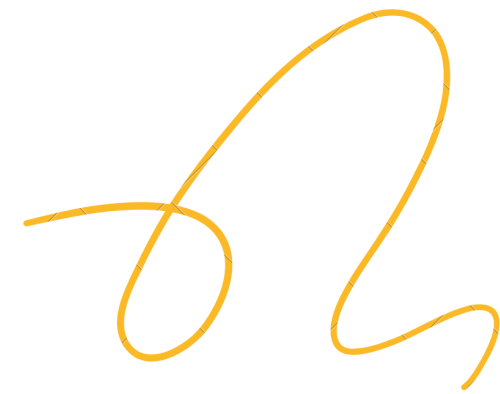
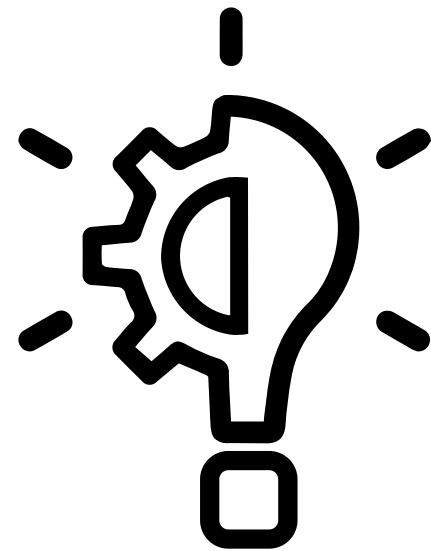


# BYTE PAIR ENCODING

NEURAL MACHINE TRANSLATION OF BARE  
WORDS WITH SUBWORD UNITS



김민주 엄지민 이재원 전예림 정상윤



# CONTENTS

## INTRODUCTION

## NEURAL MACHINE TRANSLATION

## SUBWORD TRANSLATION

- RELATED WORK
- BYTE PAIR ENCODING (BPE)

## EVALUATION

- SUBWORD STATICS
- TRANSLATION EXPREIMENTS

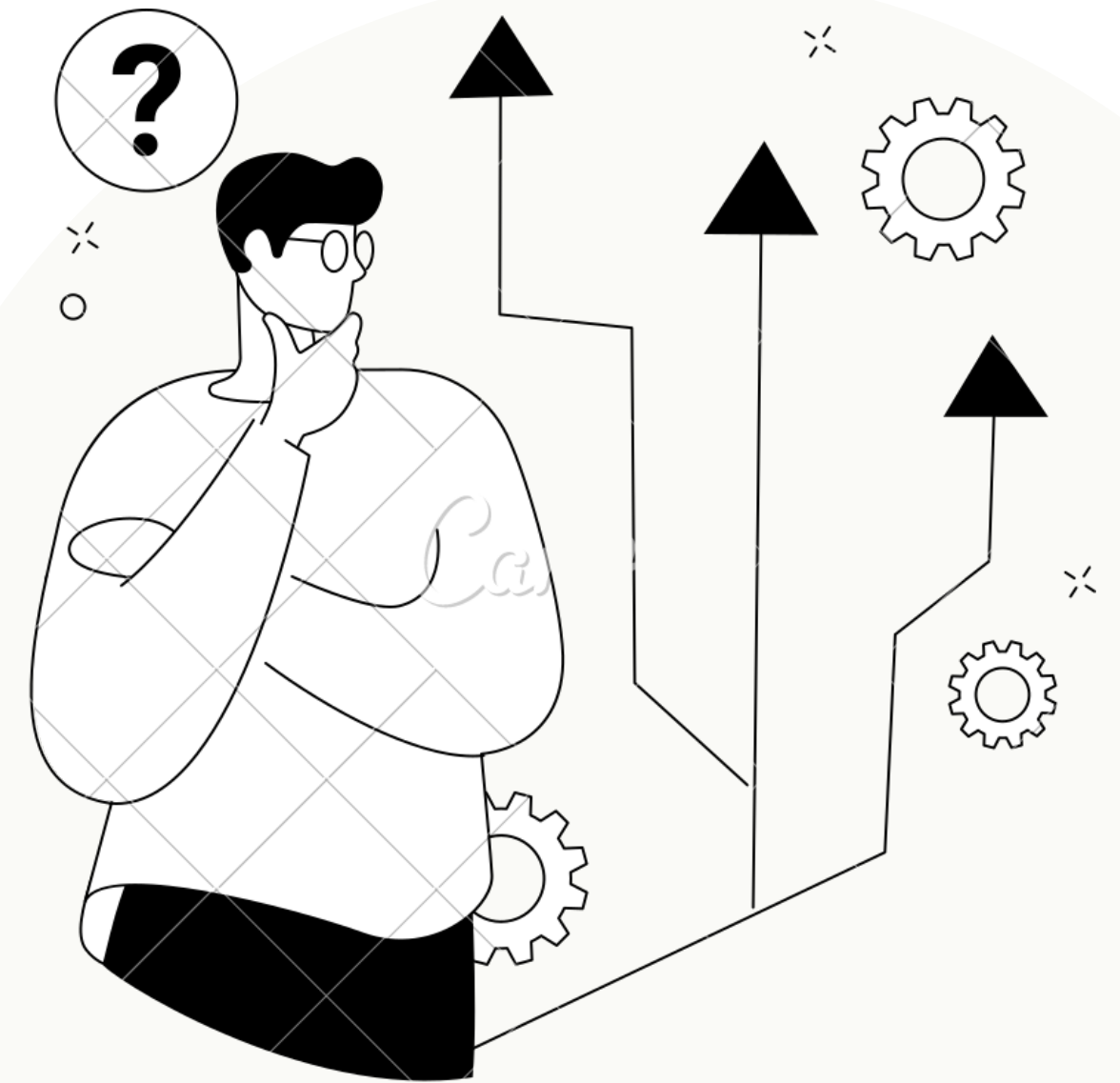
## ANALYSIS

- UNIGRAM ACCURACY
- MANUAL ANALYSIS

## CONCLUSION



# INTRODUCTION



# INTRODUCTION

기존의 **NMT** 모델은 생소한 단어를 번역하는데 취약(**OOV**)

