언어처리 표준화와 평가의 중요성: MATEC99 와 그 이후

박재득
한국전자통신연구원 지식정보연구부
jdpark@etri.re.kr

The Importance of Standardization and Evalutation of Natural Language Processing: MATEC99 and Beyond

Jay Duke Park, Knowledge Information Department,
Electronics and Telecommunications Research Institute

1.서론

언어처리의 연구가 오랫동안 계속되었음에도 많은 사람들로부터 사랑을 받는 소위 히트상품이라고 불리는 자연언어처리 응용제품이라는 것이 있는지가 의문이다. 없다면 그 원인은 어디에 있을까? 초기에 언어처리 연구에 뛰어드는 사람들은 대부분 언어처리 기술 자체가 물리학에서의 아인슈타인처럼 몇몇 천재적 영감으로 훌륭한 이론이 발견되고, 이 이론에 입각한 프로그램을 개발되므로써 조만간 해결될 수 있으리라는 원대한 비젼을 품고 시작하였다. 이러한 매력적인 꿈이 혼자만의 힘으로 조만간 실현될 것이라는 희망은 점점 퇴색되고 있으며, 멋있는 이론적 작업보다는 이보다 훨씬 중요하고 많은 노력이 필요한 방대한 양의 데이터 구축 및 정련작업이 필요한 3D 분야로 점점 변모하고 있는 것 같다.

이와같이 자연언어처리가 매우 어렵고 다루기 까다로운 분야라는 인식이 최근에 일반화되기 시작하여 언어처리의 특수성을 모두들 공감하고 있다. 즉, 언어처리는 다른 일반적인 소프트웨어들과는 달리 훌륭한 알고리즘의 개발도 중요하지만 언어데이타 또는 언어지식이 훨씬 더 중요하고 구축도 훨씬 힘들다는 사실이다.

그리하여, 언어처리 전문가들은 대용량의 언어자료가 필요하다고 말하고 있지만 직접 구축하거나 구하는 것이 쉽지가 않다. 그리고, 원시 말뭉치만 가지고 어떤 통계적 처리를 할 때는 비용은 적게 들지만 획득되거나 처리된 정보의 품질을 보장할 수 없기 때문에, 대부분 태그가 부착된 말뭉치를 사용하고자 한다. 이러한 자료는 구축에 많은 비용과 노력이 소요되며 일정한 품질을 유지하기가 매우 힘들다.

이러한 제약적 상황에서, 각자가 산발적이고 영세적으로 각기 다른 철학과 방식으로 이러한 자료를 구축한다면 호환성과 공유가 어렵기 때문에 당분간은 자료 빈곤(data sparseness)의 문제를 치유하기 힘들 것이므로, end-user 를 감동시키는 성능을 가진 제품이 탄생하는 것을 기다리기 힘들 것이며, 이러한 상태에서의 기술수준은 소위 '도토리 키재기'식의 상대적 우월성밖에 보여주지 못할 것이다.

2. 자연언어처리 기술 표준화의 중요성

이러한 상황인식하에서 우리가 갈망하는 자연언어처리 분야의 히트상품을 탄생시키기 위해서, 경쟁 이전에 협조하여 공유와 호환을 통하여 가용한 자원과 기반기술의 양과 폭과 깊이를 늘이고, 이를 바탕으로 각자 응용제품을 개발하여 히트시키는 것으로 경쟁을 해야할 것이다. 즉, 협동적 경쟁(copetition)이 이 분야 전문가들의 상생과 공존을 위해서 불가피한 선택이 되어버렸다고 해도 과언이 아니라고 생각한다.

잘 아시다시피, 이러한 협동적 경쟁과 공유와 호환을 도모하기 위한 공통의 프로토콜을 도출하는 것이 곧 표준화이다. 이러한 측면에서, 언어처리 표준화는 반드시 수행되어야 할 중요한 과제의 하나라고 할 수 있다.

