ACL 2016
AUGUST 7 – 12 | BERLIN, GERMANY

On the Role of Seed Lexicons in Learning Bilingual Word Embeddings

You-Dong Yun, Chanhee Lee

BLP Lab

INDEX

1. Introduction
2. Learning SBWES using Seed Lexicons
3. Experimental Setup
4. Result and Discussion
5. Conclusions and Future Work

- Main Issue of this Research

Main Issue

- 본 연구에서는 서로 다른 각각의 벡터 공간에서 하나의 **Shared Bilingual Word Embedding Space (SBWES)**를 유도하는 과정에서 **seed lexicons의 역할과 중요성에 대해** 분석한다.

- 분석에 기초해 **desirable properties (P1), (P2)를 모두 만족하는** 간단하지만 효과적인 **HYbrid Bilingual Word Embedding(HYBWE)** 모델을 제안한다.

Motivation

- 기존 연구에서는 seed lexicon에 대해 **체계적인 연구 없이 high-quality라고 가정**하였다.

- 기존 연구에서는 본 연구에서 제시한 **desirable properties (P1), (P2)를 모두 만족하는 모델이 존재하지 않았다**.

P1) 풍부한 monolingual training data를 활용*하고, bilingual signal을 이용하여 각 언어를 연결할 수 있을 것*

P2) *다양한 언어 및 분야로 연구를 확장할 수 있도록* ***최소한의 bilingual signal로도 SBWES의 학습이 가능****할 것*

1. Introduction

- **Word**에 대한 dense 실수 벡터인 **Word Embedding(WE)**은 NLP 분야에서 인기를 얻고 있는 모델로, 다양한 NLP 분야에 성능 향상을 불러왔다 (Turian et al., 2010; Collobert et al., 2011; Chen and Manning, 2014).

- 널리 사용되는 **Skip-Gram model with Negative Sampling(SGNS)**은 가장 우수한 WE 모델로 간주되며, 다양한 semantic tasks에서 단순하고 빠른 training으로 견고하고 우수한 성능을 보인다 (Tomas Mikolov et al., 2013)

Bilingual Word Embedding(BWE)

- 언어의 종류에 상관없이, 두 개의 다른 언어로부터 **하나의 shared space**에 **단어를 embedding**하는 것
(Zou et al. Bilingual Word Embeddings for Phrase-Based Machine Translation, 2013.)

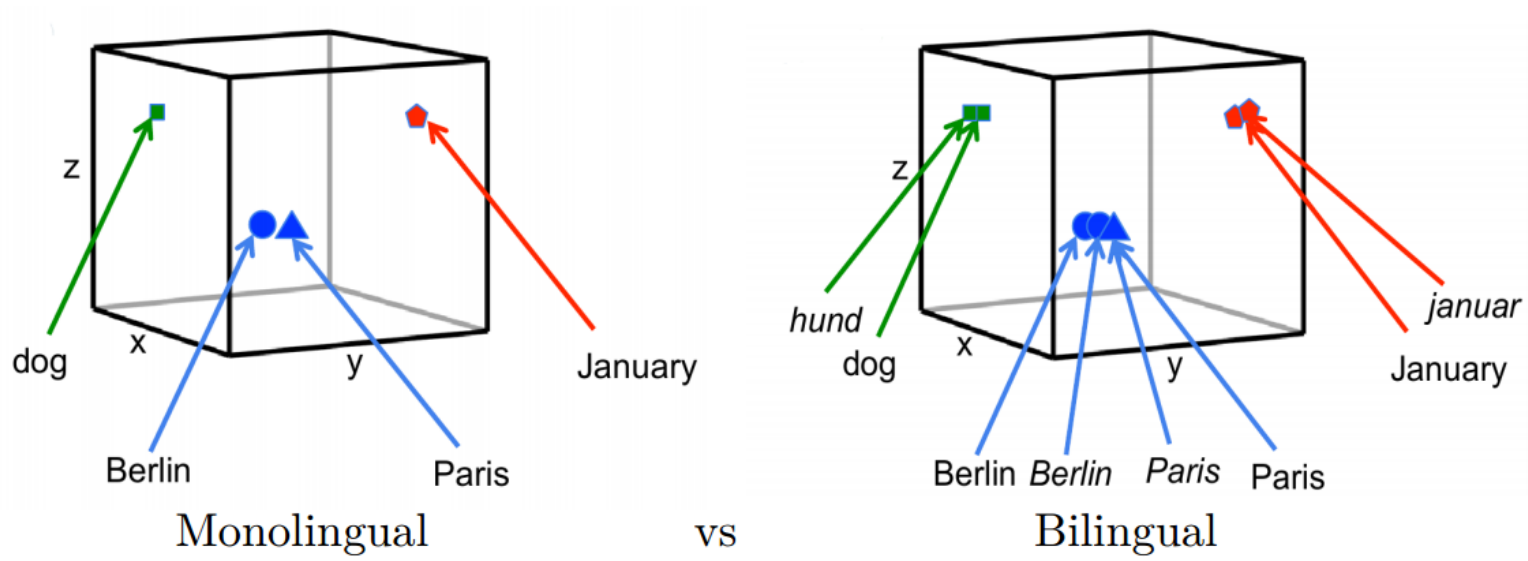
- 이전의 연구에서는 중국어-영어 간의 Bilingual training을 위해 사용되었다.

1. Introduction

BWE with SBWES

- 최근 **Bilingual Word Embedding(BWE)**에 대한 관심이 증가하고 있다.

- **BWE learning model**은 **Shared Bilingual Word Embedding Space(SBWES)**의 유도에 초점을 맞추며, 언어에 무관하게 유사한 의미의 단어가 SBWES 상의 유사한 공간에 mapping되도록 한다.



1. Introduction

Fields Related to SBWES

- **Computing cross-lingual/multilingual semantic word similarity** (Faruqui and Dyer, 2014)
  - *여러 언어에 존재하는 단어에 대한 유사성 계산*
- **Learning bilingual word lexicons** (Mikolov et al., 2013a; Gouws et al., 2015; Vulic et al., 2016)
  - *두 가지 다른 언어에 대해 word lexicons을 학습*
- **Cross-lingual entity linking** (Tsai and Roth, 2016)
  - *문서 내 등장하는 특정 단어가 다른 언어의 어떤 의미로 사용됐는지 추론*
- **Parsing** (Guo et al., 2015; Johannsen et al., 2015)
  - *입력된 문장의 구조를 분석하는 과정*
- **Machine translation** (Zou et al., 2013)
  - *인간이 사용하는 자연 언어를 컴퓨터를 사용하여 다른 언어로 번역*
- **Cross-lingual information retrieval** (Vulic and Moens, 2015; Mitra et al., 2016)
  - *사용자가 원하는 다른 언어의 정보를 검색도구를 활용하여 찾아내는 과정*

1. Introduction

Two Desirable Properties for BWE

- **P1) 풍부한 monolingual training data를 활용**하고, bilingual signal을 이용하여 각 언어를 연결할 수 있을 것
- **P2)** 다양한 언어 및 분야로 연구를 확장할 수 있도록 **최소한의 bilingual signal로도 SBWES의 학습이 가능**할 것

Type 4 Operates as Follow

Approach 01

두 개의 별도로 준비 된 **non-aligned monolingual embedding spaces**는 **monolingual WE learning model**을 사용하여 유도하게 된다 (일반적인 선택은 SGNS).

