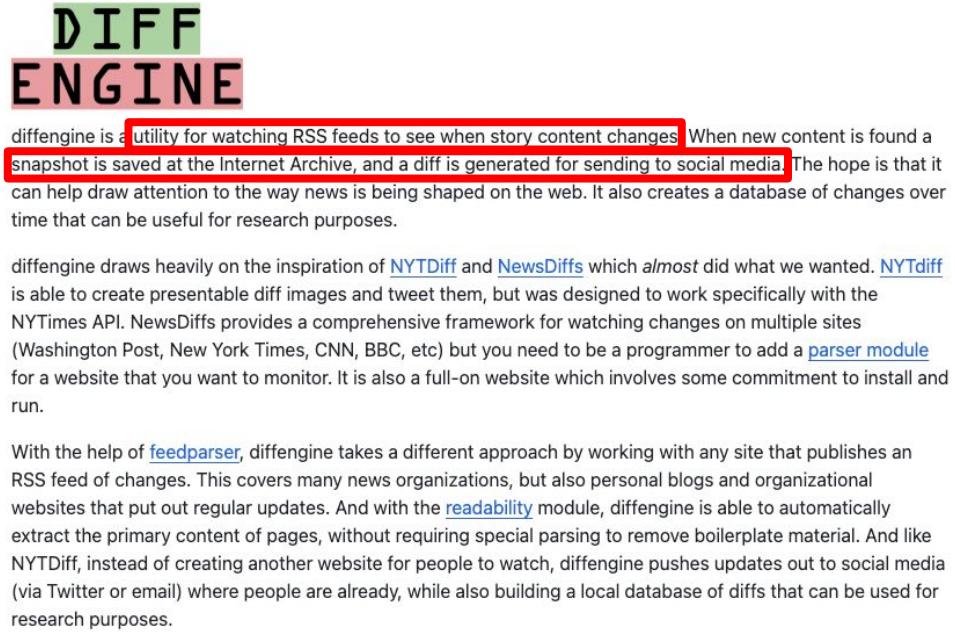
Sydney's New Year's Eve celebrations appear set to enter full swing after rail unions and the state government struck a last-minute deal to tone down industrial action authorities had warned could force them to cancel public events.

The New South Wales government had lodged an application at the Fair Work Commission to have the combined rail unions' planned actions – which included limitations on the distances train crews could operate over the new year period – suspended on the grounds they would pose safety and economic risks on the evening of 31 December.

Lawyers for the government and the rail unions appeared at the commission on Tuesday morning in what was expected to be a full-day hearing, with a range of witnesses providing evidence.

But lawyers informed the commission that the parties were close to reaching an agreement. They emerged from discussions about 10am, informing the FWC's deputy president, Bryce Cross, that the unions had agreed to withdraw bans that would have affected services on and in the lead-up to New Year's Eve.

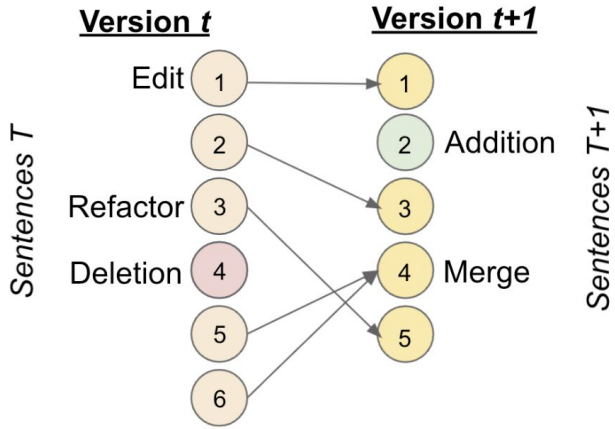
# NewsEdits 데이터셋 개요 (DiffEngine)



