엑소브레인 자연어 분석 및 심층질의응답 기술

2020. 6. 18.

김 현 기

언어지능연구실 / 한국전자통신연구원

발표 순서

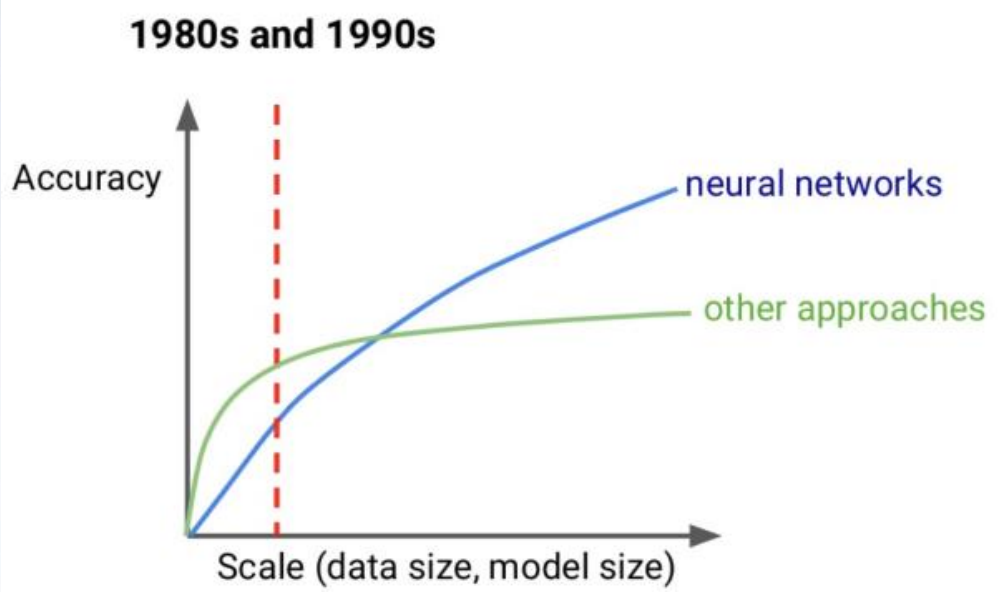
I 자연어 처리 개요

II 딥러닝 언어 모델

III 엑소브레인 한국어 분석/질의응답 기술

딥러닝의 부상

- 대표적인 성공 분야
  - 이미지 인식
  - 음성 인식
  - 기계 번역
  - 게임 플레잉

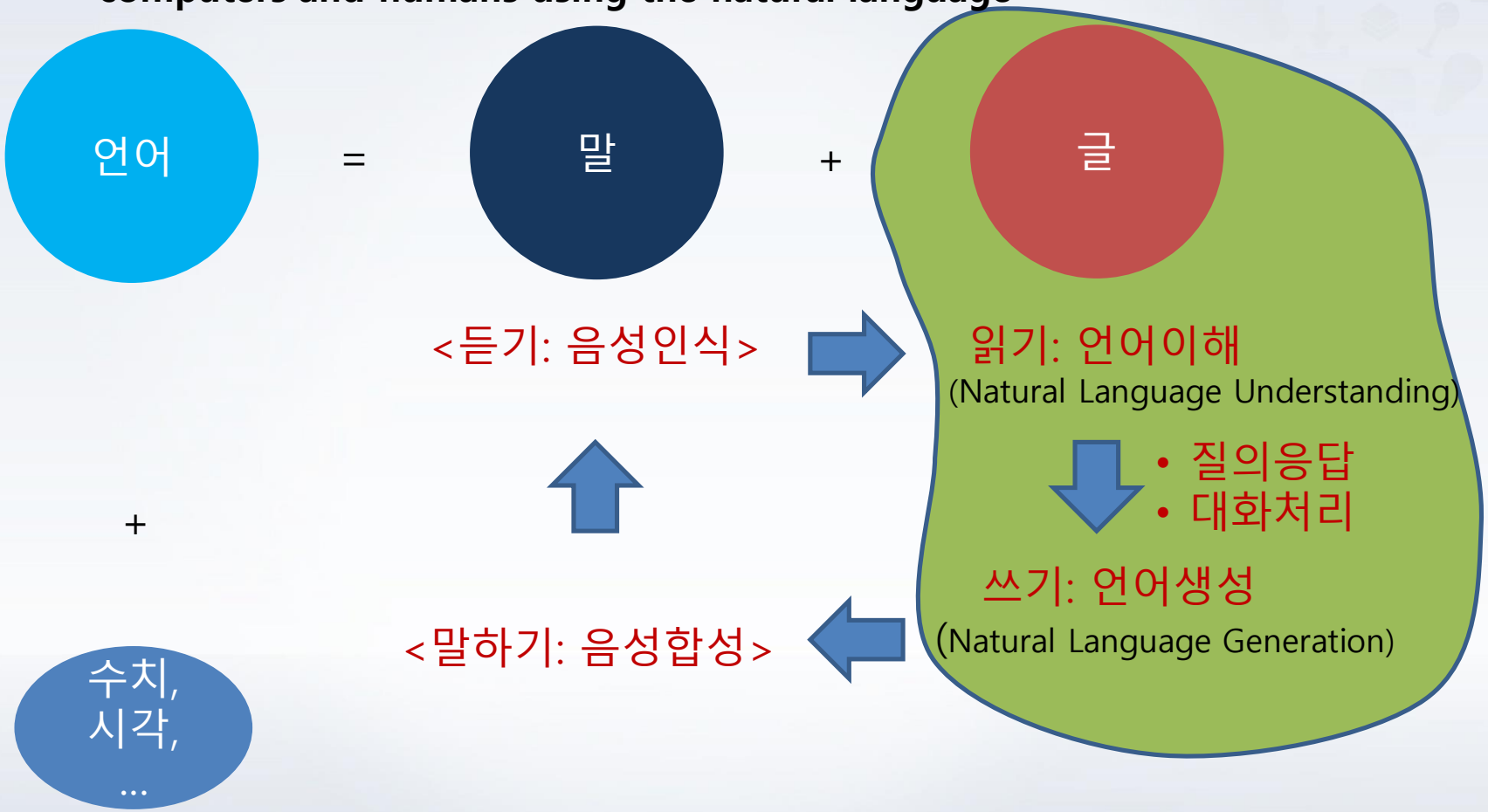


- 타 분야보다 늦게, 자연어 처리는 최근 5년 동안 발전을 이룸
  - BERT(2018, Google) 등장 이후에 비약적인 발전을 이룸

Copyright© 2020 by ETRI

자연어 처리(Natural Language Processing) 기술

- **자연어**: 자연 언어, 일반 사회에서 자연히 발생하여 쓰이는 언어
  - 말과 글은 구어와 문어로 구분, 구어는 주로 대화가 주임
- **자연어 처리 (NLP)** = NLU + NLG
  - a branch of artificial intelligence that deals with the **interaction between computers and humans using the natural language**



Copyright© 2020 by ETRI

자연어 처리가 어려운 이유(1/2)

- 언어의 다양성, 불확실성
  - 전 세계에 7,111개의 언어가 사용됨
  - 영어를 모국어로 사용하는 비율은 10% 이하

<표> 5천만 명 이상이 제1모국어로 사용하는 세계 언어 순위(2019.2.21.)

순위	언어	주요 국가	사용 국가 수	사용자 수(백만 명)
1	중국어(Chinese)	중국	39	1,311
2	스페인어(Spanish)	스페인	31	460
3	영어(English)	영국	137	379
4	힌디어(Hindi)	인도	4	341
5	아랍어(Arabic)	사우디아라비아	59	319
6	벵골어(Bengali)	방글라데시	4	228
7	포르투갈어(Portuguese)	포르투갈	15	221
8	러시아어(Russian)	러시아연방	19	154
9	일본어(Japanese)	일본	2	128
10	란다어(Lahnda)	파키스탄	6	119
11	마라티어(Marathi)	인도	1	83
12	텔루구어(Telugu)	인도	2	82
13	말레이어(Malay)	말레이시아	20	80
14	터키어(Turkish)	터키	8	79
15	한국어(Korean)	대한민국	6	77.3

- 학습데이터 부족
  - ImageNet: 14 million
  - 영어 학습데이터의 경우에도, 대부분 십만 건 이하 수준

Copyright© 2020 by ETRI

자연어 처리가 어려운 이유(2/2)

- AI's Language Problem (MIT Technology Review by Will Knight, 2016)
  - Machines that truly understand language would be incredibly useful.
  - But we don't know how to build them.

- 문법 분석: 문어체는 언어학자 수준 분석 가능
  - 품사 분류, 구문 분석 등
  - 신조어나 학습이 안된 단어는 취약

<조삼모사>

1. 「아침에 세 개, 저녁에 네 개」라는 뜻
2. 간사한 꾀로 남을 속임
3. 신조어

- 의미 분석: 전 세계적으로 연구 초기 단계
  - Meaning: 중의, 함의, 은유, 반의, 축약, 생략
  - Context: 문맥, 맥락, 상황
  - Reasoning: 상식과 상식적 추론 능력 필요
    - 살인죄의 공소시효는?
    - (형사소송법 제253조의2) 사람을 살해한 범죄

< 문자는 유한, 의미는 무한 >

1. 언어는 문맥에 따라 의미가 결정됨
2. 그리고 세계에 대한 배경지식이 필요

Copyright© 2020 by ETRI

자연어의 의미란?

- 의미 (meaning)

국어사전 단어 1-5 / 185건

의미[2] (意味) ★★

[명사]

1. 말이나 글의 뜻.
2. 행위나 현상이 지닌 뜻.
3. 사물이나 현상의 가치.

[유의어] 의지[6], 의의[4], 감각[2]

Merriam-Webster SINCE 1828

GAMES | BROWSE THESAURUS | WORD OF THE DAY | WO

meaning

DICTIONARY THESAURUS

mean·ing | \ ˈmē-niŋ \

Definition of *meaning* (Entry 1 of 2)

1 a : the thing one intends to convey especially by language : PURPORT
// Do not mistake my *meaning*.

b : the thing that is conveyed especially by language : IMPORT
// Many words have more than one *meaning*.

2 : something meant or intended : AIM
// a mischievous *meaning* was apparent

signifier (symbol) ⟺ signified (idea or thing)

- Distributional Semantics
  - "You shall know a word by the company it keeps" (J. R. Firth 1957: 11)
  - "나는 맛있는 사과[5]를 먹었다." 사과[5] (沙果/砂果)「명사」 사과나무의 열매.
  - "나는 싫지만 사과[8]를 했다. 사과[8] (謝過)「명사」 자기의 잘못을 인정하고 용서를 빎.

