도메인 적응 기술 기반 질문 문장에 대한 의미역 인식 연구

임수종○, 김현기
한국전자통신연구원, 자동통역인공지능연구센터
{isj, hkk}@etri.re.kr

A Study of Semantic Role Labeling using Domain Adaptation Technique for Question

Soojong Lim○, Hyunki Kim
Automatic Speech Translation and AI research Center, ETRI

요 약

기계학습 방법에 기반한 자연어 분석은 학습 데이터가 필요하다. 학습 데이터가 구축된 소스 도메인이 아닌 다른 도메인에 적용할 경우 한국어 의미역 인식 기술은 10% 정도 성능 하락이 발생한다. 본 논문은 기존 도메인 적응 기술을 이용하여 도메인이 다르고, 문장의 형태도 다를 경우에 도메인 적응 알고리즘을 적용하여, 질의응답 시스템에서 필요한 질문 문장 의미역 인식을 위해, 소규모의 질문 문장에 대한 학습 데이터 구축만으로도 한국어 질문 문장에 대해 성능을 향상시키기 위한 방법을 제안한다. 한국어 의미역 인식 기술에 prior 모델을 제안한다. 제안하는 방법은 실험결과 소스 도메인 데이터만 사용한 실험보다 9.42, 소스와 타겟 도메인 데이터를 단순 합하여 학습한 경우보다 2.64의 성능향상을 보였다.

주제어: 질문 문장 의미역 인식, 도메인 적응 기술, prior 모델

1. 서론

의미역 인식(Semantic Role Labeling)이란 자연어 문장에서 'Who does what to whom'을 인식하는 기술로, 문장의 서술어를 중심으로 서술어에 대한 의미적인 역할(예를 들어, 행위자, 경험자, 대상격, 도구격 등) 을 하는 문장의 부분을 인식하는 것을 말한다. 지식을 처리하는 응용 서비스가 발달함에 따라서 형태소 분석, 개체명 인식 같은 어절 단위 자연어 분석 기술 이외에도 의미역 인식 같은 문장 단위 의미 분석 기술에 대한 수요도 점점 늘어나고 있는 추세이다.

영어권에서는 CoNLL-2004를 시작으로 의미역 인식에 관한 연구가 활발히 진행되고 있는데, Out of Domain을 다루기 시작한 CoNLL-2005에서는 구구조 기반 학습 데이터가 구축된 소스 도메인(WSJ corpus)이 아닌 다른 타겟 도메인(Brown corpus)에 적용할 경우 10% 이상의 성능 하락 현상이 일어났고[1], 의존 구분 분석 결과 기반 의미역 인식을 수행한 CoNLL-2008 Shared Task에서도 마찬가지로 성능 하락 현상이 발생하였으며[2], 한국어를 대상으로 한 연구에서도 뉴스 도메인과 위키피디아 도메인에서 약 15% 정도의 성능 하락 현상이 발생하였다[3].

이러한 성능 하락 현상을 극복하는 방법은 타겟 도메인에 대해서도 소스 도메인만큼의 학습 데이터를 구축하여 타겟 도메인에서도 같은 시스템을 새롭게 구축하는 것이지만, 이는 시간과 비용적인 측면에서 장애가 되는 요소이다. 도메인 적응(Domain Adaptation) 기술은 이러한 문제를 적은 양의 타겟 도메인의 학습 데이터 구축으로도 소스 도메인에 비해 급격한 성능 하락을 방지하기 위해서 제안되었고, 많은 연구가 진행되어, 도메인 이식성을 높이는데 기여하고 있다[4,5,6,7,8].

심층학습을 적용할 경우, 단순한 자질 사용으로 인해 형태소 분석, 구문 분석과 같은 이전 단계 기술에 대한 의존도를 줄여 오류의 여지를 줄이기 때문에 기존 방법에 비해 도메인 적응성이 올라간다는 연구도 있으나, 이는 기존 방법과 비교한 것으로 도메인 적응 알고리즘을 적용할 경우와 비교하지는 않았다[9].

일반적으로 의미역 인식을 위한 학습 데이터는 정답 후보가 되는 평서문을 중심으로 구축되어 있다. 질의응답 시스템에서는 정확한 정답을 찾기 위해서 사용자의 의도가 담겨있는 질문 문장을 정확히 분석하는 것이 중요하지만, 학습 데이터로 사용할 수 있는 질문 문장이 부족한 것이 현실이다.

