김일성종합대학학보
정보과학
주체109(2020)년
제66권 제2호

망환경에서 기계번역용자료기지의 실시간통합방법

김광혁, 김룡혁

개별적인 개발자들이 문장번역실험을 진행하면서 구축한 숙어번역패턴[1, 2]자료기지를 통합하는것과 같은 경우 대상분할의 독립성을 보장할수 없는것으로 하여 결과통합과정에 반드시 대상충돌문제가 제기된다.

론문에서는 대규모의 기계번역용자료기지구축의 속도와 정확성, 효률성을 높이기 위하여 많은 개발자들이 망환경에서 자료기지에 대한 수정결과를 실시간적으로 통합하기 위한 한가지 방법을 제안하였다.

1. 자료기지의 실시간통합문제

기계번역체계에서 번역의 정확성을 개선하기 위한 개발과정에는 여러가지 종류의 사전을 비롯한 대규모의 자료기지들을 검사수정하거나 새로 구축하여야 할 필요성이 자주 제기된다. 입력언어의 형태단어사전으로부터 시작하여 두 언어병렬코퍼스에 이르기까지 그러한 자료기지의 종류는 다종다양하다. 그 공통적인 특징은 규모가 방대한것으로 하여 많은 번역전문가들이 참가할것을 요구한다는것이다. 따라서 대규모자료기지구축과정에는 개별적인 개발자들에게 작업대상을 분할해주고 그들의 구축결과를 통합하는 문제가 필수적으로 제기된다.

자료기지구축내용이 형태론적정보만을 입력하는것으로 이루어진 입력언어의 형태단어사전구축과 같은 경우 대상이 입력언어의 형태단어들이므로 입력언어의 단어사전모임을 개별적인 개발자들에게 겹치지 않도록 분할해주면 결과통합에서 아무러한 문제도 제기되지 않는다. 그러나 개별적인 개발자들이 문장번역실험을 진행하면서 구축한 숙어번역패턴[1－3]자료기지를 통합하는것과 같은 경우 대상분할의 독립성을 보장할수 없다.

한편 세계적으로 놓고볼 때 인터네트를 리용하여 많은 개발자들을 대규모자료기지구축에 인입시키는것이 하나의 효과적인 자료기지구축방식으로 되고있다.

자료기지는 일반적으로 유일식별할수 있는 대상과 그것에 대응하는 내용으로 이루어진 기록들의 모임으로 모형화할수 있다.

$$\mathrm{DATABASE} = \{\mathrm{record}_i \mid \mathrm{record}_i = (\mathrm{key}_i, \mathrm{content}_i),$$
$$i = 1, |\mathrm{DATABASE}|; \forall i, j; i \neq j, \mathrm{record}_i \neq \mathrm{record}_j\}$$

일반적으로 망환경에서 자료기지의 실시간통합과정은 망봉사기에 존재하는 하나의 자료기지를 두명이상의 자료기지구축자들이 동시에 수정하는 경우 개별적구축자들의 수정결과를 정확히 통합하는 문제이다.

망환경에서 자료기지의 실시간통합체계의 일반적인 구성방식을 그림 1에 보여주었다. 그림 1에서 $\forall i = 1, \cdots, n$, $\mathrm{DOWN}_i \subset \mathrm{DATABASE}$는 구축자$_i$가 요구한 기록자료모임이고 UP_i는 구축자$_i$가 수정한 기록자료모임이며 MERGE_i는 자료기지에 통합할 기록자료모임이다.

자료기지수정 및 통합대행체는 물리적으로 놓고볼 때 망봉사기측에 위치할수도 있

고 의뢰기측에 위치할수도 있다. 그리고 최종적으로 구축하려고 하는 자료기지는 망봉사기측에 의하여 관리되며 개별적인 의뢰기들에는 존재할수도 있고 존재하지 않을수도 있다.

