제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

개체명 공기 정보를 이용한 이벤트 문장의 단문 구조 분석

임수종, 김태현, 황이규, 윤보현
한국전자통신연구원 지식처리연구팀
e-mail : isj@etri.re.kr

Clausal Segmentation for Event Sentences Using Named Entity Co-occurrence Information

Soojong Lim, Tae-Hyun Kim, Yi-Gyu Hwang, Bo-Hyun Yun
Knowledge Processing Research Team
Electronics and Telecommunications Research Institute

요 약

정보추출이란 자연어로 작성된 문서 집합에서 원하는 정보를 선택하여 구조화된 표현으로 생성하는 것을 말한다. 문장 단위로 정보 추출 작업을 수행할 때 추출되는 정보를 보유한 문장을 이벤트 문장이라고 정의하고 이러한 이벤트 문장의 구조를 분석하여 최종적으로 유용한 정보를 추출하기 위해서는 이벤트 문장의 구조를 파악하기 위해 이벤트 문장을 단문으로 분할하여 구조를 분석한다. 본 연구에서는 단문 구조 분석을 위해 일반적인 한국어 문장의 특성과 용언의 조사 정보를 이용하고 이러한 정보로 분석할 수 없는 문장에 대해서는 공기 정보를 사용한다. 사용되는 공기 정보는 개체명이 많이 사용되는 이벤트 문장의 특성을 이용하기 위하여 개체명으로 확장된 명사(개체명)-조사-용언의 공기 정보를 구축하여 사용한다. 개체명 확장된 공기 정보는 일반 공기 정보에 비해 이벤트 문장에서 F-Measure 기준으로 약 2%의 성능향상을 보인다.

1. 서론

최근 컴퓨터를 이용한 지식 관리가 일반화 되면서 인터넷등을 이용하여 접근할 수 있는 지식의 양이 늘어나고 있다. 그러나 이런 정보가 늘어날수록 사용자가 특정 정보에 접근하기 위해서 많은 노력을 필요로 한다. 정보 추출(Information Extraction)이란 자연어로 작성된 문서 집합에서 원하는 정보를 선택하여 구조화된 표현으로 생성하는 것을 말한다[8].기존의 정보 추출 기법은 도메인 제한적인 지식을 사용하여 수행되어 왔으나 이식성의 문제로 인하여 도메인 제한적인 지식을 최소화하기 위해 연구가 진행되고 있다.

문서에서 정보 추출의 대상이 되는 '인명', '조직명', '장소', '시간' 등의 개체명이 포함된 문장을 이벤트 문장이라 정의하고 이러한 문장을 대상으로 정보 추출을 하였다. 그러나 자연어 문장의 특성상 이러한 문장들은 한 개 이상의 동사를 포함하는 복문의 형태로 되어 있고 이러한 문장 안에는 정보 추출에서 대상이 되지 않는 부분을 포함하고 있기 때문에 복문을 단문으로 분할하여 구조를 분석해야 할 필요성이 제기 된다.

단문 구조 분석은 아래의 예처럼 한 문장에 중심어인 용언이 복수개인 경우 용언을 중심으로 문장을 나누어 용언의 필수 정보에 해당하는 명사상당어구를 인식한다. 단문 구조 분석은 부분적인 자연어 문장에 대한 이해로 응용이 가능한 문서요약, 질의응답, 정보 추출 분야의 기본 기술이다.

> 사고 항공기는 이날 오전 9 시 37 분 베이징을 출발해 오전 11 시 35 분경 김해공항에 도착할 예정이었다 .
>
> (항공기는, 출발하다) (베이징을, 출발하다)
> (김해공항에, 도착하다)

본 연구에서는 복문의 단문 구조 분석에서 일반적인 한국어 특성을 고려한 분석 규칙과 정보 추출의 대상이 되는 이벤트 문장을 특성을 고려한 공기정보를 사용하여 이벤트 문장의 단문 구조 분석 방법을 제안한다.

2. 관련연구

기존의 단문 구조 분석은 복잡한 문장을 단순화하여 문장의 구문 분석의 효율을 높이는 과정으로 연구가 진행되어 왔다. 한국어에 대한 단문 구조 분석은 한국어의 구문 특성

및 의미 정보를 이용하는 방법과 통계정보를 이용한 방법으로 나눌 수 있다.