그러나, 표준화가 각 시스템의 특수성이나 특장점을 살리지 못하는 단점도 있을 수 있다. 그래도, 전반적으로 언어처리 분야의 성능 향상에 기여할 것이 기대된다면 이러한 단점을 보완하면서 반드시 추진해야 할 것이다.

표준화는 공유의 공통분모인 태그부착 말뭉치와 이의 구축을 위한 원시말뭉치 수집 및 구축, 태그 세트, 태깅방식의 표준화부터 시작된다고 할 수 있다. 언어데이타라고 하지만, 이 품사부착 말뭉치는 많은 의사결정과정이 필요하므로, 그 자체가 프로그램 로직의 일부분이라고 해도 과언이 아닐 정도이며, 단순히 관찰되거나 수집된 자료와는 차원이 다른 것이다.

3. 표준화와 평가의 관계 및 중요성

공유를 위한 표준화를 위해서는 표준안의 도출과정에 각 연구자들의 대안 제시에 경쟁이 있을 수 있다. 이 대안의 경쟁에서 우열을 가리고, 또한 나중에 공유기술로 각자 개발한 응용 제품의 객관적 성능 비교를 하기 위해서는 성능 평가의 기준과 방법이 마련되어 있어야 한다. 이러한 관점에서 표준화와 평가는 밀접한 관계를 갖고 있으며, 평가방법에 있어서도 표준화가 필요하다.

평가에 있어서 표준화는 평가에 참여하는 참가자들에게 공정하고 합리적인 게임의 룰을 도출하는 것이다. 이 룰은 평가대상자들의 장단점을 반영하는데에 있어서 객관적이고 공정하며 합리적인 방식으로 합의를 거쳐 도출되어야 하는 것이다.

이러한 인식하에 외국에서는 이미 TREC, MUC, SUMMAC 등의 평가대회가 매년 개최되고 있어(Wilks 1999), 우리는 다소 늦은 감이 있지만, 그래도 영어, 독어 등 이외의 언어처리에 비해서 늦지는 않은 것 같다.

음성 인식기술의 평가대회의 주요한 성과는 매년 전반적으로 괄목할 음성인식 성능의 향상을 가져왔다는 것이다. 이렇듯 평가는 단순히 우열만 가리자는 경쟁뿐 아니라, 동분야의 기술의 전반적 향상을 도모하는 상생의 결과를 가져다 준다.

이러한 평가의 또 다른 부수적인 효과로 다음과 같은 것을 들 수 있을 것이다.

- 언어처리 기술의 현재 수준의 전반적 파악
- 수준 및 현황 파악에 바탕한 향후 연구방향 설정
- 소비자에 대한 구매 정보로써 제품의 가격대비 성능 및 장단점 비교 자료 제시
- Q마크와 같은 언어처리 소프트웨어의 품질 인증제도 도입

4. MATEC99의 성격

이러한 필요성에 따라서 그동안 진행되어오던 자연어 정보처리 기술 표준안(ETRI,1999)을 바탕으로 품사 부착 말뭉치를 구축하고 이를 배포하였고 형태소 분석기 및 품사태거와 명사추출기 등의 성능의 평가하기 위한 평가기준과 방법도 도출하였다. 그리고, 1999년 국내 처음으로 MATEC99라는 평가대회를 ETRI 주최로 여러 참가팀의 적극적 참여에 힘입어 개최할 수 있었다.

그러나, 이 1회 대회는 평가방법이나 평가를 위한 준비상태가 미흡하지만 더 이상 미룰 수 없다는 판단하에서, 시행착오를 통하여 구체적 문제점을 발견하고 개선한다는 생각을 가지고 개최하였다. 그 대신에 절대적 또는 상대적 평가 결과를 공개하지 않는다는 원칙을 가지고 많은 참가팀들의 호응을 얻을 수 있었다. 향후 있을 본격적인 평가대회를 위한 연습경기의 성격으로서 평가에 있어서 개선해야될 구체적인 문제점을 많이 발견하고 일부 개선방안도 고려할 수 있었던 것이 무엇보다도 중요한 성과라고 볼

수 있다.