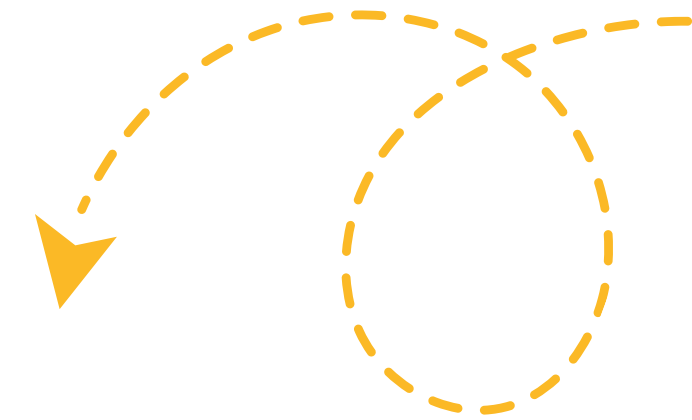
**RARE WORDS**에 대해 **BACK-OFF MODEL**이 필요하지 않은 채로 **OPEN-VOCABULARY** 번역을 수행해내는 **NMT** 모델을 만들어내는 것이 목표

**BPE(BYTE PAIR ENCODING)**을 단어 분할 작업에 적용  
생소한 단어(**RARE WORDS**)는 서브 워드(**SUBWORDS**)로 인코딩

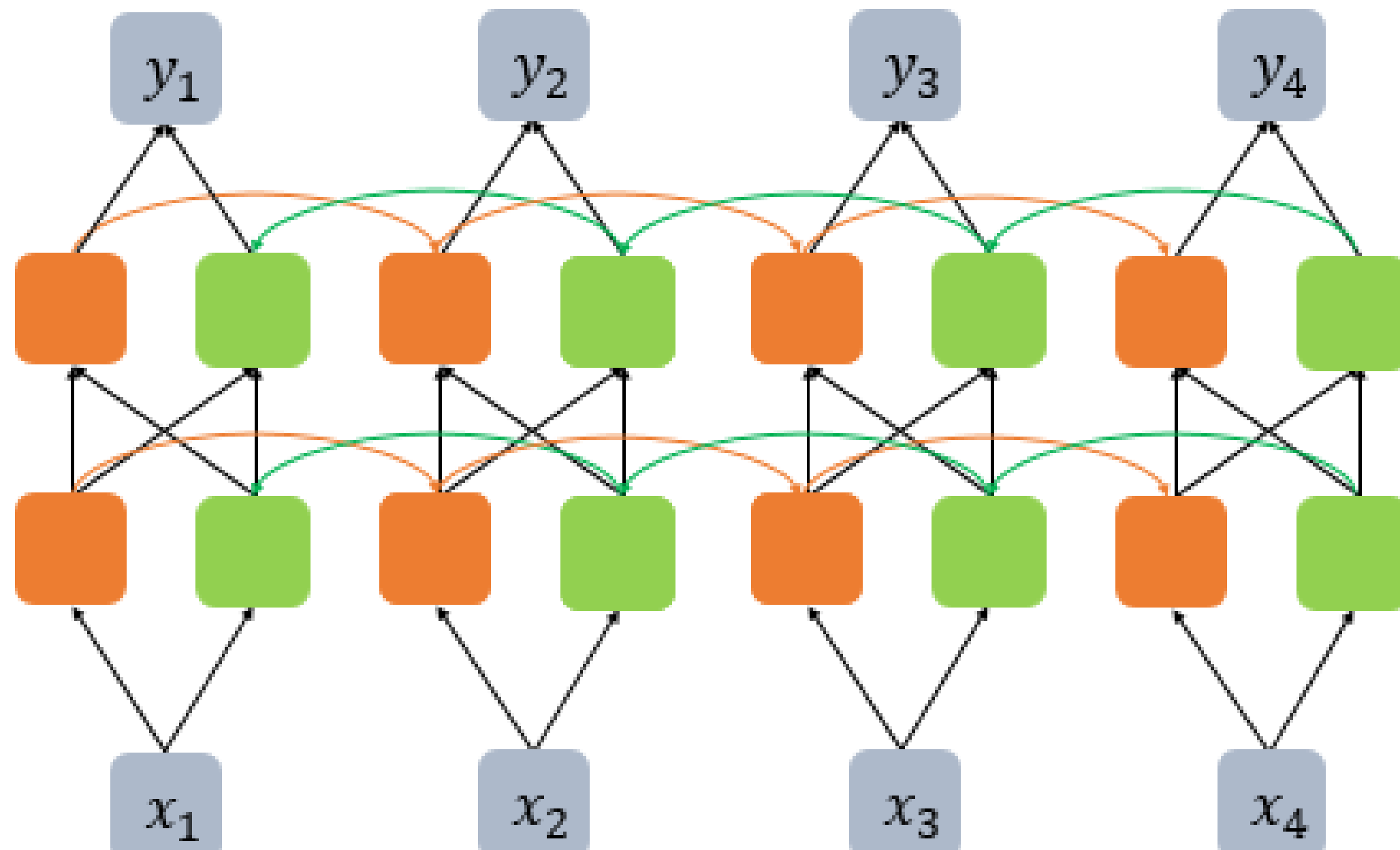
**SUBWORD MODEL**은 **LARGE-VOCABULARY** 모델과 **BACK-OFF DICTIONARY**보다 학습 하지 않았던 새로운 단어들도 더 정확하게 만들어내는 모습을 보이게 됨

