(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0049403
(43) 공개일자 2022년04월21일

(51) 국제특허분류(Int. Cl.)
G06F 40/58 (2020.01) *G06F 40/51* (2020.01)
G06N 20/00 (2019.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06F 40/58 (2020.01)
G06F 40/51 (2020.01)
(21) 출원번호 10-2020-0132988
(22) 출원일자 2020년10월14일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
서정연
서울특별시 서초구 신반포로16길 15-20, 103동 2702호 (반포동, 반포힐스테이트)
허광호
서울특별시 마포구 백범로 35, 리찌과학관 R908호(신수동) 학교)
(74) 대리인
유미특허법인

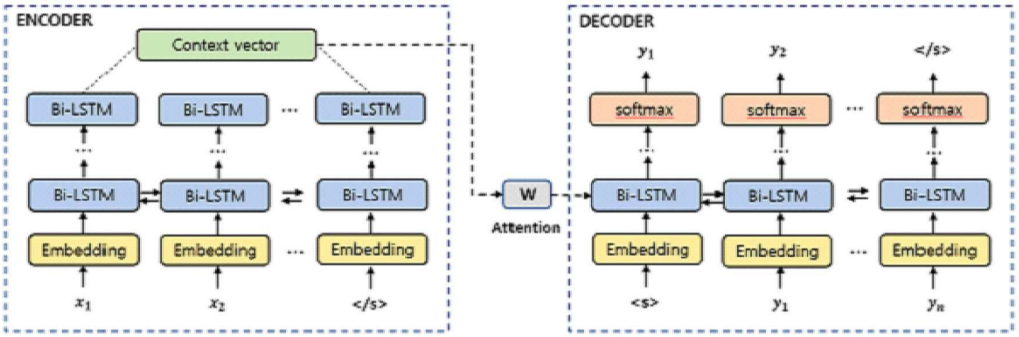
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **가상 병렬 코퍼스의 노이즈 필터링을 이용한 기계 번역 성능 향상 방법**

(57) 요 약

상기의 과제를 해결하기 위한 본 발명은 역-번역을 이용하여 합성한 가상 병렬 코퍼스의 노이즈를 필터링하는 방법으로 기계 번역기의 성능을 향상하는 것에 주력하였고 교차-언어(Cross-lingual) 단어 임베딩을 이용한 병렬 문장에 대한 노이즈 필터링 방법에 대한 것이다.

대 표 도 - 도1



(52) CPC특허분류
G06N 20/00 (2021.08)
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-00255
과제번호	202039008
부처명	과학기술정보통신부
과제관리(전문)기관명	IITP 정보통신기획평가원
연구사업명	정보통신.방송 연구개발
연구과제명	(지능정보-총괄/1세부) 자율지능 디지털 동반자 프레임워크 및 응용 연구개발
기 여 율	1/1
과제수행기관명	전자부품연구원
연구기간	2020.01.01~2020.12.31

명 세 서

청구범위

청구항 1

역-번역을 이용한 신경망 기반의 기계 번역 및 교차-언어 단어 임베딩 학습 방법과 병렬 문장 임베딩의 코사인 유사도를 이용한 병렬 품질 측정 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 가상 병렬 코퍼스의 노이즈 필터링을 이용한 기계 번역 성능 향상 방법에 대한 것이다.

배 경 기 술

[0002] 기존의 역-번역 기반의 기계 번역은 주로 단일 언어 코퍼스를 가상 병렬 코퍼스로 합성 후 노이즈 필터링 과정을 거치지 않고 바로 기계 번역기의 학습 데이터로 사용하였다. 이러한 방식은 다음과 같은 문제점이 있다.

[0003] 역-번역으로 합성한 가상 병렬 코퍼스에는 노이즈가 포함되며 이를 직접 학습 데이터에 포함할 경우 기계 번역기의 성능 저하를 초래한다.