평가는 그 목적에 따라 적합성 평가, 오류진단 평가, 성능 평가 등의 3가지로 크게 나눌 수 있다고 한다(Cole 1996). 현재의 MATEC99와 같은 평가대회의 주목적은 각 시스템의 성능향상에 기여하자는 것과 성능의 전반적 수준을 파악하고자 있으므로 오류진단 및 성능평가에 주 목적이 있다고도 할 수 하겠다. 향후에는 평가기술도 진보함에 따라 절대적 우열을 가리는 대회로 변화할 것으로 전망한다.

5. MATEC 그 이후

외국에서는 평가기술은 언어처리 기술 못지않게 중요한 논문연구의 대상으로 간주되고 있다(Cole, 1999). 한국어의 자연언어처리 기술의 발전 및 평가 기술의 발전을 위해서도 이러한 평가에 관한 논문활동이 국내에서도 많이 전개되기를 희망한다.

MATEC99에서는 품사태거 및 형태소 분석기 평가를 먼저 시작하였지만, 약간 늦은 감이 없지 않다. 그리고, 이러한 기술들은 아직 개선이 여지는 있지만 어느 정도 성숙 단계에 들었다고 판단한다. 그래서, 다음에는 다른 레벨의 기반 기술 및 응용 기술의 평가로 범위를 넓혀, 구문-의미 분석기, 담화분석기, 기계번역, 정보검색, 대화, 분류, 요약 등과 같은 응용 시스템의 평가로 나아가서 표준화와 평가가 기술의 발전을 선도하도록 해야할 것이다.

그런데, 성능 평가에 지나치게 많은 노력을 경주하는 것은, 좋은 점수를 얻는 것이 좋은 연구를 하는 것보다 중요하게 되는 위험-회피 위주의 전략의 남용을 초래하는 위험이 있다. "연구가 생명력을 유지하려면, 반드시 평가는 위험을 감수하는 모험적인 연구에 대해 어떤 식으로든 보상을 해줄 필요가 있다(Cole, 1999)"는 말을 염두에 둘 필요가 있다.

그리고, 평가 주최측에서는 평가에 참여하여 표준안에 맞추기 위하여 기존 시스템의 방식의 변경등에 필요한 소모적인 작업의 양을 최소화할 수 있도록 참가자들을 지원하여야 한다. 또한 대회 참가자들의 보다 열띤 참여를 유도하기 위해서는 괄목할 정도의 상금이나 인센티브를 내걸 필요도 있다고 본다. 아울러, 평가대회 참가자들은 각자의 시스템이나 기술의 특장점을 평가에 잘 반영시킬 수 있도록 평가대회 이전에 평가의 결과에 영향을 미치는 표준안에 자신들의 대안을 반영시키려는 노력을 경주하여야 할 것이다.

이러한 공동의 노력이 결실을 맺어 이 분야의 히트상품을 창출시켜, 한국어처리 분야의 눈부신 발전을 이루어 이 분야연구자들의 긍지와 자부심을 함양하는 견인차가 되었으면 하는 바람이다.

참고문헌

[1] R.A.Cole, J. Mariani, et al., 1996, Survey of the State of the Art in Human Language Technology,

[2] Y. Wilks, "Book Review on Karen Sparck Jones' Evaluating Natural Language Processing Systems: An Analysis and Review", 1999, V.107, pp.165-170, Artificial Intelligence.

[3] ETRI, 1999, 품사 부착 말뭉치 구축 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.

[4] ETRI, 1999, 전자사전 표제어 선정 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.

[5] ETRI, 1999, 품사 부착 말뭉치 구축을 위한 품사 태그 세트 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.