문맥(context)

Copyright© 2020 by ETRI

Word Vectors

- 주어진 문서 집합에서, 각 단어에 대해 비슷한 문맥에 나타난 단어의 벡터와 유사한 벡터를 갖도록 만들어진 결과
  - 자연어의 기호를 기계가 처리할 수 있는 숫자의 나열인 벡터로 표현
  - word vectors = word embeddings = word representations

Dimensions

Word vectors				
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

- 문법적/의미적 정보의 임베딩
  - 단어, 문장, 단락 간의 유사도 계산 가능
  - 다양한 태스크에 Fine-tuning을 통한 전이학습 적용 가능

Copyright© 2020 by ETRI

Language Model

- Language Model
  - 주어진 문맥에서 다음 단어를 예측하는 태스크를 수행하는 시스템

More formally: given a sequence of words $x^{(1)}, x^{(2)}, \ldots, x^{(t)}$, compute the probability distribution of the next word $x^{(t+1)}$:

$$P(x^{(t+1)} | \, x^{(t)}, \ldots, x^{(1)})$$

where $x^{(t+1)}$ can be any word in the vocabulary $V = \{w_1, \ldots, w_{|V|}\}$

- 구글 검색의 언어모델 사용 예

Copyright© 2020 by ETRI

Word2Vec

- (Tomas Mikolov, 2013) Efficient Estimation of Word Representations in Vector Space
  - two model architectures for computing continuous vector representations of words from very large data sets.
    - CBOW (Continuous Bag of Words) : Predict center word from (bag of) context words
    - Skip-gram : Predict context ("outside") words (position independent) given center word

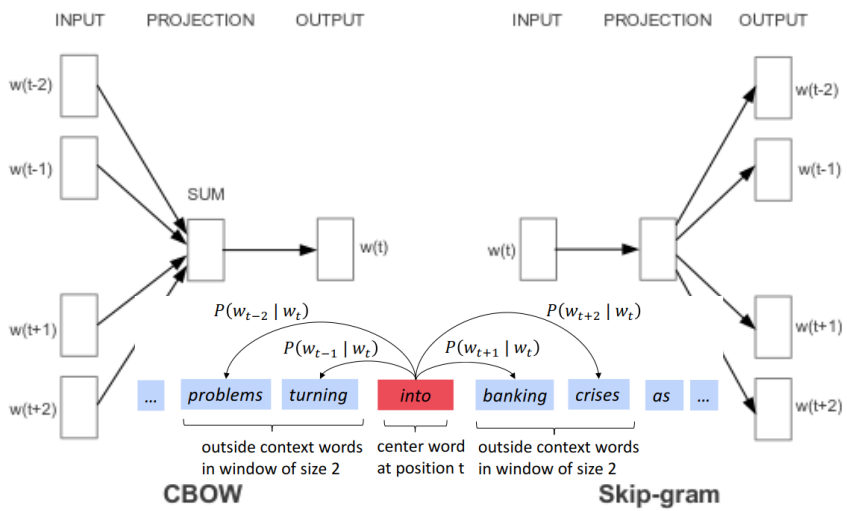


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Method	Time [min]	Syntactic [%]	Semantic [%]	Total accuracy [%]
NEG-5	38	63	54	59
NEG-15	97	63	58	**61**
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
The following results use 10^{-5} subsampling				
NEG-5	14	61	58	60
NEG-15	36	61	61	**61**
HS-Huffman	21	52	59	55

Table 1: Accuracy of various Skip-gram 300-dimensional models on the analogical reasoning task as defined in [8]. NEG-k stands for Negative Sampling with k negative samples for each positive sample; NCE stands for Noise Contrastive Estimation and HS-Huffman stands for the Hierarchical Softmax with the frequency-based Huffman codes.

Copyright© 2020 by ETRI

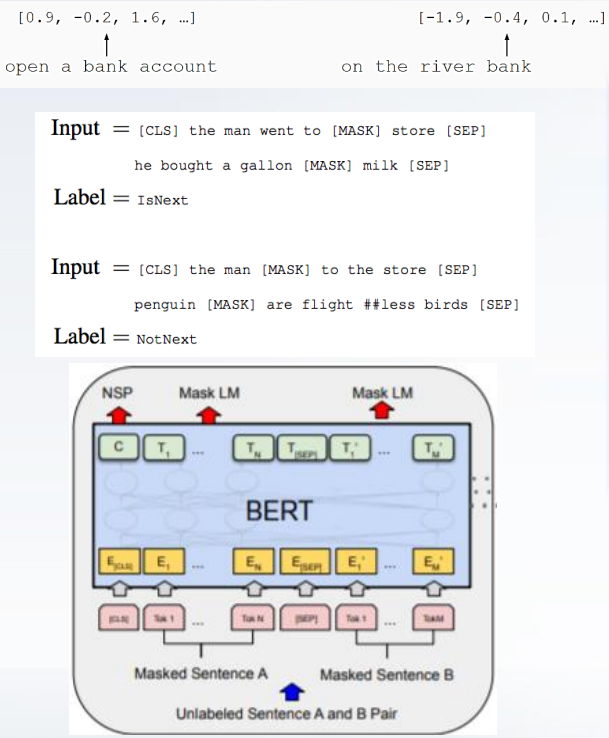
BERT: 자연어 처리의 혁신적 패러다임 시작 ('18.~)

- **Pretrain then Finetune**: A Transfer Learning Paradigm in NLP
  - 방법론: ELMO, OpenAI GPT, BERT, XLNet, ERNIE, RoBERTa, ALBERT, …

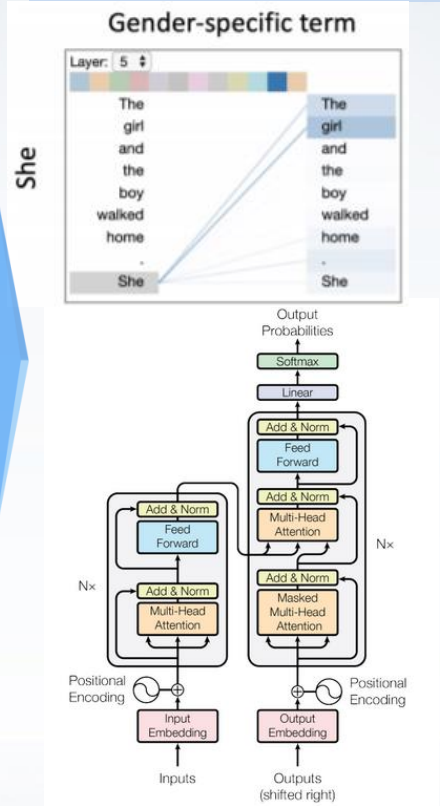
Contextual Language Model
- Masked Language Model
- Next Sentence Prediction

Multi-Head Attention
- 12~16 attention heads

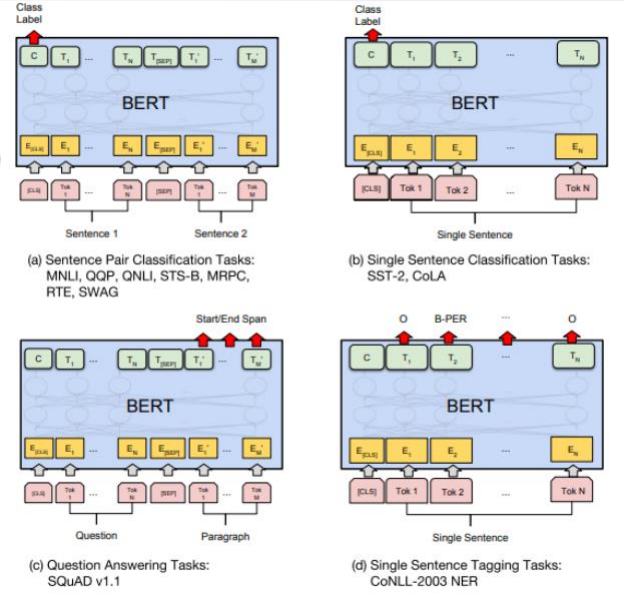
Transfer Learning from Supervised Data
- Easy Training (2-4 epochs)
- Less Data
- Good Results



〈Pretraining Language Model〉



〈Transformer Model〉



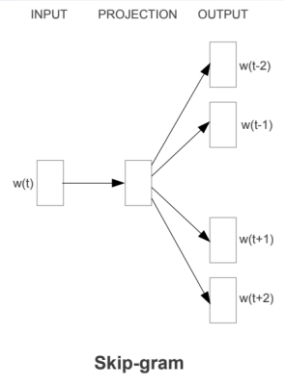
〈Finetuning Language Model〉

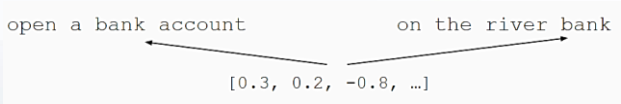
* Language Model: 문장(단어 시퀀스)의 확률을 예측하는 모델

Copyright© 2020 by ETRI

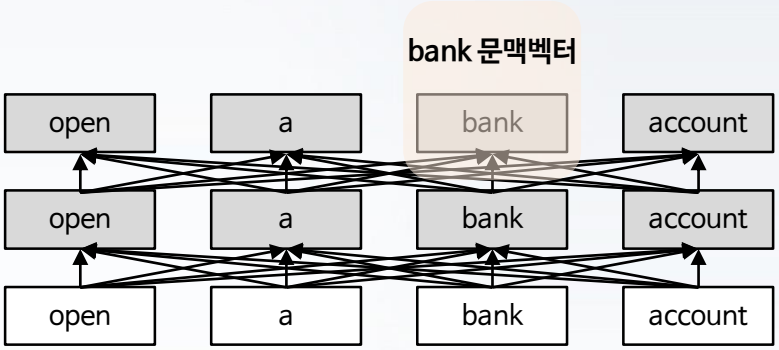
BERT 언어모델

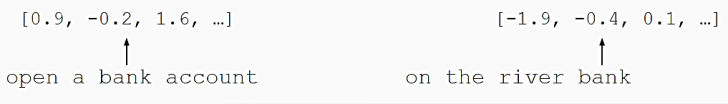
- BERT에서 해결하고자 **하는 문제** = contextual representation
  - 문제점
    - 기존 워드 임베딩 접근 방법은 문맥을 고려하지 못하는 한계를 지님
  - 접근 방법
    - 입력 문장(N개 단어)에 대해서, 뉴럴넷을 적용한 결과(N개 출력)를 단어의 문맥 반영 벡터로 활용





〈기존 워드임베딩 접근방법 〉
(context free manner)



