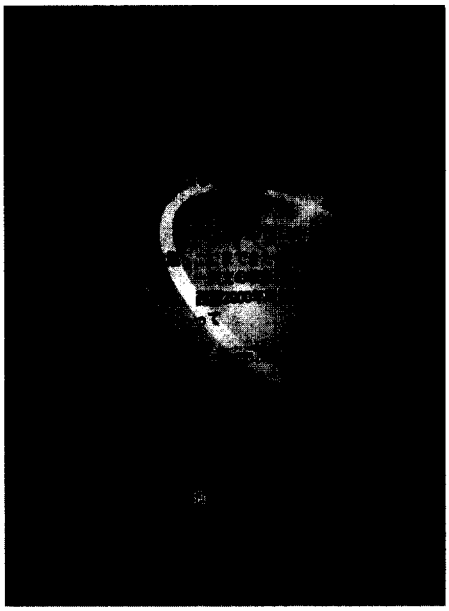
□IPSJ 요약□

자연언어처리 현황 및 서기 2000년 문제 회고

- IPSJ Magazine, Vol. 41, No. 7, July 2000 -

수원대학교 이명원*



일본 정보처리학회지 2000년 7월호에서는 특집으로 「여기까지 온 자연언어처리 - 예문의 수집과 이용」과 소특집으로 「서기 2000년 문제를 회고하며」를 중심으로 기술하고 있다. 전체 내용을 목차, 특집 요약, 소특집 요약, 기타 내용으로 요약한다.

1. 목차

1) 특집: 여기까지 온 자연언어처리 - 예문의 수집과 이용

- 편집에 즈음하여: Yoich Tomiura(Kyushu U.) and Hideo Watanabe(IBM Research)
- 예문을 사용하여 문장을 해석합시다: Kentaro Inui(Kyushu Institute of Technology) and Kiyoaki Shirai (Toyko Institute of Technology)
- 코퍼스(Corpus)가 먼저인가, 파서(Parser)가 먼저인가: Sadao Kurohashi (Kyoto Univ.)
- 거대한 코퍼스를 공유합시다: Hozumi Tanaka(Tokyo Institute of Technology), Shin-ichiro Kamei(NEC Corp.), Minoru Moriguchi(Sharp Corp.) and Yasuhiko Kato(The National Language Research Institute)
- 어떻게 하면 데이터 공유를 성공시킬 수 있는가 - 음성인식 분야에서의 사례: Katunobu Itou(ETL), Tatsuya Kawahara (Kyoto Univ.) and Kazuya Takeda (Nagoya Univ.)
- 언어 코퍼스를 보다 유효하게 사용하기 위해서: Takehito Utsuro(Toyohashi Univ. of Technology)
- 코퍼스를 기반으로 한 언어처리의 한계와 전망: Yuji Matsumoto(Nara Institute of Science and Technology) and Takenobu Tokunaga(Tokyo Institute of Technology)

* 종신회원, Email:mwlee@mail.suwon.ac.kr

2) 연재

- 방송과 정보처리 - 통신위성을 이용한 광대역 인터넷 서비스: Yoshihiko Tanaka and Hiromi Komatsu(NTT Satellite Communications Inc.)
- 인터랙티브 에세이 - 이대로 괜찮은가? 일본의 수퍼컴퓨터 / 힘낼테다 일본의 수퍼컴퓨터 / 앞으로는 클러스터로 충분하지 않을까 - 소프트웨어 하우스의 독백 / 일본의 수퍼컴퓨터, 나쁜 것은 자네가 아니야 / 그건 역시 대역폭: Taisuke Boku(Univ. of Tsukuba), Tadashi Watanabe(NEC), Satoshi Matsuoka(Tokyo Institute of Technology), Satoshi Sekiguchi(Electrotechnical Lab.) and Taisuke Boku (Univ. of Tsukuba)

3) 소특집: 서기 2000년 문제를 회고하며

- 서기 2000년 문제를 회고하며 - 소특집의 편집에 즈음하여: Mikio Aoyama (Niigata Institute of Technology) and Tatsuya Omata(Japan Information Service Industry Association)
- 일본에 있어서 서기 2000년 문제 대책을 회고하며: Kaoru Nakamura(Government of Japan)
- 정보서비스 산업에 있어서 서기 2000년 문제 대응을 회고하며: Ryuichi Kono (INTEC Inc.)
- 서기 2000년 문제의 법률적 측면을 회고하며: Kazuno Otani(The Japan Research Institute, Ltd.)
- 유저 기업으로부터 본 서기 2000년 문제: Akira Harada(The Mitsubishi Trust and Banking Corp.)
- 전기통신 분야에 있어서 서기 2000년 문제 대응을 회고하며: Yasuki Hayashi(NTT Communicationware Corp.)