[0004] 기계 번역기 학습을 위한 병렬 데이터가 부족할 경우 역-번역을 수행하는 기계 번역기의 성능이 제한되며 이를 이용하여 합성한 가상 병렬 코퍼스에는 더 많은 노이즈가 포함될 수 있다.

[0005] 또한, 일부 연구에서는 가상 병렬 코퍼스에 포함된 노이즈를 제거하기 위하여 합성된 문장을 다시 한 번 번역한 후 기존 문장과의 번역 정확도 (BLEU 점수)를 이용하여 가상 병렬 코퍼스의 각 문장들의 품질을 측정하는 방식으로 노이즈를 제거하였다.

[0006] 이 방법은 가상 병렬 코퍼스의 노이즈 측정을 위해 X→Y', Y'→X' 두 방향에 대한 기계 번역기가 각각 필요하다. 그리고 노이즈 측정을 위한 기계 번역기의 성능이 낮을 경우 측정은 왜곡될 수 있다.

발명의 내용

해결하려는 과제

[0007] 따라서, 본 발명은 역-번역을 이용하여 합성한 가상 병렬 코퍼스의 노이즈를 필터링하는 방법으로 기계 번역기의 성능을 향상하는 것에 주력하였고 교차-언어(Cross-lingual) 단어 임베딩을 이용한 병렬 문장에 대한 노이즈 필터링 방법을 제안한다.

과제의 해결 수단

[0008] 상기의 과제를 해결하기 위한 본 발명은 역-번역을 이용한 신경망 기반의 기계 번역 및 교차-언어 단어 임베딩 학습 방법과 병렬 문장 임베딩의 코사인 유사도를 이용한 병렬 품질 측정 방법으로 구성된다.

발명의 효과

[0009] 본 발명은 교차-언어 단어 임베딩을 학습하고 같은 임베딩 벡터 공간에서 두 병렬 문장에 대한 문장 임베딩 유사도가 결정되면 일정한 임계값 t를 이용하여 병렬 품질이 낮은 문장들을 가상 병렬 코퍼스에서 제거하고 학습 데이터의 전반적인 품질을 향상한다.

도면의 간단한 설명

[0010] 도 1은 본 발명의 실시예에 따른 양방향 LSTM 신경망 기반의 기계 번역 모델의 구조도이다.

도 2은 본 발명의 실시예에 따른 역-번역을 이용한 신경망 기반의 기계 번역의 예시도이다.

도 3은 본 발명의 실시예에 따라 교차-언어 단어 임베딩을 이용한 노이즈 필터링 방법을 적용한 역-번역 기계

번역기 학습 과정을 나타낸 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0011] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0012] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0013] 도 1은 본 발명의 실시예에 따른 양방향 LSTM 신경망 기반의 기계 번역 모델의 구조도이다.

[0014] 도 1에 대해 설명하기 앞서, 용어를 설명하면 다음과 같다.

[0015] 병렬 코퍼스: 2개 국어 이상의 번역된 문장 혹은 문서의 집합을 병렬 코퍼스(parallel corpus)라고 부른다. 본 연구에서 사용한 병렬 코퍼스는 여러 개의 문장 쌍(sentence pair)으로 구성되어 있으며 각 문장 쌍은 같은 의미를 가지고 2개 국어로 구성되어 있다. 한국어-영어 병렬 코퍼스를 수식으로 표현하면 다음 수학식 1과 같다.

수학식 1

[0016] $$D_{ko-en}=\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$$

[0017] 여기서 한국어-영어(ko, en은 한국어, 영어의 약자를 나타낸다.) 병렬 코퍼스 D_{ko-en}는 N개의 문장 쌍 (X, Y)로 구성되고 x는 소스 문장, y는 타깃 문장이라고 하고 두 문장은 각각 한국어와 영어로 구성된 같은 의미를 나타내는 문장이다. 병렬 코퍼스 내의 각 문장 쌍은 독립적으로 존재한다. 즉 문장 쌍 사이의 의미적 연관성이 없다고 가정한다.