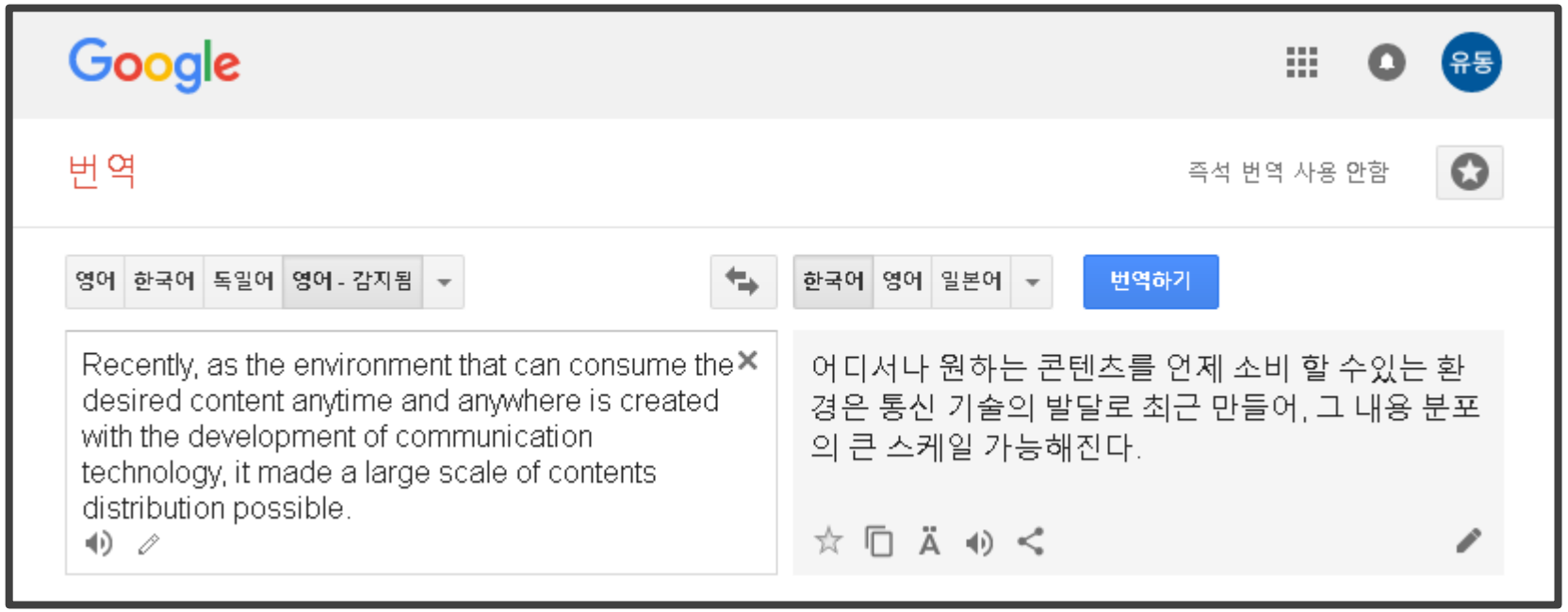
Approach 02

트레이닝을 위한 **bilingual signal**은 단어 번역 쌍을 seed lexicon으로 사용하고, 이를 통해 두 개의 monolingual spaces를 하나로 묶는 mapping function을 학습함으로써 **SBWES**를 구성한다.

1. Introduction

Problem of Existing Research

- 기존의 연구에서는 **“High-quality training seed lexicons”** 가 존재한다고 가정

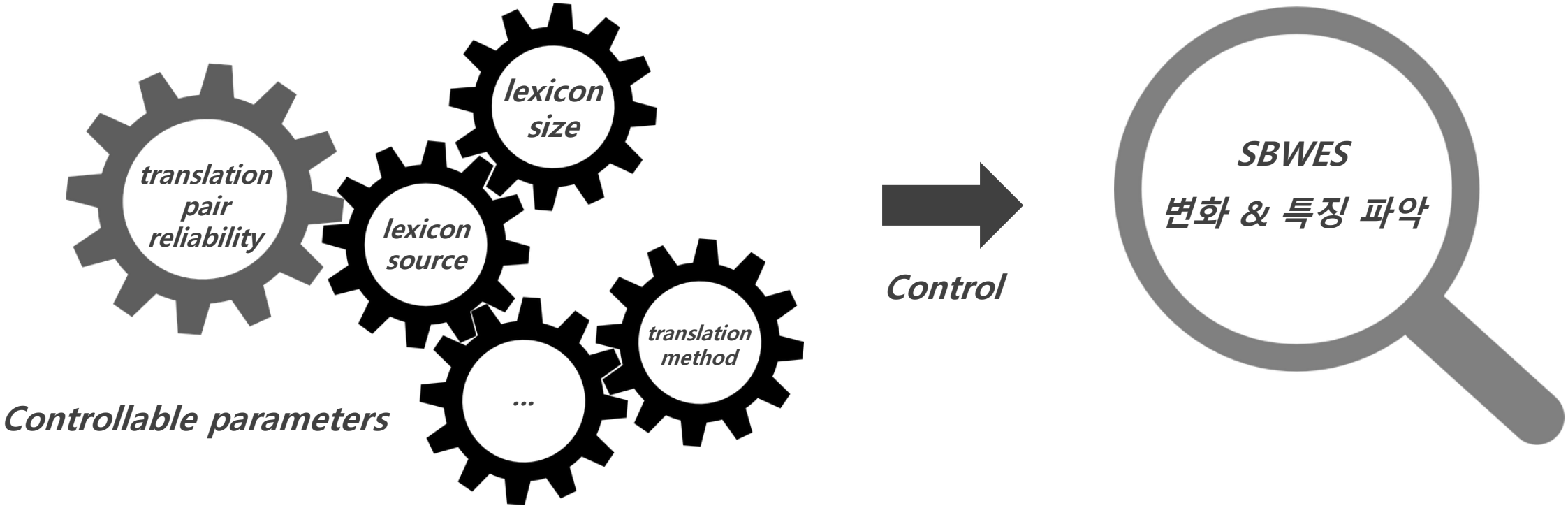


High-Quality?

1. Introduction

Goal

- **Controllable parameters**의 변화에 따른 **SBWES**의 특징을 파악하여 "**seed lexicon**"을 조금 더 지능적으로 선택할 수 있도록 한다.



1. Introduction

Contributions

C1

Monolingual WE spaces 사이의 mapping function 학습을 위한 **seed lexicon의 중요성에 대해 체계적인 연구를** 제시하였다.

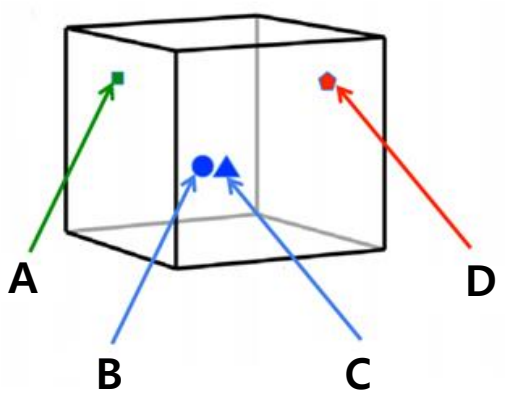
C2

(P1), (P2)를 만족시키는 간단하면서도 효과적인 **HYbrid BWE model(HYBWE)**의 제안하였다.

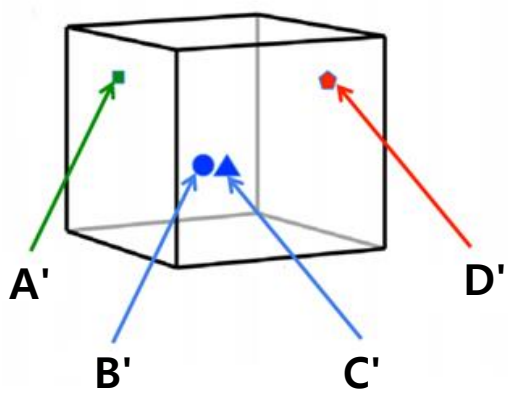
C3

Seed lexicon 구성 시 단어 쌍을 신중하게 선택함으로써 Bilingual Lexicon Learning task에서 **HYBWE**가 기존의 **BWE model들의 성능을 능가**함을 보였다.

2. Learning SBWES using Seed Lexicons



Source WES



Target WES

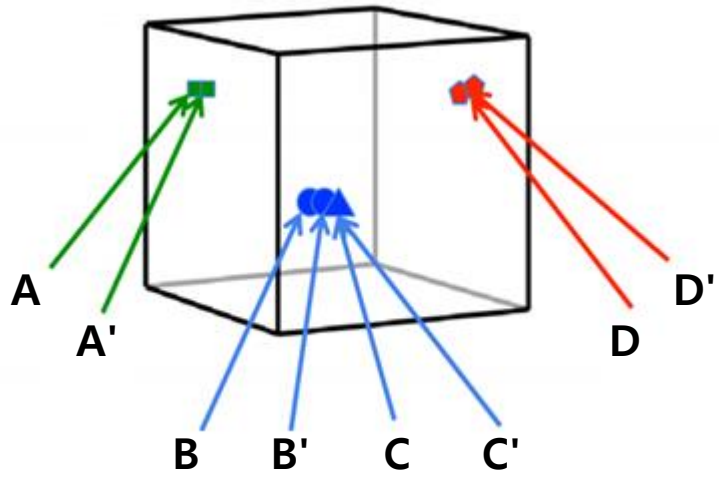
- 모든 BWE model은 주어진 source와 target 언어의 어휘 $w \in V^S \cup V^T$을 SBWES 상의 d 차원의 실수 벡터 $[f_1, \ldots, f_d]$로 mapping하는 것을 목표로 함
- 두 단어 w, v 사이의 Semantic similarity $sim(w, v)$는 **similarity function(SF)**를 적용하여 계산한다.