한국어의 내포문을 단문으로 분리하기 위해서 안긴 문장을 통제할 수 있도록 용언을 유형별로 분류하고 명사의미 표지로부터 용언의 하위범주 정보를 이용한 방법이 있고[1], 구문 패턴이나 문형을 이용하는 방법이 있다[3, 7].그리고 기구축된 하위범주 정보와 명사 의미 정보를 이용한 후 구문적인 정보와 기타 문법적인 지식을 사용하여 의미 정보를 보완하는 방법이 있다.[4]

통계적인 방법으로는 단문 분할을 통해 명사구 색인에 대한 연구가 있다. 말뭉치에서 자주 나타나는 표현이 보다 자연스럽게 받아들여질 수 있다는 사실에 기반하여 단문 분할을 위해서 조사의 격과 용언간의 공기 정보를 이용하여 가장 높은 수치를 나타내는 용언을 선택하여 단문 분할을 한다[6]. 단문 분할의 목적은 아니지만 구문 분석을 위해서 명사-조사-용언 공기 정보를 이용한 방법도 있다.[5]

의미 정보를 이용하는 경우 [1]의 경우처럼 연구를 위한 소량의 의미 정보만을 구축하는데 그친다. [3]는 기구축된 의미 정보를 이용하였는데 한국어의 특성상 용언의 하위범주 정보는 거의 완벽하게 구축할 수 있지만 고유명사를 포함한 개체명에 대한 명사의 의미 정보까지 포함하는 의미 정보와 한국어에서 빈번하게 일어나는 명사파생동사까지 망라하는 하위범주 정보를 구축하는 것은 불가능하기 때문에 한계가 있다.

통계정보를 이용하는 경우 의미 정보를 구축하는 문제점을 해소할 수 있으나 통계정보의 정확성과 자료 희귀성 문제를 해결하여야 한다. 의미 정보와 마찬가지로 공기 정보 역시 본 연구에서 대상으로 삼고 있는 이벤트 문장에 빈번하게 출현하는 개체명에 대한 어휘 자료를 충분히 구축하기 어렵다.

이러한 문제를 해결하기 위해서 본 논문에서는 기존의 연구에서 사용된 구문적인 정보와 문법적인 지식, 그리고 의미 정보를 사용하고 이러한 정보로 파악할 수 없는 단문 구조 분석에 대해서는 공기 정보를 사용한다. 공기 정보의 자료 희귀성 문제를 극복하기 위해서는 공기 정보를 구축할 때 어휘 정보와 더불어 개체명에 해당하는 경우 개체명 정보를 추가하여 자료 희귀성 문제를 완화하고자 한다.

3. 단문 구조 분석

한 문장에서 출현하는 각각의 용언들에 대한 필수 성분을 조사 정보나 문형을 이용하여 찾아내고 이 정보만으로 필수 성분을 찾을 수 없는 경우에는 미리 구축된 공기정보를 이용하여 필수 성분을 찾아낸다. 본 논문에서 필수성분은 품사적으로 일반 명사, 대명사, 개체명을 말한다.

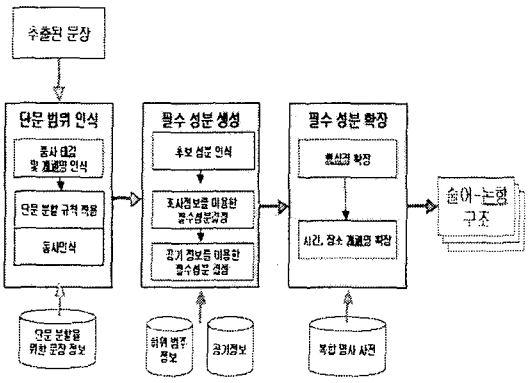

그림 1 단문구조분석 시스템 구성도

3.1 단문 분할

한국어의 복문 중에는 문장 형태나 사용된 연결 어미에 의해 명확하게 단문으로 구분할 수 있는 형태가 있다.

3.1.1 문장 형태를 이용한 단문 분할

아래와 같이 특별한 기호를 사용하여 문장의 형태만으로 단문 분할이 가능한 경우를 처리한다.