4) 해설

- 인증기술의 현재와 미래: Hitoshi Sakano and Itsukazu Nakamura(NTT DATA Corp.)
- 멀티미디어 홈 컴퓨팅의 미래 - 제1회 가정정보화의 주역은 무엇인가: Takahiko Kamae(Labs. of Image Science and Technology)
- 차세대 인터넷 연구개발의 최전선 1. 애드혹 네트워크(Adhoc Networks) 구축 기술 - WIDE 합숙 네트워크를 예로 해서: Shoko Mikawa(Keio Univ.), Hiroshi Takada(NEC Corp.) and Kenjiro Cho (Sony Computer Science Labs., Inc.)
- 동영상의 실시간 포맷 변환 - 모바일 멀티미디어 통신의 용도를 넓히는 동영상 처리 기술: Tsutomu Uenoyama(Matsushita Electric Industrial Co., Ltd.) and Takeshi Yukitake(Matsushita Communication Industrial Co., Ltd.)
- LEDA: 복잡한 알고리즘도 간단하게 프로그램화할 수 있는 마법의 도구: Tetsuo Asano, Koji Obokata(Japan Advanced Institute of Science and Technology) and Kurt Mehlhorn(Max Planck Institute fur Informatik)
- 설계자가 원하는 설계지원시스템의 개발: Yotaro Hatamura and Masayuki Nakao(The Univ. of Tokyo)

5) 칼럼

- 정보기술의 신시대를 향해 - 텔레콤 문화와 컴퓨터 문화의 충돌: Yutaka Matsushita (Keio Univ.)
- 진정한 인터넷을 향해 - 라이프 라인(Life Line)으로서의 인터넷: Masataka Ohta (Tokyo Institute of Technology)
- 난세(亂世)의 액세스 네트워크: 라인 공유의 실현으로 본격화하는 ADSL에 의한 고속 데이터 통신: Toru Murase(Sumimoto Electric Industries, Ltd.)

- SE의 지혜 주머니 - I Love You(컴퓨터 바이러스)에 커다란 함정: Minoru Senoh (Nagoya Univ. of Commerce and Business Administration) and Tetumasa Shingai(Nikko Shoji Co., Ltd.)
- 미국 인터넷 사정 - 네트워크 주식의 급락 속에서도 신장하는 벤처기업: Toru Maegawa(Waseda Univ.)
- 현대·컴퓨터 시장 - 일본이 부활하는 Non-PC 시대: Norio Shishido(Tera Media Inc.)

6) 길잡이
- XML의 이것저것: Naohiko Uramoto (IBM Research)

7) 회의보고서
- NETWORLD+INTEROP 2000 Tokyo

8) 핫타임
- 기업에 SI 부문은 필요한가(상)
- 확대되는 선택형 서비스
- IT 산업에 있어서의 연구 - 커뮤니티의 변용과 전망

2. 특집 요약

본 특집에서는 최근의 언어 코퍼스(Corpus)를 이용한 자연언어 처리 기술의 발전과 앞으로의 가능성을 소개한다.

급속한 컴퓨터 하드웨어의 진보와 언어 코퍼스(예문집)을 기반으로 한 방법의 도입으로 1990년대에 들어서 자연언어 처리 기술이 큰 진전을 보이고 있다. 자연언어 처리 기술을 이용한 응용으로는 기계번역, 텍스트 요약, 검색 등 인터넷 상의 대량의 정보 중에서 필요한 정보를 획득하기 위한 것이 많다. 이러한 응용의 기반 기술은 구문해석 기술이며, 번역이나 요약의 질, 검색의 적중율 등의 정확도가 구문해석 시스템의 정확도에 크게 의존한다.