A stylized illustration of a person sitting on the floor, holding a smartphone. The person is wearing a white long-sleeved shirt and black pants. Surrounding the person are various icons: a large envelope, a play button, gears, and several user profile cards. A large orange arrow points downwards from the top right towards the person. The entire scene is set against a light gray circular background.

# NEURAL MACHINE TRANSLATION



양방향 순환 신경망



- $X$ : 입력 벡터
- 주황색: **FORWARD STATES**
- 초록색: **BACKWARD STATES**
- $Y$ : 출력 벡터



# SUBWORD TRANSLATION

- RELATED WORK
- BYTE PAIR ENCODING

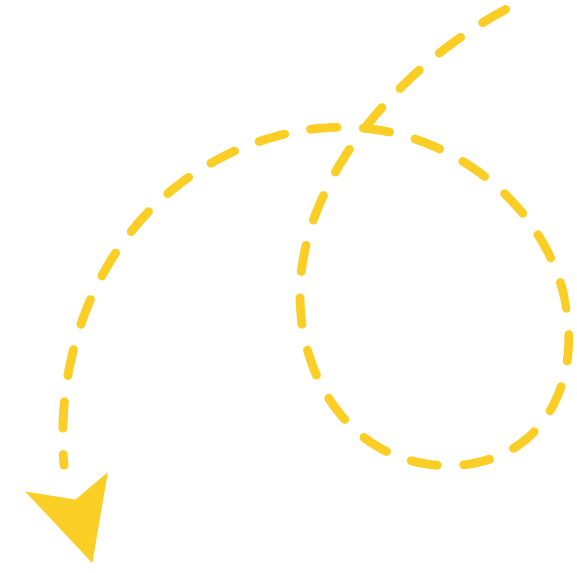
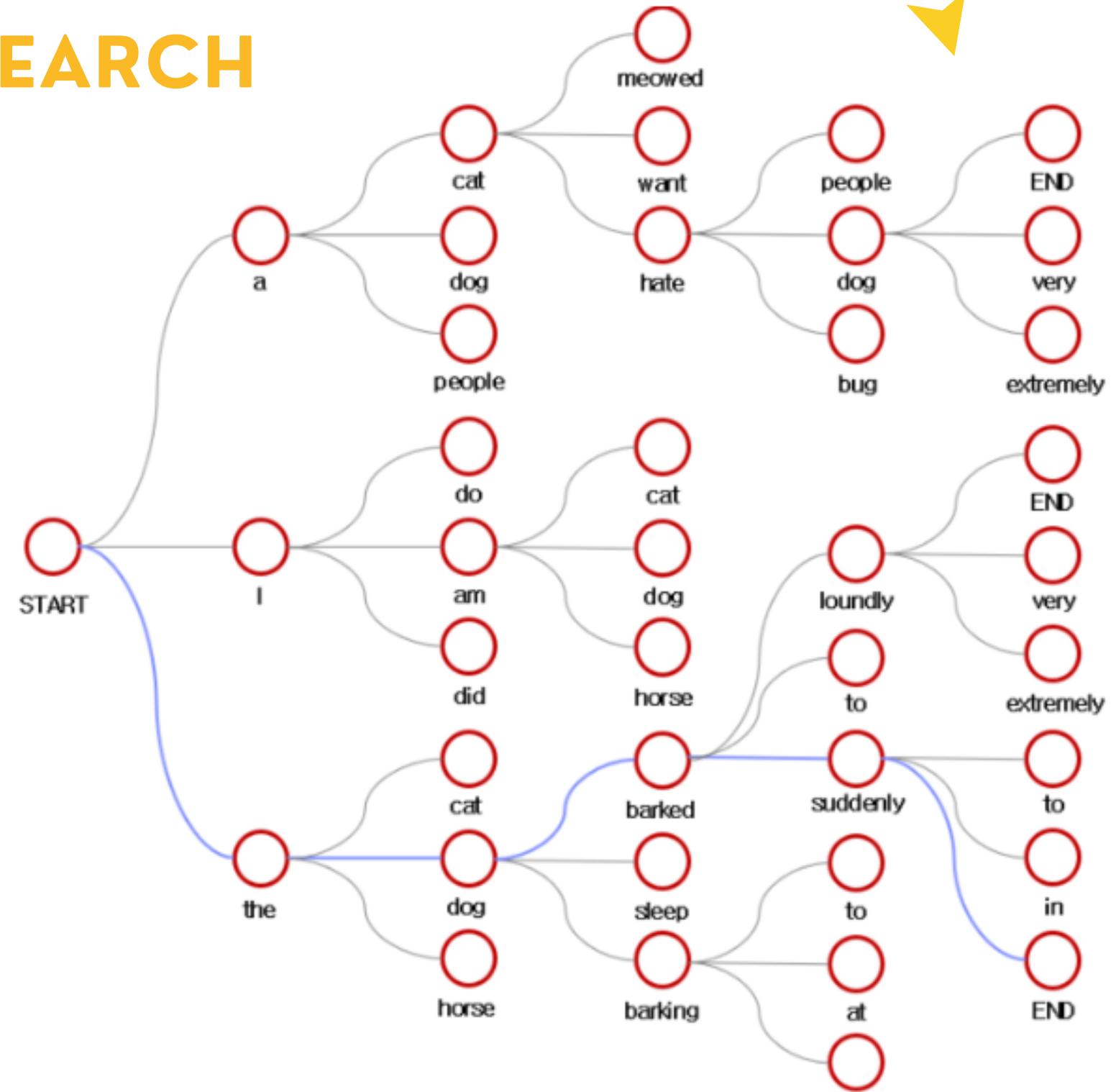
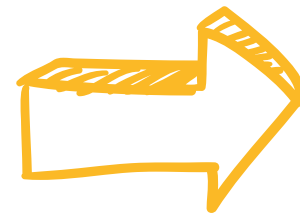
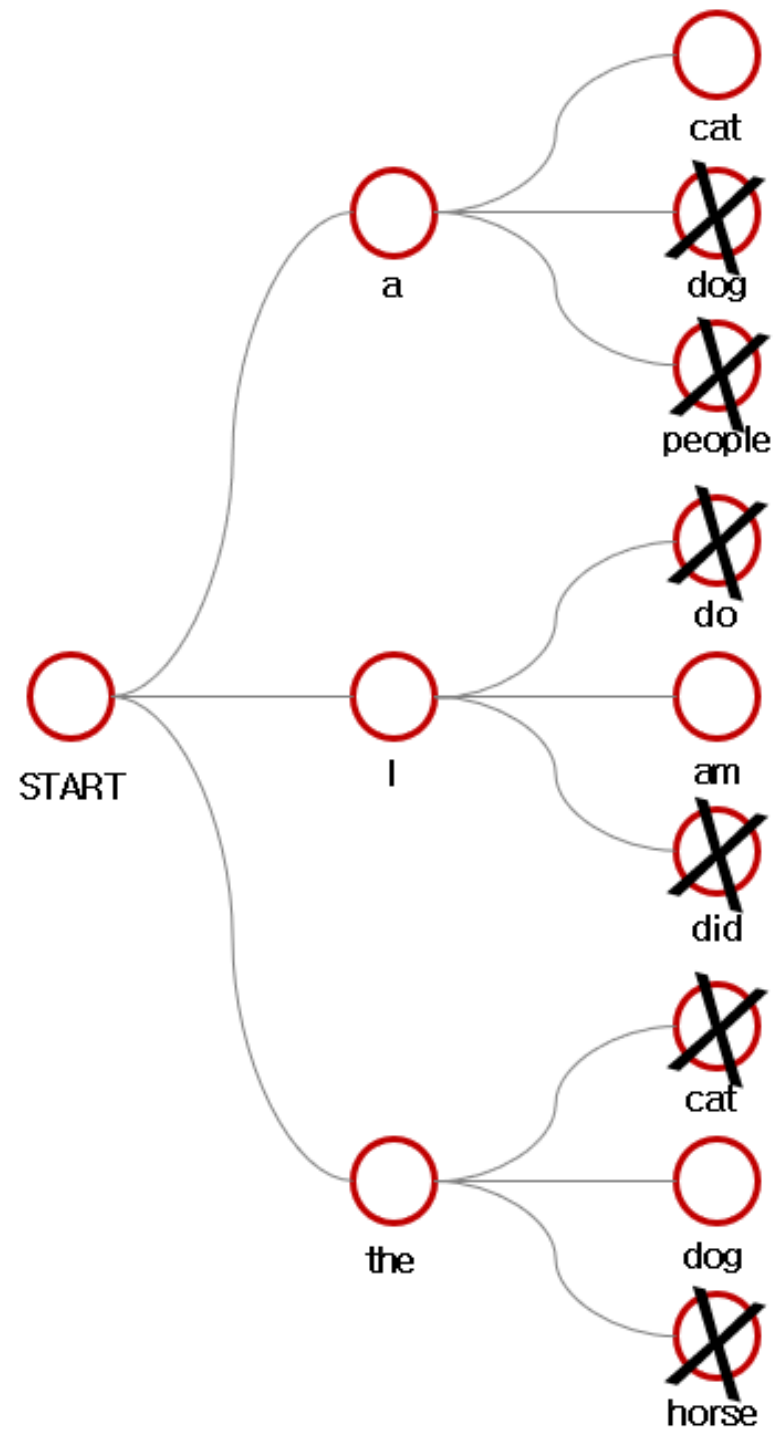
# SUBWORD TRANSLATION

- 이 논문은 유능한 번역자에게만 해석될 정도로, 혹은 번역자도 해석 못할 만큼 생소한 단어도 어간이나 음운 서브워드로 번역이 가능하다고 말한다.
- 또한, 다음과 같은 종류의 언어들이 번역 가능하다고 말한다.
- **COGNATE**: 어원이 서로 같은 단어
- 예) **HAUS**(독일어), **HAUSE**(영어) : 어원이 같다.(게르만어)
- **LOANWORDS** : 외래어 - 외국으로부터 들어와 모국어처럼 쓰이는 것
- 예) 챌린지, 버스, 고무 등
- **MORPHOLOGICALLY COMPLEX WORDS**: 복합어
- 예) 과학자 : 과학(명사) + 자(접미사)