김성진 청와대 부대변인이 " 김대중 대통령과 알렉산더 크바스니예프스키 폴란드 대통령이 오는 4 일 정상회담을 갖는다" 고 1 일 발표했다.
-김성진 청와대 부대변인이 1 일 발표했다.
-김대중 대통령과 알렉산더 크바스니예프스키 폴란드 대통령이 오는 4 일 정상회담을 갖는다.

3.1.2 규칙을 이용한 단문 분할

대등 접속 어미로 연결되면서 각각의 부분에 주어, 동사가 출현하는 경우도 규칙을 사용하여 단문으로 분할하였다. 사용된 대등 접속 어미는 다음과 같다.

고, 으며, 며, 으면서, 면서, 고서, 은데, ㄴ데, 던데, 는데, 거니와, 으나, 나, 으나마, 나마, 어도, 지만, 으되, 되, 건만, 느니, 거나, 든지

3.1.3 용언을 중심으로 한 단문 분할

위의 2 가지 규칙으로 단문 분할이 불가능할 경우에는 용언을 기준으로 다음 용언이 나오기 전까지를 단문 단위로 분할하였다. 용언은 본 용언만을 대상으로 삼았고, 이벤트 문장의 특성상 실제적인 단문구조 분석은 동사만을 대상으로 하였다.

그러나 한국어의 구조적 특징에 의해 관형형 어미(-는, -은, -을, -ㄹ, -ㄴ, -던)가 부착된 용언의 경우에는 오른쪽에 필수 성분이 존재하여 오른쪽에 출현하는 용언과 필수 성분을 공유하지만 관형절의 종류에 따라 단문 분할을 확장해야 할 경우와 아닌 경우로 나뉜다.

관형절의 종류를 보면

1)필수 성분 중 하나가 탈락되어 용언의 오른쪽에 존재하는 관계 관형절
2)관형형 어미가 부착되었음에도 불구하고 용언의 오른쪽에 필수성분이 모두 존재하면서 ' 사실, 소식, 보도, 사건, 냄새, 소문, 결심,..' 등의 보문명사의 수식을 받는 동격 관형절
3)동격 관형절과 마찬가지로 용언의 오른쪽에 필수성분이 모두 존재하면서 용언의 오른쪽에는 의존명사가 존재하는 의존 관형절

로 구분하여 단문 분할의 범위를 확대할 필요가 없는 동격/의존 관형절이 아닌 경우는 관계 관형절로 간주하고 단문의 범위를 확장한다.

3.2 필수 성분 결정

분할된 단문에 대해 필수 성분을 찾기 위해 본 논문에서는 용언에 대한 조사 정보를 이용하고 조사 정보가 존재하지 않는 경우에 대해서 공기 정보를 사용한다.

3.2.1 조사정보

각 용언에 해당하는 조사 정보를 얻기 위해 ETRI 에서 구축한 하위범주 사전의 일부를 이용하였다. 하나의 용언에 대해서 여러가지 형태의 조사 정보가 존재하고 또 이런 조사가 문장에서 부분적으로 쓰이기 때문에 순서나 개수를 제한적으로 적용하지 않고 부분적으로 존재여부만을 사용하였다.

동사 ' 태우다' 를 예로 들면 4 개의 하위범주 정보가 있지만 조사 정보로 제한시키면 ' -이 - 을 - 에 태우다' 의 한가지 형태로 축소시킬 수 있다.

3.2.2 공기정보

말뭉치에서 자주 나타나는 표현이 실제로 유효하다는 사실에 기반하여 공기정보는 하위범주 사전에 등록되지 않아 조사 정보를 얻을 수 없는 경우에 한하여 구축된 공기정보를 적용하기 위하여 (명사, 조사, 용언)의 정보를 수집하였고 수집 방법은 단문 분할 규칙을 적용하여 단문 분할 후 용언의 왼쪽 어절이 명사+조사의 형태인 경우 수집하였다. 그러나, 공기정보 수집의 목적이 하위범주 사전에 등록되지 않은 용언에 대한 정보를 수집하기 위함이기 때문에 저빈도의 용언이나 명사 파생 동사이기 때문에 자료희귀성(data sparseness) 문제가 발생하기 때문에 (조사, 용언) 쌍을 수집하였고 단문 분할의 대상 문장이 ' 인명' , ' 조직명' , ' 장소' , ' 시간' 등의 개체명이 존재하는 이벤트 문장이기 때문에 (명사, 조사, 용언)의 정보를 수정하여 개체명으로 인식된 명사는 해당 개체명의 범주를 사용하였다.