Using seed lexicon

V^S = *A, B, C, D*

V^T = *A', B', C', D'*

SBWES



d-dimensional real-valued vector

Mapping

2.1 Related work : BWE Models and Bilingual Signals

- BWE model들은 학습에 사용되는 **bilingual signal 및 특성 (P1), (P2)에 따라** 네 가지 유형으로 분류가 가능하다.

  - **Type 1 : *Parallel-Only***
  - **Type 2 : *Joint-Bilingual Training***
  - **Type 3 : *Pseudo-Bilingual Training***
  - **Type 4 : *Post-Hoc Mapping with Seed Lexicons***

(Type 1) Parallel-Only

- Data source로 문장 and/or 단어 정렬된 parallel data에 의존함
- 학습 시 풍부한 monolingual dataset들을 사용하지 않음 **(not satisfying P1)**
- 많은 bilingual signal을 필요로 함 **(colliding with P2)**

2.1 Related work : BWE Models and Bilingual Signals

(Type 2) Joint-Bilingual Training

- 이 모델은 학습 시 cross-lingual regularizer로 cross-lingual objective 와 monolingual objective를 동시에 최적화함

- **Objective function**: $\gamma(\mathrm{Mono_S} + \mathrm{Mono_T}) + \delta\mathrm{Bi}$

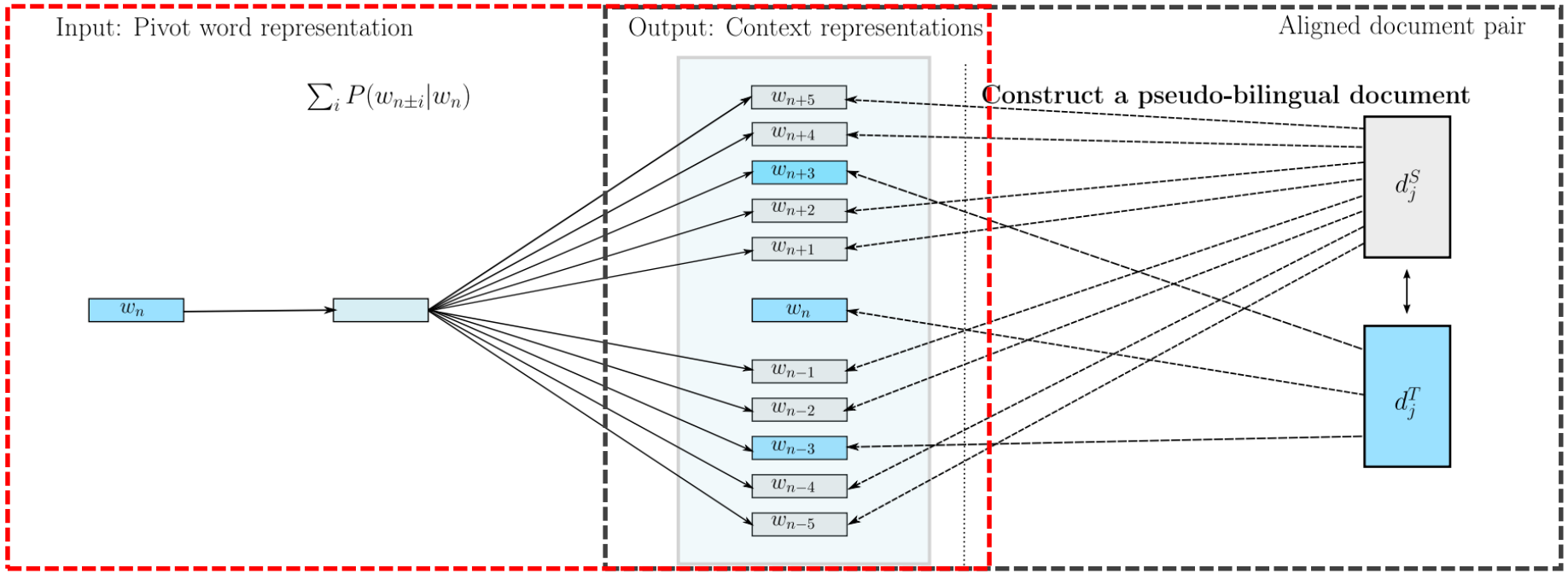
- Monolingual objective는 각 언어에서 유사한 단어에 유사한 embedding 할당을 유도함과 동시에 각 언어의 semantic 구조 학습을 목표로 하고, cross-lingual objective는 두 언어 간 유사한 단어에 유사한 embedding이 할당되도록 유도함
- 이 모델은 SBWES에 두 개의 monolingual spaces를 동시에 유도함 **(satisfying P1)**
- Bilingual signal로 구하기 어려운 parallel data를 필요로 함 **(colliding with P2)**

2.1 Related work : BWE Models and Bilingual Signals

(Type 3) Pseudo-Bilingual Training

- 이 모델에서는 SBWES를 유도하는 bilingual signal로 문서 쌍을 사용함
- 문서 쌍을 important local information를 보존하는 방식으로 병합하여 pseudo-bilingual training 데이터를 만들고, 이 데이터로 word2vec 중 SGNS을 training하는 데 사용
- 비교적 얻기 쉬운 문서 번역 쌍을 bilingual signal로 사용함 – Parallel data를 사용하지 않음 **(satisfying P2)**
- Training에 monolingual corpus를 이용할 수 없음 **(unlike Type2, Type4 ; colliding with P1)**

SGNS



2.1 Related work : BWE Models and Bilingual Signals

What is Comparable Corpus?

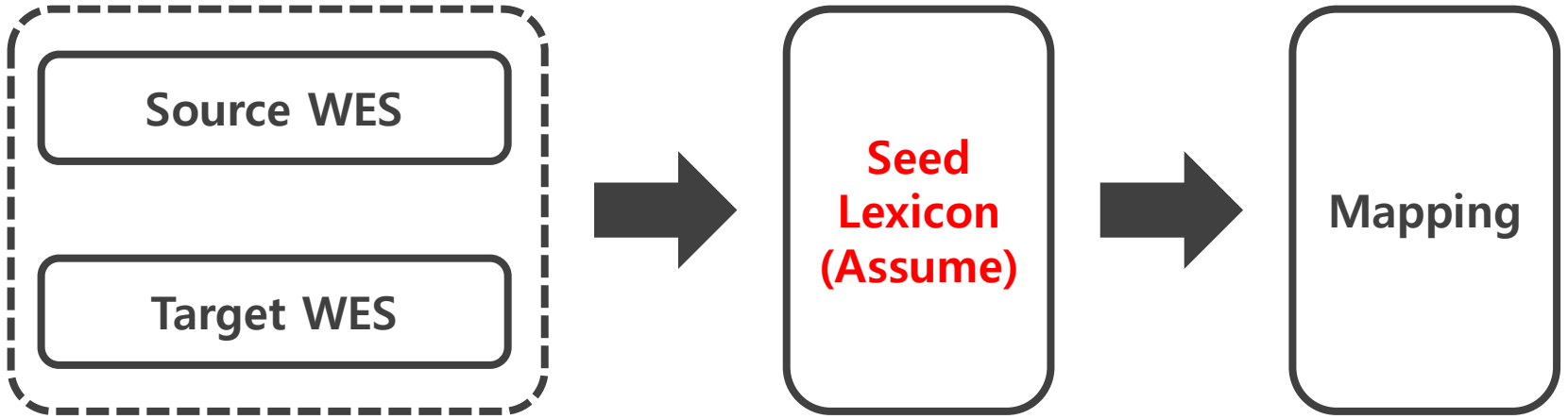
- **Comparable Corpus**란, **언어는 다르지만 내용은 유사한 문서 쌍들의 집합**을 말한다 (문서의 유사도에 대한 명확한 기준은 없음).

- 문장 단위의 정확한 변역보다는 문서의 문맥이나 상황에 집중함으로써, 문장 단위 번역에서 생길 수 있는 내용의 왜곡을 줄이는 것이 목적이다.

2.1 Related work : BWE Models and Bilingual Signals

(Type 4) Post-Hoc Mapping with Seed Lexicons

- 이 모델은 두 개의 다른 언어에 대해 개별적으로 유도 된 monolingual WE spaces 사이의 사후매핑 function을 학습함
- 기존의 Type 4 model은 Google Translation(GT)등과 같은 쉽게 사용할 수 있는 seed lexicon에 의존함(가정)
- 이미 seed lexicon이 존재한다고 가정 **(colliding with P2)**, 그러나 조건 **P1**은 충족시킬 수 있음



monolingual training set를 서로 연결(satisfying P1)

2.2 Post-Hoc Mapping with Seed Lexicons

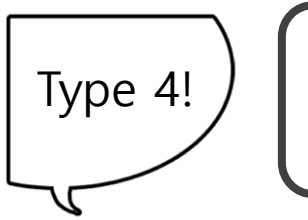
Key Intuition

- Type 3 model을 통해 seed lexicon을 생성하고, 이를 Type 4 model 학습에 사용함으로써 **안정적인 translation pairs을 사용**하게 되며, 이러한 방법을 사용한 Type-hybrid 모델은 다음의 요구 사항을 충족시킬 수 있다.

