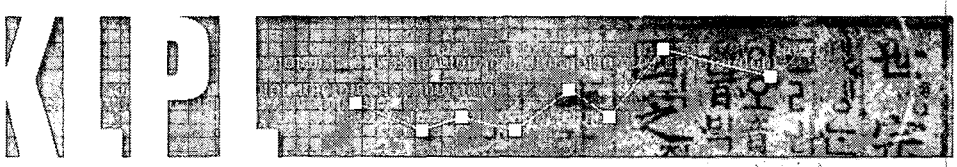
인터넷환경에서 언어정보처리 기술의 응용과 발전 방향



● 부산대학교 전자전기정보컴퓨터공학부

권혁철

정보과부하

인터넷을 통한 정보 유통혁명이 가져온 정보 과부하 해결

세계화에 따른 언어장벽 극복
- 기계번역, 다국어정보검색

정보제어의 필요성
- 인터넷을 통한 매춘, 포르노와 유언비어, 스팸메일 유포 등 역작용 극복
- 그러면서 사생활과 정보유통 자유 보장

새로운 시장 탄생

전자상거래, eCRM, 인터넷을 활용한 교육, 지식검색, KMS, Call-Center

기계번역, 자동통역

- 유럽, 일본, 중국

가전제품처럼 쉬운 대화형 인터페이스

- 유비쿼터스, 텔레매틱스, 로봇제어

시맨틱웹

정보 과부하, 정보제어 불가능의 해결

- 시맨틱웹(Semantic Web): 내용에 기반을 둔 웹으로의 진화
- 영역지식의 온톨로지화와 이를 통한 추론
- 사람이 기계가 알 수 있는 방법으로 정보에 의미를 붙이는 것은 어렵고 비용이 많이 듦

▦ 이해에 기반한 언어처리

정보과부하의 해결 방법은 언어처리 고도화에 있음

이해에 기반한 언어처리만이 기계번역, 자동통역, 문서요약, 문서분류 등 문제 해결

정보 생성-유통-해석-통합의 지원

언제 가능한가?

그러면 그 동안은?

▦ 개발기술의 요구사항

◘ 언어 중심으로 인터넷 정보처리가 가능할 것

◘ 다양한 응용시스템에서 상업적으로 활용할 수 있을 것

◘ 장기간 지속적으로 성능개선이 가능할 것

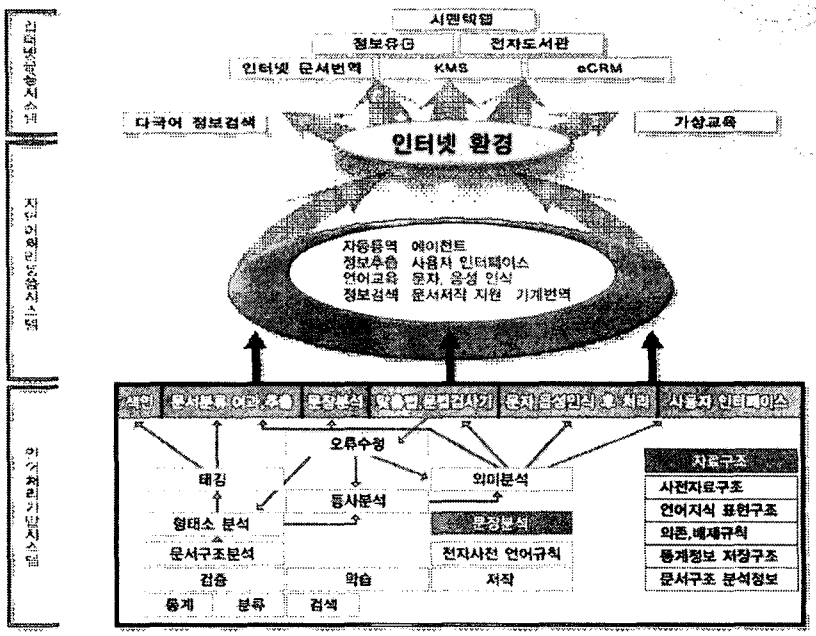
단계적 지능화(Scalable)

언어처리기술의 응용환경 변화

내용	지금까지	지금부터	원인/예
문서의 질	깨끗함	지저분함 (띄어쓰기오류, 구어체, 미등록어, 오용어)	문서의 종류에 제한 없음 (홈페이지, 채팅)
대상문서 양	소량 (매뉴얼 등)	초대용량 (수천만~수십억 페이지의 문서)	인터넷 일반화, eCRM
정보 밀도	높음	매우 낮음	대용량, 무제한 문서
사용자 요구	관련 자료 모두	필요한 것만, 요약하여	정보 과부하
응용분야	기계 번역, 정보 검색	정보 통합, 요약, 분류, 추출, 인터넷 데이터마이닝, 의미기반 번역, 대화형 명령	언어 처리 응용 분야의 확대, 기술수요 확대
대상 문서/영역	문어체/제한적	구어체/무제한적	인터넷, 전자출원, 지식검색, 질의/응답
문장	완전	불완전, 비문법적	일반인이 정보 제공자, 음성언어
구성요소	문자 위주	문자, 테이블, 멀티미디어 자료 통합	인터넷 문서, 정보의 디지털화