표 1 은 이벤트 문장과 비이벤트 문장의 평균 어절 수, 문장당 개체명 수, 용언 수를 비교하였다.

표 1 이벤트, 비이벤트 문장 비교

문장종류	평균어절	문장당 개체명	문장당 용언	용언당 어절
이벤트	27.68	7.58	4.15	6.66
비이벤트	18.97	1.51	3.01	6.36

표에서 보여지듯이 이벤트 문장으로 분류된 문장은 길이가 길고 하나의 용언과 관계되는 어절도 많다. 그러나 문장에서 평균 7.58 개의 개체명이 발생하기 때문에 어휘 단위의 공기정보보다는 개체명 확장된 공기 정보를 사용하는 것이 이벤트 문장을 단문 구조 분석하는데 도움이 될 것이다.

공기 정보를 이용하여 대상 명사-조사-용언에 대한 공기값은 다음의 식 (1)을 이용하고 조사가 생략된 경우에는 식 (2)를 이용한다.

$$Co(v,n,p) = \lambda_1 P(n,p \mid v) + \lambda_2 P(p \mid v) + d \quad (1)$$

$$Co(v,n) = \lambda_1 P(n \mid v) + d \quad (2)$$

(단, d 는 좌우거리 1 어절에 대해서 상수 0.1)

4. 단문 구조 분석 과정

본 논문에서는 단문 구조 분석 과정은 그림과 같다.이벤트 문장을 입력 받아 형태소 분석과 개체명 인식을 수행한 결과에 다음의 과정을 적용하여 단문 구조 분석을 한다.

1. 단문 범위 결정
   ㄱ) 문장 형태나 분할 규칙을 사용하여 단문 범위 결정
   ㄴ) 문장 형태나 규칙에 해당하지 않는 경우 용언 단위로 단문 범위 결정. 단, 용언에 관형형 어미가 부착된 경우 관형절의 종류를 판단하여 범위 결정
2. 필수 성분 결정
   ㄱ) 용언에 해당하는 단문 범위에서 필수 성분의 후보를 선택
   ㄴ) 용언에 해당하는 조사 정보를 확보한 후 해당 조사를 사용한 필수 성분을 후보 중에서 결정
   ㄷ) 필수 성분 중 모호성이 발생하거나 조사 정보가 없는 경우는 공기 정보를 사용하여 필수 성분을 결정
3. 필수 성분의 확장
   ㄱ) 결정된 필수 성분에 대해 왼쪽의 어절의 형태소가 일반 명사, 복합 명사, 개체명의 일부인 경우는 확장한다.

아래의 문장을 제안하는 알고리즘에 의해 단문 구조 분석하는 과정은 다음과 같다.

예문 1) 225 명의 승객과 승무원을 태우고 대만을 떠나 홍콩으로 가던 대만의 중화항공 여객기가 25 일 오후 대만해협에 추락했다.

•단문 범위 결정 결과
1-1. 225 명의 승객과 승무원을 태우고
1-2. 대만을 떠나
1-3.홍콩으로 가던 대만의 중화항공 여객기가
1-4.대만의 중화항공 여객기가 25 일 오후 대만해협에 추락했다.

•필수 성분 결정 결과
2-1. (승무원을, 태우다)
2-2. (대만을, 떠나다)
2-3. (홍콩으로, 가다), (여객기가, 가다)
2-4. (여객기가, 추락하다) (대만해협에, 추락하다)

•필수성분 확장 결과
3-1. (225 명의 승객과 승무원을, 태우다)
3-2. (대만을, 떠나다)
3-3. (홍콩으로, 가다), (대만의 중화항공 여객기가, 가다)
3-4. (대만의 중화항공 여객기가, 추락하다), (대만해협에, 추락하다)

예문을 단문 분할 하기 위해 단문 범위를 결정하는 과정에서 문장 형태나 규칙을 이용할 수 없고 용언을 인식하여 범위를 결정한다. 예문에서는 ' 태우다' , ' 떠나다' , ' 가다' , ' 추락하다' 4 개의 용언을 중심으로 왼쪽에 있는 어절로 단문의 범위를 결정한다. 그러나 가던(가/pv+던/em)은 용언에 부착된 관형형 어미에 의해 관형절로 판명되고 보문 명사나 의존명사가 오른쪽 어절에 존재하지 않기 때문에 관계 관형절로 간주하여 1-3 과 같이 오른쪽 어절에서 격조사가 나오는 부분까지를 ' 가다' 의 단문 범위로 설정한다.