본 논문에서는 기존 한국어 의미역 인식 시스템과 도메인 적응 기술을 활용하여, 한국어 질문 문장에 대한 의미역 인식 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 개선하고자 하는 기존 시스템 및 본 논문에서 제안하는 도메인 적응 기술에 대해 설명하며, 4장에서는 실험 및 결과를 분석하고, 마지막으로 5장에서 결론을 기술한다.

2. 관련 연구

도메인 적응 기술을 적용하는 과정은 그림 1[8]과 같다. 기본적으로 도메인 적응 기술을 위해서는 기계학습 기반의 소스 도메인 시스템과 타겟 도메인에 대한 학습 데이터가 소량이나마 필요하다. 소스 도메인의 데이터와 알고리즘(Algorithm 1)을 이용하여 구축된 소스 모델($\mathbf{w}_{src}$)를 입력받아, 타겟 도메인에서 구축된 학습 데이터와 도메인 적응 알고리즘(Algorithm 2)를 이용하여 최종적으로 타겟 도메인에 최적화된 타겟 모델($\mathbf{w}_{tgt}$)을 구축하는 도메인 적응 기술을 적용하는 과정을 보여준다.

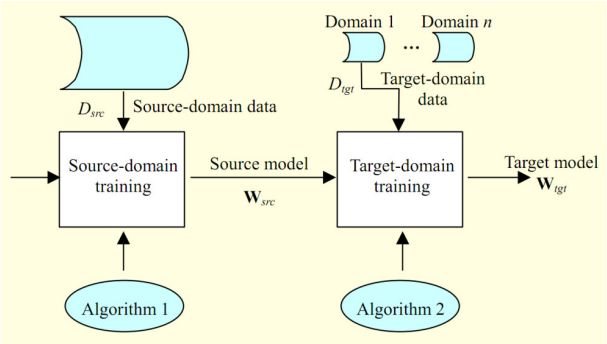

그림 1. 도메인 적응 시스템 구축 과정

이러한 도메인 적응 기술은 다양한 분야의 기술에 적용하기 위해서 제안되어 왔으며, Daume and Marcu[5]는 다음과 같이 도메인 적응 기술을 분류하였다.

- source only(SRC-only): 소스 도메인의 학습 데이터만을 사용하는 것으로 도메인 적응 기술의 베이스라인으로 간주
- target only(TGT-only): 타겟 도메인의 학습 데이터만을 사용
- All and weighted model: 소스, 타겟 도메인의 학습 데이터를 모두 사용하지만 데이터의 비율이 다름을 감안하여 가중치를 적용
- PRED: SRC-only 방법으로 구축된 기술을 이용하여 타겟 도메인의 학습데이터를 분석하고 그 결과를 타겟 도메인의 모델을 구축할 때 자질로 사용
- Linearly interpolation: SRC-only, TGT-only 방법으로 각각 모델을 구축하고 이를 선형보간법(linearly interpolation)을 적용하여 하나의 모델로 통합하는 방법으로 다음의 수식을 이용

  LININT Model
  = λ* Source Model + (1-λ) * Target Model

- Feature Augmentation(FA): 공통적으로 사용 가능한 자질, 소스 도메인에 특화된 자질, 타겟 도메인에 특화된 자질과 같이 3가지로 분류하여 각각의 자질을 이용하여 모델을 독립적으로 구축.
- prior 모델: SRC-only 모델의 가중치 벡터(weight vector)를 타겟 도메인 기술을 구축시 이용. 이를 분류 문제로 변환한 개념은 그림 2와 같다. 타겟 가중치 벡터를 찾기 위해 소스 가중치 벡터를 시작점으로 참조하여 타겟 도메인 학습데이터로 학습.

본 논문은 뉴스 도메인에서 구축된 한국어 의미역 인식 시스템을 도메인도 다르고 문형도 다른 질문 문장에 적용할 때 발생하는 성능 하락을 최소화하기 위해 소규모로 구축된 위키피디아에 대한 질문 학습 데이터를 도메인 적응 기술인 prior 모델을 적용하여 백과사전 도메인의 질의응답 시스템에서 질문 문장에 특화된 한국의 의미역 인식 시스템을 제안한다.

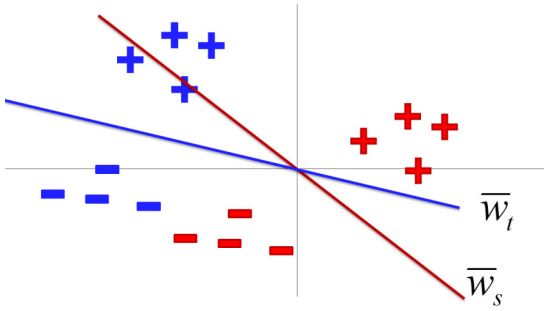

그림 2. prior 모델 적용 예

3. 질문 문장 의미역 인식

3.1 질문 문장

본 논문에서 대상으로 하고 있는 질문 문장의 형태는 다음과 같다.