P1

Type 1, Type 3과는 달리 **monolingual data로부터 학습할 수 있고**, 신뢰도가 높은 translation pairs를 사용하여 두 개의 monolingual space를 묶을 수 있다.

P2



Type 1, Type 2와는 달리 **parallel data를 필요로 하지 않으며**, 요구되는 유일한 bilingual signal은 문서 번역 쌍이다. 따라서, 본 연구의 초점은 이 새로운 Type 4 model의 변형이다.

2.2 Post-Hoc Mapping with Seed Lexicons

Key Intuition

Type 3 ▶ **Make seed lexicon**

The role of seed lexicon = The role of Google translation

Seed lexicon from Type 3 ▶ **Type 4**

Overview – Standard Learning Setup

- 각각의 monolingual embedding spaces $\mathbb{R}^{d_S}$와 $\mathbb{R}^{d_T}$를 CBOW(Continuous Bag Of Words) 또는 SGNS와 같은 **표준 monolingual WE model을 사용하여 유도**한다. (d_S와 d_T는 monolingual spaces의 차원의 수)

- **Bilingual seed lexicon** = 단어 쌍 (x_i, y_i), where $x_i \in V^S, y_i \in V^T$

2.2 Post-Hoc Mapping with Seed Lexicons

Learning Objectives

- Training은 multivariate regression problem으로 볼 수 있으며, 이는 training data(seed lexicon)을 사용하여 **source language vectors를 target language vector로 mapping하는 function의 학습**을 의미한다.

- 일반적인 접근 방식으로는 **L2-regularized least-squares error objective**를 사용하여 mapping matrix W를 학습하는 것이다. Linear map이 $W \in R^{d_S \times d_T}$ 이며 , map은 다음의 최적화 문제를 해결함으로써 학습한다.

$$\min_{\mathbf{W} \in \mathbb{R}^{d_S \times d_T}} \|\mathbf{XW} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2$$

W : *mapping function*

OUTPUT

INPUT

X : Source 언어의 word 벡터 ***Y :*** *Target 언어의 word 벡터*

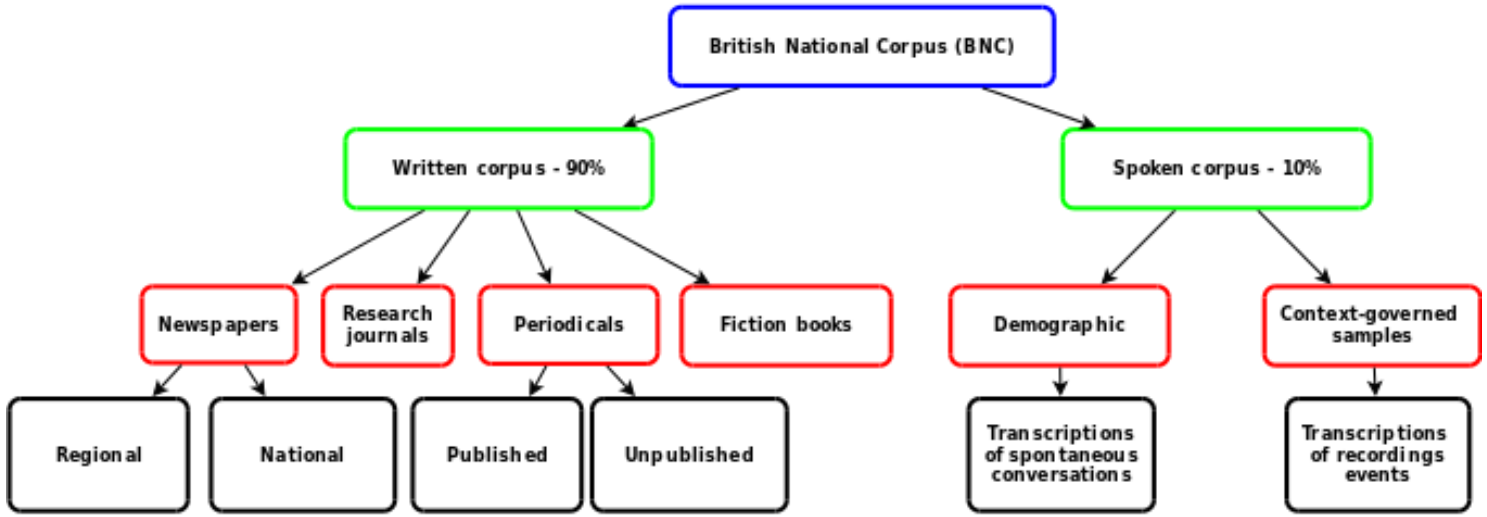
X, Y : *Seed lexicon*

2.2 Post-Hoc Mapping with Seed Lexicons

Seed Lexicon Source and Translation Method

- 이전의 연구는 post-hoc mapping에 사용할 seed lexicon으로 높은 빈도로 나타난 영어 단어를 **번역 시스템을 이용하여** Czech, Spanish, Italian 등의 다른 언어로 번역하여 생성했다.

- 이 방법은 높은 **품질의 외부 번역 시스템이 존재함을 전제**로 하며, 이 실험에서는 BNC 단어 빈도 목록 [Kilgarriff(1970)] 중 가장 빈번히 등장한 기본형 영어 단어 6,318개 (BNC list)를 사용했다.



2.2 Post-Hoc Mapping with Seed Lexicons

Another Option Proposed in this Paper

- 본 연구에서는 다른 BWE model을 사용하여 "1차" **SBWES-1을 학습**하고, BNC list의 각 x_i에 대하여 SBWES-1상에서의 가장 가까운 cross-lingual word $y_i \in V^T$ 를 얻는다.

- 단어 쌍 (x_i, y_i)는 type 4 모델에서 monolingual spaces 간의 mapping을 학습하는데 사용할 seed lexicon이 된다. 그리고 이 seed lexicon은 최종 **SBWES-2를 유도**하는데 사용된다.

- 본 연구에서는 SBWES-1 학습에 **document-level Type 3 BWE induction model에 의존한다**(Type 1,2제외)**.**

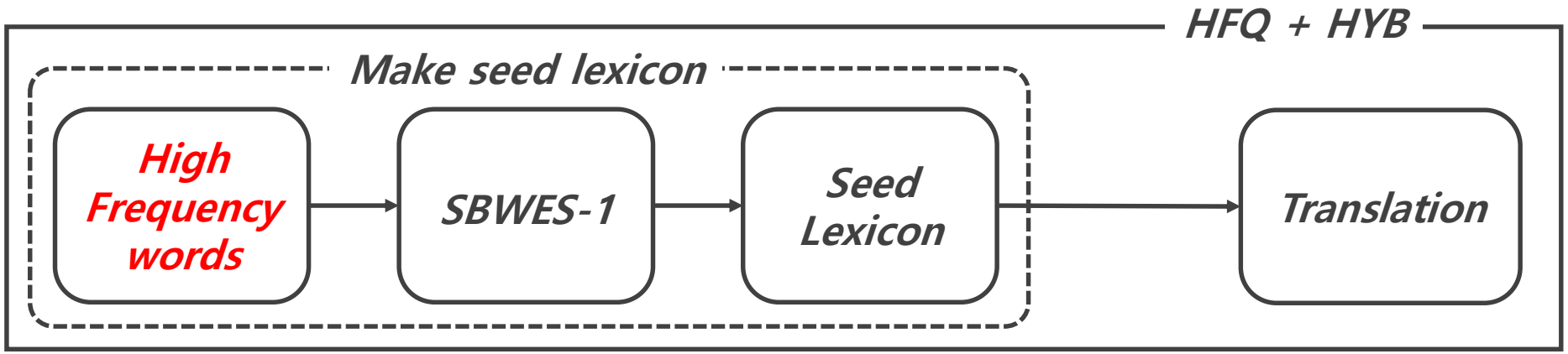
- HYBWE는 Type 3 model(SBWES-1) 및 Type 4 model(SBWES-2)을 결합하며, 이 seed lexicon 및 BWE 학습의 변형은 **BNC + HYB**로 칭하기로 한다.