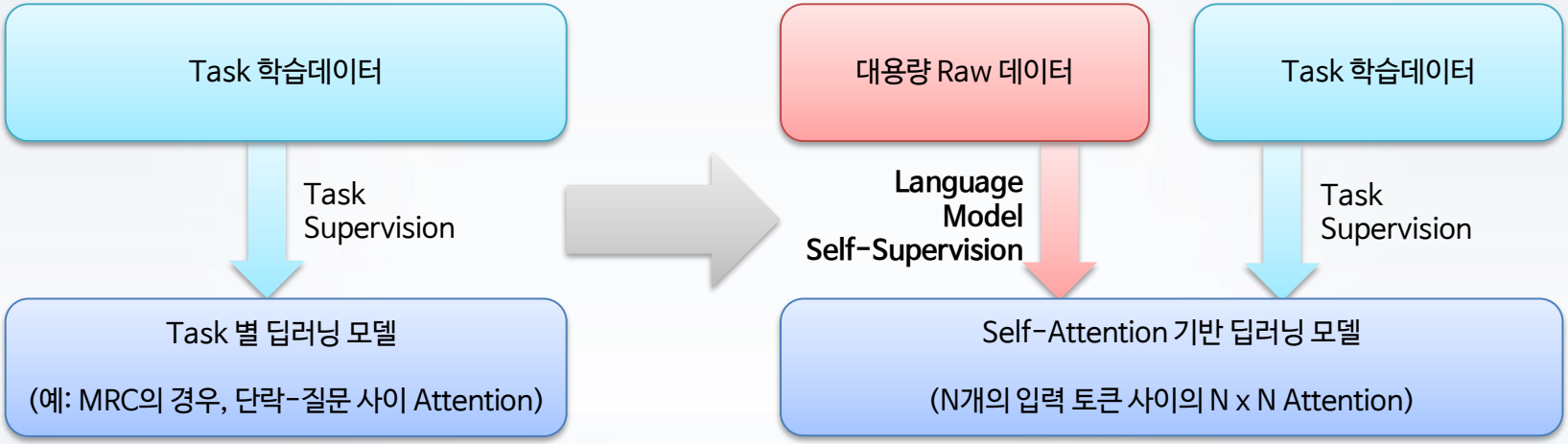


〈BERT 접근방법 〉
(contextual representation)

Copyright© 2020 by ETRI

BERT 언어모델

- Contextual Representation 적용 전/후
  - 적용 전
    - 응용 태스트 별 별도의 딥러닝 모델과 학습 데이터를 이용하여 모델 개발
  - 적용 후
    - 1) 대용량 Raw 데이터로부터 공통 언어모델(Language Model) 미리 학습(pre-train)
    - 2) 학습된 공통 모델을 응용 태스크 별로 재학습(fine-tuning)하여 적용



〈기존 접근방법〉 〈BERT 접근방법〉

Copyright© 2020 by ETRI

BERT 언어모델

- **BERT 모델 Pre-training: Masked LM + NSP Task**
  - 1) 공백 단어 예측
    - 입력 문장 중, 15%의 단어를 masking 후 해당 단어를 맞추는 태스크

```
                                store                gallon
                                  ↑                    ↑
the man went to the [MASK] to buy a [MASK] of milk
```

양방향 기반 단어 예측을 위하여 BERT에서 제안한 학습방법

    - BERT 특징: 양방향 정보를 이용한 단어 예측
      - 기존 연구: 단어 예측 시, 단일 방향의 정보만 고려

  - 2) 문장 선후관계 예측
    - 임의의 두 문장에 대해, 두 문장이 선/후 관계가 맞는지 맞추는 태스크

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

  - (비고) 두 태스크 모두 별도의 정답 말뭉치 없이 대용량 원시 말뭉치로부터 자동으로 생성 가능 (→ 용어: self-supervised learning)

Copyright© 2020 by ETRI

한국어 언어모델 KorBERT 개발

한국어에 최적화된 언어 모델 개발 (구글 대비 4.5% 우수)

* 한국어 = 내용어(명사, 동사 등)와 기능어(조사, 어미 등)이 결합하여 어절을 구성하는 교착어

※ BERT 언어모델에서 언어 단위의 중요성

1. 입력 언어 단위 = 출력 언어 단위 ➔ 언어 단위 별 응용 태스크 활용
   - (구글방식) "백설공주가" 에서 "주가"를 한 단위로 인식 → "주가" 단위로 개체명인식 활용
2. BERT 언어모델 = 입력 언어 단위 사이의 관계 학습 모델
   - (구글방식) 주식 문서의 "주가"와 "백설공주가"의 "주가" 사이의 모호성(ambiguity) 증가

	구글 배포 모델 언어 단위	ETRI KorBERT 언어 단위
접근 방법	영어*의 언어 단위 구성 방법 적용 * 내용어와 기능어가 공백으로 구분된 언어 (굴절어)	한국어* 의미의 최소 단위인 형태소 기반 언어단위 * 내용어와 기능어가 어절로 결합된 언어 (교착어)
언어 단위 구축 방법	말뭉치에서 통계적으로 추출한 음절 단위 n-gram 적용 방법 (BPE)	형태소 분석 이후, 형태소 단위에 대해서 BPE 적용 방법
언어 단위 적용 결과	한국 ##어 단 ##어는 형태 ##소로 구 ##성된 ##다.	한국어/NNP_ 단어/NNG_ 는/JX_ 형태소/NNG_ 로/JKB_ 구성/NNG_ 되/XSV_ ㄴ다/EF_ ./SF_
동일 학습데이터 평가결과	Google wordpiece 모델 - 기계독해: 90.68% - 단락순위화: 66.3%	KorBERT 형태소 모델 - 기계독해: 95.02% (+4.3%) - 단락순위화: 73.7% (+7.4%)

Copyright© 2020 by ETRI

딥러닝 한국어 언어모델 KorBERT의 국내 배포

http://aiopen.etri.re.kr/ 사이트에서 배포 중

(1) 한국어 기계독해 챌린지 62개팀 중 1위 (LG CNS 주관, 2019.10.25)

(2) 구글 BERT 모델 대비 평균 4.5% 우수 (5개 태스크 대상 비교 평가)

(3) 산업계/학계 주요 딥러닝 언어모델로 자리매김 (KSC 2019, HCLT 2019 논문)

KorQuAD 1.0

The Korean Question Answering Dataset

Leaderboard

KorQuAD 1.0의 Test set으로 평가한 Exact Match(EM) 및 F1 score 입니다.

Rank	Reg. Date	Model	EM	F1
-	2018.10.17	**Human Performance**	80.17	91.20
1	2019.10.25	KorBERT-Large v1.0 ETRI ExoBrain Team	**87.76**	**95.02**
2	2019.06.26	LaRva-Kor-Large+ + CLaF (single) Clova AI LaRva Team	86.84	94.75
3	2019.06.04	BERT-CLKT-MIDDLE (single model) Anonymous	86.71	94.55

평가 항목

[아래 5개 태스크를 평가 항목으로 선정]

- 의미역 인식(Semantic Role Labeling): 문장 내에서 술어에 의해 기술되는 사건에 대한 개체들의 역할을 인식
- 기계 독해(Machine Reading Comprehension): 주어진 단락에서 질문이 요구하는 정답을 찾음
- 단락 순위화(Passage Ranking): 검색결과 집합에서 질문에 찾는 정답이 들어있는 단락 순위화
- 문장 유사도 추론(Natural Language Inference): 2개 문장 간 의미가 동일한 지 여부를 분류
- 문서 주제분류: 문서의 주제를 기정의된 54개의 클래스 중 하나로 분류

평가 결과

구분	의미역인식	기계독해	단락순위화	문장유사도추론	문서주제분류
평가데이터 및 규격	Korean Propbank, 학습: 19,302 문장 평가: 3,773 문장	KorQuAD 데이터, 학습: 60,406건 평가: 5,773건 (dev셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	학습: 10,874문장쌍 평가: 1,209문장쌍 (이진 분류체계: 유사, 무관)	학습: 9,307 평가: 1,035 (54개 분류체계)
평가 방법	F1[2]	Exact Match[3] / F1	Precision@Top1	Accuracy	Accuracy
(Google) Word Piece[4]기반 한국어 언어모델	81.85%	80.82% / 90.68% (정답 경계 구분을 위해 후처리 수행)	66.3%	79.4%	91.1%
(엑소브레인) Word Piece 기반 한국어 언어모델	85.10%	80.70% / 91.94% (정답 경계 구분을 위해 후처리 수행)	70.5%	82.7%	93.4%
(엑소브레인) 형태소 기반 한국어 언어모델	85.77%	86.40% / 94.18%	73.7%	83.4%	93.7%

[2] F1: 정확률(Precision, 시스템이 결과가 정답인 비율)과 재현률(Recall, 실제 정답을 시스템이 맞춘 비율)의 조화평균
[3] Exact Match: 시스템이 제시한 결과와 정답이 완전히 일치하는 비율
[4] Word Piece: 하나의 단어를 내부 단어(Subword Unit)들로 분리하는 단어 분리 모델

전자신문 2019년 06월 12일 수요일 027면 전국 20.4 x 13.6 cm

한국말 잘 알아듣는 '코버트'…구글 '버트' 넘어섰다

(KorBERT) (BERT)

ETRI, 최첨단 한국어 언어모델 공개 문장 유사도 추론 등 정확도 4.5%↑

한국전자통신연구원(ETRI·원장 김명준)이 최첨단 한국어 언어모델을 개발해 공개했다. 이 모델은 한국어 특성을 반영해 구글 모델보다 우수한 성능을 자랑한다. 향후 인공지능(AI) 에서, AI 질의응답, 지능형 검색 등 한국어를 활용한 서비스 고도화에 쓰인다.

ETRI는 한국어 언어모델 "버트(KorBERT)"를 자체 운영하는 "AI개방플랫폼 서비스 포털"에 10일 공개했다. 언어모델은 단어가 나타나는 유형을 분석해 모은 덩어리다. 사전에 자료 학습 단계에 필수 요소다. 기존에는 구글이 개발한 다국어 모델 "버트(BERT)"가 주로 쓰였다.

코버트는 버트에 비해 5가지 성능 평가 항목에서 우수하다. 문장 유사도 추론, 주제 분류, 기계독해 등 영역에서 정확도 성능이 평균 4.5% 우수했다. 특히 정확한 단어를 순위매김하는 "단락 순위화" 평가에서는 7.4% 높았다.

한국어 특성을 반영한 결과다. 버트는 단어가 모두 띄어져 있는 고립어 기반인데, 코버트는 어근에 접사가 붙는 교착어 기반이다. 한국어는 대표적인 교착어다. ETRI는 고유어에 맞게 코어 파스 학습인 형태소까지 고려하는 모델을 구현, 성능도 높였다.

학습 데이터도 늘렸다. 버트가 사용한 데이터는 40여만건 위키백과 문서다. ETRI는 여기에 추 요리가 하여 5천여 문서를 더한 신문기사·백과사전 정보를 더해 45여개 형태소를 학습시켰다.

코버트는 AI 서비스 경쟁에서 늦어는에 활용가능성이다. 주요 질문에 프리뷰워크인 파이토치나 텐서플로우 환경에서 모두 사용할 수 있다. 이미 지난 3월부터 한전KDN 지식검색 네트워크에 활용되고 있다. ETRI는 "현도고 도서 도움 챗봇의 방법 등에 도과하여 AI개발 속가능 개하고, 유사 계산 지능형 문서기술도 출시할 계획이다. 더 많은 언어 데이터를 제시할 수 있도록 모델 고도화도 준비하고 있다.

김명기 ETRI 센서는 "영어에 최적화된 언어모델을 만들었다"며 "한국어 분석, 지식추론, 질의응답 등 다양한 한국어 입력에 기술 고도화가 가능할 것"이라고 강조했다.