- 최고점의 해발고도가 약 39m인 평평한 섬으로 우리나라 최남단에 있는 이 섬은 어디일까?
- 고려(高麗) 제25대 왕이면서 원나라 세조(世祖) 쿠빌라이의 사위는?
- 이것은 어떤 일이 시작될 때 있었던 아주 작은 차이가 전체에 막대한 영향을 미칠 수 있음을 설명하는 데 '카오스 이론'의 토대가 되기도 한 이것은?

평서문과는 다르게 용언으로 문장이 종료되지 않고, '이것', '저것'과 같은 지시 대명사나 질문의 초점이 되는 명사(위의 문장에서는 '사위')로 종료된다. 즉, 겹문장과 관형절로 안은 문장 형태가 주류를 이루게 된다.

3.2 기존 의미역 시스템

본 논문에서 채택한 기존 베이스라인 시스템은 Structural SVM을 이용한 순차적 레이블링 방법[10]이며, prior 모델을 적용하려는 기존 알고리즘은 그림3과 같다.

1: **Input**: S, λ, T, k

2: **Initialize**: $\mathbf{w}_1 = \mathbf{0}$

3: **For** $t = 1,2,\ldots,T$ **do**

4: Choose $A_t \subseteq S$, where $|A_t| = k$

5: Set $A_t^+ = \{(\mathbf{x}, \mathbf{pr}, \mathbf{y}) \in A_t : l(\mathbf{w}_t; (\mathbf{x}, \mathbf{pr}, \mathbf{y})) > 0\}$

6: $\forall (\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+$:

$\mathbf{y}_{ij}^* = \text{argmax}\{L(\mathbf{y}_{ij}, \mathbf{y}) - \mathbf{w}_t^T \Psi(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y})\}$

7: $\eta_t = 1/\lambda t$

8: $\mathbf{w}_{t+1} = (1 - \eta_t \lambda)\mathbf{w}_t + \frac{\eta_t}{k} \sum_{(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+} \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}^*)$

9: **Output**: $\mathbf{w}_{T+1}$

그림 3. 순차적 레이블링 의미역 인식 알고리즘[10]

3.3 도메인 적응을 위한 prior 모델

본 논문에서는 자질 값을 단순화하는 방법과 함께 알고리즘을 이용하여 타겟 도메인에서 한국어 의미역 기술을 개발하도록 시도하였다. 여러 가지 도메인 적응 알고리즘 중에서 Chelba and Acero[6]이 제안한 prior 모델을 변형하여 structural SVM에 적용한 연구[8]를 참조하여 한국어 질문 문장에 대한 의미역 기술에 적용하였다.

prior 모델은 기계학습의 목적인 최적화된 가중치 벡터 결정을 위해 상대적으로 소규모인 타겟 도메인 학습 데이터만을 이용하기보다, 소스 도메인에서 학습된 가중치 벡터를 참조하여 효과적으로 타겟 도메인의 최적 가중치 벡터를 탐색하는 개념이다.

소스 도메인 한국어 의미역 인식 기술에 prior 모델을 적용하기 위해서 목적 함수를 아래와 같이 수정한다.

$$f(\mathbf{w}; A_t) = \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_{src}\|^2 + \frac{1}{k} \sum_{(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t} l(\mathbf{w}; (\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}))$$

$$\textit{where} \quad l(\mathbf{w}; (\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij})) = \max\{0, \max_{\mathbf{y}} \{L(\mathbf{y}_{ij}, \mathbf{y}) - \mathbf{w}^T \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y})\}\}$$

$$\textit{and} \quad \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}) = \Psi(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) - \Psi(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}).$$

위 식에서 $\mathbf{w}_{src}$는 소스 도메인에서 학습된 가중치 벡터이고, $\mathbf{x}_i$는 학습 데이터의 i번째 문장 벡터, $\mathbf{pr}_{ij}$는 i번째 문장의 j번째 서술어(predicate), $\mathbf{y}_{ij}$는 i번째 문장의 j번째 서술어에 대한 의미역 인식 결과 벡터, A_t는 학습 데이터에서 임의로 선택된 부분집합이다.

