KIPS Tr. Software and Data Eng.
Vol.3, No.7 pp.271~276 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.7.271

A Normalization Method of Distorted Korean SMS Sentences for Spam Message Filtering

Kang Seung-Shik†

ABSTRACT

Short message service(SMS) in a mobile communication environment is a very convenient method. However, it caused a serious side effect of generating spam messages for advertisement. Those who send spam messages distort or deform SMS sentences to avoid the messages being filtered by automatic filtering system. In order to increase the performance of spam filtering system, we need to recover the distorted sentences into normal sentences. This paper proposes a method of normalizing the various types of distorted sentence and extracting keywords through automatic word spacing and compound noun decomposition.

Keywords : Spam Message, SMS Filtering, Sentence Normalization, Automatic Word Spacing, Keyword Extraction

스팸 문자 필터링을 위한 변형된 한글 SMS 문장의 정규화 기법

강 승 식†

요 약

휴대폰에서 문자 메시지 전송 기능은 현대인들에게 매우 편리한 새로운 형태의 의사소통 방식이다. 반면에 문자 메시지 기능을 악용한 광고성 문자들이 너무 많이 쏟아져서 휴대폰 사용자들은 스팸 문자 공해에 시달리는 심각한 부작용을 낳게 되었다. 광고성 문자를 발송하는 사람들은 문자 메시지가 자동으로 차단되는 것을 회피하기 위해 한글 문장을 다양한 형태로 변형하거나 왜곡시키고 있으며, 이러한 문자 메시지를 자동으로 차단하기 위해서는 변형되거나 왜곡된 문장들을 정상적인 한글 문장으로 정규화하는 기술이 필수적이다. 본 논문에서는 변형되거나 왜곡된 광고성 문자 메시지를 정상적인 문장으로 정규화하고 정규화된 문장으로부터 자동 띄어쓰기 및 복합명사 분해 과정을 거쳐 키워드를 추출하기 위한 방법을 제안하였다.

키워드 : 스팸 문자, SMS 필터링, 문장 정규화, 자동 띄어쓰기, 키워드 추출

1. 서 론

휴대폰으로 문자를 주고받는 SMS 메시지 전송 기능은 현대인들에게 새로운 형태의 의사소통 방식을 제공하고 있다. 그럼으로써 모바일 환경에서 새로운 문화를 만들어 내기도 하고 현대 생활을 영위하는 데 매우 편리한 수단으로 자리를 잡고 있다[1]. 이러한 편의성이 있는 반면에 휴대폰의 문자 메시지 기능을 악용하여 광고성 문자들을 불특정 다수에게 보내는 일이 도를 지나치는 일들이 발생하고 있으며, 휴대폰 사용자들은 스팸 문자의 공해에 시달리는 부작용이 사회적인 문제로 등장하게 되었다[2].

† 종신회원 : 국민대학교 컴퓨터공학부 교수
Manuscript Received : March 19, 2014
First Revision : June 20, 2014
Accepted : June 21, 2014
* Corresponding Author : Kang Seung-Shik(sskang@kookmin.ac.kr)

스팸 메일과 스팸 문자, 뉴스 기사 및 블로그 등에 대한 광고성 댓글은 수많은 사람들에게 피해를 주고 있으며 이로 인해 지불해야 하는 사회적 비용은 매우 크다[3]. 스팸 메일을 차단하는 연구는 이미 심도있게 진행되어 왔고 자동으로 스팸 메일을 차단 시스템이 개발되어 메일 서버는 스팸 메일로 분류된 메일들을 별도로 관리를 하고 있다[4,5,6]. 스팸 메일과 정상적인 메일을 구분하는 방법으로는 문서 분류 기법을 이용한다. 스팸 메일 인식 시스템은 모든 이메일들을 두 개의 범주로 구분하는 문제이기 때문이다. 따라서 스팸 메일을 인식하는 기법으로는 SVM, Naive Bayse, 최대 엔트로피 등 일반적인 기계 학습 이론을 이용하고 있다[7,8,9].

스팸 메일은 여러 문장으로 구성되는 문서를 스팸 문서와 스팸이 아닌 정상적인 문서로 분류하는데 비해 휴대폰에서 스팸 문자 메시지는 한두 문장으로 구성되는 짧은 문서이

다. 이는 문자 메시지의 특성상 통상적으로 80바이트(한글 40 음절) 이내의 짧은 문장으로 구성되기 때문이다. 또한, 광고성 스팸 문자를 작성하는 사람들은 스팸으로 자동 차단되는 것을 피하기 위해 불필요한 기호들을 추가하거나 문자를 변형, 왜곡시키는 기법을 사용한다. 따라서 스팸 문자인지를 자동으로 판단하는 것은 스팸 메일 분류 시스템보다 어려운 문제이다. 스팸 문자 메시지의 예는 아래와 같다.