2.2 Post-Hoc Mapping with Seed Lexicons

Another Option Proposed in this Paper

- SGNS와 같은 통계적 모델의 특성상 **높은 빈도로 등장한 단어의 의미가 더 정확히 학습되어 있다는 가정**에 따라 hybrid model의 seed lexicon source로 SBWES-1 training data에서 가장 자주 나타나는 단어를 사용하는 것을 고려하기로 한다.

- 이 고 빈도 단어들을 SBWES-1을 통해 번역하여 seed lexicon pairs (x_i, y_i) 를 얻으며, 이 seed lexicon을 사용한 모델은 **HFQ + HYB**로 칭하기로 한다.



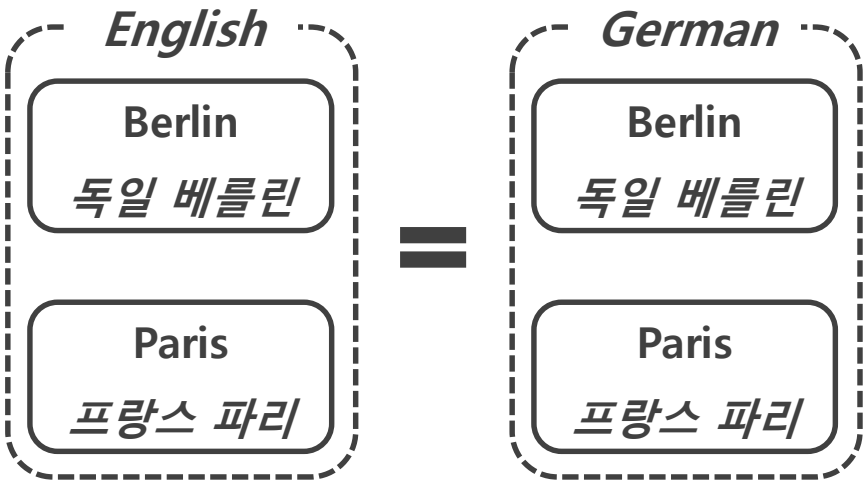
2.2 Post-Hoc Mapping with Seed Lexicons

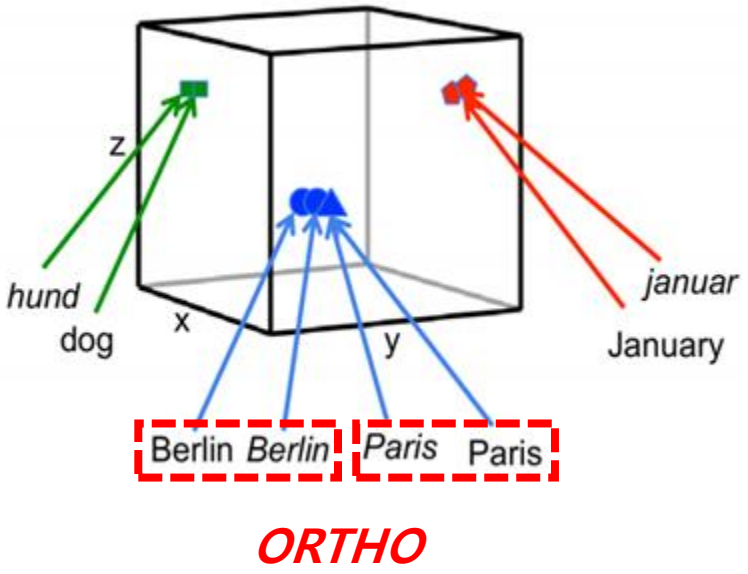
Another Option Proposed in this Paper

- 최근 제안된 새로운 모델로, mapping function을 학습하기 위해 **두 개의 언어 사이에서 공유되는 단어들을 사용**하는 방법이 있다. 본 연구에서는 이러한 접근 방법의 성능과 한계 또한 실험해보기로 한다.

- Seed lexicon pairs: (x_i, x_i) where $x_i \in V^S$ and $x_i \in V^T$

- 이러한 seed lexicon을 사용하는 변형은 **ORTHO**로 칭하기로 한다.





2.2 Post-Hoc Mapping with Seed Lexicons

Seed Lexicon Size

- 기존의 연구에서는 **제한된 seed lexicon size에 대한 결과만 보고되어 왔다**(일반적으로 1K, 2K, 5K 개). 본 연구에서는 다음 두 가지 중요한 질문에 대한 대답을 찾기 위해 **더 극단적인 설정**도 실험하였다.

(1)

Type 4 SBWES는 오직 몇 백 개의 단어 쌍만 사용 가능한 **제한된 환경에서** 유도될 수 있는가?

(2)

더 많은 seed lexicon pairs를 사용한다면 Type 4 model은 더 좋은 성능을 보이는가?

2.2 Post-Hoc Mapping with Seed Lexicons

Translation Pair Reliability

- **SBWES-1을 통해** seed lexicon을 구축할 때, **신뢰할 수 있는 단어 쌍의 사용**이 더 우수한 **SBWES-2**로 이어질 수 있다는 점에 착안하여 seed lexicon 구축 시 번역의 신뢰도를 제어할 수 있는 방법을 제안한다.

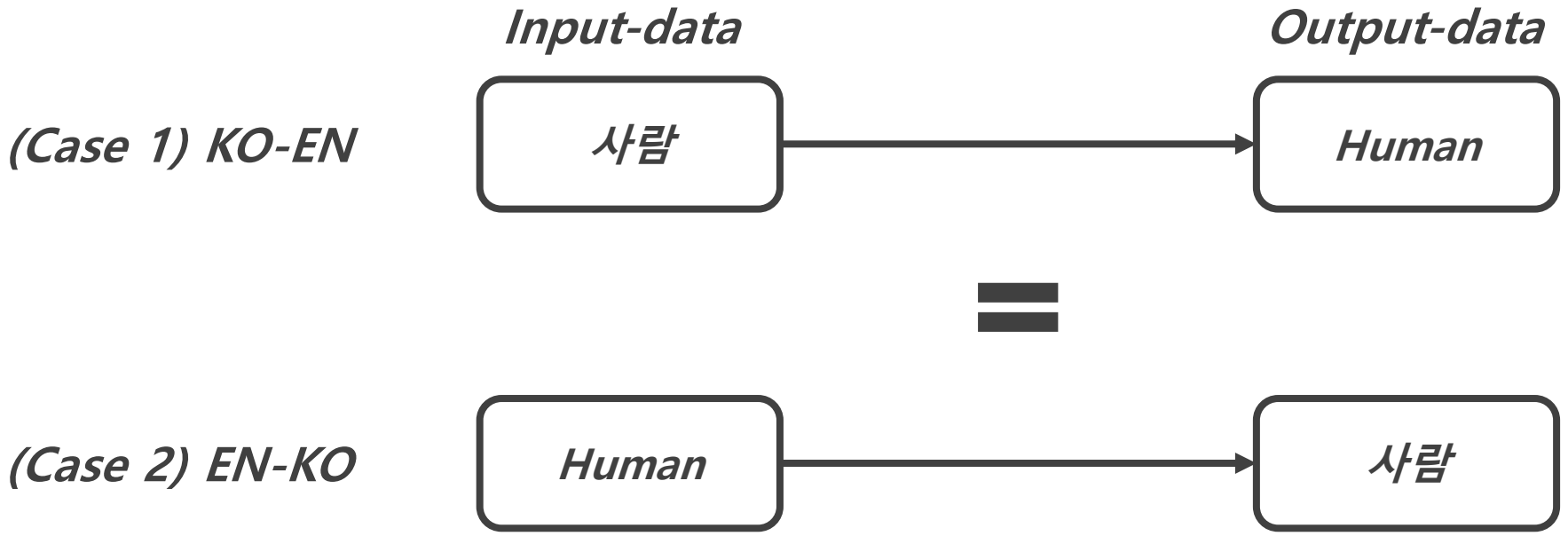
- 번역 쌍을 위한 간단하지만 효과적인 신뢰도 제어 방법은 **symmetry constraint**이다.

- SBWES-1 상에서 두 단어의 벡터가 **"서로" 가장 가까운 이웃인 경우에만** 두 단어 $x_i \in V^S, y_i \in V^T$를 seed lexicon으로 사용한다.

- 이러한 제약 조건을 추가한 seed lexicon을 사용한 모델은 각각 BNC+HYB+**SYM**와 HFQ+HYB+**SYM**로 칭하기로 하며, symmetry constraint이 없는 모델은 BNC+HYB+**ASYM**와 HFQ+HYB+**ASYM**로 칭하기로 한다.

2.2 Post-Hoc Mapping with Seed Lexicons

What is Symmetry Constraint?