대전=김병준기자 kyb@dailynews.com

※ KorBERT 활용 논문 (서강대,국립국어원 주관 국어정보처리시스템 경진대회 대상 수상)

4. 실험

본 논문에서는 ETRI[10]에서 사전 학습한 BERT base 모델을 사용하였다. BERT를 사용하기 위한 형태소 분석으로는 ETRI 형태소 분석기를 사용하였다. ELMo는 4GB의 대용량 뉴스 데이터를 이용하여 사전 학습하였다. 의존 구

표 1. 사전 학습 데이터에 따른 성능 비교(%, dev)

Model	EM	F1
BERT-google multi-lingual (baseline)	77.69	89.98
BERT-ETRI	84.82	92.74

표 1. 상호참조해결 모델 별 성능 비교(자동 분석 형태소 이용)

DEV				
Model	**MUC**	**B^3**	**CEAFe**	**CoNLL F1**
규칙기반 [5]	44.25	45.21	50.91	46.79
Ptr-net [7]	59.07	55.16	53.58	55.94
Ptr-net + char + ELMo [7]	60.63	56.79	55.46	57.63
BERT-coref$_{rand}$ (ours)	49.62	47.79	46.29	47.90
BERT-coref (ours)	**70.23**	**68.01**	**65.66**	**67.97**
TEST				
Model	**MUC**	**B^3**	**CEAFe**	**CoNLL F1**
Ptr-net + char + ELMo [7]	52.70	48.93	46.82	49.48
BERT-coref (ours)	**65.98**	**64.03**	**63.64**	**64.55**

Copyright© 2020 by ETRI

도메인에 따른 언어모델(KorBERT) 확장 시 성능 개선 확인

(1) 언어모델 학습 고도화를 위한 **자가학습 및 가중치 학습 방법 고도화**

(2) 전이학습 기반 **도메인 특화 법률 • 특허 언어모델 개발 및 성능 개선**

[일반분야 언어모델 구축]

대용량 한국어 텍스트

- 위키백과 및 신문기사 23.5 GB (약 15년 분량)
- 47억개 형태소 수집

학습 대상 문서 선별

형태소 기반 의미화 (47억개 형태소)

학습 시 동적 문제 생성 (신규 문제, 난이도 상향)

언어 모델 예측

대량 예제 기반 올바른 가중치 학습 (학습 안정성 향상)

KorBERT 언어모델

※ 자가학습 태스크 및 가중치 학습 방법 비교

KorBERT 모델	구글 배포 모델
Whole word masking	Single token masking
Dynamic masking	Static masking
Flexible difficulty level (15% → 20%)	Fixed difficulty level
Robust Optimization (8,000 example base)	Regular optimization (256 example base)

※ 학습 대상 예문 집합 중, "눈/명사"가 모두 eye 의미 경우,
- Snow 의미에 대한 올바른 가중치 학습 실패 가능성 증가
- 256개 예문 기반 학습 시, 데이터 편향 위험성 증가
- 8,000개 예문 기반 학습 시, 올바른 가중치 학습 가능

[법률/특허 도메인 언어모델 구축]

일반분야 KorBERT + 법률분야 텍스트 (186MB)

특허분야 원시 텍스트 (XML 53GB)

학습 대상 문서 선별

형태소 기반 의미화

법률/특허 분야 문제 생성

언어 모델 예측

대량 예제 기반 올바른 가중치 학습

KorBERT-Legal 언어모델

KorBERT-Patent 언어모델

법률 기계독해 적용 → 2.01% 개선

특허 분류 적용 → 6.25% 개선

Copyright© 2020 by ETRI

자연어처리 + 딥러닝

- 한국어 자연어처리 문제 예
  - 언어분석
    - 형태소분석 / 개체명인식 / 어휘의미분석
    - 의존구문분석 / 의미역인식
    - 생략어복원 / 상호참조해결
  - 문장추론
    - 패러프레이즈 분류 / 문장유사도 인식 / 함의 추론
  - 정보검색/질의응답
    - 기계독해 / 단락 순위화
  - 그 외 (not covered)
    - 문서주제분류
    - 정보추출
    - 기계번역
    - 자연어 생성 (i.e. 요약생성)

Copyright© 2020 by ETRI

- 한국어 자연어처리 문제 예
  - [유형#1] 미리 정의된 출력 클래스 중 하나로 분류하는 문제

태스크	출력 클래스 집합	입력 예	출력 예
형태소분석	형태소 태그셋 NNG, NNP, JX, etc (BIO 태그 결합 후 사용)	백 설 공 주 가	B-NNP I-NNP I-NNP I-NNP B-JKS
개체명인식	개체명 태그셋 PS_NAME, CV_OCCUPATION, etc (BIO 태그 결합 후 사용)	백설공주/NNP 가/JKS 독/NNG 사과/NNG 를/JKO	B-PS_NAME OTHER B-PT_FRUIT I-PT_FRUIT OTHER
의미역인식	의미역 집합 ARG0, ARG1, ARGM-LOC, ETC (BIO 태그 결합 후 사용)	백설공주/NNP + 가/JKS 독/NNG + 사과/NNG + 를/JKO	B-ARG0 B-ARG1
패러프레이즈 분류	O / X 이진 분류	(문장1) 백설공주가 독사과를 먹었다. (문장2) 백설공주가 먹은 사과에 독이 들어 있었다.	O / X
문서 주제 분류	N개의 주제 클래스 (데이터 별 상이)	소비자원 “프랜차이즈 치킨 만족도 1위 페리카나” …	경제

Copyright© 2020 by ETRI

- 한국어 자연어처리 문제 예
  - [유형#2] 입력 단위 중, 하나를 선택하는 유형

softmax over input sequence

태스크	유형	예제
의존구문분석	각 어절에 대해서, 입력 어휘 열 중, 자신의 지배소 선택	예문: 백설공주가 독이 든 사과를 먹었다. 피지배소: 백설공주가 지배소 후보: 독이 / 든 / 사과를 / 먹었다. / ROOT 정답 지배소: 먹었다.
기계독해	입력된 질문-단락에 대해서, 단락 내 정답의 경계(시작/끝) 인식	단락: 백설공주가 독이 든 사과를 먹었다. 질문: 백설공주가 먹은 것은? 정답 시작/끝 후보: - 백설공주 / 가 / 독 / 이 / 드 / ㄴ / 사과 / 를 / 먹 / 었 / 다 / . 정답 경계 시작: 독 정답 경계 끝: 사과 (정답 = 독이 든 사과)
상호참조해결	문서 내에서 동일 대상을 지칭하는 명사구 인식	문서: 백설공주가 독이 든 사과를 먹었다. 그녀는 바로 잠이 들었다. 상호참조해결 대상: 그녀 동일 대상 후보: 백설공주 / 독 / 사과 / 독이 든 사과 상호참조 정답: 백설공주

Copyright© 2020 by ETRI

- 한국어 자연어처리 문제 예
  - [유형#3] 특정 클래스가 아닌 연속 값을 찾는 유형 (regression)

태스크	유형	예제
문장 유사도 인식	입력된 두 문장 사이의 유사도 점수 인식 (i.e. 0점 ~ 5점 사이)	(문장1) 백설공주가 독사과를 먹었다. (문장2) 백설공주가 먹은 사과에 독이 들어 있었다. 유사도 점수: 5점
질문-단락 재순위화	질문 및 검색 단락 집합에 대해, 정답을 포함한 순위로 검색 단락 재순위화 (정답 포함 단락: 1.0, 정답 미포함 단락: 0.0)	질문: 헌법상 대통령의 임기는 얼마인가? 검색단락#1: 대통령의 임기가 만료된 때에는… 검색단락#2: 원장은 대통령이 임명하고, 그 임기는 … 검색단락#3: 대통령의 임기연장 또는 중임변경을 … 검색단락#4: <u>대통령의 임기는 5년으로 하며</u>, 중임 … 〈재순위화 결과〉 검색단락#1: <u>대통령의 임기는 5년으로 하며</u>, 중임 … 검색단락#2: 원장은 대통령이 임명하고, 그 임기는 … 검색단락#3: 대통령의 임기연장 또는 중임변경을 … 검색단락#4: 대통령의 임기가 만료된 때에는…

Copyright© 2020 by ETRI

• 한국어 언어분석 예

소피스트란 그리스어로 지혜로운 자 또는 지혜를 만들어내는 사람이라는 뜻으로, BC 5~4세기의 그리스의 철학자들을 말한다. 이들은 아테네 사람들을 대상으로 하였고, 수사학과 웅변술을 가르쳤다.

어휘

형태소 분석

소피스트/NNG+란/JX 그리스/NNP+어/XSN+로/JKB 지혜롭/VA+ㄴ/ETM 자/NNB 또는/MAG 지혜/NNG+를/JKO 만들/VV+어/EC+내/VX+는/ETM 사람/NNG+이/VCP+라는/ETM 뜻/NNG+으로/JKB+,/SP BC/SL 5/SN+~/SO+4/SN+세기/NNP+의/JKG 그리스/NNP+의/JKG 철학/NNG+자/XSN+들/XSN+을/JKO 말/NNG+하/XSV+s다/EF+./SF 이/NP+들/XSN+은/JX 아테네/NNP 사람/NNG+들/XSN+을/JKO 대상/NGG+으로/JKB 하/VV+았/EP+고//EC+,/SP 수사/NNG+학/XSN+과/JC 웅변/NNG+술/XSN+을/JKO 가르치/VV+었/EP+다/EF+./SF

개체명 인식

<CV_OCCUPATION:소피스트/NNG>+란/JX <CV_LANGUATE:그리스/NNP+어/XSN>+로/JKB 지혜롭/VA+ㄴ/ETM 자/NNB 또는/MAG 지혜/NNG+를/JKO 만들/VV+어/EC+내/VX+는/ETM 사람/NNG+이/VCP+라는/ETM 뜻/NNG+으로/JKB+,/SP <DT_DURATION:BC/SL 5/SN+~/SO+4/SN+세기/NNP>+의/JKG <LCP_COUNTRY:그리스/NNP>+의/JKG 철학/NNG+자/XSN+들/XSN+을/JKO 말/NNG+하/XSV+s다/EF+./SF 이/NP+들/XSN+은/JX <LCP_CAPITALCITY:아테네/NNP> 사람/NNG+들/XSN+을/JKO 대상/NGG+으로/JKB 하/VV+았/EP+고//EC+,/SP <FD_ART:수사/NNG+학/XSN>+과/JC <FD_ART:웅변/NNG+술/XSN>+을/JKO 가르치/VV+었/EP+다/EF+./SF

어휘의미 분석

소피스트란 그리스어로 지혜로운 자__18_0000/NNB 또는 지혜__02_0001/NNG+를 만들__00_0101/VV+어내는 사람__00_0001/NNG+이라는 뜻__00_0002/NNG+으로, BC 5~4+세기__03_0002/NNG+의 그리스__02_0000/NNP+의 철학자들을 말하__00_0101/VV+ㄴ다.