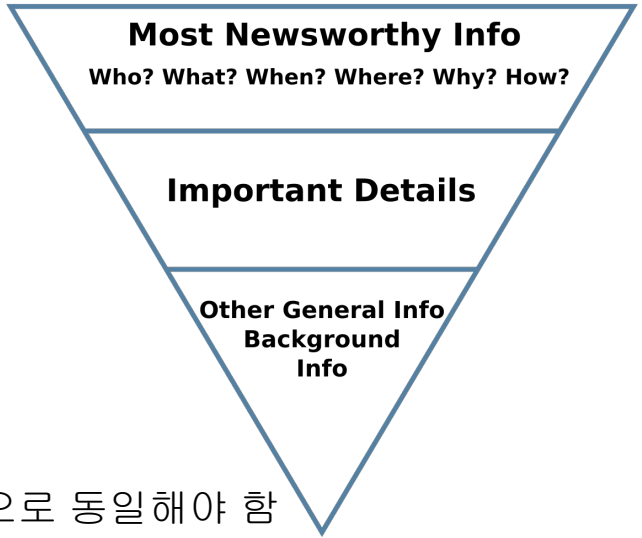

- DiffEngine - 웹사이트의 변경 사항을 추적하고 기록하는 유틸리티
- RSS 피드를 활용하여 웹사이트 변경 사용 추적
- 새로운 내용 발견 시 웹사이트의 아카이브를 저장할 수 있는 Internet Archive에 저장 후, 소셜 미디어로 보낼 수 있는 diff가 생성됨

# NewsEdits 데이터셋

- 120만개의 뉴스 기사와 460만개의 수정 버전들
- 저자들은 기사가 버전마다 어떻게 차이가 있는지에 중점을 두고 연구를 진행함
  - 단어 수정이 아닌, 문장이 수정되는 것에 집중함 (document-level actions)
  - 문장이 단순히 수정되는 것에 그치지 않고, 추가되거나 삭제되는 현상을 파악하는 것은, 편집이 기사에 얼마나 큰 변화를 가져오는지(편집의 강도) 연구하는데 도움이 됨
- 뉴스 기사의 변경 사항은 다음 4가지로 분류
  - **Addition, Deletion, Edit, Refactor**



# NewsEdits 데이터셋




- **Additions**
  - 이전 버전에서는 없는 새 정보를 포함하는 것
- **Deletions**
  - 이전 버전의 내용이 삭제되는 것
- **Edits**
  - 이전 버전과 비교했을 때, 문장의 내용(의미)은 실질적으로 동일해야 함
  - 수정 내용은 문체, 표현, 시제 등이 될 수 있음
- **Refactor**
  - 문장 자체의 내용은 동일하지만, 문장의 위치가 기사 내에서 이동하는 경우
  - 기사의 구조나 정보의 흐름을 조정하기 위해서 수행됨
  - Inverse Pyramid에 의하면 기사에서 위쪽에 있는 글이 더 중요함

# NewsEdits 데이터셋

- *Maximum alignment* metric의 비대칭 버전
- 두 개 이상의 시퀀스를 정렬 할 때 정렬의 품질을 평가하기 위한 지표

$$\mathrm{Sim}_{asym}(x,y) = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_{j}\phi(x_i,y_j)$$

- $\phi(x_i,y_j)$ : x문장의 i번째 단어와, y문장의 j번째 단어의 유사도
- x_i에 대해서 y에서 가장 유사한 단어를 찾고, 그 유사도 값을 취함
- 문장 x의 모든 단어에 대해 위 과정을 반복 후 유사도 값을 합산
- 합산된 유사도 값을 문장 x의 단어 개수로 나누어 정규화

# NewsEdits 데이터셋

$$\mathrm{Sim}_{asym}(x,y) = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_{j}\phi(x_i,y_j)$$

- $\phi(x_i,y_j)$ - 유사도 함수로 Lexical Overlap과 Word Embeddings
- Lexical Overlap: lemma(x_i)=lemma(y_j) 가 같으면 1, 아니면 0
- Word Embeddings: Emb(x_i)·Emb(y_j) → TinyBERT 사용

# NewsEdits 데이터셋 Insight

1. **기사의 수정, 삭제, 추가 시점과 위치는 속보성 뉴스의 패턴과 역피라미드형 기사 구조를 반영함**
    - 기사 작성 초기에는 추가와 편집이 발생하고 주로 앞부분에서 발생 (새로운 정보 추가, 내용 수정)
    - 시간이 지날수록 삭제의 비중이 증가하고 주로 뒷부분에서 발생 (불필요 및 오래된 정보 제거)
    - 이는 속보의 전형적인 형태를 보여줌 → 초기에는 정보 수집 및 추가에 집중, 후에는 내용 다듬기
2. **수정되거나 삭제된 문장들은 변경되지 않은 문장들에 비해 속보성 뉴스와 관련된 사실 패턴(인용구, 사건, 핵심 아이디어)을 포함할 가능성이 더 높음**
    - 추가/삭제된 문장은 유지된 문장에 비해 인용문, 사건, 주요 아이디어를 포함하는 비율이 높음
    - 추가와 삭제는 단순한 문체 수정 이상의 의미를 지니고, 기사의 핵심 내용을 추가하거나 제거하는 (편집 강도가 높은) 변화