- 일반적으로는 Case 1만 만족시켜도 seed lexicon pair로 사용하지만(e.g. GT)
  symmetry constraint의 경우 Case 1, Case 2를 **모두 만족하는 경우** seed lexicon pair로 사용한다.

2.2 Post-Hoc Mapping with Seed Lexicons

Translation Pair Reliability

- 더욱 까다로운 신뢰도 제어 방법으로, 단어 x_i의 번역의 모호성을 수치화하는 방법을 제안한다.

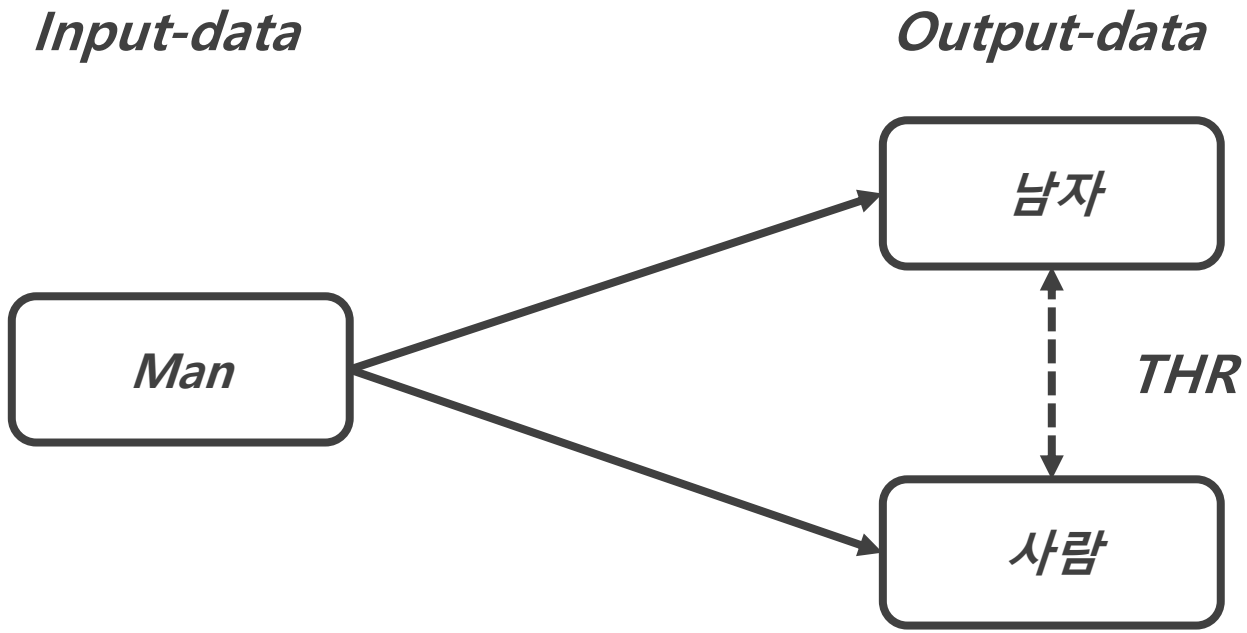
- Symmetry constraint with a threshold: 두 단어 x_i, y_i가 **SBWES-1상에서 서로 가장 가까운 이웃인 동시에 다음도 만족시킬 경우** seed lexicon pair (x_i, y_i)로 포함시킨다.

$$sim(x_i, y_i) - sim(x_i, z_i) > THR$$
$$sim(y_i, x_i) - sim(y_i, w_i) > THR$$

- 단어 $z_i \in V^T$는 x_i의 차선 번역이고, $w_i \in V^S$는 y_i의 차선 번역이며, THR은 최선 번역과 차선 번역 사이 거리의 임계값을 의미한다 **(최소 거리 값)**.

2.2 Post-Hoc Mapping with Seed Lexicons

What is Symmetry Constraint?



- 두 가지 이상의 번역 결과가 나타날 때의 **모호성이 존재할 수 있다.** 때문에, 각 번역의 similarity를 계산하여 일정 수치(THR) 내에 들어오면 **모호성 문제를 일으킬 수 있다고 감지하고 제거한다.**

3. Experimental Setup

Task : Bilingual Lexicon Learning (BLL)

- Source 언어의 단어 n개 $x_{u1, \dots,} x_{un}$에 대하여 SBWES를 이용하여 각 단어의 target 언어로의 번역 t를 찾는 것을 목표로 한다. t는 유도 된 SBWES 내의 source 언어의 단어 x_u에 **가장 가까운 target 언어의 단어**로서, cross-lingual nearest neighbor라고도 한다.

- 이렇게 얻어진 n쌍의 단어 (x_u, t)를 정답 BLL test set과 비교하여 정확도를 계산한다.

- 표준 절차에 따라, 정확도 계산에 사용되는 모든 단어 $x_{u1, \dots,} x_{un}$들은 Type 4 model에서의 **학습용 seed lexicons에서 제거된다.**

➡ ***BLL 정확도 계산에 사용되는 단어를 seed lexicon으로 사용할 경우,***
학습된 모델의 성능을 공정하게 평가할 수 없기 때문에

3. Experimental Setup

Test Sets

- 본 연구에서는 세 개의 언어 쌍으로 이루어진 일대일 정답 번역 1,000쌍을 통해 BLL task를 평가한다:
  {Spanish (ES)-, Dutch (NL)-, Italian (IT)}-English (EN)

- 이 dataset은 일반적으로 non-parallel data로부터 학습된 BLL model의 성능을 평가하는 표준 테스트로 사용된다.

Evaluation Metrics

- BLL 성능 측정에는 표준으로 사용되는Top 1 accuracy **(Acc1)**를 사용한다.
  (모델의 출력 중 **가장 가능성 높은 1개만** 정답과 비교)

3. Experimental Setup

Baseline Models

- SBWES-1을 유도하기 위해 Vulic and Moens (2016)의 **document-level embedding (Type 3) 를 사용**

- ***BiCVM*** = *Type 1의 대표 model*

  BilBOWA = *Type 2의 대표 model*

  BWESG = *Type 3의 대표 model*

  BNC+GT = *Type 4의 대표 model*

3. Experimental Setup

Training Data and Setup

- 본 연구에서는 모든 모델에 대해 각각의 언어에서 **가장 빈번하게 사용된 단어 100K** 개만 사용하였다.

- **Monolingual WE spaces의 유도에는 SGNS 알고리즘과 Stochastic Gradient Descent(SGD), global learning rate 0.025**를 사용하였으며, 학습 데이터는 정리 및 토큰화 된 Wikipedia text를 사용하였다.

- **BilBOWA** 모델의 재현은 원 논문과 저자의 제안에 따라 **SGD, global learning rate 0.15**를 사용했으며, window size는 2-16사이에서 2 단위로 바꿔가며 실험하고 성능은 **가장 우수한 모델의 것만을 표시**하였다. 학습 데이터는 Europarl.v7의 첫 500K 개의 문장을 사용하였다.

- **BWESG** 모델의 재현에는 **SGD, global learning rate 0.025**를 사용했으며, 학습 데이터는 정리 및 토큰화 된 Wikipedia 번역 쌍을 이용하여 생성한 **pseudo-bilingual documents**를 사용하였다.

3. Experimental Setup

Training Data and Setup

- **BiCVM** 모델의 재현에는 원 저자가 공개한 툴을 사용하였으며, 각 언어 쌍에 대하여 전체 Europarl.v7을 학습 데이터로 사용하였다. 모든 BiCVM 모델은 200번의 반복된 훈련을 거쳤다.

- 모든 모델에 대하여, 40, 64, 300, 500 차원의 BWE 모델들을 제작하였으나, 모두 유사한 특성을 보임에 따라 **300 차원 BWE들에 대한 결과만을 표에 나타냈다.**

- 다른 매개 변수 : 15 epochs, 15 negatives, 1e-4의 subsampling rate

4. Result and Discussion

Exp. Ⅰ: Standard BLL Setting

- 기존 Type 4 model에 사용했던 BLL test와 같이 source 언어의 단어를 고빈도 순으로 5K 단어 쌍을 선택하여 seed lexicon으로 사용하였다.