타겟 도메인의 최적 가중치 벡터 $\mathbf{w}$를 결정하기 위해서 소스 가중치 벡터 $\mathbf{w}_{src}$를 선행 정보로 이용하여 타겟 도메인의 목적 함수를 위와 같이 정의한다. 목적 함수가 최소값이 되는 최적 가중치 벡터를 구하기 위해서 subgradient 함수는 다음과 같이 정의하였다.

$$\nabla f(\mathbf{w}; A_t) = \lambda(\mathbf{w} - \mathbf{w}_{src}) - \frac{1}{|A_t|} \sum_{(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+} \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}^*)$$

$$\textit{where} \quad \mathbf{y}_{ij}^* = \arg\max_{\mathbf{y}} \{L(\mathbf{y}_{ij}, \mathbf{y}) - \mathbf{w}^T \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y})\}$$

$$\textit{and} \quad A_t^+ = \{(\mathbf{x}, \mathbf{pr}, \mathbf{y}) \in A_t : l(\mathbf{w}; (\mathbf{x}, \mathbf{pr}, \mathbf{y})) > 0\}$$

수정된 두가지 수식을 이용하여 그림 3의 알고리즘에 prior 모델을 적용하면 그림 4와 같은 알고리즘이 된다. 수정된 알고리즘에서 D_{tgt}는 타겟 도메인의 학습데이터 집합, λ는 정규화 상수, T는 반복횟수, k는 sub-gradient를 계산하기 위해 한 번에 사용하는 학습데이터 수, w_{src}는 소스 도메인 최적 가중치 벡터이다.

Inputs: $D_{tgt}, \lambda, T, k, \mathbf{w}_{src}$

1: $\mathbf{w}_1 = \mathbf{0}$ // Initialization.

2: For $t = 1,2,\ldots,T$ do

3: Choose $A_t \subseteq D_{tgt}$, where $|A_t| = k$

4: Set $A_t^+ = \{(\mathbf{x}, \mathbf{pr}, \mathbf{y}) \in A_t : l(\mathbf{w}_t; (\mathbf{x}, \mathbf{pr}, \mathbf{y})) > 0\}$

5: $\forall (\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+$:

$\mathbf{y}_{ij}^* = \text{argmax}\{L(\mathbf{y}_{ij}, \mathbf{y}) - \mathbf{w}_t^T \Psi(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y})\}$

6: $\eta_t = 1/\lambda t$

7: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \lambda (\mathbf{w}_t - \mathbf{w}_{src}) + \frac{\eta_t}{k} \sum_{(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+} \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}^*)$

8: Return $\mathbf{w}_{T+1}$ as output

그림 4. prior 모델을 적용한 의미역 인식 알고리즘

4. 실험 및 결과

본 논문에서는 Korean Propbank(KPB)[11]에서 군대 도메인인 Virginia 말뭉치를 제외한 뉴스 도메인 말뭉치 4,882 문장을 소스 도메인의 학습 데이터로 사용하였다. 한국전자통신연구원(ETRI)에서 구축한 WiseQA[12] 2차 표준 평가셋의 질문 852문장은 타겟 도메인의 학습 데이터로 사용하였다. 평가셋으로는 질의응답 시스템 평가를 위한 GS(Gold Standard) 3.0에 의미역 태깅을 한 엑소브레인 언어분석 말뭉치를 사용하였는데, 기존 GS 2.0의 50셋을 제외하고 GS 3.0 중 새로 추가한 66셋 중에서 질문 107문장을 평가용으로 사용하였으며, KPB에서 정의한 24개 의미역을 사용하였다.

기계학습을 위한 형태소 분석, 어휘의미 분석, 개체명 인식, 구문 분석 정보는 ETRI 언어분석기[13]를 사용하여 자동으로 분석된 결과를 이용하였다. 따라서 자동으로 추출된 학습 자질에는 오류가 포함되어 있을 수 있다. 한국어 논항 인식/분류(Argument identification and classification)에 대해서만 성능을 측정하였고, 정확율(Precision), 재현율(Recall)에 기반하여 계산하는 F1-score를 성능 척도로 사용하였다.

소스 도메인 시스템을 베이스라인으로 하여, 4가지 종

류의 실험을 진행하였고, 실험결과는 표1과 같다. 첫 번째, 소스 도메인인 뉴스 데이터를 그대로 사용하여 학습 모델을 구축하고 이를 질문 문장으로 평가를 하였다(SRC-only). 소스 도메인 데이터로 학습한 경우 실험 결과 중 가장 나쁜 성능을 보였으며, 도메인이 다를 경우 일반적으로 보여지는 성능 하락 현상과 동일하였다.

