

사용자에 의한 문형 확장 방식을 이용한 다국어 채팅 시스템의 설계 및 구현

The Design and Implementation of Multilingual Chatting System

Using Expansion of Sentence Patterns By User.

(주)두레소프트. 박홍원 (waits123@hitel.net)

DooReSoft Corporation. Hongwon, Park

[논문요약]

본 연구에서는 정해진 주제내에서 높은 번역율과 빠른 처리 시간을 동시에 수용할 수 있는 효과적인 다국어 채팅 시스템을 구현하기 위해 사용자가 어절 단위로 단어를 입력하거나 선택하여 이미 구축되어 있는 문형에 접근하도록 유도하는 사용자 문형확장 방식을 제안하였다.

사용자 문형확장 방식을 사용하여 다국어 채팅 시스템을 구현할 경우 사용자 입력과 동시에 구문분석, 변환, 생성등 일련의 번역과정을 최소한의 처리시간으로 처리할 수 있으므로 매우 용이하게 실시간 번역 시스템을 구현할 수 있다는 장점이 있다.

사용자 문형확장 방식과 더불어 이와 함께 사용될 수 있는 통합 문형코드와 통합 품사체계도 제안하였다. 이는 번역의 대상이 되는 한국어, 영어, 일본어 각각에 대해 문형코드와 품사코드를 따로 설정하지 않고 통일된 하나의 코드체계를 적용함으로써 기계번역에서의 변환과정을 최소화하기 위해 고안하였다.

1. 서론

1-1. 연구의 배경 및 필요성

인터넷과 게임 수요가 급속히 증가함에 따라 많은 게임 개발자들이 온라인 게임 개발에 열을 올리고 있다. 온라인 게임의 특성상 게임 중에 외국인 게이머와 채팅을 하게 되거나 특정 주제에 대해 대화를 나누어야 할 때가 많다. 특히 최근에 개발되고 있는 온라인 게임의 경우 대화 그 자체가 게임을 진행해 나가는 데 중요한 역할을 하는 경우도 있다. 그러나, 영어나 일본어에 익숙하지 않은 게이머라면 게임을 즐기기가 어려운 것이 사실이다.

한편, 기존의 기계번역 시스템을 온라인 게임에 접목시켜 번역 기능을 지원하는 다국어 채팅 시스템을 구현할 수도 있으나 구문분석, 사전검색, 역문생성 등에 적지 않은 시간이 소요되어 온라인 상에서의 채팅 시스템에서 사용되기에는 부적합한 것으로 보인다.

온라인 게임에서 사용될 다국어 채팅 시스템을 구현하기 위해서는 처리시간의 최소화, 정해진 주제내에서의 높은 번역율, 빠른 처리 시간 등의 기본조건을 충족시켜야만 한다. 그러나, 기존의 기계번역 방식에서 사용하고 있는 구문분석 방식을 온라인 채팅에서 사용할 경우 처리시간이 많이 소요되고 구문분석에 실패할 경우 부분적인 문장번역조차 이루어지지 않아 정상적인 번역 기능을 수행하기 어려운 면이 있다.

이에 본 연구에서는 문형정보와 문형패턴 등에 대한 연구를 다국어 채팅 시스템에 접목시켜 대화에 참여하고 있는 사용자(게이머)가 스스로 문장의 구성요소를 선택하게 하여 미리 정의된 문형코드에 접근하도록 유도하는 사용자에 의한 문형 확장 방식을 제안하였다. [1][2][3] 또한, 이 방식을 사용하여 한국어 - 영어 - 일본어의 상호 번역이 가능한 <두레 다국어 채팅 시스템(이하 DMCS)>을 구현하였다. 참고로 이 시스템은 현재 (주)두레소프트에서 개발하고 있는 다국어 온라인 미팅 시뮬레이션 게임 <밀레니엄 미팅>에서 사용되고 있다.

1-2. 연구 방법과 범위

본 논문은 (주)두레소프트에서 개발중인 다국어 기반의 실시간 미팅 시뮬레이션 게임인 <밀레니엄 미팅>에 사용되고 있는 다국어 채팅 시스템에 대한 연구를 요약·정리한 것으로 주로 설계 부분과 구현 원리에 대해 다루었다.

<밀레니엄 미팅> 다국어 버전의 발표를 앞두고 현재 문형 추가작업을 계속하고 있다. 최대 처리가능한 어절은 12어절이며, 문형은 기본문형과 확장문형을 포함하는 1수준 문형이 170개, 전치사구와 부사구를 포함하는 2수준 문형이 80개이며, 등록되어 있는 표제어는 영어 2000여개, 일본어 3400여개, 한국어 2700여개에 이른다. (99년9월 현재)

한편, 본 연구에서 제안하는 다국어 채팅 시스템(DMCS)은 범용적으로 사용될 수 있는 기계번역 시스템이나 대화번역 시스템과는 달리 몇가지 제약 사항이 있다. 첫째, 특정 분야(온라인 게임)에서만 사용될 수 있다. 둘째, 영어, 한국어, 일본어에 대해서만 가능하며 다른 언어가 추가될 경우 설계상의 수정이 필요할 수 있다. 셋째, 대화문은 최대 12어절(영어기준)까지만 지원한다. 넷째, 사용자의 입력(선택)이 필수적이므로 대화문 입력시간이 길어질 수 있다.