- 아래 표는 결혼을 뜻하는 스페인어 "casamiento"를 input 값으로 넣었을 때, 각 모델 별로 **가장 가까운 단어 7개**를 나타낸 예시로서, Type 3 모델을 통해 간접적으로 생성한 seed lexicon으로도 **기존의 type 4 모델과 비슷하거나 더 우수한 SBWES**를 유도할 수 있음을 시사한다. (단, ORTHO 제외)

BNC+GT	BNC+HYB+ASYM	BNC+HYB+SYM	HFQ+HYB+ASYM	HFQ+HYB+SYM	ORTHO
casamiento	*casamiento*	*casamiento*	*casamiento*	*casamiento*	*casamiento*
marriage	marry	*marriage*	*marriage*	*marriage*	maría
marry	*marriage*	marry	marry	marry	señor
marrying	marrying	marrying	betrothal	betrothal	doña
betrothal	wed	wedding	marrying	marrying	juana
wedding	wedding	betrothal	wedding	wedding	noche
wed	betrothal	wed	daughter	wed	amor
elopement	remarry	marriages	betrothed	elopement	guerra

4. Result and Discussion

Exp. Ⅰ: Standard BLL Setting

Model	ES-EN	NL-EN	IT-EN
BICVM (TYPE 1)	0.532	0.583	0.569
BILBOWA (TYPE 2)	0.632	0.636	0.647
BWESG (TYPE 3)	0.676	0.626	0.643
BNC+GT (Type 4)	0.677	0.641	0.646
ORTHO	0.233	0.506	0.224
BNC+HYB+ASYM	0.673	0.626	0.644
BNC+HYB+SYM (3388; 2738; 3145)	0.681	**0.658***	0.663*
HFQ+HYB+ASYM	0.673	0.596	0.635
HFQ+HYB+SYM	**0.695***	0.657*	**0.667***

[세 언어 쌍에서의 각 모델의 정확도]

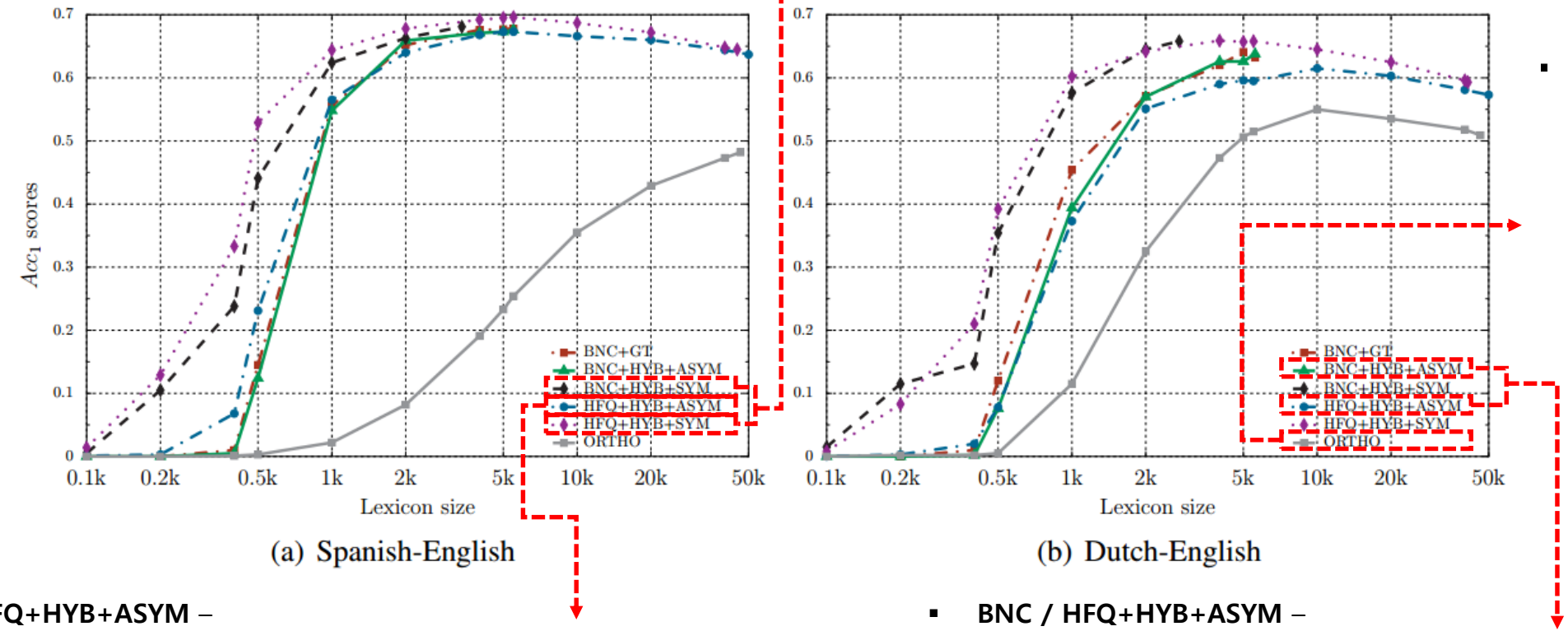
Results

- BNC/HFQ + HYB +SYM 모델은 **모든 실험에서 Type 1, 2, 3 모델의 성능을 상회했다**.
- 특히, BNC+ HYB +SYM의 결과로 볼 때, **신뢰도가 높은 단어 쌍을 신중히 선택**하면 적은 단어 쌍으로도 우수한 성능을 이끌어낼 수 있음을 알 수 있다.

4. Result and Discussion

Exp. II: Lexicon Size

- **BNC+HYB+SYM** and **HFQ+HYB+SYM** –
  symmetric pairs의 set에 의존하며, 이 모델들은 **모든 lexicon size에서 최고의 성능**을 보인다.



(a) Spanish-English

(b) Dutch-English

- **ORTHO** –
  성능에서 경쟁력은 부족하나, **가장 쉽게 얻을 수 있는 bilingual signal**에 의존함에도 불구하고 **합리적인 BLL 점수**를 보여주었다.

- **HFQ+HYB+ASYM** –
  Type 4 model, document-level Type 3, 번역 신뢰도 제어를 조합한 간단한 HYBWE model은 **전반적으로 강력한 성능**을 보여준다.

- **BNC / HFQ+HYB+ASYM** –
  외부의 lexicon or translation system을 BLL 성능에 **손실 없이** document-level embedding model로 대체할 수 있음을 보여준다.

4. Result and Discussion

Exp. Ⅱ: Lexicon Size

- **Symmetry constraint**을 적용한 두 모델은 **단어 쌍의 수가 부족**할 때 상대적으로 **강한 성능**을 보였다.

- **더 많은 단어 쌍의 추가는 BLL 성능의 향상으로 이어지지 않았다.** 이는 SBWES-1에 등장 빈도가 낮고, 따라서 의미의 정확도가 떨어지는 단어는 오히려 SBWES-2의 학습을 방해할 수 있음을 시사한다.

Quantity < Quality

4. Result and Discussion

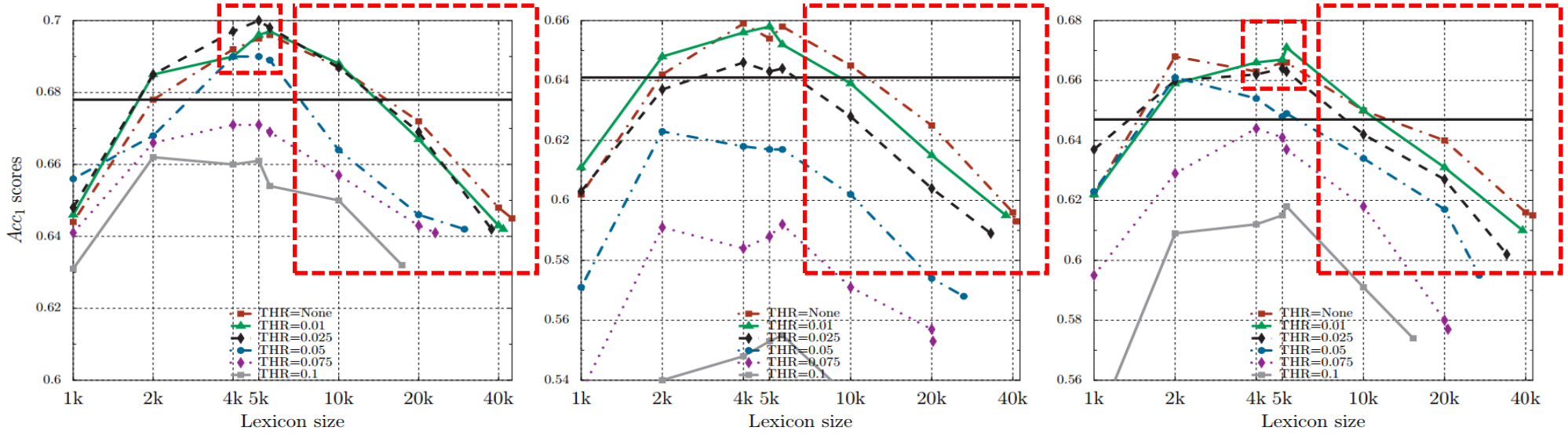
Exp. Ⅲ: Translation Pair Reliability

- **실험 내용**

  HFQ+HYB+SYM 모델에서 THR 임계 값을 변화시킨다.