[0018] 예시: X_1: 저는 학생입니다.

[0019] Y_1: I am a student.

[0020] X_2: 나는 사과를 먹었다.

[0021] Y_2: I ate an apple.

[0022] 신경망 기반의 기계 번역 모델 구조

[0023] 본 발명에서 사용하는 신경망 기반의 기계 번역기는 양방향 장단기 기억장치(Bi-LSTM: Bi-directional long short term memory)기반의 인코더-디코더 (Encoder-Decoder) 모델을 사용한다. 그림에서 제시한 바와 같이 인코더(좌측)는 입력된 소스 문장을 의미 벡터(Context vector)로 변환하고 디코더(우측)는 의미 벡터를 이용하여 번역된 타깃 문장을 생성한다.

[0024] 도 1에 도시된 바와 같이, 인코더는 1개 입력 층, 1개 임베딩 층과 다중 계층 Bi-LSTM으로 구성되어 있다. 인코더의 연산 과정을 수식으로 표현하면 다음과 같다. N개의 문장 쌍으로 구성된 학습 병렬 코퍼스를 $D=\{(X_i, Y_i)\}_{i=1}^{N}$ 라고 하고 소스 문장 $X=x_1, \cdots, x_m$ 와 타깃 문장 $Y=y_1, \cdots, y_n$ 은 각각 m, n개의 하위 단어 시퀀스로 구성되었다고 한다.

[0025] 인코더의 입력 층은 문자열 형태의 하위 단어 시퀀스를 개의 One-hot 벡터 $x_i \in R^{|V|}$ 로 변환하는데 여기서 $|V|$ 는 하위 단어 사전 V의 크기를 나타낸다.

[0026] 임베딩 층은 단어 사이의 유사도 계산을 위하여 One-hot 벡터를 저차원의 임베딩 벡터 $x_i \in R^{d_e}$로 변환하는데 여기서 d_e는 임베딩 벡터 차원을 나타낸다.

[0027] 다중 계층 Bi-LSTM의 각 계층은 정 방향(Forward) LSTM과 역 방향(Backward) LSTM으로 구성된다. LSTM은 순환 신경망(RNN: Recurrent Neural Network) 구조를 가지며 Sepp Hochreiter와 JÜrgen Schmidhuber이 처음 제안하였다. LSTM은 다음 수학식 2에 의하여 t번째 입력 벡터 x_t를 이용하여 이전 단계 은닉 벡터 h_{t-1}를 h_t로 업데이트 한다.

수학식 2

[0028]
$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

[0029] 여기서 f, i, o는 각각 Forget, Input, Output 게이트의 연산 결과이고, W, U는 LSTM의 가중치 메트릭스(Weight Matrix)를 나타낸다. 제안 시스템에서 LSTM은 개의 하위 단어 시퀀스에 해당하는 임베딩 벡터 시퀀스 $x_1, \ldots x_m$를 입력으로 받아 은닉 벡터 시퀀스 $h_1, \ldots h_m$을 출력한다.

[0030] 다중 계층 LSTM은 하위 계층의 LSTM 연산 결과를 상위 계층 LSTM의 입력 값으로 사용하여 새로운 은닉 벡터 시퀀스를 연산하는 역할을 수행하는데 다양한 문장 내의 복잡한 의미들을 파악하는데 유리하다. 인코더의 Context vector는 다음 수학식 3과 같이 Bi-LSTM 은닉 벡터 시퀀스로 표현할 수 있다.

수학식 3

[0031]
$$S = h_1, \ldots h_m$$

[0032] 디코더는 1개의 입력 층, 1개의 임베딩 층, 다중 계층 LSTM과 출력 층으로 구성되어 있다. 디코더의 입력 층, 임베딩 층은 타깃 문장의 하위 단어 시퀀스를 임베딩 벡터 시퀀스로 변환하며 처리 과정은 인코더와 동일하다.