구문해석이란 간단하게 말하면 술부(述部)는 무엇이며 그 목적은 무엇인가와 같이 문장의 구조를 해석하는 것이다. 일반적으로 하나의 문장에 대해 복수의 구문 구조의 후보가 얻어지며, 의미적으로 이상한 구문 구조도 다수 포함된다. 예를 들어 「빨간 이탈리아의 자동차」나 「빨간 모자의 어린이」와 같이 〈형용사〉〈명사〉「의」〈명사〉의 형을 갖는 명사구를 생각해보자. 『연체형의 형용사는 명사를 수식한다』라는 단순한 문법 규칙에 따르면 이 명사구의 선두에 있는 형용사는 어느쪽 명사에도 관련될 수 있다. 그러나, 의미를 생각하면 「빨간 이탈리아의 자동차」에서는 「빨간」이 「자동차」를 수식하고, 「빨간 모자의 어린이」에서는 「빨간」이 「모자」를 수식하는 것이 맞다. 따라서, 어떻게 해서 맞는 구문구조를 선택하는가(구문 구조의 애매함을 해소하는가)가 구문 해석에 있어서의 큰 문제이다.

종래의 구문해석 시스템은 경험 등을 기반으로 애매함을 해소하는 규칙을, 사람의 손으로 준비하여 구축하였기 때문에 규칙 작성자의 언어적 센스에 강하게 의존하게 되어서 보수나 확장성의 관점에서 문제가 있었다. 이에 비해 최근 언어 코퍼스를 이용하여 학습한, 통계적 언어 모델 기반의 애매함 해소법이나 언어 코퍼스 중에 유사 사용예를 기반으로 한 애매함 해소법 등의 코퍼스 기반의 방법이 한창 연구되고 있다. 이러한 코퍼스 기반의 방법은 종래 방식에서 개발되었던 것과 거의 대등한 정확도를 얻고 있다.

제1편에서는 구문해석의 중요성, 구문구조의 애매함을 해소하기 위한 고전적인 방법과 한계에 대해 기술하고, 언어 코퍼스를 이용한 구문해석의 개요를 해설한다.

제2편에서는 양질의 코퍼스를 작성하는 문제에 관해서 교또(京都) 대학의 코퍼스 작성을 통해 얻은 식견을 들어본다. 코퍼스에 부여하는 구문구조는 정확하고 일관성이 있어야 한다. 이것을 전부 사람 손에 의존하는 것은 곤란하며, 구문해석기(파서)를 이용하는 것이 유효하다. 한편 정확도가 높은 파서 개발에 있어서는 대규모의 구문구조를 포함한 코퍼스가 필요하게 된

다. 이러한 「닭이 먼저냐 계란이 먼저냐」와 같은 문제에 대한 한 가지 해결법을 제공한다.

제3편에서는 언어자원의 공유화 동향에 대해 해설한다. 언어 코퍼스의 개발은 방대한 노력을 필요로 하기 때문에 대규모의, 동시에 양질의 언어 코퍼스를 해결하는 방법으로 다양한 연구기관에서 개발한 언어 코퍼스를 공유하는 것이 생각되고 있다.

제4편에서는 자연언어 처리와 밀접한 관계에 있고 코퍼스를 이용한다는 점에서 긴 역사를 갖고 있는 음성인식 분야의 상황을 소개한다. 충분한 성능을 갖는 시스템을 개발하는데 필요한 음성 코퍼스, 텍스트 코퍼스의 양과 질에 대해 고찰하고 코퍼스를 정비하기 위한 취급 방법에 대해 기술한다.

제5편에서는 제1편의 내용을 깊이 파고들어서, 한정된 코퍼스를 유효하게 이용하는 방법(모델에 도입하는 정보의 선택 방법이나 파라미터의 추정 방법 등)에 대해 해설한다.

제6편에서는 본 특집의 결론으로서 코퍼스 기반의 언어처리의 배경, 적용 범위, 현상에 대해 개관하고 코퍼스 기반의 언어 처리의 이해득실과 평가에 있어서의 문제점에 대해 기술하고 앞으로의 전망에 대해 논의한다.