문장

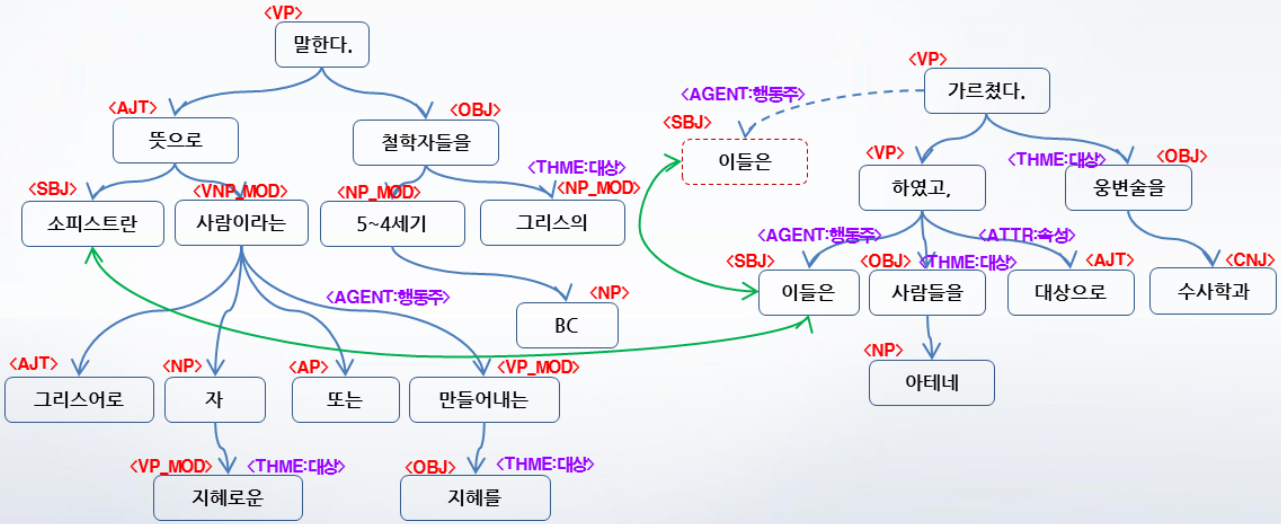
의존구문 분석

의미역 인식

문맥

상호참조 해결

생략어 복원

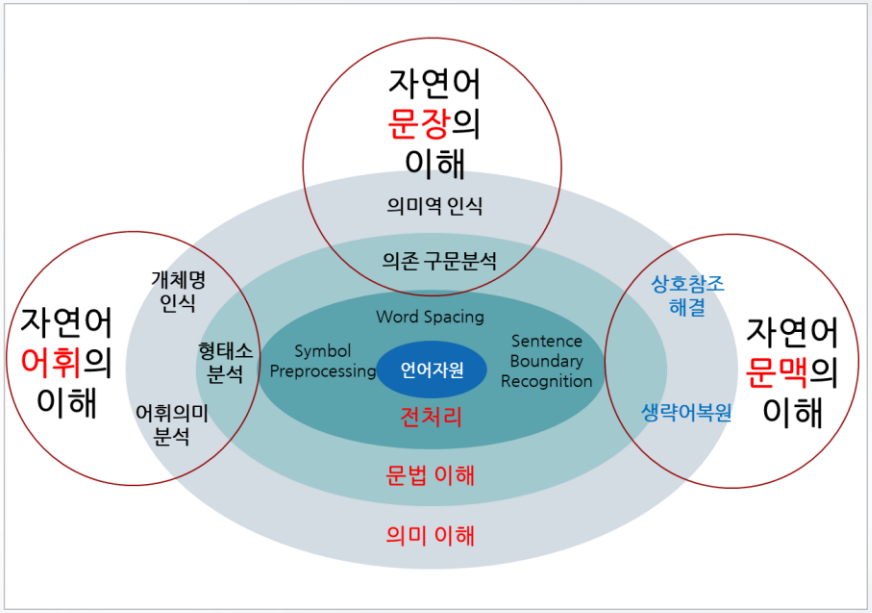


Copyright© 2020 by ETRI

엑소브레인 한국어 분석 기술

- 한국어 문장 또는 문서를 입력 받아 아래의 정보를 자동 분석
  - 문장: 행정안전부는 주민참여예산 활성화를 위해 전국 52개 자치단체를 대상으로 주민참여 예산 자문을 실시한다.
    - 형태소: 행정안전부 → 행정/일반명사+안전/일반명사+부/접미사
    - 개체명: 행정안전부 → 정부기관
    - 구문분석: {행정안전부는}/주어 {주민참여예산 활성화를}/목적어 {위해}/동사
    - 의미역할: {행정안전부는}/행위자 {주민참여예산 활성화를 위해}/목적 … {실시한다}/술어

형태소 분석	문장을 분해 가능한 의미 최소단위로 분리하고 품사를 태깅
개체명 인식	문장에서 개체명을 추출하고 추출된 개체명의 범주를 결정
동음이의어 분석	문장의 문맥에 맞는 동음이의어(글자의 음은 같으나 뜻이 다른 어휘)를 표준국어대사전의 의미번호에 부합되게 분석
다의어 분석	문장의 문맥에 맞는 다의어를(하나의 낱말이 여러가지 뜻으로 쓰이는 어휘)를 표준국어대사전의 의미번호에 부합되게 분석
의존구문분석	문장의 구문구조를 "지배소와 피지배소"의 의존관계로 분석(지배소, 의존관계, 피지배소)
의미역 인식	문장 내에서 서술어에 의해 기술되는 사건에 대한 개체들의 역할을 인식



Copyright© 2020 by ETRI

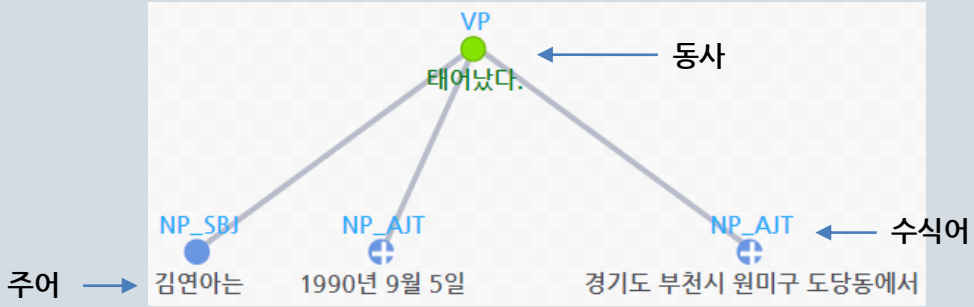
자연어 어휘 이해

- **(문법) 형태소 분석**
  - 독립적 의미를 가진 가장 작은 언어표현 단위
    - ✓ 명사(예: 하늘, 학교, 한국), 용언(예: 높, 예쁘, 달리), 조사(예: 은, 는, 이, 가), 어미(예:ㄴ, ㄹ, 고, 게), 부사 등…
- **(의미) 개체명 인식**
  - 인명, 지명, 기관명 등의 개체명을 문서에서 추출하고 종류를 결정
    - ✓ 15개 대분류 및 146개 세부분류 개체명 태그 사용
- **(의미) 동음이의어/다의어 분석**
  - 동음이의어: 형태는 동일하나 의미들이 전혀 다른 단어
    - ✓ 예제: 배 → '배_01(신체부위)', '배_02(기계, 운송수단)', '배_03(열매)'
  - 다의어: 포괄적인 의미는 동일하지만 어원적으로 다른 의미를 가지고 있는 단어
    - ✓ 예제: '배_01(신체부위)'
      - 배_01_00_01: 사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위
      - 배_01_00_02:절족동물, 특히 곤충에서 머리와 가슴이 아닌 부분
      - 배_01_00_03: 긴 물건 가운데의 볼록한 부분

Copyright© 2020 by ETRI

자연어 문장 이해

- **(문법) 의존 구문분석**
  - 어절 간의 지배-피지배 관계 분석을 통하여, 문장의 구조적 관계를 분석
    - ✓ 예: 김연아는 1990년 9월 5일 경기도 부천시 원미구 도당동에서 태어났다.



- **(의미) 의미역 인식**
  - 문장 내에서 서술어를 중심으로 서술어에 대한 의미적 역할을 하는 문장의 부분(논항)을 인식

Predicate	태어나 . 1
행위자격(ARG0)	
대상격(ARG1)	김연아
시작점(ARG2)	
도착점(ARG3)	
부가격(ARGM)	시간 : 1990년 9월 5일 장소 : 경기도 부천시 원미구 도당동

Copyright© 2020 by ETRI

문장 유사도 추론의 어텐션 시각화

- 리스트에 들어간 75명은 한 명도 빠짐없이 서류전형을 통과하였습니다.
- 리스트에 있는 75명은 공채 1차 서류전형에서 전원 합격하였습니다.

Copyright© 2020 by ETRI

KorBERT 기반 문장 유사도 추론

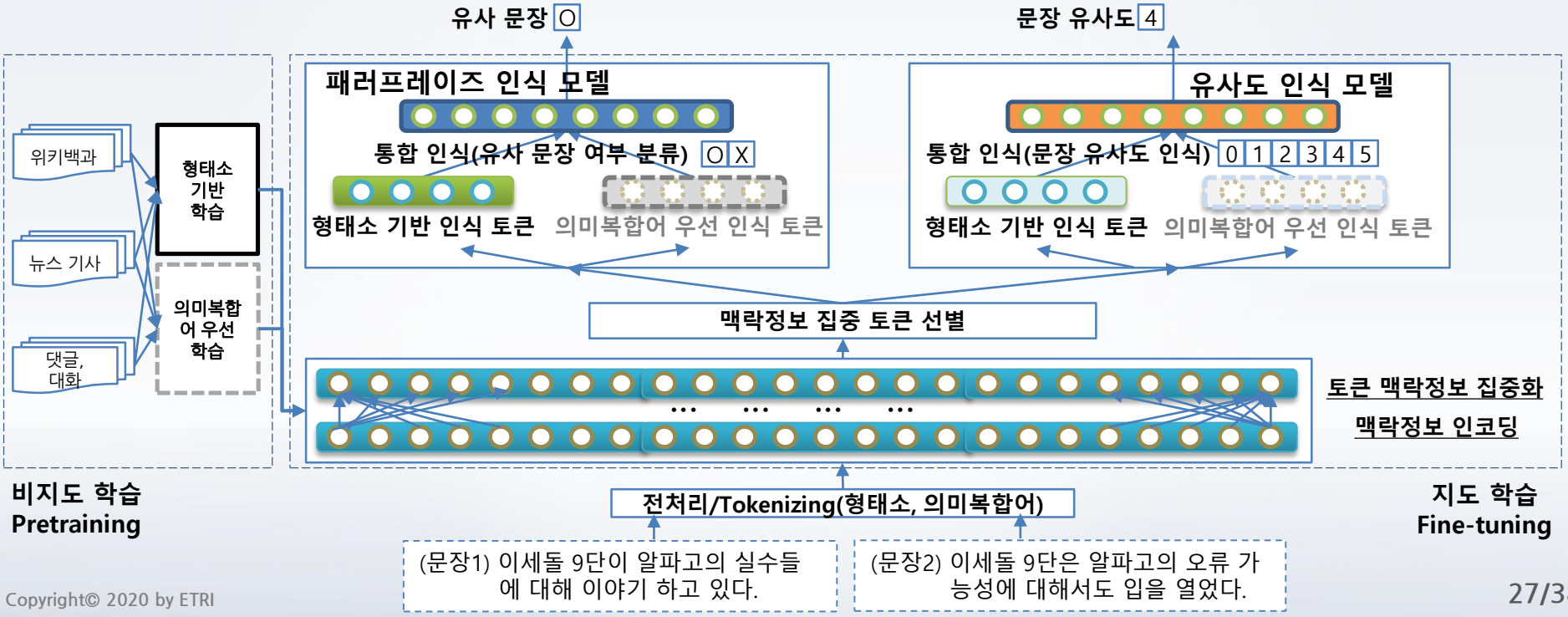
연구 목표

- 딥러닝 기반 문장 간 유사도(Natural Language Inference or Paraphrase Recognition) 추론 성능 개선
- 질문-단락 간 유사도 분석, 특허 출원서-등록특허 간 유사도 분석에 활용