[0033] 디코더의 다중 계층 Bi-LSTM은 인코더와 기본적으로 동일하지만 주의 집중(Attention) 방법을 활용한다는 점에서 차이를 보인다. 제안 시스템은 Dzmitry Bahdanau이 제안한 주의 집중 방법을 차용하였다.

[0034] 디코더의 출력 층은 softmax 함수를 이용하여 LSTM의 은닉 벡터 h_t를 사전에 등록된 모든 하위 단어들에 대한 생성 확률로 변환한다. 디코더의 단어 생성 확률은 다음 수학식 4와 같다.

수학식 4

[0035]
$$p(y_t | y_1, \ldots y_{t-1}, S) = softmax(W_o \cdot \ h_t)$$

[0036] 여기서 S는 인코더의 Context vector이고 Wo는 디코더 출력 층의 가중치 메트릭스, h_t는 디코더의 Bi-LSTM 은닉 벡터를 나타낸다.

[0037] 기계 번역 학습을 위한 병렬 코퍼스를 D라고 할 때 학습 목표 함수는 단어 생성 확률의 크로스 엔트로피(cross entropy) 손실 L로 정의하고 모델 파라미터의 학습은 이 함수를 최소화 하는 방향으로 진행된다.

수학식 5

[0038] $$L = \sum_{(X,\ Y) \in D} \sum_{t=1}^{N} \log p(y_t | y_1, \cdots y_{t-1}, S)$$

[0039] 학습 데이터는 mini-batch 방법으로 샘플링하고 최적화 알고리즘은 확률적 경사하강법(SGD: Stochastic Gradient Descent)을 사용하였고 기울기 학습률(Learning Rate)은 1.0으로 설정하고 학습이 진행됨에 따라 점진적으로 작게 조정해주는 방법을 이용하였다. 모델 및 학습과 관련된 하이퍼 파라미터 값은 표 1에 기재하였다.

표 1

[0040]

하이퍼 파라미터	Hyperparameters	값
사전 크기	Size of Vocabulary	32,000
임베딩 벡터 차원	Dimension of Embedding vector	1,024
LSTM 은닉 벡터 차원	Dimension of Hidden vector	1,024
Bi-LSTM 계층 수	Layers of Bi-LSTM	4
최적화 알고리즘	Optimization Algorithm	SGD
기울기 학습률	Learning Rate	1.0

[0041] 도 2은 본 발명의 실시예에 따른 역-번역을 이용한 신경망 기반의 기계 번역의 예시도이다.

[0042] 도 2에 도시된 바와 같이, 역-번역은 기계 번역의 모델 구조를 변경하지 않고 타깃 언어의 단일 언어(Monolingual) 데이터를 기계 번역기의 학습에 사용하는 방법이다. 병렬 코퍼스 $D_p = \{(X_p^n, Y_p^n)\}_{n=1-N}$ 와 타깃 단일 언어 코퍼스 $D_{tm} = \{Y_{tm}^m\}_{m=1-M}$ 일 때 역-번역 과정은 다음과 같다.

[0043] 첫째, 주어진 병렬 코퍼스 D_p 로 역방향 기계 번역기 NMT_{Y-X} 를 학습한다.

[0044] 둘째, 역방향 기계 번역기 NMT_{Y-X} 를 이용하여 타깃 단일 언어 코퍼스 D_{tm} 를 소스 단일 언어 코퍼스 $D_{st} = \{X_{st}^m\}_{m=1-M}$ 로 번역한다.

[0045] 셋째, 타깃 단일 언어 코퍼스 D_{tm} 와 소스 단일 언어 코퍼스 D_{st} 를 병합하여 가상 병렬 코퍼스 $D_{sun} = \{(X_{st}^m, Y_{st}^m)\}_{m=1-M}$ 를 생성한다.