3. 소특집 요약

본 소특집은 국가, 정보산업, 컴퓨터 유저 등의 입장에서 Y2K 문제에 대처해 왔던 여러 분들에 의한 경험과 교훈을 정리한 내용이다.

Y2K 문제에 대한 대응은 대규모 컴퓨터 시스템을 기업 활동의 핵심으로 개발 및 이용하고 있는 통신, 항공운수, 금융 등의 분야에서는 일본에서도 1995년경부터 착수되었다. Y2K 문제 발생에 대한 정부의 통계(원본에 나와 있슴)에서 보면 어느 정도의 건수는 발생했으나 외부에 영향을 미친 문제는 적었다고 할 수 있다. 특히, 사회 전체에 심각한 영향을 미친 문제는 없었다.

Y2K 문제는 기술면에서는 소프트웨어 보수(保守)의 하나이다. 그러나, 문제가 사회 전체에 퍼지고, 대상으로 하는 소프트웨어 규모의 크기와, 기한이 변경될 수 없다는 점에서 종래에 없었던 어려움이 있었다. 그 때문에, Y2K 문제 대응은 기술면에서 다음과 같은 영향을 가져왔다. (1) 보수에서 진화로: Y2K 문제에 의해 소프트웨어 수명의 길이가 인식되게 되었다. 특히, 소프트웨어는 개발과 가동 후에, 유저의 니즈나 환경의 변화에 따라 본질적으로 진화해야 하는 것이 인식되었다. 일본의 「소프트웨어 발전」이나 미국의 EDCS(Evolutionary Design of Complex Software) 등의 연구 프로젝트의 출현은 소프트웨어 공학의 틀 중에서 이러한 문제가 「보수」의 문제로부터 「진화」의 문제로 취급되도록 하였다는 것을 의미한다. (2) 시스템의 네트워크화: 인터넷 상에서 복수개 기업간의 응용의 연휴가 확산되고, 이와 함께 네트워크를 통한 기업간의 위험도 확산되는 문제가 제기되었다. 이에 따라 시스템을 네트워크를 통한 전체로서 취급할 필요성이 명확해졌다. (3) 프로그램 해석기술의 진화: Y2K 문제에 대응하는 기술로서 프로그램의 영향파급 해석(Impact analysis), 의존성 해석(Dependency analysis) 등의 해석기술과 지원 도구가 개발되었다. 특히 대규모 기간업무시스템에서는 사람에 의한 해석을 지원하기 위해 이러한 프로그램 해석 기술이 이용되었다. (4) 시스템 재구축의 진전: Y2K 문제를 계기로 시스템을 재구성한 유저도 적지 않았다. 소프트웨어 자산 분석(Asset analysis)으로 붕괴시킨 뒤 객체 지향에 의해 재개발하는 실천 경험이 많아졌다. 중소규모의 시스템에서는 오피스 컴퓨터로부터 PC 등을 이용하는 클라이언트/서버 시스템으로 이행되었다. 대규모의 기간시스템에서는 재개발이 곤란한 경우나 효과가 비용에 맞지 않는 경우가 있다. 그래서, 표준 인터페이스를 통해서 기존 시스템을 이용할 수 있도록 하는 기술, 인터넷 상으로 전개하는 기술, 응용내의 컴포넌트나 비즈니스 도구를 추출하는 어플리케이션 마이닝 등의 새로운 기술이 생성되었다.

Y2K 문제 중에는 사회면에서 볼 때 그 영향의 중요성이 지적되고 있다. Y2K문제는 현대사회가 「정보기술」이라는 신기술이 가져온 위험

을 인식하고 학습한 다음과 같은 프로세스였다. (1) 정보기술이 경영문제로: 기업경영에 있어서 정보기술의 중요성에 대한 인식이 높아졌다. 또한, 비즈니스의 글로벌화에 의해 국제적으로 국가로서의 대응이 문의되었다. Y2K를 기회로 해서 기업의 정보처리시스템 전체를 리엔지니어링한 기업도 있었다. 이러한 기업들에게는 정보기술이 기업 경영의 전략적 과제라는 인식이 높았기 때문이다. (2) 정보기술의 설명책임이 문의되다: 이제까지 정보기술에 대한 일반 사회의 이해는 「잘 모른다」「블랙박스」이었다. 이것은 1999년말 소위 「Y2K 문제 전문가」라고 칭하는 사람들에 의해 기술적으로 근거가 없는 풍설의 선전을 조장하였다. Y2K 문제는 정보기술이 사회의 기반이 되어 있다는 것을 일반 사람들에게 인식시키는 계기가 되었다. 특히, PC뿐 아니라 광범위한 기계에 정보기술이 침투되어 있다는 것이 인식되었다.