- **세 언어 쌍에 대한 결과**

  까다로운 기준을 적용했을 때 모든 모델에서 성능이 **감소하는 것**을 보였다.



(a) Spanish-English (b) Dutch-English (c) Italian-English

- **강화된 선택 기준(0.01, 0.025 THR)**

  5K 근처의 ES-EN, IT-EN에서 **BLL 점수 향상**에 미치는 영향을 관찰하였다.

- **HFQ+HYB+SYM과 HYBWE의 전체 성능 저하의 요인**

  높은 임계 값(THR), 과도하게 까다로운 pair 선택 기준

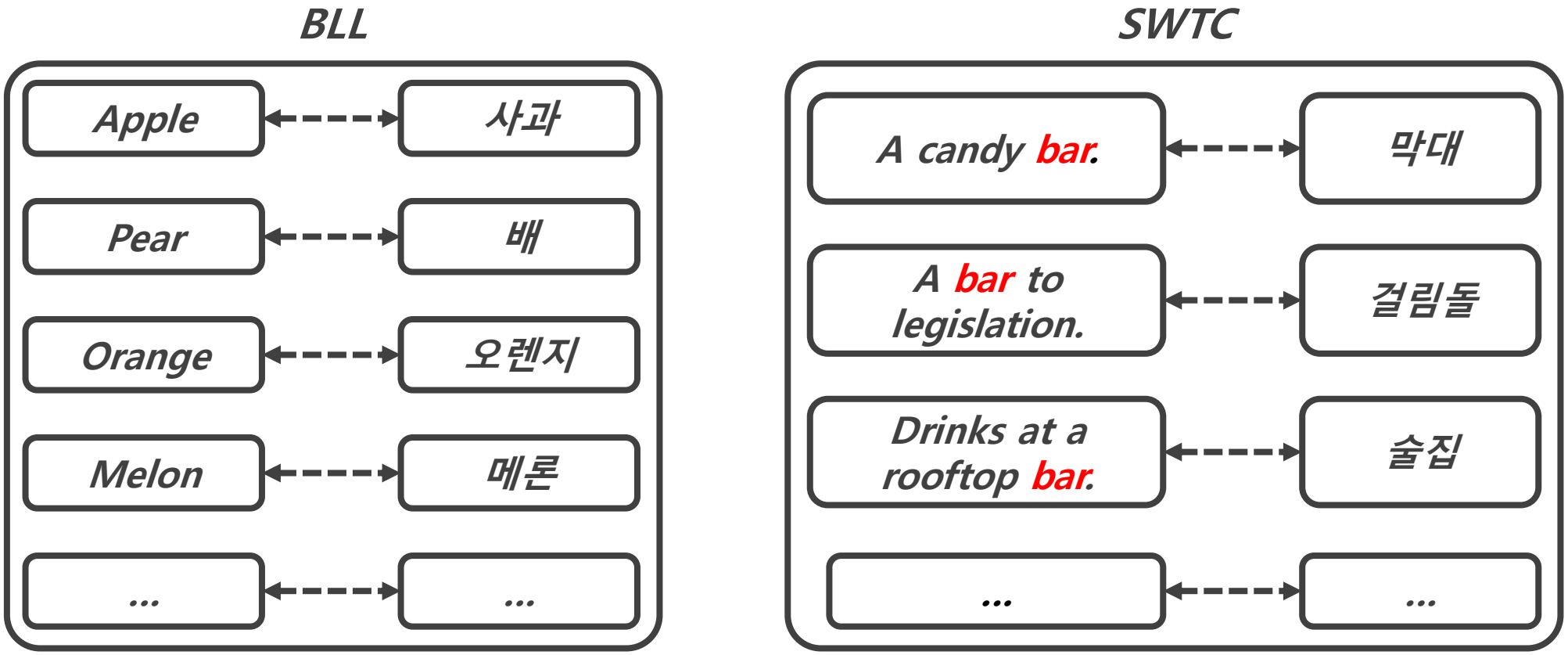
4. Result and Discussion

Exp. IV: Another Task – Suggesting Word Translations in Context (SWTC)

- Vulic and Moens (2014)에 의해 제안 된, **Suggesting Word Translations in Context (SWTC)** 실험
- 문맥 속 단어 $w \in V^S$에 여러 의미가 있을 수 있으므로, SWTC task에서는 주어진 단어들의 목록 $TC(w) = \{t_1, \dots, t_{tq}\}, TC(w) \subseteq V^T$ 중 문맥을 고려했을 때 w의 번역으로 가장 알맞은 것을 선택한다.
- BLL 실험과 달리 SWTC에서는 V^T의 전체 단어가 아닌, **몇 개의 사전에 주어진 단어들 중에서 정답을 선택한다.**

4. Result and Discussion

The Difference between BLL and SWIC



- BLL task는 **word-word**에서 translation을 수행하지만, SWTC는 **sentence-word**에서 translation을 수행한다.

4. Result and Discussion

Exp. IV: Another Task – Suggesting Word Translations in Context (SWTC)

- 본 연구에서는 SWTC 실험에서 BWE 모델의 효과를 분리시켜 검증하기 위해 모든 모델에 대해 다음과 같이 동일한 방법으로 번역 단어를 선정했다. 문맥으로 등장하는 source 언어에서 단어들의 벡터는 **다음 식을 이용하여 하나로 합친다**.

$$\mathbf{Con(w)} = \mathbf{cw}_1 + \mathbf{cw}_2 + \ldots + \mathbf{cw}_r$$

- 이후, 번역 후보로 주어진 target 언어의 단어들 중 SBWES에서 Con(w)와 **가장 가까운 단어**를 번역 단어로 선택한다. (Cosine similarity 사용)

- 평가 set는 세 개의 언어에서 각각 15개의 다양한 의미를 가진 명사를 기반으로 한 360 문장을 구성하였다.

- 다음 표에 SWTC task의 결과**(Acc1 scores)**가 요약되어 나타나 있다.

4. Result and Discussion

Exp. IV: Another Task – Suggesting Word Translations in Context (SWTC)

Model	ES-EN	NL-EN	IT-EN
NO CONTEXT	0.406	0.433	0.408
BEST SYSTEM (Vulić and Moens, 2014)	0.703	0.712	0.789
BICVM (TYPE 1)	0.506	0.586	0.522
BILBOWA (TYPE 2)	0.586	0.656	0.589
BWESG (TYPE 3)	0.783	0.858	0.792
BNC+GT (TYPE 4)	0.794	0.858	0.783
ORTHO	0.647	0.794	0.678
BNC+HYB+ASYM	0.806*	0.872	0.778
BNC+HYB+SYM	**0.808***	**0.875***	**0.814***
(3839; 3117; 3693)			
HFQ+HYB+ASYM	0.789	0.864	0.781
HFQ+HYB+SYM (THR = None)	0.792	0.869	0.786
HFQ+HYB+SYM (THR=0.01)	0.792	0.858	0.789
HFQ+HYB+SYM (THR=0.025)	0.800	0.853	0.792

Results

- **최고 성능 모델**은 BNC+HYB+SYM와 HFQ+HYB+SYM으로 나타났다.
- ASYM과 SYM의 비교에서 **symmetry constraint**을 활용한 translation pairs 필터링 작업을 통해 **성능의 향상**을 보였으나, 높은 THR, 엄격한 선택 기준은 되려 성능을 저하시켰다.
- 다양한 HYBWE 모델들은 기존 BWE model을 기반으로 개선하였으며, 이 model들은 **새로운 SWTC 성능 신기록을 달성했다.**

5. Conclusions and Future Work

Three Conclusions

- 많은 cross-lingual / multilingual NLP task에 사용될 수 있는 bilingual word embedding을 학습하는 과정에서 **seed bilingual lexicons의 중요성과 속성에 대한 자세한 분석을 제시**하였다.
- Inexpensive document-level embedding space에서 얻은 신뢰할 수 있는 symmetric 단어 번역 쌍을 사용하여 두 **monolingual embedding spaces 간의 mapping을 학습**하는 방법을 제안하였다.
- 제안된 모델을 사용하여 **Bilingual Lexicon Learning**과 **Suggesting Word Translation in Context** 테스트에서 성능 신기록을 달성하였다.

Future Work

- 데이터가 부족한 상황에서 seed lexicon으로 사용할 단어 쌍을 선정하는 방법에 대한 연구
- 본 연구에서 제안된 방법을 더 많은 언어 및 분야로 확장

ACL 2016
AUGUST 7 – 12 | BERLIN, GERMANY

Thank you for your

Attention

You-Dong Yun, Chanhee Lee

BLP Lab