# SUBWORD TRANSLATION

## BEAM SEARCH



# RELATED WORK

- **SMT:** 단어를 세분화 하는데 너무 소극적
- **LING ET AL., 2015B:** 해당 논문과 비슷한 시도, 하지만 결과가 좋지 않았다.
- 이유는 문장의 세분화가 단어 수준에서 그쳤기 때문.

|   | 야크 참새 얼룩말입니다 코뿔소 펠리칸 마 버룩 자고 말 |   |   |   |   |   |   |   |   |   |
|---|--------------------------------|---|---|---|---|---|---|---|---|---|
| 펠 | 거                              | 미 | 피 | 들 | 소 | 족 | 룩 | 새 | 리 | 고 |
| 리 | 게                              | 카 | 마 | 새 | 마 | 제 | 순 | 소 | 참 | 자 |
| 칸 | 오                              | 코 | 우 | 하 | 벌 | 비 | 오 | 치 | 고 | 말 |
| 우 | 여                              | 슴 | 뿔 | 나 | 비 | 입 | 니 | 다 | 벼 | 딩 |
| 타 | 사                              | 크 | 야 | 소 | 다 | 니 | 입 | 말 | 룩 | 얼 |
| 조 | 낙                              | 황 | 새 | 입 | 니 | 다 | 기 | 매 | 갈 | 뱀 |

# BYTE PAIR ENCODING (BPE)

## BPE

- 자주 등장하는 **BYTE PAIR** 들을 사용하지 않는 하나의 **BYTE**로 반복적으로 대체
- **CHARACTER**들을 하나의 **UNICODE**로만 인식하기 때문에 언어에 종속되지 않는다
- **OOV** 문제를 해결할 수 있다

## BPE (Byte Pair Encoding)

*Most often byte pair → replaced by a byte*

Ex1) "ABABCABCD"

→ XXCXCD : X = AB

→ XYD : Y = XC

Ex2) "aaabdaaabc"

→ ZabdZabac : Z = aa

→ ZYdZYac : Y = ab

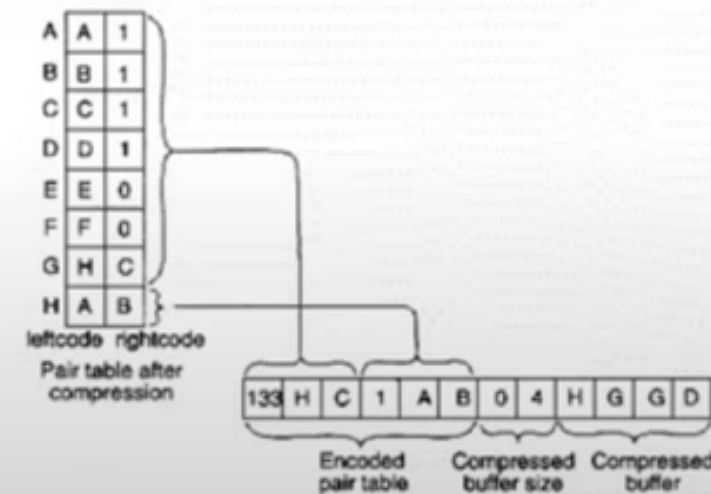
→ XdXac : X = ZY

A B A B C A B C D  
Input string

H H C H C D  
Buffer contents after first pass

H G G D  
Buffer contents after second pass

| Available Character Set |      |
|-------------------------|------|
| Character               | Code |
| 'A'                     | 0    |
| 'B'                     | 1    |
| 'C'                     | 2    |
| 'D'                     | 3    |
| 'E'                     | 4    |
| 'F'                     | 5    |
| 'G'                     | 6    |
| 'H'                     | 7    |



# BYTE PAIR ENCODING (BPE)

---

## Algorithm 1 Learn BPE operations

---


```
import re, collections

def get_stats(vocab):
    pairs = collections.defaultdict(int)
    for word, freq in vocab.items():
        symbols = word.split()
        for i in range(len(symbols)-1):
            pairs[symbols[i],symbols[i+1]] += freq
    return pairs

def merge_vocab(pair, v_in):
    v_out = {}
    bigram = re.escape(' '.join(pair))
    p = re.compile(r'(?!\S)' + bigram + r'(?!\S)')
    for word in v_in:
        w_out = p.sub(' '.join(pair), word)
        v_out[w_out] = v_in[word]
    return v_out

vocab = {'l o w </w>' : 5, 'l o w e r </w>' : 2,
         'n e w e s t </w>':6, 'w i d e s t </w>':3}
num_merges = 10
for i in range(num_merges):
    pairs = get_stats(vocab)
    best = max(pairs, key=pairs.get)
    vocab = merge_vocab(best, vocab)
    print(best)
```

---



```
('e', 's')
('es', 't')
('est', '</w>')
('l', 'o')
('lo', 'w')
('n', 'e')
('ne', 'w')
('new', 'est</w>')
('low', '</w>')
('w', 'i')
```



# EVALUATION

- SUBWORD STATICS ENCODING
- TRANSLATION EXPREIMENTS



# EVALUATION



- NMT에서 rare-word 번역을 subword 모델로 표현함으로써 향상시킬 수 있는가?
- 어휘 크기, 텍스트 크기, 그리고 번역 품질 측면에서 모델 성능 비교

- 영어 -> 독일어, 영어 -> 러시아로 번역을 실험.
- bleu: 기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여 번역에 대한 성능을 측정
- f1: 클리핑된 유니그램 정밀도와 리콜의 조화 평균으로 계산하는 유니그램
- 모든 네트워크는 1000의 은닉 레이어 크기와 620의 임베딩 레이어 크기를 가지고 있음