# NewsEdits 데이터셋 Insight

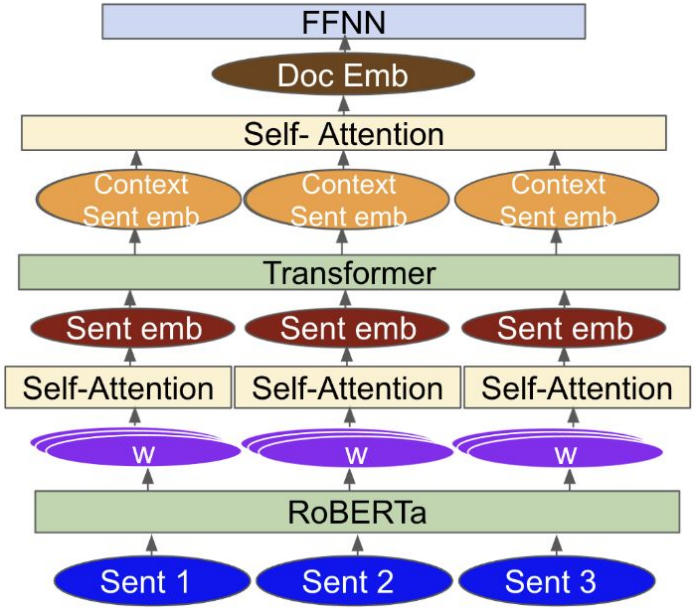
**3. 편집된 문장들은 주로 업데이트되는 사건들을 포함하고 있음**

- 편집은 문체 수정을 넘어 사건의 전개 또는 변화를 반영하는 경우가 많음
- 편집된 문장에서 추출된 사건은 (attack, killed), (injured, killed)와 같이 사건의 전개 또는 변화를 나타내는 경우가 많음
- 편집을 통해 기사가 최신 정보를 반영하고, 독자에게 정확한 정보를 제공함을 알 수 있음

# 데이터셋 기반 실험

- **Task 1: Will this document update?** (업데이트 여부 예측)
  - 주어진 기사(v)가 미래에 업데이트(v+1) 될지 여부를 예측
  - 입력: v, 출력: 업데이트 여부 (0 또는 1)
- **Task 2: How much will it update?** (업데이트 정도 예측)
  - 주어진 기사(v)가 다음 버전(v+1)에서 얼마나 업데이트 될지 예측
  - Addition, Deletion, Edit, Refactor 될 문장의 개수 예측 (개수는 [0, 1), [1, 3), [3, ∞) 로 분류)
- **Task 3: How will it update?** (업데이트 방식 예측)
  - 주어진 기사(v)의 각 문장에 대해 다음 버전(v+1)에서 해당 문장이 어떻게 변화할지 예측
  - 문장 자체가 변경될지? (Deletion, Edit, Unchanged)
  - Refactor가 나타날지? (Up, Down, Unchanged)
  - Addition이 나타날지? (Addition Above, Addition Below)

# 모델 구성




문장 임베딩을 결합하여 전체 문서를 대표하는 하나의 Document Embedding 생성

문장 임베딩에 주변 문맥 정보를 추가하여, 문맥을 고려한 표현 생성 (+Contextual)

각 단어의 중요도(attention weight)를 계산하여 문장 임베딩 생성

문장을 RoBERTa 모델에 입력하여, 각 토큰에 대한 임베딩 벡터 출력
Layer Freeze도 사용한 모델도 있음 (+Partially Frozen)