주요 성과

- KorBERT 기반 맥락정보 집중화 토큰 선별을 통한 "유사 문장 여부" 분류 모듈 개발
  - 성능(정확률): 83.37% (베이스라인 대비 +3.97%) 베이스라인(구글 다국어 모델 기반): 79.40%
- KorBERT 기반 맥락정보 집중화 토큰 선별을 통한 "Likert 0~5 Scale 문장 유사도" 인식 모듈 개발
  - 성능(Pearson상관계수 x100): 88.46 (베이스라인 대비 +4.09) 베이스라인(구글 다국어 모델 기반): 84.37

평가데이터: 뉴스문장 12,807쌍(학습:평가=9:1)



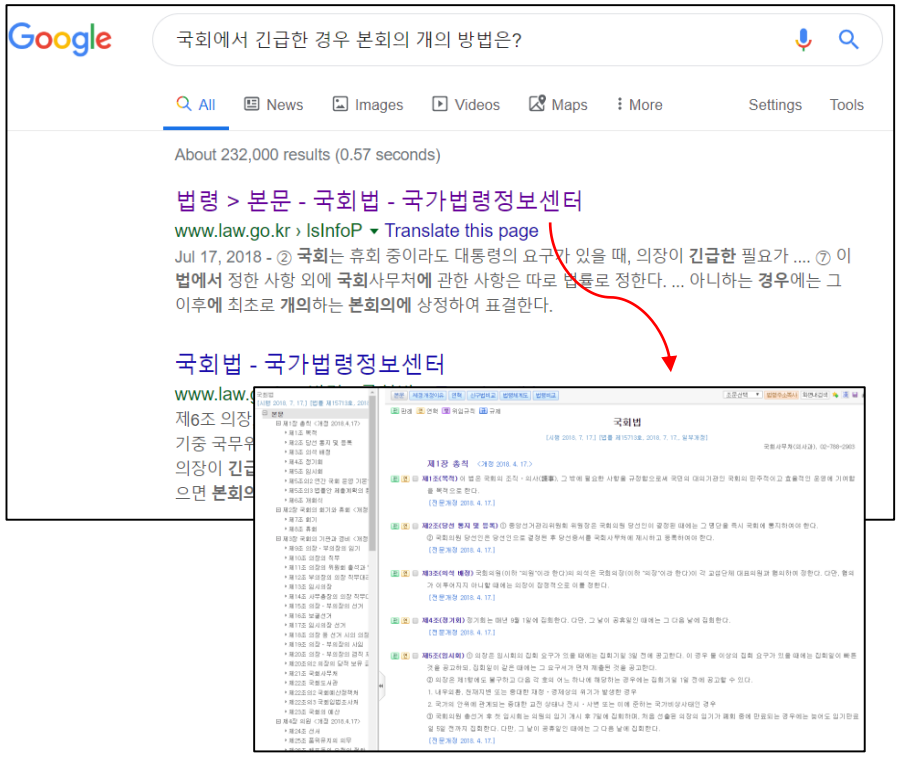
Copyright© 2020 by ETRI

자연어 질의응답 기술

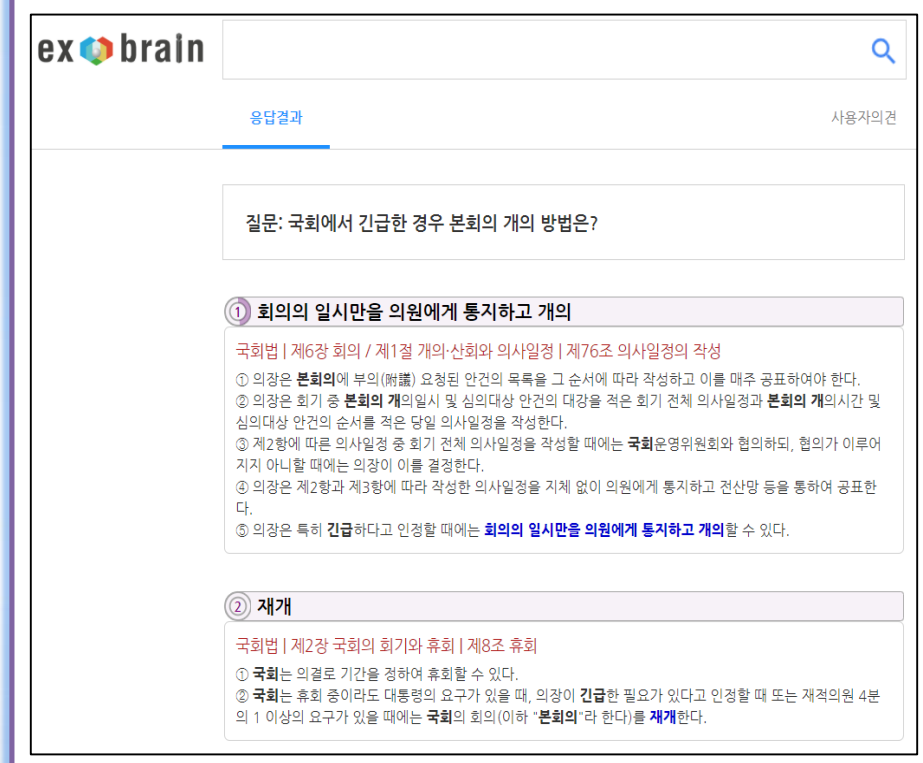
자연어로 기술된 문제의 의미를 이해하고, 정답을 추론하여 생성함

- 문제: 국회에서 긴급한 경우 본회의 개의 방법은?
- 정답: 회의의 일시만을 의원에게 통지하고 개의

정보검색: 단어 → 문서



질의응답: 질문 → 정답



Copyright© 2020 by ETRI

뉴럴-심볼릭 하이브리드 심층질의응답 시스템

문제유형 별로 뉴럴 또는 심볼릭 QA방법을 Planning하여 최적정답 생성

- 한컴오피스 2020에 지식검색으로 탑재됨 -

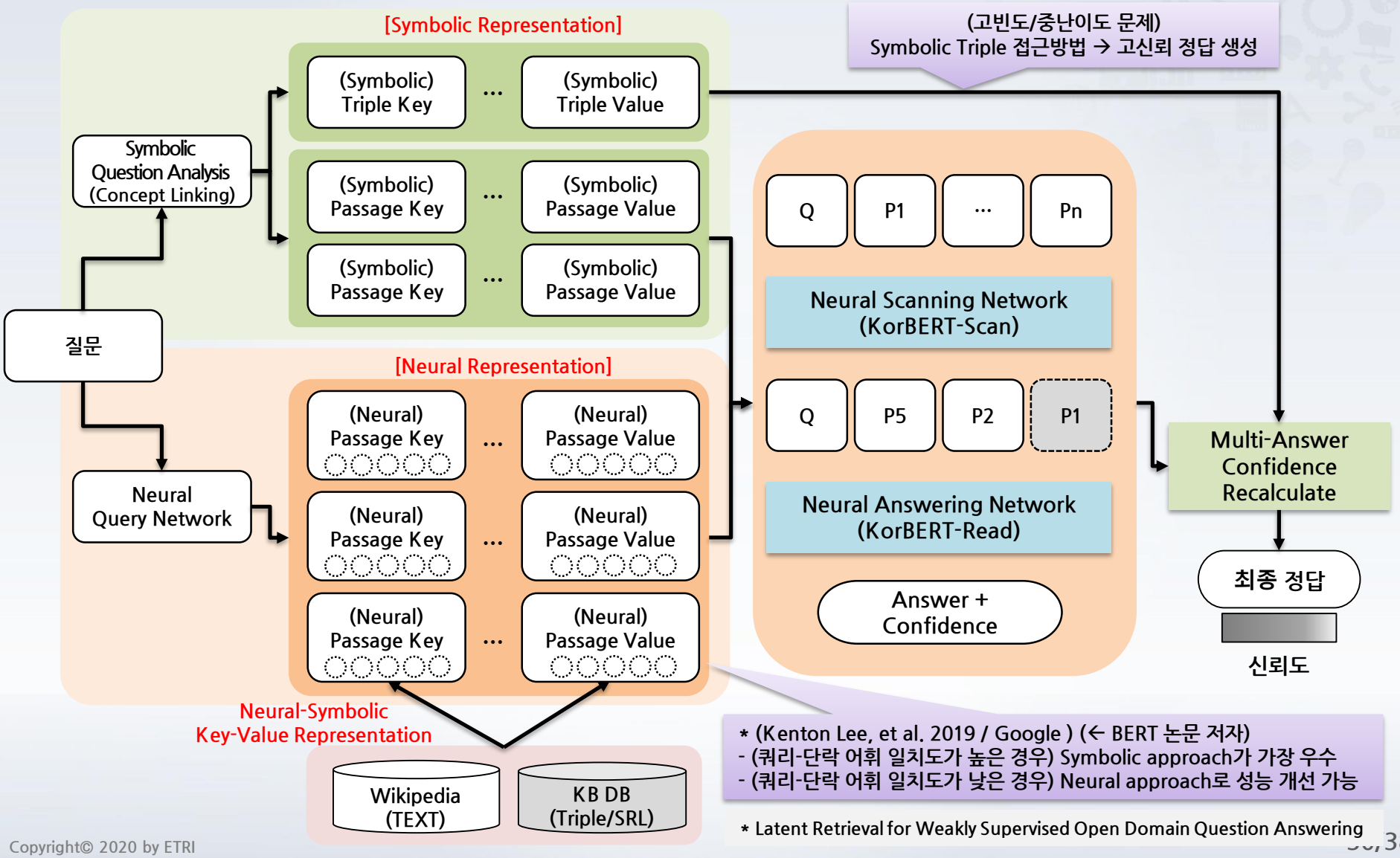
ex brain

질문	아프리카 돼지열병이 뭐야?	봉오동 전투에서 독립군의 피해는 어떻게 되나요	중국어 방은 무엇을 증명하기 위한 목적으로 제안되었나요?
질문 분석	정의형	관계형	독해형
뉴럴-심볼릭 하이브리드 QA (Cascading 방식)	**정의형 QA** 개체 인식 → 제약 인식 → 중의성 해소	**지식그래프 QA** 개체 인식 → 속성 인식 → 속성값 매칭 정답신뢰도 <= 임계치	**기계독해 QA** (KorBERT 기반 Scan&Read 모델) 정답 스캔 신경망 → 정답 독해 신경망 (KorBERT, 질문, 검색 단락 → 질문, 순위화된 단락)
		정답신뢰도 > 임계치	정답신뢰도 > 임계치
정답 및 근거	**정의형 정답** 〈정답〉 아프리카돼지열병은 아프리카돼지열병바이러스 감염으로 인한 돼지의 전염병이다. 〈출처〉 아프리카돼지열병은 아프리카돼지열병바이러스 감염으로 인한 돼지의 전염병이다. 돼지와 멧돼지의 질병으로, 사람에게는 감염되지 않는다.	**관계형 정답** 〈정답〉 전사 4명, 부상 2명 〈출처〉 (봉오동 전투) - (피해 규모) - (전사 4명, 부상 2명)	**독해형 정답** 〈정답〉 기계의 인공지능 여부를 판정할 수 없다는 것 〈출처〉 중국어 방 혹은 중국인 방(영어: Chinese room)은 존 설(John Searle)이 튜링 테스트로 기계의 인공지능 여부를 판정할 수 없다는 것을 논증하기 위해 고안한 사고실험이다.