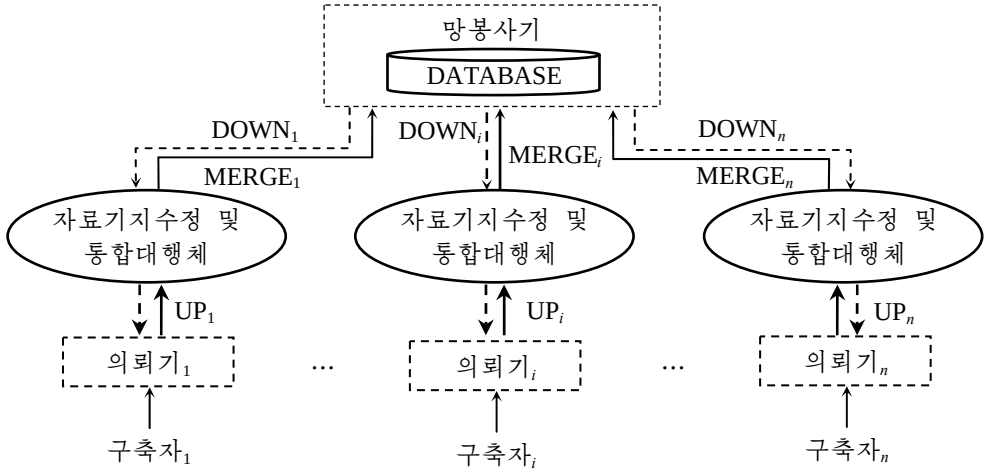

그림 1. 망환경에서 자료기지의 실시간통합체계의 일반적인 구성방식

기계번역용자료기지의 실시간통합체계도 이와 류사한 구성방식을 가지고 동작한다. 직결식자료기지통합체계에서 자료기지는 시간이 흐름에 따라 변화된다. 자료기지구축이 진행되는 과정의 어떤 시각 t의 자료기지를 $\mathrm{DATABASE}^t$라고 표시한다. 직결식자료기지통합체계에서 자료기지구축공정은 개별적인 구축자가 시각 t_{down}에 자료기지수정 및 통합대행체를 통하여 자료기지에서 내리적재한 기록모임(DOWN_i)을 일정한 시간동안 수정하여 갱신된 기록모임(UP_i)을 구축하고 시각 $t_{\mathrm{up}}(t_{\mathrm{down}} < t_{\mathrm{up}})$에 자료기지수정 및 통합대행체를 통하여 자료기지에 통합시키는 요소공정들의 련속과정으로 이루어진다.

시간구간 $[t_{\mathrm{down}}, t_{\mathrm{up}}]$을 구축자의 조작시간이라고 한다. 가입한 어떤 구축자가 자료기지에 가하는 수정조작은 본질에 있어서 개별적인 기록의 추가와 삭제 및 변경조작들의 조합으로 표시할수 있다.

$$\mathrm{DOWN}_i = \{\mathrm{downrecord}_i \mid \exists \mathrm{record}_k \in \mathrm{DATABASE}^t, \mathrm{downrecord}_i = \mathrm{record}_k\}$$

$$\mathrm{UP}_i = \{(\mathrm{uprecord}_q, \mathrm{update}_q) \mid \mathrm{uprecord}_q =$$

$$= (\mathrm{key}_q, \mathrm{contents}_q), \mathrm{update}_q \in \{\mathrm{APPEND}, \mathrm{MODIFY}, \mathrm{REMOVE}\}\}$$

$$\mathrm{update}_q = \mathrm{APPEND} \Rightarrow$$

$$\forall \mathrm{record}_k = (\mathrm{key}_k, \mathrm{contents}_k) \in \mathrm{DATABASE}^{t_{\mathrm{down}}}, \forall q, k; q \neq k, \mathrm{key}_q \neq \mathrm{key}_k$$

$$\mathrm{update}_q = \mathrm{MODIFY} \Rightarrow$$

$$\exists \mathrm{downrecord}_k = (\mathrm{key}_k, \mathrm{contents}_k) \in \mathrm{DOWN}_i,$$

$$\forall q, k; q \neq k, \mathrm{key}_q = \mathrm{key}_k, \mathrm{contents}_q \neq \mathrm{contents}_k$$

$$\mathrm{update}_q = \mathrm{REMOVE} \Rightarrow$$

$$\exists \mathrm{downrecord}_k = (\mathrm{key}_k, \mathrm{contents}_k) \in \mathrm{DOWN}_i, \forall q, k; q \neq k, \mathrm{key}_q = \mathrm{key}_k$$

여기서 p, q, k는 t시각의 기록모임 $\mathrm{UP}_i, \mathrm{DOWN}_i$의 임의의 요소를 가리키는 첨수이다.

자료기지의 실시간통합문제는 결국 개별적인 구축자들이 구축한 갱신된 기록모임 (UP_i) 들을 리용하여 통합할 기록모임 ($MERGE_i$) 을 구하고 자료기지를 변경시키는 문제이다. 만일 조작시간이 서로 겹치는 두 구축자의 갱신된 기록모임들 속에 같은 대상이 들어있는 경우에는 충돌이 발생했다고 한다.