[0046] 넷째, 기존 병렬 코퍼스 D_p 와 가상 병렬 코퍼스 D_{sun} 을 이용하여 정방향 기계 번역기 NMT_{X-Y} 를 학습한다.

[0047] 도 3은 본 발명의 실시예에 따라 교차-언어 단어 임베딩을 이용한 노이즈 필터링 방법을 적용한 역-번역 기계 번역기 학습 과정을 나타낸 예시도이다.

[0048] 도 3에 도시된 바와 같이, 제안하는 교차-언어 단어 임베딩을 이용한 가상 병렬 데이터 필터링 방법은 다음과 같다.

[0049] 입력 문장 x에 대한 문장 임베딩 벡터 s_x 는 다음 수학식 6과 같이 문장에 포함된 각 단어들의 단어 임베딩 $w_1 \sim w_{|x|}$ 의 평균값으로 표현한다.

수학식 6

[0050] $$s_x = \frac{1}{|x|}\sum_{i=1}^{|x|} w_i$$

[0051] 마찬가지로 가상 병렬 코퍼스의 두 병렬 문장 (x, y)에 대한 문장 임베딩 s_x와 s_y의 코사인 유사도는 다음 수학식 7과 같이 계산한다.

수학식 7

[0052] $$similarity(s_x, s_y) = \frac{s_x \cdot s_y}{|s_x||s_y|}$$

[0053] 두 병렬 문장은 서로 다른 언어이기 때문에 두 문장 임베딩 사이의 코사인 유사도가 유의미한 값을 가지기 위해서는 교차-언어 단어 임베딩을 이용하여 두 문장 임베딩이 같은 벡터 공간에 위치해야 한다. 교차-언어 단어 임베딩은 다음과 같은 방법으로 얻을 수 있다. 먼저 병렬 코퍼스의 두 언어에 대한 단어 임베딩 X, Y을 학습한다. 다음 소량의 이중-언어 사전을 이용하여 선형 변환 행렬 W를 학습하는데 목표 함수는 다음 수학식 8과 같다.

수학식 8

[0054] $$\underset{W}{argmin}\sum_{i=1}^{n} \| X_i W - Y_i \|^2$$

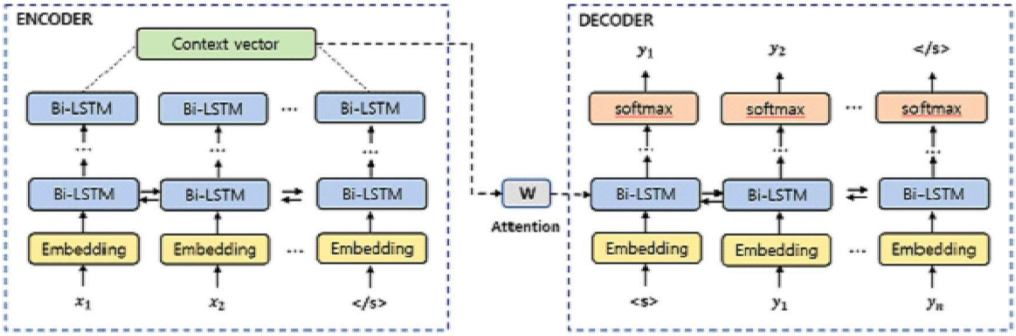
[0055] 위에서 제시한 방법으로 교차-언어 단어 임베딩을 학습하고 같은 임베딩 벡터 공간에서 두 병렬 문장에 대한 문장 임베딩 유사도가 결정되면 일정한 임계값 t를 이용하여 병렬 품질이 낮은 문장들을 가상 병렬 코퍼스에서 제거하고 학습 데이터의 전반적인 품질을 향상한다.