# SUBWORD STATISTICS

| segmentation                             | # tokens | # types   | # UNK |
|--|----------|-----------|-------|
| none                                     | 100 m    | 1 750 000 | 1079  |
| characters                               | 550 m    | 3000      | 0     |
| character bigrams                        | 306 m    | 20 000    | 34    |
| character trigrams                       | 214 m    | 120 000   | 59    |
| compound splitting <sup>△</sup>          | 102 m    | 1 100 000 | 643   |
| morfessor*                               | 109 m    | 544 000   | 237   |
| hyphenation <sup>◇</sup>                 | 186 m    | 404 000   | 230   |
| BPE                                      | 112 m    | 63 000    | 0     |
| BPE (joint)                              | 111 m    | 82 000    | 32    |
| character bigrams<br>(shortlist: 50 000) | 129 m    | 69 000    | 34    |

Table 1: Corpus statistics for German training corpus with different word segmentation techniques. #UNK: number of unknown tokens in newstest2013. <sup>△</sup>: (Koehn and Knight, 2003); \*: (Creutz and Lagus, 2002); <sup>◇</sup>: (Liang, 1983).

#TOKENS: TEXT SIZE

# TYPES: VOCABULARY SIZE

#UNK : OOV WORD의 개수

- BPE의 경우 OOV의 수가 0개
- BPE(JOINT)의 경우에도 OOV수가 32개로 준수한 성능을 보임
- TOKENS(시퀀스)의 크기가 줄었다
- VOCABULARY SIZE의 크기도 적은 편
- OOV가 확실하게 줄었다

# TRANSLATION EXPERIMENTS

| name                                     | segmentation | shortlist | vocabulary |         | BLEU        |             | CHRF3       |             | unigram F <sub>1</sub> (%) |             |             |
|--|--------------|-----------|------------|---------|-------------|-------------|-------------|-------------|----------------------------|-------------|-------------|
|  |              |           | source     | target  | single      | ens-8       | single      | ens-8       | all                        | rare        | OOV         |
| syntax-based (Sennrich and Haddow, 2015) |              |           |            |         | 24.4        | -           | 55.3        | -           | 59.1                       | 46.0        | 37.7        |
| WUnk                                     | -            | -         | 300 000    | 500 000 | 20.6        | 22.8        | 47.2        | 48.9        | 56.7                       | 20.4        | 0.0         |
| WDict                                    | -            | -         | 300 000    | 500 000 | 22.0        | 24.2        | 50.5        | 52.4        | 58.1                       | 36.8        | <b>36.8</b> |
| C2-50k                                   | char-bigram  | 50 000    | 60 000     | 60 000  | <b>22.8</b> | <b>25.3</b> | 51.9        | 53.5        | 58.4                       | 40.5        | 30.9        |
| BPE-60k                                  | BPE          | -         | 60 000     | 60 000  | 21.5        | 24.5        | <b>52.0</b> | 53.9        | 58.4                       | 40.9        | 29.3        |
| BPE-J90k                                 | BPE (joint)  | -         | 90 000     | 90 000  | <b>22.8</b> | 24.7        | 51.7        | <b>54.1</b> | <b>58.5</b>                | <b>41.8</b> | 33.6        |

[영어 → 독일어]

| name                               | segmentation | shortlist | vocabulary |         | BLEU        |             | CHRF3       |             | unigram F <sub>1</sub> (%) |             |             |
|------------------------------------|--------------|-----------|------------|---------|-------------|-------------|-------------|-------------|----------------------------|-------------|-------------|
|                                    |              |           | source     | target  | single      | ens-8       | single      | ens-8       | all                        | rare        | OOV         |
| phrase-based (Haddow et al., 2015) |              |           |            |         | 24.3        | -           | 53.8        | -           | 56.0                       | 31.3        | 16.5        |
| WUnk                               | -            | -         | 300 000    | 500 000 | 18.8        | 22.4        | 46.5        | 49.9        | 54.2                       | 25.2        | 0.0         |
| WDict                              | -            | -         | 300 000    | 500 000 | 19.1        | 22.8        | 47.5        | 51.0        | 54.8                       | 26.5        | 6.6         |
| C2-50k                             | char-bigram  | 50 000    | 60 000     | 60 000  | <b>20.9</b> | <b>24.1</b> | 49.0        | 51.6        | 55.2                       | 27.8        | 17.4        |
| BPE-60k                            | BPE          | -         | 60 000     | 60 000  | 20.5        | 23.6        | <b>49.8</b> | 52.7        | 55.3                       | 29.7        | 15.6        |
| BPE-J90k                           | BPE (joint)  | -         | 90 000     | 100 000 | 20.4        | <b>24.1</b> | 49.7        | <b>53.0</b> | <b>55.8</b>                | <b>29.7</b> | <b>18.3</b> |

[영어 → 러시아어]