2. 온라인 게임을 위한 DMCS의 설계

2-1. DMCS 구현상의 애로점

온라인 게임은 게임내에서의 사건 해결이나 게이머의 협력을 유도하기 위해 서로간의 대화(채팅)를 유도한다. 이때 게이머들이 사용하는 대화는 일반적으로 회화체이고, 게임내에서 약어와 은어를 사용하며, 이들 대화가 서로 다른 언어를 사용하는 게이머들간에 '번역'이라는 과정을 거친 후 자신의 모국어로 각각의 모니터에 표시되어야 한다는 문제점을 가진다. 이러한 구현상의 애로점에 대해 문제점과 해결방안을 정리하면 <표1>과 같다.

	DMCS에서의 처리 방안
회화체(생략 등)	회화체 문형도 코드설계시 문형정보에 등록시켜 코드화
약어·은어 사용	별도의 DB를 통해 일반 표제어와 동일한 방식으로 처리
빠른 번역 처리	사용자 문형확장 방식 채용. 역문생성까지 처리 시간 최소화

<표1> 온라인 게임에서의 DMCS 구현상의 애로점

2-2. DMCS 구현에서의 고려사항

DMCS에서 반드시 고려되어야 할 3가지 요소는 ①자유로운 대화(불특정한 문장에 대한 처리), ②높은 번역율, ③빠른 처리시간이라고 할 수 있을 것이다. 일반적인 기계번역 시스템의 경우 ①②는 매우 중요시되는데 반해 ③은 비교적 덜 중요시된다. 한편, DMCS의 경우는 특정 주제내에 있으므로 ②③에 비해 게임내에서 충분히 대화하기에 지장이 없는 수준의 대화라면 오히려 ①의 중요성이 떨어질 수 있다. 반면, ②③은 매우 중요하다. 특히 ③의 경우 처리시간이 길어지면 채팅 시스템으로서 가치가 떨어진다. 이를 정리하면 <표2>와 같다.

	기계번역 시스템	다국어 채팅 시스템
①대화의 자유도	매우 중요함	별로 중요치 않음
②높은 번역율	매우 중요함	비교적 중요함
③빠른 처리시간	별로 중요치 않음	매우 중요함

<표2> DMCS 구현에서의 고려사항별 중요도 비교 분석

이상에서 살펴본 바와 같이 DMCS에서는 게임내에서의 특정 대화를 전제로 하고 있기 때문에 대화의 자유도 보다는 높은 번역율과 빠른 처리시간을 요하고 있다. 따라서, 기존의 기계번역 시스템에서 사용하는 [형태소분석-구문분석-변환-생성-후처리] 등의 복잡한 처리 과정으로 인한 시간 손실을 초래할 수 없다.[4]

3. 사용자 문형확장 방식

3-1. 사용자 문형확장 방식의 기본개념

사용자 문형확장 방식은 빠른 처리시간과 높은 번역율을 동시에 유지해야 하는 DMCS를 위해 고안된 방식으로 기존의 기계번역 시스템에서 볼 수 있는 구문분석, 변환, 생성등 일련의 번역 과정을 포함하고 있다.

이는 기존의 연구에서 제안된 격틀을 이용한 번역이나 예문을 이용한 번역시스템과 번역 방식에 있어서는 유사한 부분이 있다.[5][6] 그러나 대상어의 처리범위와 개수, 사용자 입력시 해당문형을 참조하여 사용자로부터 문형을 선택할 수 있게 한다는 점, 생성시간을 최소화하기 위해 다양한 기법을 사용하였다는 점에서는 차이를 보인다.

이 방식은 대화에서 주로 사용되는 문형을 정의하여 DB로 코드화시켜 놓고 사용자의 입력문장을 어절단위로 입력받아 코드화된 문형에 매칭되도록 유도하여 문형을 결정한 후 코드화된 상대언어의 문형을 참조하여 역문을 생성하는 방식이다.

즉, 사용자 스스로 자신이 입력하고자 하는 문장의 문형을 미리 구축된 문형정보의 범위내에서 확장해가며 최종문형을 결정하게 되므로 기계번역 시스템에서의 구문분석의 과정을 대체한다고 볼 수 있다. 한편, 문형이 결정되면 상대언어의 문형은 이미 준비된 코드에 의해 자동으로 참조되어 역문을 생성하게 되므로 변환과 생성의 과정을 사실상 대체하는 역할을 한다. <표3>에서 상세히 정리하였다.