[0056] 이처럼 상기와 같이 본 발명의 실시한 예에 대하여 상세히 설명하였으나, 본 발명의 권리범위는 이에 한정되지 않으며, 본 발명의 실시한 예와 실질적으로 균등의 범위에 있는 것까지 본 발명의 권리범위가 포함되는 것은 당연하다.

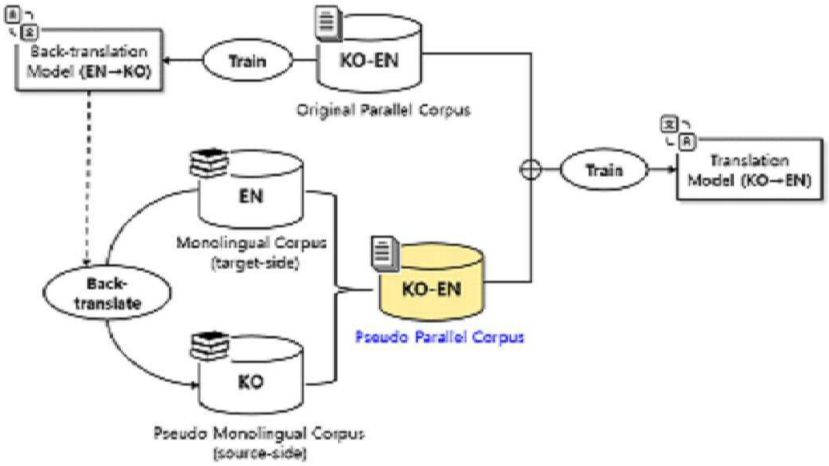
[0057] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

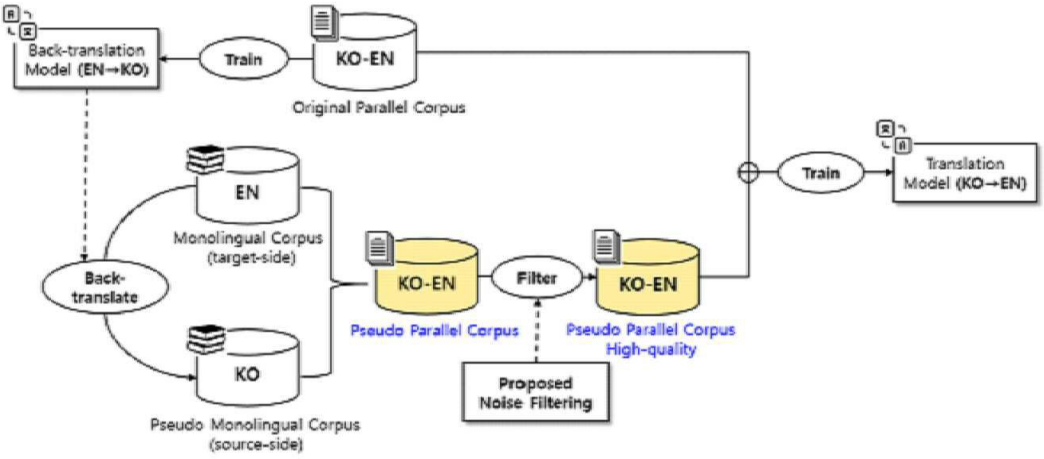
도면1



도면2


Back-translation
Model (EN→KO)
Train
KO-EN
Original Parallel Corpus
Train
Translation
Model (KO→EN)
EN
Monolingual Corpus
(target-side)
Back-
translate
KO-EN
Pseudo Parallel Corpus
KO
Pseudo Monolingual Corpus
(source-side)

도면3


Back-translation
Model (EN→KO)
Train
KO-EN
Original Parallel Corpus
Train
Translation
Model (KO→EN)
EN
Monolingual Corpus
(target-side)
Back-
translate
KO-EN
Filter
KO-EN
Pseudo Parallel Corpus
Pseudo Parallel Corpus
High-quality
KO
Proposed
Noise Filtering
Pseudo Monolingual Corpus
(source-side)