Conflict(UP_i, UP_j) =

$$= \begin{cases} 1, \ [t^i_{\text{down}}, t^i_{\text{up}}] \cap [t^j_{\text{down}}, t^j_{\text{up}}] \neq \Phi \\ \quad \exists(\text{uprecord}_p, \text{update}_p) \notin UP_i, \exists(\text{uprecord}_q, \text{update}_q) \in UP_j, \forall p, q; \ p \neq q, \text{key}_p = \text{key}_q \\ 0, \ \text{기타} \end{cases}$$

자료기지의 실시간통합에서 해결하여야 할 기본문제는 충돌을 해소하는것이다.

2. 기계번역용자료기지의 실시간통합절차

기계번역용자료기지의 실시간통합에서 통합의 기본단위는 기록이다. 자료기지수정 및 통합대행체에서는 기록을 단위로 하여 개별적인 자료기지구축자들의 수정결과사이 충돌을 검출하고 통합을 진행한다.

실시간통합체계의 자료기지수정 및 통합공정을 단계별로 서술하면 다음과 같다.

걸음 1 자료기지 ($\text{DATABASE}^{t_{\text{down}}}$) 로부터 수정하려고 하는 기록들의 모임 ($\text{DOWN}_i$) 을 내리적재한다.

걸음 2 기록들에 대한 수정을 진행한데 기초하여 수정한 기록들의 모임 (UP_i) 을 추출한다.

걸음 3 수정한 기록들의 모임 (UP_i) 에 대하여 자료기지 ($\text{DATABASE}^{t_{\text{up}}}$) 로부터 대응하는 기록들의 모임 ($\text{DOWN}_{i'}$) 을 다시 내리적재하여 충돌이 일어나지 않는 기록들의 모임 (MERGE_i) 과 충돌이 일어난 기록들의 모임 (ERROR_i) 을 분리시킨다.

걸음 4 충돌이 일어나지 않는 기록들의 모임 (MERGE_i) 을 자료기지에 통합시킨다.

걸음 5 충돌이 일어난 기록들의 모임 (ERROR_i) 이 비여있으면 통합절차를 끝내고 비여있지 않으면 그것에 대한 재수정을 진행하여 수정된 기록모임 (CORRECT_i) 을 얻어낸다. 다음 걸음 3으로 이행한다.

$$\text{DOWN}'_i = \{(\text{key}_i, \text{contents}'_i) \mid \text{key}_i \in \text{DOWNKEY}_i, \exists(\text{uprecord}_q, \text{update}_q) \in \text{UP}_i, (\text{key}_i, \mathbf{c}\text{contents}'_i) \in \text{DATABASE}^{t_{\text{up}}}\}$$

$$\text{DOWNKEY}'_i = \{\text{key}_i \mid (\text{key}_i, \text{contents}_i) \in \text{DOWN}_i\}$$

$$\text{MERGE}_i = \left\{ (\text{uprecord}_p, \text{update}_p) \middle| \begin{array}{c} (\text{uprecord}_p, \text{update}_p) \in \text{UP}_i, \text{uprecord}_p = (\text{key}_p, \text{contents}_p), \\ \exists(\text{key}_i, \text{contents}_i) \in \text{DOWN}_i, (\text{key}_i, \text{contents}'_i) \in \text{DOWN}'_i, \\ \text{key}_p = \text{key}_i, \text{contents}_i = c\text{contents}'_i \end{array} \right\}$$

$$\text{ERROR}_i = \left\{ (\text{uprecord}_q, \text{update}_q) \middle| \begin{array}{c} (\text{uprecord}_q, \text{update}_q) \in \text{UP}_i, \text{uprecord}_q = (\text{key}_q, \text{contents}_q), \\ \exists(\text{key}_i, \text{contents}_i) \in \text{DOWN}_i, (\text{key}_i, \text{contents}'_i) \in \text{DOWN}'_i, \\ \text{key}_q = \text{key}_i, \text{contents}_i = \text{contents}'_i \end{array} \right\}$$

직결식자료기지통합체계의 동작과정을 그림 2에 보여주었다. 직결식자료기지통합체

계의 동작과정에 실시간통합체계가 자동적으로 통합될수 없는 경우가 존재한다.