	DMCS에서의 해당부분이나 처리방안
사전검색	필요. 주로 게임내에서 사용되는 단어 수록.
형태소분석	여타의 기계번역 시스템과 동일.
구문분석	사용자 문형확장 방식에 의한 문형선택.
변환	선택된 문형이 자동으로 참조.
생성	참조된 문형코드가 이미 역문생성정보를 포함하고 있는 상태이므로 별도의 처리 불필요.
후처리	코드화시키는 문형은 완전한 문장이므로 더 이상의 처리는 불필요함.

<표3> 기계번역 시스템의 처리과정과 비교 분석

3-2. 통합 문형코드

3-2-1. 통합 문형코드의 기본개념

통합 문형코드는 사용자 문형확장 방식을 구성하는 가장 핵심적인 요소이다. 이는 영어, 한국어, 일본어에 대해 각각 DB로 구축하지 않고 하나의 코드체계로 통합하여 DB한 것이다. DMCS는 이러한 통합된 코드체계에 의해 기존의 기계번역 시스템에서 사용하고 있는 구문분석, 변환, 생성을 사실상 대체하고 있다.

DMCS에서 번역 대상어가 되는 3개국어에 대한 문형 모두를 하나의 통합된 문형코드로 설정하여 DB화시킨 것은 구문분석, 변환, 생성의 과정을 하나의 코드내에 통합시켜 처리시간을 최소화하기 위한 것이며 통합 문형코드를 사용함으로써 DMCS는 사용자의 입력과 동시에 역문을 출력할 수 있게 된다.

3-2-2. 통합 문형코드의 체계

통합 품사코드는 모두 8자리로 구성된다. 1번째 자리는 원문의 언어, 2번째 자리는 문장종류, 3번째 자리는 영어의 문장형식, 4번째 자리는 한국어의 문장형식, 5번째 자리는 일본어의 문장형식, 6번째부터 8번째 자리는 구문번호로 구성되어 있다.

따라서, 문형이 1개 추가될 경우 상대언어가 되는 나머지 2개국어의 문형도 추가해 주어야 한다. 만약 EQ222101이라는 새로운 문형을 추가할 경우 KQ222101, JQ222101도 추가해 주어야 한다. 즉, 언어 구분코드인 1번째 자리만 다른 코드 3개가 항상 동시에 추가되어야만 정상적으로 번역기능을 수행할 수 있다.

3-3. 통합 품사코드

3-3-1. 통합 품사코드의 기본개념

통합 품사코드는 통합 문형코드를 사용하기 위해 필수적으로 사용되어야 하는 통합 품사체계이다. DMCS에서의 통합 품사체계는 문장내에서의 어휘의 기능성에 초점을 둔 것으로 통사적으로 동일하거나 비슷한 자질을 가지는 품사에 대해서는 1개의 품사코드를 사용하도록 하였다.[7][8]

3-3-2. 통합 품사코드의 체계

통합 품사코드는 10개의 품사코드만을 사용하는 DMCS의 최적화된 품사코드체계이다. 통합 품사코드체계는 기본적으로 번역 대상어인 영어, 한국어, 일본어의 품사코드를 최소화시킨다는데 그 의의를 두고 있으므로 각각의 언어가 가지는 언어학적인 특수성은 상당부분 무시되었다.

영어의 경우 대명사의 소유격을 관형사(한국어)와 연체사(일본어)와 같은 통사적 자질을 가진 것으로 보고 동일한 코드를 부여하였으며, 소유대명사도 명사와 동일한 코드를 부여하였다. 한편, 자질이 비교적 다른 전치사와 접속사에 대해서는 별도의 코드를 부여하였다.

한국어의 경우 대명사, 수사(수사)는 사실상 명사와 통사적으로 별다른 차이점이 없으므로 명사로 통합시켜 처리하였다.

일본어의 경우 “조동사”를 “활용에 의한 분류”에 따라 분류하고 활용형에 기준하여 형용사, 형용동사, 동사로 구분하여 일반의 용언과 동일하게 취급하였다. 즉, 조동사를 별

도의 품사로 처리하지 않고 입력시 사용자에게 형용사, 동사, 형용동사의 활용형으로 선택하는 방법을 사용하도록 하였다. 한편, 형용동사에 대해서도 형용사와 동일한 코드를 사용하도록 하였다.