# 모델 실험 결과

| | F1 | | F1 |
|---|---|---|---|
| Most Popular | 56.6 | Baseline | 60.8 |
| Random | 50.6 | +Partially Frozen | 66.0 |
| Human | **80.1** | +Contextual | 61.7 |
| | | +Version | **77.6** |

| | Num. Additions | | Num. Deletions | | Num. Edits | | Num. Refactors | |
|---|---|---|---|---|---|---|---|---|
| | Mac F1 | Mic F1 | Mac F1 | Mic F1 | Mac F1 | Mic F1 | Mac F1 | Mic F1 |
| Most Popular | 19.8 | 25.0 | 25.6 | 47.8 | 21.9 | 32.0 | 39.2 | 64.5 |
| Random | 32.5 | 33.9 | 30.2 | 36.4 | 31.7 | 35.1 | 25.8 | 35.1 |
| Baseline ($n = 30,000$) | 22.1 | 27.9 | 25.6 | 46.5 | 21.4 | 30.6 | 35.2 | 64.5 |
| ($n = 150,000$) | 29.7 | 36.3 | 25.7 | 48.1 | 22.4 | 32.8 | 39.2 | 64.6 |
| +Partially Frozen | **52.2** | 54.0 | 44.8 | 59.0 | 49.3 | 53.1 | 44.3 | **65.6** |
| +Contextual | 50.7 | 52.2 | 41.0 | 57.4 | **50.8** | **54.8** | **45.0** | 64.3 |
| +Version | 52.0 | **54.5** | **45.3** | **59.8** | 49.9 | 53.7 | 43.8 | 63.1 |
| +Multitask | 46.7 | 50.2 | 28.2 | 48.4 | 42.1 | 49.5 | 40.3 | 55.1 |
| Human | **66.4** | **69.3** | **64.6** | **67.5** | **65.9** | **75.6** | **71.3** | **70.7** |

| | Additions | | Sentence Operations | | Refactors | |
|---|---|---|---|---|---|---|
| | Above (F1) | Below (F1) | Mac. F1 | Mic. F1 | Mac. F1 | Mic. F1 |
| Most Popular | 0.0 | 0.00 | 18.1 | 20.2 | 34.7 | 53.3 |
| Random | **11.8** | **14.4** | 28.0 | 38.3 | 24.7 | 34.7 |
| Baseline | 8.3 | 0.1 | **36.5** | **61.9** | 35.2 | 54.2 |
| +Partially Frozen | 3.5 | 0.0 | 35.4 | 60.9 | 35.4 | 54.6 |
| +Version | 0.1 | 0.0 | 30.3 | 59.0 | **41.6** | **57.2** |
| +Multitask. | 0.0 | 0.0 | 27.5 | 57.8 | 39.5 | 54.8 |
| Human | **38.6** | **46.7** | **63.8** | **63.5** | **45.6** | **91.5** |

**Task 1: Will this document update?** (업데이트 여부 예측)

- 이전 버전 번호를 추가한 +Version이 좋은 성능
- 세 가지 중 가장 쉬운 Task

**Task 2: How much will it update?** (업데이트 정도 예측)

- +version과 +contextual이 상황에 따라 더 좋은 결과를 줌

**Task 3: How will it update?** (업데이트 방식 예측)

- Addition에서 가장 낮은 성능
- Multitask에서 큰 향상은 없음

# 정리

- 뉴스 기사의 수정 이력을 체계적인 수집을 통한 첫 대규모 데이터셋
- 뉴스 기사의 수정 과정을 4가지로 정의 - Addition, Deletion, Edit, Refactor
- 3가지 Task
    1. 문서 업데이트 여부 예측
    2. 문서 업데이트 범위 예측
    3. 문장별 편집 유형 예측
- RoBERTa 모델도 좋은 성능을 보였으나, 훈련된 전문가(Human)과는 격차가 존재함

# 소감

- 뉴스 기사의 수정 이력을 체계적인 수집을 통한 첫 대규모 데이터셋
- 뉴스 기사를 수정할 때 나타날 수 있는 유형을 4가지 카테고리 (Addition, Deletion, Edit, Refactor)로 명확하게 정의하고, 이를 구분할 수 있는 알고리즘 연구
- Edit과 Refactor의 명확한 구분 (둘은 혼동하기 쉬울 수 있음)

- Task 2에서 Addition, Deletion, Edit, Refactor 될 문장의 개수를 [0, 1), [1, 3), [3, ∞)로 나누는 것에 대한 정당성 부족
- 전반적으로 모델링과 관련된 설명이 부족해보임

# Open Question

- 이 데이터셋은 영어와 프랑스 기사로 구성됐는데 다른 언어에서도 이 논문에서 발견한 특징은 유효할까?
- 언론사별 특징도 이 데이터셋과 모델로 알아낼 수 있을까?