- 카@드@결@제,연@체@자금때문머리아프시죠, 저희가 결제해드립니다,전화주세요.
- 귀빈을[카_지]노 *특별회원 으로* 모십니다 [현금3만원]드림 www.tuu33.com
- ▶df8282.ⓒom 국_내&최고>ㅋr지ㅣ노<설&명절/이벤트당첨/2만>회원가입후고객센터
- 병원처방20정/곽32정/할인행사/후불/전문상담!정K품!비K아~시K알문의는문자예약!

본 연구는 휴대폰에서 대량으로 발송되는 광고성 스팸 문자를 차단하기 위하여 문자 메시지의 내용을 분석함으로써 왜곡되거나 변형된 한글 문장을 정상적인 문장으로 복원하는 시스템을 설계하고 구현하는 방법에 관한 것이다. 이 연구에서 변형된 한글 문장을 복원하기 위하여 한글 자모와 유사한 문자를 한글로 변환하거나, 한글 자모로 구성된 문자열을 음절로 결합하는 방법을 적용한다.

2. SMS 문장 정규화의 필요성

한글 SMS 문장은 여러 가지 유형의 자모 변형 및 음절 변형, 축약 등이 포함되어 있어서 각 변형 유형에 따라 변형된 부분을 복원하는 과정이 필요하다. 또한, 광고성 SMS 문장의 특징은 띄어쓰기를 무시하거나 불필요한 기호를 삽입하기 때문에 불필요한 기호를 제거하고 자동으로 어절을 구분해 주는 과정을 거쳐서 SMS 문장에서 주요 키워드를 추출해 주어야 한다.

위 예제는 실제 문자 메시지에서 다양한 형태로 표현되는 어휘들의 실제 예제의 일부이다. 이 데이터를 살펴보면, '대출'이라는 어휘가 스팸 어휘로 등록되어 차단되는 것을 회피하기 위하여 '대'와 '출' 사이에 다양한 특수 기호들을 삽입하는 방법을 사용하고 있다. 그 이유는 특수 기호를 삽입하더라도 의미를 전달하는 데 아무런 문제가 없기 때문이다. 그런데 특수 기호를 삽입하는 방식은 단순히 기호만 제거하면 해당 문구를 차단할 수가 있기 때문에 스팸 어휘로 차단되는 것을 피하기 위한 다양한 방법들이 고안되었고 'ㄷㅐ'와 같이 자모를 분리한다든지, 'ㄷHㄹ1'와 같이 분리된 자모를 유사한 형태의 영문자, 숫자로 대치하여 스팸으로 차단되지 않도록 하는 교묘한 방법들이 사용되고 있다.

또한, 문자 메시지는 한글 띄어쓰기 규칙을 지키지 않을 뿐만 아니라 스팸 메시지의 경우에 "대 리 운 전"과 같이 고의로 공백들을 삽입하거나 "대.리.운.전"의 예처럼 각 문자들 사이에 문장부호 등 특수문자를 삽입하는 형태로 왜곡하기도 한다. 이와 같이 자모를 분리하거나 유사한 문자로 대치하는 방법에 의하여 하나의 어휘에 대해 매우 다양한 형태의 변형된 문자열 조합이 가능하다. 따라서 이러한 모든 조합의 변형된 문자열을 차단 어휘집에 수록하는 것은 불가능하며, 변형된 문자열을 정규화함으로써 정규화된 어휘가 차단 어휘집에 등록되어 있으면 모든 변형된 문자열을 차단할 수 있도록 할 필요가 있다.

본 연구에서 한국어 문장 복원 시스템은 단순히 문자열 일치 방식에 의해 "대리운전"이 포함된 문자 메시지가 스팸 문자로 차단되는 것을 피하기 위해 "㉦ㅐㄹI우ㄴㅈㅓㄴ"과 같이 다양한 형태로 단어를 변형하더라도 '대리운전'으로 문장을 복원해 줌으로써 매우 다양한 형태로 변형된 문자열들을 모두 차단 문구로 등록하지 않더라도 해당 문자열이 포함된 문자메시지가 스팸 문자로 차단될 수 있도록 하기 위한 것이다.