언어처리 연구사항

연구내용과 결과의 활용



최근의 변화

국제화(globalization)와 인터넷 활성화에 따른 언어처리 기술의 수요증가

- 2003년 EU는 공식문서 130만 쪽 번역, 통역·번역 전문가는 4천명, 인건비 10억 유로
- 인터넷 활성화에 따라 eCRM, KMS가 일반화하고, 정보과부하를 해결하기 위해 텍스트마이닝 기술이 널리 쓰이며, Call Center 자동화 필요성 증대

한국어 맞춤법·문법 검사기에 대한 사용자 요구의 변화

- 실수에 의한 불필요 오류 교정 요구[전자신문에서 맞춤법/문법검사기 도입 과정]
- 기계번역(특허 홈페이지 등), 문장분석, 의미분석에서 자동으로 원문 오류를 교정해주는 시스템 개발을 요구 [ETRI 등 기계번역 시스템 연구팀의 요구]

차이가 큰 언어 간 기계번역의 성공 및 특허문서 한/영 기계번역 시스템 요구

- 일본 특허청(http://www.jpdl.inpit.go.jp)이 80% 이상 정확도의 일·영 특허 번역기 개발 (한·영 기계번역 정확도: 40%)
- 한·중·일 3국의 특허공동인정을 위한 전단계로 한·영 특허번역 시스템 개발을 ETRI와 특허청이 공동으로 시도
- 번역률 향상을 위해서는 특허문서의 맞춤법·문법 오류를 특허문서 작성과정이나 작성 후 (자동)교정

인터넷의 정보과부하 해결을 위한 언어이해기술의 중요성 증대

- 최근 의미에 기반을 둔 정보접근을 위해 인터넷 자원을 온톨로지에 기초하여 태깅하는 시맨틱웹 연구가 활발
- 모든 인터넷 자원을 인간이 의미로 태깅하기는 불가능하므로 이해를 바탕으로 문장을 분석하여 지식을 추출하는 기법의 중요성이 커짐

문서에 있는 오류의 자동교정과 불필요하거나 중복된 내용 제거 기술의 필요성 증가

- 인터넷 텍스트마이닝 과정에서 원문서의 오류 및 중복 또는 불필요한 정보 제거 필요
- 기계번역, 언어이해 시스템 구축, 음성·문자인식의 전처리와 후처리에서도 오류교정이 자동으로 이루어져야 함

규칙과 통계 [1/3]

규칙에 의한 접근

장점
- 특수한 부분에 대한 조건까지 제시해 줄 수 있으므로 중의성을 해결하는 성능이 우수함

단점
- 구축하고 유지 관리하는데 부담이 큼
  - 시간과 비용이 많이 들며 규칙의 제어가 어려움
  - 규칙 관리자의 능력이 중요하므로 전문가가 필요

띄어쓰기 모델
- 휴리스틱(heuristics) 사용
  - 단어 정보나 특정 통사적 패턴 및 음절 정보를 띄어쓰기 단서로 이용
- viable prefix를 이용한 최장 일치 기법 사용
  - 어절에 대한 최장의 유효한 앞부분을 찾아서 형태소 분석이 되는지 검사
- '접두 명사' 및 접두사와 이웃하는 명사 간의 조합 규칙 이용
- 형태소 분석 결과 이용
  - 형태소 분석정보를 바탕으로 가중치 별로 띄어 쓸 위치를 정하고 주로 분석해야 할 형태소 범위 설정

규칙과 통계 [2/3]

통계에 의한 접근

장점
- 단순하고 계산적으로도 부담이 적음
- 언어의 생산성(language productivity)에 잘 대처할 수 있음
- 영역지식의 쉬운 활용

단점
- 자료 부족 문제 (대용량 말뭉치 필요)
- 통계 정보를 추출한 말뭉치 분야에 의존적
  - 유사 분야에 대해 우수한 성능을 보이지만 분야에 따라서는 적용이 힘듦

규칙과 통계 [3/3]

통계에 의한 접근

띄어쓰기 모델

- 어절 n-gram 통계, 음절 n-gram 및 형태소 n-gram 등의 통계 정보를 띄어쓰기 모델의 매개 변수로 사용 (n 값이 클수록 메모리를 많이 차지하지만 성능이 더 우수한 것은 아님)

$$\arg\max_{S} \prod_{k=1}^{n} P(W_k) \frac{P_{inner5}(\text{LS of } W_k, \text{FS of } W_{k+1})}{1 - P_{inner5}(\text{LS of } W_k, \text{FS of } W_{k+1})}$$