결정된 범위 안에서 각각의 용언에 대해 하위범주정보에서 조사 정보를 획득한다. 용언 ' 태우다' 는 ' - 이 - 을 - 에 태우다' 는 조사 정보를 갖기 때문에 ' 승무원/nc+을/jc' 을 필수 성분으로 채택한다.

4 개의 동사 중에서 명사파생동사인 ' 추락하다' 는 하위범주사전의 엔트리로 등록되어 있지 않기 때문에 공기 정보를 사용하여야 한다.

5. 실험

5.1 실험 방법

본 논문에서 제안하는 방법에 대한 실험의 준비사항은 다음과 같다.

하위범주사전 ETRI 에서 구축한 하위범주 사전으로 약 1

만여개의 동사만을 대상으로 대표 조사 정보를 이용한다

공기정보 신문 기사에 이벤트성 문장이 많이 존재하기 때문에 자체 수집한 신문 기사 약 1200 만 어절을 대상으로 하였고 일반적인 방법으로 구축된 명사-조사-동사는 76 만쌍이 추출되었으며 평균 빈도는 약 1.8 이다. 개체명으로 구성된 명사(개체명)-조사-동사는 평균빈도 약 2.4 이다. 조사-동사는 평균 26.7 이다.

실험 문장 실험을 위해서 이벤트 문장 추출 평가를 위해 구축해 놓은 문장을 사용하였다[2]. 각 문서집합에 있는 모든 문장들을 대상으로 도메인과의 관련성과 이벤트 관련 정보의 포함 정도에 따라 0-4 의 점수가 부여되었는데 4 점으로 부여된 144 문장을 이벤트 문장이라 가정하였고, 0 과 1 점이 부여된 문장 중에서 복문 144 문장 만을 골라 비이벤트 문장이라 가정하고 실험을 하였다.

실험은 조사 정보, 조사 정보와 명사-조사-동사 공기 정보, 조사 정보와 명사(개체명)-조사-동사 공기정보를 각각 사용하여 이벤트 문장에 대해서 문장에서 (필수성분, 용언)쌍을 추출하였으며 개체명 공기 정보의 효용성을 검증하기 위해서 이벤트 문장과 비이벤트 문장을 사용하여 실험 2 와 실험 3 을 수행하였다.

4.2 실험 결과와 분석

실험의 평가 척도로는 정확률(P)와 재현률(R), 그리고 F-Measure 를 사용하였다. 각각의 계산 수식은 다음과 같다.

$$P = \frac{\text{정답(필수성분,용언)쌍의개수}}{\text{시스템이찾은 (필수성분,용언)쌍의개수}}$$

$$R = \frac{\text{정답(필수성분,용언)쌍의개수}}{\text{총(필수성분,용언)쌍의개수}} \qquad F = \frac{2*P*R}{P+R}$$

위와 같은 척도를 사용하여 실험한 결과는 다음과 같다.

표 2 단문구조분석 실험결과

		비행기사고	교통사고	재해	합
하위 범주	P	84.7	85.2	89.0	85.9
	R	83.8	78.6	90.2	83.2
	F	84.2	81.7	89.6	84.5
공기정보 (어휘)	P	84.9	81.4	86.3	83.9
	R	96.3	86.4	93.0	91.5
	F	90.2	83.8	89.5	87.5
공기정보 (개체명)	P	86.2	84.6	89.4	86.4
	R	97.9	88.4	95.1	93.2
	F	91.6	86.4	92.2	89.7