실례를 들어 어떤 구축자가 시각 t_1에 내리적재하여 수정한 기록(key, contents)을 시각 t_2에 올리적재하려고 한다고 하자. 이때 시각 $t_3(t_1 < t_3 < t_2)$에 갱신자료를 올리적재한 다른 구축자에 의하여 자료기지에서 대상 key의 기록이 삭제되였다면 이 기록의 갱신은 충돌에 의한 수동적인 통합조작으로 해결할수 있다.

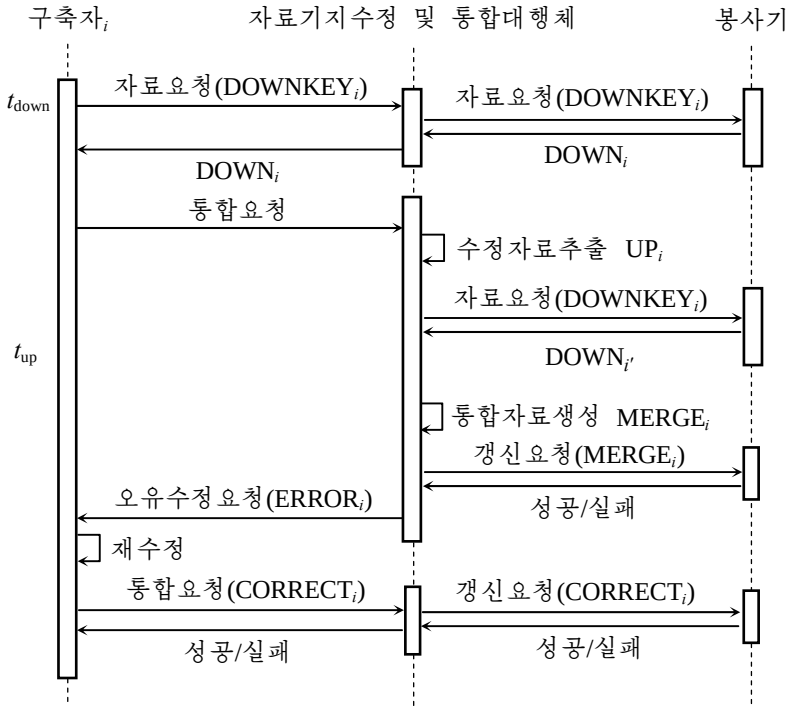

그림 2. 직결식자료기지통합체계의 동작과정

3. 실현 및 평가

론문에서 제안한 방법을 도-조기계번역체계개발을 위한 도-조숙어번역패턴자료기지구축에 도입하였다.

도-조숙어번역패턴의 구조[2]는 다음과 같다.

exp = {Ger-Exp, KOR-EXP}, KOR-EXP = {Kor-pattern$_1$, …, Kor-pattern$_n$}

여기서 exp는 도-조숙어번역패턴, Ger-exp는 도이췰란드어표현, KOR-EXP는 조선어번역패턴실례모임, Kor-pattern$_i$는 조선어번역패턴실례이다.

도이췰란드어단어별로 사전화되여있는 도이췰란드어표현구를 자료기지기록의 유일식별대상(key)으로 설정하고 조선어번역패턴실례들의 동일성검사를 충돌판정조건으로 리용하여 실시간통합체계를 실현하였다.

실시간통합체계를 구축하고 도입한 결과 망환경에서 여러 개발자들이 문장번역실험과정에 진행하는 도-조숙어번역패턴자료기지에 대한 수정작업을 실시간적으로 통합하면서 개발을 진행할수 있다는것을 검증하였다.

맺 는 말

기계번역용자료기지의 실시간통합방법은 개별적인 자료기지들의 특성에 따라 유일식별대상설정과 충돌판정조건을 세분화시켜 적용할수 있는 효률적인 방법이다.

참 고 문 헌

[1] 김성존; 영조기계번역론, 과학기술출판사, 113～132, 주체99(2010).

[2] 김광혁, 김룡혁; **김일성**종합대학창립 70돐기념 전국부문별과학토론회 론문집(정보, 자동화), 16, 주체105(2016).

[3] Alexander Fraser et al.; Computational Linguistics, 39, 1, 58, 2013.

주체109(2020)년 2월 5일 원고접수

A Method of Real Time Integrating of Machine Translation Database in Network Environment

Kim Kwang Hyok, Kim Ryong Hyok

In this paper, we supposed a method for allowing many developers to integrate the modification results on database in the network environment and implemented by online integrating system.

Keywords: machine translation database, online integrating system, network environment