Copyright© 2020 by ETRI

뉴럴-심볼릭 하이브리드 심층질의응답 시스템 구성도

뉴럴과 심볼릭 기법을 융합하여 단답형•서술형 정답을 추론 가능한 심층질의응답 시스템



* (Kenton Lee, et al. 2019 / Google) (← BERT 논문 저자)
- (쿼리-단락 어휘 일치도가 높은 경우) Symbolic approach가 가장 우수
- (쿼리-단락 어휘 일치도가 낮은 경우) Neural approach로 성능 개선 가능

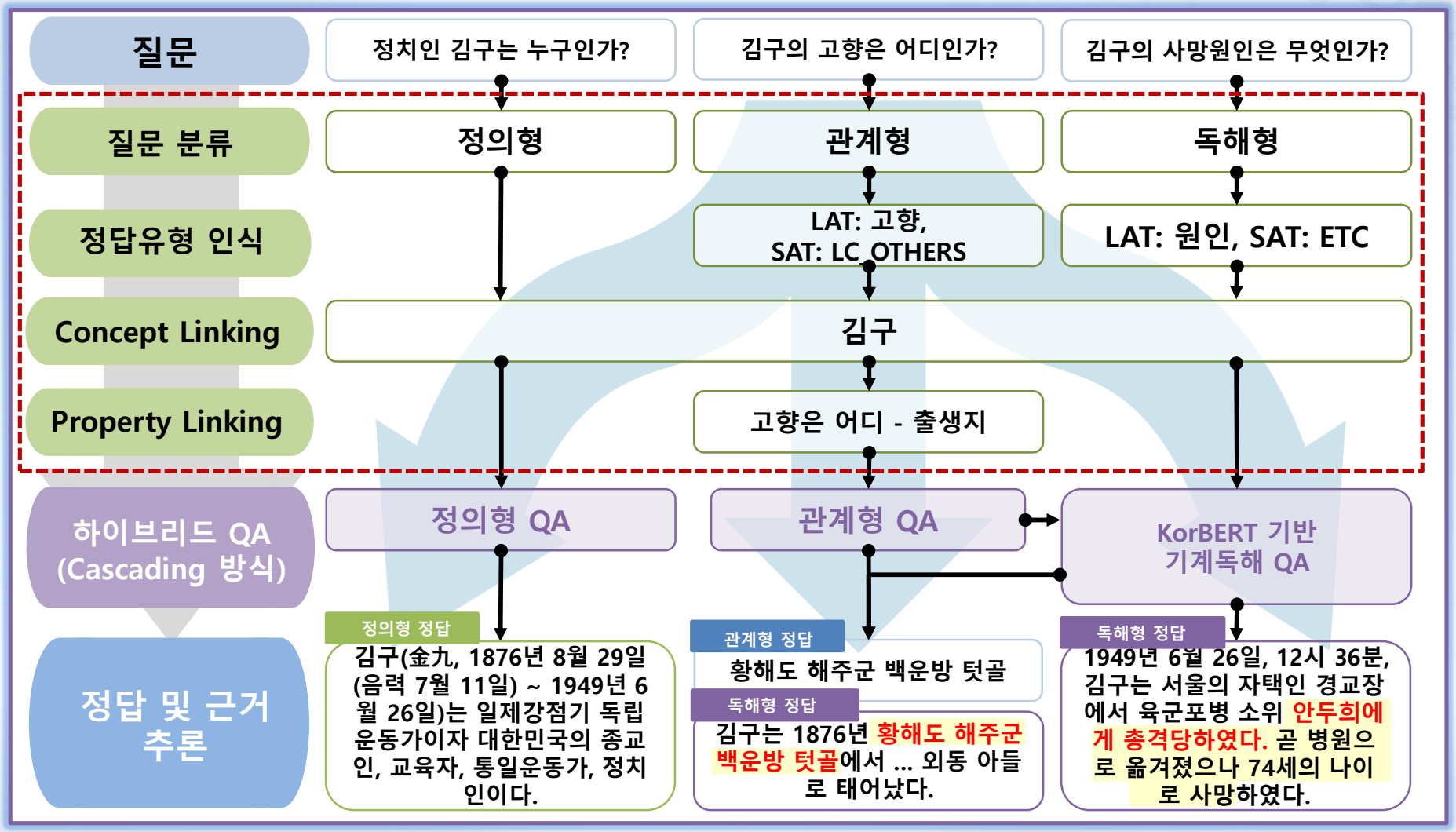
* Latent Retrieval for Weakly Supervised Open Domain Question Answering

Copyright© 2020 by ETRI

자연어 질문이해 기술

자연어 질문으로부터 질문 유형, 정답 유형, 핵심 개념어 등을 분석하는 기술

- 어휘정답유형(LAT): 89.94%, 의미정답유형(SAT): 85.34%, Concept Linking: 83.27% -



* LAT(Lexical Answer Type): 어휘정답유형, SAT(Semantic Answer Type): 의미정답유형

Copyright© 2020 by ETRI

Scan and Read 기계독해 모델

검색된 다수 단락으로부터 상용화 수준(속도/성능)의 정답 독해 기술 개발

1. Scan 모델 : 정답 추출 확률이 높은 순으로 단락 순위화, 정답 추출 확률이 낮은 단락 제거
2. Read모델 : 질문과 단락으로부터 단답형 또는 서술형 정답의 경계(시작위치~끝위치) 인식

* 구글도 BERT 모델을 적용하여 검색결과를 재순위화한 서비스 발표 ('19.10)

[KorBERT-Scan 모델]

1) 학습 목표
- P(정답존재여부 | 질문, 단락)

2) 학습 데이터
- 정답/오답 학습데이터 자동 생성

3) 결과 활용
- 정답 존재 확률이 높은 순으로 재순위화
- 정답 추출 확률이 낮은 단락 제거

4) 예시

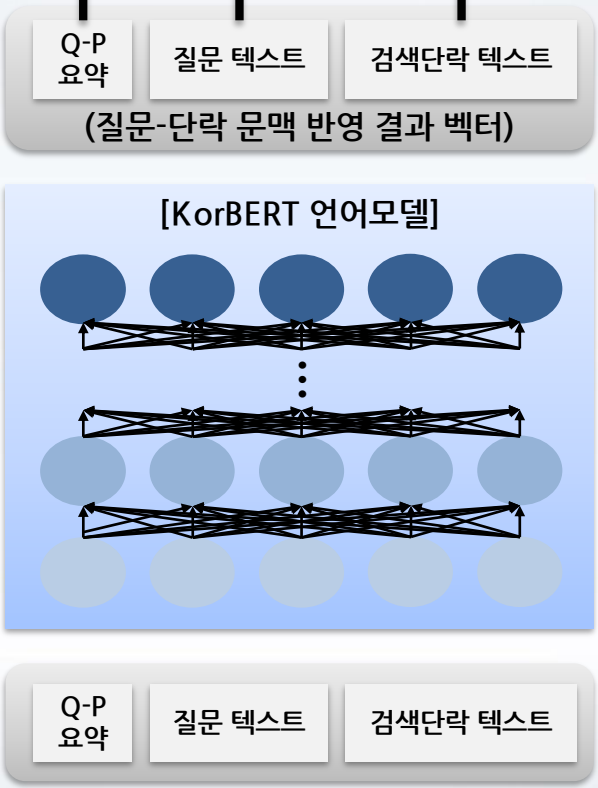
질문: 헌법상 대통령의 임기는 얼마인가?

검색단락#1: 대통령의 임기가 만료된 때에는…
(no-answer prediction: 제거)

검색단락#2: 원장은 대통령이 임명하고, 그 임기는 …

검색단락#3: 대통령의 임기연장 또는 중임변경을 …
(no-answer prediction: 제거)

검색단락#4: 대통령의 임기는 5년으로 하며, 중임 …



[KorBERT-Read 모델]

1) 학습 목표
- P(정답시작, 정답끝 | 질문, 단락)

2) 학습 데이터
- 질문 / 단락 / 정답경계 학습 데이터

3) 결과 활용
- 단락 내 올바른 정답 경계 인식

4) 예시

질문: 헌법상 대통령의 임기는 얼마인가?

<1순위 정답> 5년
근거: 대통령의 임기는 **5년**으로 하며, 중임할 수 없다.
출처: 대한민국헌법 | 제4장 정부 / 제1절 대통령 …

<2순위 정답> 4년
근거: … 그 임기는 **4년**으로 하며, …
출처: 대한민국헌법 | … 제4관 감사원…

Copyright© 2020 by ETRI

Scan and Read 기계독해 모델 성능

구글 보다 조기에 KorBERT를 적용한 검색 성능 개선 및 단답형•서술형 기계독해 개발

[Scan 모델] Top1 재현율 82.60% 달성 (검색 대비 +5.67%)

- 학습데이터 자동생성 기술을 이용한 839,381건 학습데이터 생성 (평가데이터 1,500건)
- 검색 단락 필터링 정확률 85.65% 달성 (검색 단락 필터링 비율 67.71%)

[Read 모델] Iterative Finetuning을 이용한 성능 개선 (+0.53% F1 개선)

- MRC 일반화 태스크 연구: 학습셋과 유사한 평가셋(예: KorQuad 학습 후, KorQuad 평가)이 아닌 임의의 질문/단락 사이 정답 인식 성능 개선 추진

[일반분야 KorBERT-Scan 모델 성능]

이진 재현율	Top1	Top2	Top3	Top4	Top5
시맨틱 검색	76.93%	81.93%	84.07%	85.33%	87.60%
구글 WordPiece 모델	67.67%	81.73%	86.67%	88.33%	89.67%
KorBERT WordPiece 모델	74.53%	84.47%	87.20%	88.67%	89.33%
KorBERT 형태소 모델	82.60% (+5.67%)	86.67%	88.47%	89.07%	89.60%

[일반분야 KorBERT-Read 모델 성능]

평가셋	특징	Exact Match	F1
KorQuAD-dev셋 (학습 60,406건, 평가 5,773건)	단답형 MRC 평가셋	88.11%	95.46%
KorQuAD-test셋 (평가 3,898건)	단답형 MRC 평가셋	87.76%	95.02%

Copyright© 2020 by ETRI

[참고] 구글, 검색서비스에 BERT 적용('19.10.25)

- 길이가 길고 문맥정보가 분명한 10% 정도 질의어에 대해 BERT 기반 재순위화로 검색 서비스 개선
  - FAQ: All about the BERT algorithm in Google search
    - https://searchengineland.com/faq-all-about-the-bert-algorithm-in-google-search-324193

Does Google use BERT to make sense of all searches?