표 2 는 각각 하위범주 정보, 어휘 단위의 공기 정보, 개체명 확장된 공기 정보를 이용, 3 개 분야의 이벤트 문장을 대상으로 실험한 결과이다. 표에서도 볼 수 있듯이 한국어 문장 특성과 하위범주 정보만을 사용한 단문 구조 분석 결과보다 공기 정보를 이용한 결과는 정확률은 하락하고 재현율은 상승함을 알 수 있다. 이것은 조사 정보를 이용한 경우 하위범주 사전에 조사 정보가 없는 경우 답을 제시하지 않기 때문이다. F-Measure 기준으로 일반 어휘 공기 정보를 사용하는 경우가 약 3-4% 향상됨을 볼 수 있다. 개체명 확장 공기 정보를 사용하는 경우는 약 2% 안팎의 F-Measure 가 상승함을 알 수 있다. 이러한 결과 조사정보를 이용하는 경우 하위범주 사전에 의존하기 때문에 존재하지 않는 경우 단문 구조 분석이 불가능하지만 이런 용언에 대해서 공기 정보가 올바른 정보를 제공해준다는 것을 알 수 있다.

표 3 는 개체명 확장 공기 정보를 사용하여 이벤트 문장과 비이벤트 문장을 단문 구조 분석한 결과이다. 비이벤트 문장에서는 개체명이 빈번하게 발생되지 않기 때문에 공기 정보의 적용에 문제점이 있어서 재현율은 저하되지 않지만 정확율이 많이 저하되는 현상을 보인다. 위와 같은 사실은 개체명 확장 공기 정보를 일반적으로 쓰기에는 개선점이 있음을 보여준다.

표 3 이벤트, 비이벤트 문장 단문구조 분석결과

	이벤트 (A)	비이벤트(B)	A-B
P	84.8	70.9	13.9
R	91.5	86.5	5.0
F	88.0	77.9	10.1

5. 결론

본 연구에서는 인명, 조직명, 시간, 장소등의 정보를 갖고 있는 이벤트 문장을 단문으로 분할하여 용언 단위로 필수 성분을 분석하는 방법을 제안하였다. 제안된 방법은 일반 문장과는 다른 특성을 갖는 이벤트 문장을 단문 구조 분석함에 있어 일반적인 한국어 구문 특성과 조사 정보를 이용하였고 이러한 정보로 구조 분석이 불가능한 경우에는 명사(개체명)-조사-용언 공기 정보를 이용하여 단문 구조 분석을 시도하였다. 제안한 방법은 개체명 확장된 공기 정보를 이용하여 이벤트 문장을 구조 분석시 일반 공기 정보보다 성능이 향상됨을 알 수 있다. 그리고 인명이나 조직명등의 개체명을 갖고 있지 않은 일반 문장에 대해서는 개체명 확장된 공기 정보를 사용하거나 일반 공기 정보를 사용하거나 크게 변화가 없음을 알 수 있다.

본 연구에서 사용된 공기 정보는 기존의 구축된 공기 정보에 비해 상대적으로 소량이기 때문에 공기 정보의 양을 늘릴 필요가 있으며 단문 구조 분석의 기준이 되는 용언의 여러가지 형태에 대한 체계적인 분류가 필요하다.

참고문헌

[1] 김광진, 송영훈, 이정현, " 한국어 내포문을 단문으로 분리하는 시스템의 구현" , 제 5 회 한글 및 한국어 정보처리 학술대회, pp.25-34, 1993

[2] 김태현, 임수종, 윤보현, " 이벤트 문장 추출" , 제 14 회 한글 및 한국어 정보처리 학술대회 게재예정

[3] 박성배, " 문장분할을 이용한 한국어 분석" , 서울대학교 컴퓨터공학과 석사학위논문, 1996

[4] 박현재, 우요섭, " 의미 정보를 이용한 이단계 단문 분할" , 한국정보처리학회 논문지 제 7 권 제 9 호, pp.2876-2884, 2000

[5] 윤준태, " 공기관계 기반 어휘 연관도를 이용한 한국어 구문 분석" , 연세대학교 컴퓨터과학과 박사학위 논문, 1998

[6] 이현아, 이종혁, 이근배, " 단문 분할을 통한 명사구 색인 방법" , 한국정보과학회 논문지 제 24 권 제 3 호, pp.302-311, 1997

[7] 이현영, 황이규, 이용석, " 문형과 단문 분할을 이용한 한국어 구문 모호성 해결" , 제 12 회 한글 및 한국어 정보처리 학술대회, pp.116-123, 2000

[8] Ralph Grishman, "Information Extraction: Techniques and Challenges", In Proceedings of the 7th Message Understanding Conference(MUC-7), Columbia, MD, April 1998.