Y2K 문제의 경험은 기술, 법률, 사회의 각 방면에서 앞으로 활용되어야 할 것이다. 기술면에서는 소프트웨어 개발, 보수 및 진화의 실천 기술의 향상에 있다. 앞으로, 정보시스템은 법 개정이나 비즈니스의 변화로의 대응이 끊임없이 요구될 것이므로, Y2K 문제는 일시적인 문제가 아니라 소프트웨어 개발, 보수의 기본 과제라고 말할 수 있다. Y2K 문제 대응에서는 하나의 바위와 같은 설계때문에 변경에 따르는 광범위한 파급현상과, 설계 표준의 결핍으로 해석이나 변경이 곤란하였다. 실제의 소프트웨어 설계에 있어서는 정보은폐 등의 기본적인 설계기술을 적용해서 독립성이 높은 모듈 구조로 하는 등, 보수 및 진화에 용이한 시스템을 구현해야 한다.

그리고, 연구와 실천의 차이를 없애는 노력이 필요하다. 예를들어 COBOL이나 Visual Basic은 널리 이용되고는 있으나 연구 대상으로서는 취급되는 일이 적은 언어로 지적되고 있다. 이러한 점이 Y2K 문제에 있어서 연구 집단의 참가를 좁혔다라고도 말할 수 있다. 법률면에서 보면, 소프트웨어 개발, 구입, 이용에 있어서 계약의 명확성과 계약을 기본으로 하는 거래 관행이 바람직하다. 사회면에서는 사회 전체로 정보기술에 대한 이해촉진을 촉구해야 한다. 일반 사람들에게 정보기술에 대한 이해를 촉구하는 일은 하루아침에 이루어질 수 있는 일은 아니다. 앞으로 학회를 비롯한 업계, 단체 및 대학은 일반 사람들이 정보기술을 올바르게 이해할 수 있도록 하는 활동을 해야한다고 생각한다. 예를 들어 일본자동차공업회의 웹페이지에는 「자동차와 환경」「자동차와 안전」「초등학생을 위한 알기 쉬운 자동차백과」 등 일반 사람들을 위한 정보가 공개되고 있다.

4. 기타 내용

특집 이외의 부분에서 다루고 있는 주제로는 다음 두 가지 내용을 소개하고자 한다.

해설 「인증기술의 현재와 미래」에서는 「인증한다는 것이 무엇인가」「전자인증 방식」「물리세계와의 접점 - 바이오메트릭스」「얼굴 화상의 인식 기술」및 「인증기술의 장래」에 대해 기술하고 있다. 네트워크 사회, 전자화 사회에 있어서 신청 혹은 상거래 등을 안전하게 할 수 있게 하는 지동인증 기술을 개략적으로 소개하고 있다. 장래의 네트워크 상의 인증기술로는, 국가 CA (Certification Autority) 혹은 관청을 연결하는 브릿지 CA를 루트로 한 계층 CA가 이용되고, 물리적인 세계에서는 복수의 바이오메트릭스가 적재적소에 이용되는 형태가 가능성 높은 미래의 모습으로 보고 있다.

현재, 텔레비전 방송, DVD, 텔레비전 전화 등의 많은 용도로 디지틀 영상이 이용되고 있다. 동영상 통신이 가능한 모바일 단말기도 곧 실용화될 예정이다. 이음새 없는 네트워크로의 발전으로 이러한 영상정보를 서로 이용하고자 하는 요구도 커지고 있다. 이를 실현하기 위해서는 동영상의 압축 포맷을 서로다른 포맷으로 변환하는 포맷 변환이 필요하게 된다. 해설 「동영상의 실시간 포맷 변환」에서는 모바일 텔레비전 전화 단말기를 겨냥한 기술을 중심으로 동영상 압축 포맷과 변환 기술에 대해 해설한다.