코드	영어	한국어	일본어
A	형용사	형용사	형용(동)사 (조동사)
C	접속사		
D	부사	부사	부사
I	감탄사	감탄사	감동사
J		조사	조사
M	(대명사의 소유격)	관형사	연체사
N	명사 (대명사주격, 소유대명사)	명사 (대명사, 수사)	명사
P	(대명사의 목적격)		
R	전치사		
V	동사	동사	동사(조동사)

<표4> 통합 품사코드 체계

3-4. 통합 기계사전

DMCS도 변형된 형태이기는 하지만 분명히 번역시스템의 기능을 하고 있으므로 기계사전이 필요하다. 앞에서 언급한 바와 같이 <밀레니엄 미팅>에서 사용되는 DMCS에서는 영어 2000개, 일본어 3400개, 한국어 2700개의 단어를 탑재하고 있으며 원문으로 사용되는 언어에 대해 우선적으로 검색기능을 부여한다. 사전구성은 <표5>에서 보는 바와 같다.

원어(영어)			제1대응어(한국어)			제2대응어(일본어)		
표제어	품사	정보	표제어	품사	정보	표제어	품사	정보
your	M		당신의	M		あなたの	M	
mine	N		내것	N		私のもの	N	
have	V		가지다	V		持つ	V	
small	A		작다	A		小さい	A	

<표5> 통합 기계사전의 구성예

4. 두레 다국어 채팅 시스템(DMCS)의 구현

4-1. 사용자 문형확장 방식의 응용

사용자 문형확장 방식은 높은 번역율과 빠른 처리시간을 위해 고안된 구문분석, 변환, 생성등 일련의 번역 과정을 포함하는 구문변환 방식이다. 사용자 문형확장 방식은 게임과 같이 특정 주제에서는 매우 유용한 번역방식으로 사용될 수 있다. 특히 온라인 게임에서 다국어 채팅 시스템으로 응용하기에 용이하다.

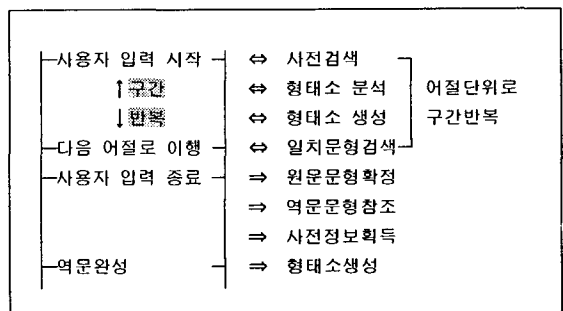
사용자 문장입력: I
I collect
I collect foreign
I collect foreign coins.

원문의 문형코드: ES3431300: N[1] V[1] M[1] N[2].
역문의 문형코드: KS3431300: N[1](은)는 M[1] N[2] V[1](8)다.
최종 역문 생성 : 나는 외국 동전을 수집한다.

<그림1> 원문·역문 문형코드의 예

4-2. DMCS의 구성

DMCS는 사용자 문형확장 방식을 기반으로 한 문형정보 DB, 표제어 및 역어를 포함하는 기계사전, 형태소 분석기, 형태소 생성기로 구성된다. 본 연구에서는 번역 처리시간을 최소화하기 위해 사용자가 문장 입력을 종료하는 시점을 기준으로 그 이후 시점에 위치하는 과정을 최대한 줄이기 위해 노력하였다.



<그림2> DMCS의 구성

사용자 문형확장 방식과 더불어 이와 함께 사용될 수 있는 통합 문형코드와 통합 품사체계도 제안하였다. 이는 번역의 대상이 되는 한국어, 영어, 일본어 각각에 대해 문형코드와 품사코드를 따로 설정하지 않고 통일된 하나의 코드체계를 적용함으로써 기계번역에서의 변환과정을 최소화하기 위해 고안하였다.

참 고 문 헌

- [1] 이충철, “문형정보를 사용한 FSA방식의 구문분석 및 의미해석 통합 모델”, 서강대학교 석사학위논문, 1996.
- [2] 박찬곤, “문장패턴을 이용한 한국어 구문분석에 관한 연구”, 청주대학교 산업과학연구 제15권, 1997.
- [3] 원진숙, “서술어의 결합가를 중심으로 한 한국어 문형 분류”, 고려대학교 어문논집 32, 1993.
- [4] 이경순, “영-한 기계번역에서 분석-변환 규칙의 자동확장”, 한국과학기술원 석사학위논문, 1996.
- [5] 강원석, 황도삼, “예문을 이용한 한·영 번역지원 시스템”, 한국 정보과학회 학술발표논문집, 제24권 2호, 1997.
- [6] 김재훈, “중간언어방식을 이용한 기계번역에서의 한국어 격조사 생성을 위한 한국어 격틀 설정”, 한국과학기술원 석사학위논문, 1987.
- [7] 송재관, 박찬곤, “한·영 기계번역을 위한 한국어 품사 분류”, 한국 정보과학회 학술발표논문집, 제25권 2호, 1998.
- [8] 우요섭, “자연언어처리를 위한 전자사전 시스템의 구축”, 인천대학교 논문집 제20집, 1995.