두 번째로 타겟 도메인의 질문 문장 만을 사용하여 학습하였다(TGT-only). 학습 데이터가 소스 도메인에 비해 부족함에도 불구하고 베이스라인 실험 결과보다 6.72이 향상되었다. 학습데이터가 소스 도메인에 비해 17.5% 정도에 불과한 점을 고려하면, 같은 도메인의 유사한 문장 형태로 학습할 경우에는 학습 데이터 양이 적더라도 보다 성능이 높다는 것을 나타낸다.

세 번째로, 사용 가능한 모든 데이터를 이용하여 학습한 경우(ALL)에는 작게나마 성능이 향상되었다. 늘어난 학습 데이터에 비해 성능 개선 폭이 작은 것은, 무조건 데이터를 확장하는 것이 성능 개선에 도움이 되지 않을 수도 있다는 것을 보여준다.

마지막으로 본 논문에서 제안하는 방법의 경우 소스 도메인에서 학습된 결과를, prior 모델을 적용여 9.42의 성능향상을 보였고, 모든 데이터를 단순히 합하여 사용한 결과보다 2.64 앞선다. 이러한 결과는 도메인 적응 알고리즘의 경우 도메인이 다르면서, 문장의 형태가 다른 경우에도 성능 개선에 도움이 된다는 사실을 입증한다.

표 1 실험 결과

실험방법	정확율	재현율	F1-Score
SRC-only	76.21	63.56	69.32
TGT-only	78.33	73.89	76.04
All	83.01	70.24	76.10
제안 방법	82.02	75.71	78.74

5. 결론

본 논문에서는 질의응답 시스템에서는 사용자 의도를 파악하기 위해서는 정교한 분석을 하는 것이 중요하지만, 일반적으로는 소량만 구축되는 질문 문장에 대한 한국어 의미역 인식 기술을 개선하기 위해서 도메인 적응 기술을 적용하는 방법에 대해서 제안하였다. 제안한 방법은 도메인과 문장 형태가 다른 소스 도메인의 데이터를 사용하여 구축된 학습 모델에, 본 논문에서 제안한 방법을 적용할 경우 높은 성능 개선 효과를 보였다. 이러한 실험 결과를 통해 도메인 적응 기술이 새로운 도메인에 대한 이식성 뿐 아니라 문장의 형태가 다른 경우에도 효과가 있음을 입증하였다.

향후 연구로는 상호참조해결, 무형대용어 복원 같은 상대적으로 학습 데이터가 소량 구축되어 있는 기술에도 도메인 적응 기술을 적용하는 것을 방향으로 하고 있다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] X. Carreras and L.Marquez, " Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling," Proc. CoNLL-2005, pp.152-154, 2005

[2] M. Surdeanu et al., "The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies,"Proc. CoNLL-2008, pp.159-177, 2008

[3] 임수종, 배용진, 김현기, 나동렬, "도메인 적응 기술을 이용한 한국어 의미역 인식," 정보과학회 논문지, 제42권, No4, pp.475-482, 2015.

[4] J.Blitzzer and Hal Daume, "Domain Adaptation,", ICML tutorial, 2010.

[5] H. Daume and D. Marcu, "Domain Adaptation for Statistical Classifiers,", J. Artif. Intell. Res., vol.26, no.1, pp.101-126, 2006.

[6] C. Chelba and A. Acero, "Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot,", Comput. Speech Language, vol.20, no.4, pp.382-399, 2006.

[7] C.Lee and M.Jang, "A Prior Model of Structural SVMs for Domain Adaptation,", ETRI Journal, vol.33, no.5, pp.712-719, 2011.

[8] S. Lim et al., "Domain-Adaptation Technique for Semantic Role Labeling with Structural Learning," ETRI Journal, vol. 36, no. 3, June, pp. 429-438, 2014.

[9] Thien Huu Nguyen and Ralph Grishman, "Event Detection and Domain Adaptation with Convolutional Neural Networks", In Proc. 53rd ACL, pp.365-371, 2015.

[10] Wang, Zhen, et al. "Chinese Semantic Role Labeling with Bidirectional Recurrent Neural Networks.", In Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP), pp. 1626-1631, 2015.

[11] S.Lim, C.Lee and D.Ra, "Dependency-based Semantic Role Labeling Using Sequence Labeling with a Structural SVM," in Pattern Recognition Letters, vol.34, pp.696-702, 2013

[12] M.Palmer et al., "Korean Proposition Bank," Linguistic Data Consortium, Philadelphia, 2006.

[13] http://exobrain.re.kr/onedintro

[14] 임준호, 윤여찬, 배용진, 김현기, 이규철, "지배소 후위 제약을 적용한 트렌지션 시스템 기반 한국어 의존 파싱 모델," 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.