No, not exactly. BERT will enhance Google's understanding of about one in 10 searches in English in the U.S.

"Particularly for longer, more conversational queries, or searches where prepositions like 'for' and 'to' matter a lot to the meaning, Search will be able t understand the context of the words in your query," Google wrote in its blog post.

However, not all queries are conversational or include prepositions. Branded searches and shorter phrases are just two examples of types of queries that may not require BERT's natural language processing.

Copyright© 2020 by ETRI

법령 심층질의응답 기술

법 조항에 관해서 단답형/서술형 질문에 정답을 제공하는 질의응답 기술

법령 질문 중 50%는 정답길이가 10글자 이상인 고난도 서술형 문제임

법령에는 전문용어와 한자가 많아 질문-조문 간 매칭이 어려움

(질문: 살인죄의 공소시효는?, 원문: (형사소송법 제253조의2) 사람을 살해한 범죄)

질문: 단말기유통법에서 이동통신서비스는 무슨 뜻이야

① 「전파법」에 따라 할당받은 주파수를 사용하는 기간통신역무

이동통신단말장치 유통구조 개선에 관한 법률 | | 제2조 정의

이 법에서 사용하는 용어의 뜻은 다음과 같다.
1. "**이동통신서비스**"란 **「전파법」에 따라 할당받은 주파수를 사용하는 기간통신역무**를 말한다.
2. "**이동통신**사업자"란 **이동통신서비스**를 제공하는 전기통신사업자를 말한다.
3. "이용자"란 **이동통신서비스**를 제공받기 위하여 **이동통신**사업자와 **이동통신서비스**의 이용에 관한 계약을 체결한 자를 말한다.
4. "**이동통신**단말장치"란 이용자가 **이동통신서비스**를 이용하기 위하여 필요한 단말장치를 말한다. …

<정의형 문제>

질문: 국회법상 예산결산특별위원회의 위원수는 몇 명인가요?

① 50명

국회법 | 제5장 교섭단체·위원회와 위원 | 제45조 예산결산특별위원회

① **예산**안, 기금운용계획안 및 **결산**(세입세출**결산**과 기금**결산**을 말한다. 이하 같다)을 심사하기 위하여 **예산결산특별위원회**를 둔다.
② **예산결산특별위원회**의 **위원** 수는 **50명**으로 한다. 이 경우 의장은 교섭단체 소속 의원 수의 비율과 상임**위원회 위원** 수의 비율에 따라 각 교섭단체 대표의원의 요청으로 **위원**을 선임한다.
③ **예산결산특별위원회 위원**의 임기는 1년으로 한다. 다만, 국회의원 총선거 후 처음 선임된 **위원**의 임기는 선임된 날부터 개시하여 의원의 임기 개시 후 1년이 되는 날까지로 하며, 보임되거나 개선된 **위원**의 임기는 전임자 임기의 남은 기간으로 한다.
④ **예산결산특별위원회**의 **위원**장은 **예산결산특별위원회**의 **위원** 중에서 임시의장 선거의 예에 준하여 본회의에서 선거한다.
⑤ **예산결산특별위원회**에 대해서는 제44조제2항 및 제3항을 적용하지 아니한다. …

<단답형 문제>

질문: 국회 휴회 중에는 언제 회의를 재개할 수 있나요?

① 대통령의 요구가 있을 때, 의장이 긴급한 필요가 있다고 인정할...

국회법 | 제2장 국회의 회기와 휴회 | 제8조 휴회

① **국회**는 의결로 기간을 정하여 **휴회**할 수 있다.
② **국회**는 **휴회 중**이라도 **대통령의 요구가 있을 때, 의장이 긴급한 필요가 있다고 인정할 때 또는 재적의원 4분의 1 이상의 요구가 있을 때**에는 **국회**의 **회의**(이하 "본**회의**"라 한다)를 **재개**한다.

<서술형 문제>

질문: 국회에서 산회가 선포되면 다시 회의를 시작할 수 있나요?

① 산회를 선포한 당일에는 회의를 다시 개의할 수 없다

국회법 | 제6장 회의 / 제1절 개의·산회와 의사일정 | 제74조 산회

① 의사일정에 올린 안건의 의사가 끝났을 때에는 의장은 **산회**를 **선포**한다.
② **산회를 선포한 당일에는 회의를 다시 개의할 수 없다**. 다만, 내우외환, 천재지변 또는 중대한 재정·경제상의 위기, 국가의 안위에 관계되는 중대한 교전 상태나 전시·사변 또는 이에 준하는 국가비상사태로서 의장이 각 교섭단체 대표의원과 합의한 경우에는 그러하지 아니하다.

<서술형 문제>

Copyright© 2020 by ETRI

자연어 질의응답 기술동향

- 단락 대상 단답형 정답 추출에서 다수단락/문서 대상 고난도 독해 시작 단계

챌린지	정답 탐색 범위	성능
Stanford SQuAD 2.0	단락	Model / EM / F1 Human Performance (Stanford University, Rajpurkar & Jia et al. '18): 86.831, 89.452 ALBERT + DAAF + Verifier (ensemble) (PINGAN Omni-Sinitic): 90.002, 92.425
Google Natural Questions	문서 (웹 검색로그에서 질문 선정)	F1: 0.61302 Precision: 0.67612 Recall: 0.56069
CMU/Stanford/ Google HotpotQA	위키피디아 전체 (Explainable Multi-hop QA)	Ans: EM **60.04**, F_1 **72.96** Sup: EM 49.08, F_1 76.41 Joint: EM 35.35, F_1 **61.18**

- Google NQ와 HopotQA 문제를 복합한 문제의 해결이 필요

구분	SQuAD 2.0		Google NQ		HotpotQA	
	Train Set	Dev. Set	Train Set	Dev. Set	Train Set	Dev. Set
총 질문 개수	130,319	11,873	10,000	1,000	90,447	7,405
평균 질문 길이 (어휘토큰 개수)	12.13	12.31	10.00	10.01	22.44	19.59
질문 당 평균 단락 개수	1	1	39.89	38.18	9.95	9.94
평균 단락 길이 (어휘토큰 개수)	151.62	163.54	197.35	198.19	118.78	123.11
질문과 정답단락의 어휘토큰 매칭률(%)	61.12	63.21	56.77	54.51	53.36	58.46

Copyright© 2020 by ETRI

엑소브레인 오픈 API 서비스 및 데이터 보급

- 중소·벤처 기업, 학교, 개인 개발자 등의 다양한 응용 개발 촉진과 관련 산업의 생태계 활성화 및 인력 양성 지원 목적
  - http://www.etri.re.kr/aiopen

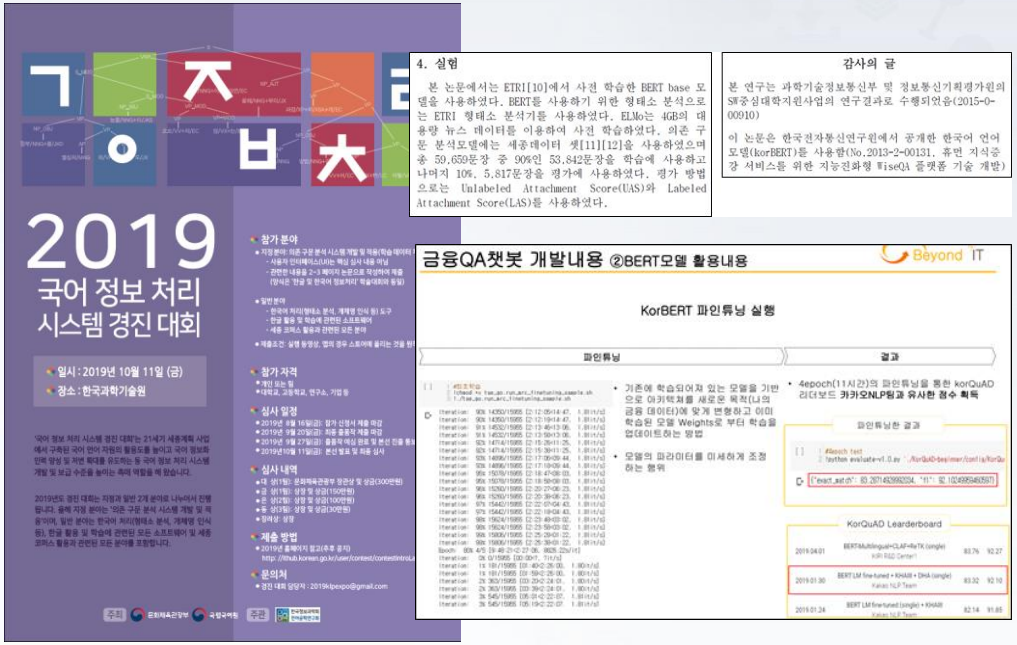
1,130개 기관
(삼성, SKT, 네이버 등)

1천 9백만건
(일 평균 2만5천건)

<엑소브레인 오픈API 사용통계>

365회
('19.6.10~8.31)

<한국어 BERT 언어모델 사용통계>



<국립국어원 주관 국어 정보처리 시스템 경진대회>

- KorBERT 활용하여 서강대는 의존구문분석 분야 대상 수상
- KB금융지주는 일반분야 수상

Copyright© 2020 by ETRI

Future of NLP

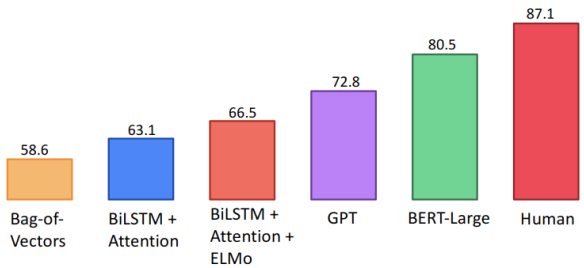
- 방향성
  - 대용량 학습데이터 필요 → 비지도 학습 기반 언어모델
    - Language Model as a Knowledge Base
  - 도메인 확장의 어려움 → 전이학습 기반 자연어 처리 기술
  - 신뢰할 만한 결과 제공 → 판단결과와 근거를 제공가능한 QA 개발

  - 장기적으로 비지도 학습 기술로 발전
    - OpenAI GPT-3 175B Parameters

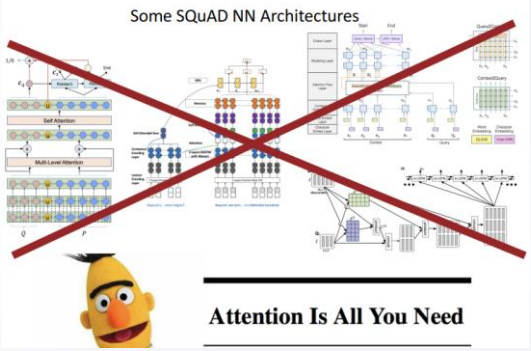
Training Huge Models

Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B
Honey Bee Brain	~1B synapses

GLUE Benchmark Results



The Death of Architecture Engineering?



<출처: C. Manning, Stanford University>

Copyright© 2020 by ETRI