# ANALYSIS

- UNIGRAM ACCURACY
- MANUAL ANALYSIS

# UNIGRAM ACCURACY

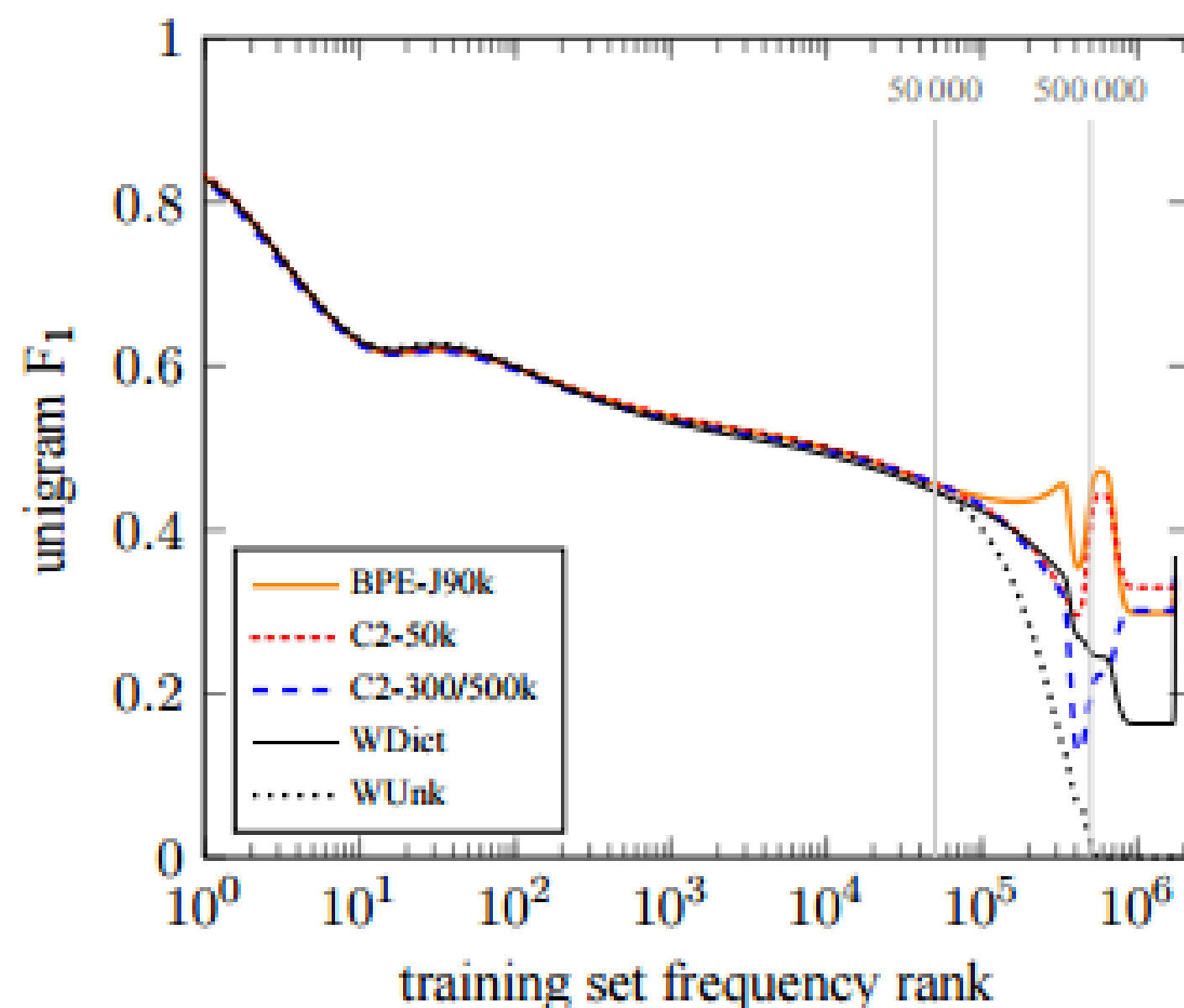


Figure 2: English→German unigram  $F_1$  on newstest2015 plotted by training set frequency rank for different NMT systems.