If $k = n$, then $p(\text{LS of } W_k, \text{FS of } W_{k+1}) = 0.5$.
W_k = k^{th} word; FS = First syllable;
LS = Last syllable

[통계 모델 간 메모리 사용량 및 정확도 비교]

	전체 메모리		어절 단위 띄어쓰기 정확도
	총 어절	중복제거 어절	
음절 bigram	593MB	4.1MB	71.22%
음절 bigram + 음절 trigram	593MB + 256MB = 849MB	4.1MB + 25.1MB = 29.2MB	93.36%
음절 trigram	773MB	63.7MB	93.06%

(테스트 데이터: ETRI 품사 태깅 말뭉치)

비교 예 [1/4]

규칙으로 잘 처리할 수 있는 문제

경매에 붙이는 (x)
해서 시간을 때우려 생각해낸 (x)

시부모를 모시는걸 효도라 하는데 (x)
총리감에 부족함이 없는 인물일수 있다. (x)
그이상의 (x)
않았을까 하는 생각을 가져본다. (x)
이제는 나의 사연이 되버렸다. (x)
많은 사람들이 (x)

[문장부호 오류]
사람입니다 그래서 ⇒ 사람입니다. 그래서
공간이다. 그렇다면 최소한 ⇒ 공간이다. 그렇다면, 최소한
군.검찰 ⇒ 군검찰

비교 예 [2/4]

통계적 기법으로 잘 처리할 수 있는 문제

(1)

[원문장]	아버지가방에들어가신다.
[맞춤법 검사기 결과: 규칙]	아버지가방에 들어가신다.
[통계정보를 이용한 띄어쓰기 결과]	아버지가 방에 들어가신다.

(2)

[원문장]

그렇지만질병을대하는의사들의행위가역설적인성격을띠게되는까닭은될수있으면질병을객관화시키고거리를두어아무런지식의척도도없는텅빈공간속에서질병의윤곽이잡히도록해야하기때문이다.

[규칙/통계정보를 이용한 띄어쓰기 결과]

그렇지만 질병을 대하는 의사들의 행위가 역설적인 성격을 띠게 되는 까닭은 될 수 있으면 질병을 객관화시키고 거리를 두어 아무런 지식의 척도도 없는 텅 빈 공간 속에서 질병의 윤곽이 잡히도록 해야 하기 때문이다.

비교 예 [3/4]

규칙과 통계정보를 이용한 띄어쓰기 비교

(1)

[원문장]

보통사람들과차이가없는데여기서개거품무는넘들은도대체뭔가

[맞춤법 검사기 결과: 규칙]

보통사람들과 차이가 없는데 여기서 게거품을 무는 놈들은 도대체 뭔가

[통계정보를 이용한 띄어쓰기 결과]

보통 사람들과 차이가 없는데 여기서 개 거품 무는 넘들은 도대체 뭔가

(2)

[원문장]

그들은더이상학생이아니다학생이란그자체도가면이다

[맞춤법 검사기 결과: 규칙]

그들은 더 이상학생이 아니다 학생이란 그 자체도 가면이다

[통계정보를 이용한 띄어쓰기 결과]

그들은 더 이상 학생이 아니다 학생이란 그 자체도 가면이다

바교 예 [4/4]

아직 해결하지 못하는 문제

모자라서 그럼검니다 기본도 (△)
비행기 자리가 마련되는대로 갈 것 한국으로 (△)
공통 분모 - 공통분모 (△)
일반 가전제품은 한번 구입해 보니
10만원씨 받으며
학 교에 가서
우리 돈 퍼다 백인
경매에 붙는 (?)
얼마 지나지 않아 더 산 가격에 경매 시장에 다시 나오고 있는 실정이다.
밥 먹어 주나?
그 이 상의 , 그 이 상 의
부강 한 나라

단기적 접근

적용분야	
영역의 제한	특허문서번역
문장의 제한	Controlled vocabulary
응용분야의 제한	통계적 접근, 온톨로지

요소기술	
Hidden Markov Model	Tagging, 음성인식, 어의 중의성 제거
kNN, SVM	분류, 필터링, 요약
규칙	문장분석, 의미분석, 개념추출

※ 미래의 접근방법

이해에 기반한 언어처리

규칙과 통계의 결합

WordNet(일반목적), 대용량 전문용어사전

대용량 의미처리용 언어지식베이스

대용량 말뭉치 및 효과적 학습기법

인공지능기술(지식표현기술)

상황인식기술