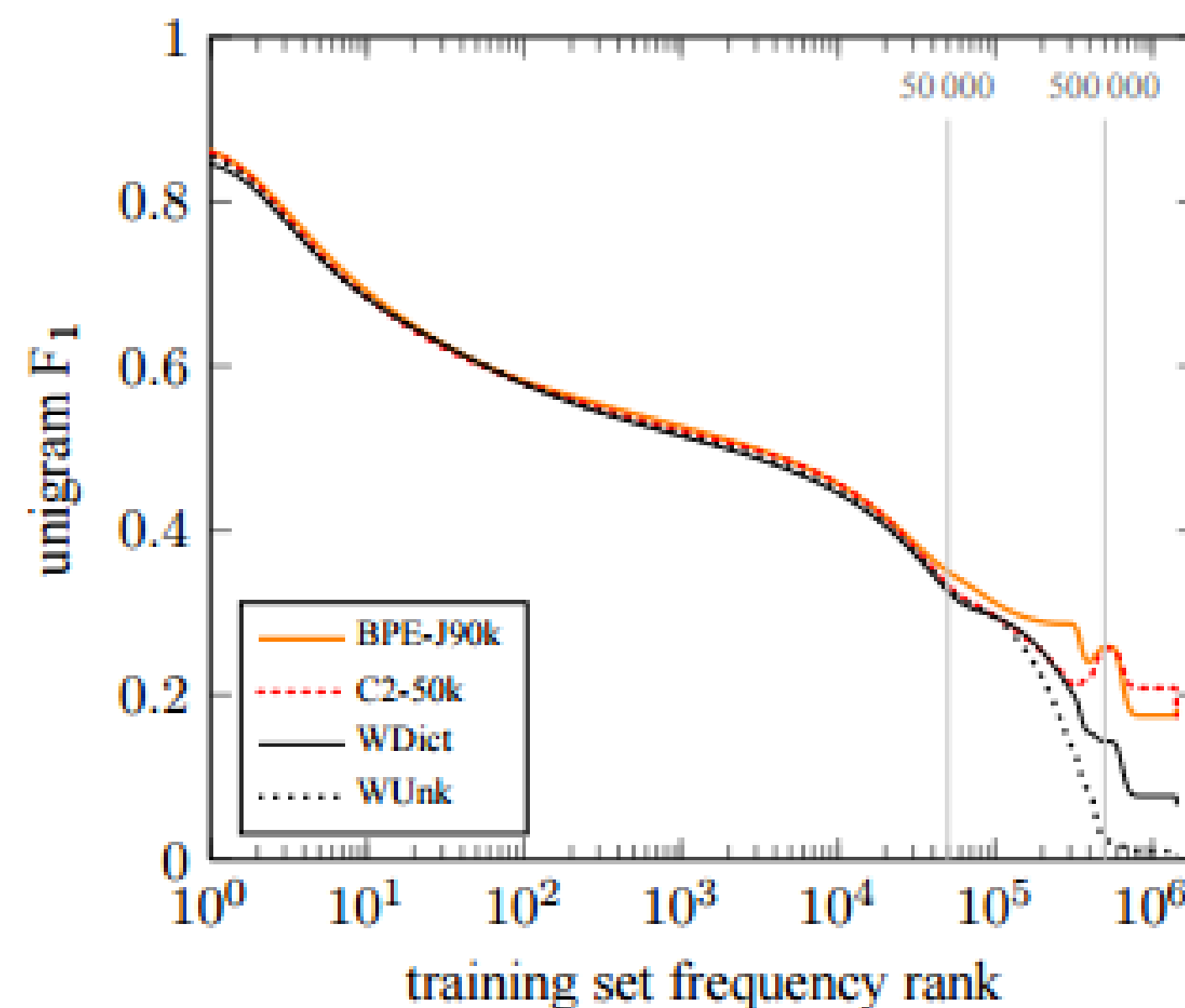


Figure 3: English→Russian unigram  $F_1$  on newstest2015 plotted by training set frequency rank for different NMT systems.

# MANUAL ANALYSIS



| system    | sentence                                     |
|-----------|--|
| source    | health research institutes                   |
| reference | Gesundheitsforschungsinstitute               |
| WDict     | Forschungsinstitute                          |
| C2-50k    | Fo rs ch un gs in st it ut io ne n           |
| BPE-60k   | Gesundheits forsch ungsinstitu ten           |
| BPE-J90k  | Gesundheits forsch ungsin stitute            |
| source    | asinine situation                            |
| reference | dumme Situation                              |
| WDict     | asinine situation → UNK → asinine            |
| C2-50k    | as in ine situation → As in en si tu at io n |
| BPE-60k   | as in ine situation → A in line- Situation   |
| BPE-J90K  | as in ine situation → As in in- Situation    |

Table 4: English→German translation example.  
“|” marks subword boundaries.



| system    | sentence                                  |
|-----------|---|
| source    | Mirzayeva                                 |
| reference | Мирзаева (Mirzaeva)                       |
| WDict     | Mirzayeva → UNK → Mirzayeva               |
| C2-50k    | Mi rz ay ev a → Ми рз ае ва (Mi rz ае ва) |
| BPE-60k   | Mirz ayeva → Мир за ева (Mir za eva)      |
| BPE-J90k  | Mir za yeva → Мир за ева (Mir za eva)     |
| source    | rakfisk                                   |
| reference | ракфиска (rakfiska)                       |
| WDict     | rakfisk → UNK → rakfisk                   |
| C2-50k    | ra kf is k → ра кф ис к (ra kf is k)      |
| BPE-60k   | rak fisk → пра ф иск (pra fisk)           |
| BPE-J90k  | rak fisk → рак ф иска (rak fiska)         |

Table 5: English→Russian translation examples.  
“|” marks subword boundaries.

# CONCLUSION



# CONCLUSION

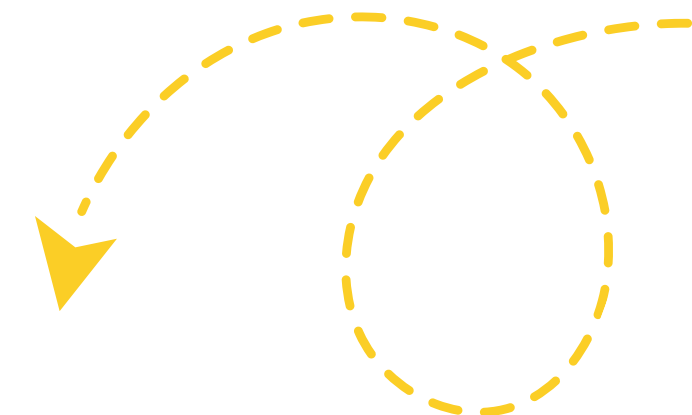
서브워드 유닛을 사용한 **NMT**모델이 **RARE WORD**에 대한 번역 성능이 좋다.

작업에서 어휘 크기의 선택은 어느 정도 임의적이며, 주로 이전 연구와의 비교에 의해 동기부여되었다.



서브워드 세분화는 대부분의 언어 쌍에 적합하며 대규모 **NMT** 어휘나 백오프 모델이 필요하지 않도록 할 수 있다.

향후 연구의 한 방향은 번역 작업에 대한 최적의 어휘 크기를 학습하는 것이며, 이는 언어 쌍 및 훈련 데이터 양에 따라 달라질 것으로 기대된다.



**THANK YOU.**  
**Q&A**