3. SMS 문장의 정규화 및 키워드 추출

3.1 전처리 및 문자 변환

그림 1은 SMS 문자열 입력으로부터 어휘 정규화 과정을 거쳐 키워드를 추출하는 과정을 기술한 것이다. 전처리 과정은 공백 문자, 특수 기호 등 스팸 차단에 불필요한 기호들을 제거하는 것이고, 2바이트 기호 변환은 괄호 문자, 원 문자 등 2바이트 기호들 중에서 한글 자모를 대치하여 사용한 기호를 해당 한글 자모로 변환하는 과정이다. 영문자/숫자 변환은 영문자 중에서 h/H와 같이 모음 'ㅐ'로 대치될 수 있는 것, 'i/I'와 같이 모음 'ㅣ'로 대치될 수 있는 것 등을 한글 자모로 변환하는 과정이다. 기호 변환 과정이 끝나면 자모로 분해되어 있는 단어를 음절로 조합한다. 정규화된 문장은 공백이 무시되었으므로 자동 띄어쓰기 모듈을 적용한 후에 복합명사 분해 과정을 거쳐서 키워드가 추출된다.

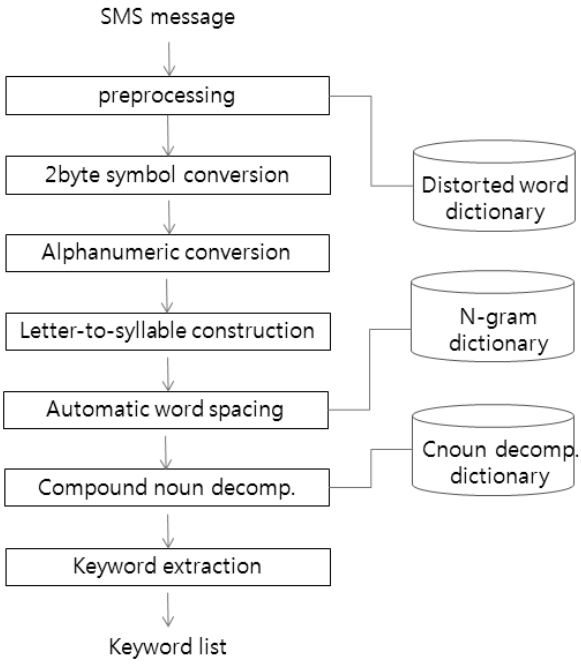

Fig. 1. Keyword extraction by normalization

3.2 SMS 문장의 문구 보정

SMS 문장을 복원하는 첫 번째 단계는 생략 또는 통상적으로 사용되는 변형 문구를 복원하는 것이다. 예를 들어, '캐피탈'을 'ㅋ피탈'이라고 모음 'ㅐ'를 누락시킴으로써 '캐피탈'이라는 키워드에 의해 문자 메시지가 차단되는 것을 회피하려고 한다. 다른 예로는 '거부080'을 '거080', 'x080'과 같이 표현하여 '거부'라는 키워드로 필터링 되지 않도록 하는 경우이다. 이러한 변형 문구들은 사전에 등록하여 변형 문자열 사전에 등록된 문자열을 복원한다.

변형 문자열 사전을 참조하여 입력 문자열을 복원한 후

```
typedef struct sms_char {
    int tag1; // tag for input character
    int tag2; // tag for normalized character

    int ch1;  // input character
    int ch2;  // normalized character
} SMS_CHAR;

typedef struct sms_message {
    char str1[SMS_SIZE]; // input string
    char str2[SMS_SIZE];// normalized string

    int n1; // number of characters in 'str1'
    int n2; // number of characters in 'str2'

    SMS_CHAR sms[SMS_SIZE];
        // internal structure for input
} SMS_TEXT;
```

Fig. 2. Data structure for SMS string normalization

문자열 복원을 위한 다음 단계를 시작하기 전에 입력 문자열을 16비트 정수 형태로 문자열을 초기화한다. 그 이유는 한글 SMS 문자열이 문자 코드집합에 따라 아스키 코드는 1바이트, 한글/한자는 2바이트로 저장되어 문자열 조작하는데 불편한 점들이 있기 때문이다. 그림 2의 구조체는 SMS 문자열 보정을 위한 입력 문자열과 보정된 문자열, 그리고 내부적으로 16비트 정수로 표현하는 데 필요한 자료구조이다. 구조체 SMS_TEXT의 'str1'은 입력 문자열이고, 'str2'는 최종적으로 문구 보정이 완료된 문자열을 저장하기 위한 것이다. SMS 문자열을 내부적인 표현인 SMS_CHAR sms[]에 저장하는 것은 여러 단계에서 문자열 보정을 하는 과정에서 문구 보정 처리의 편의성을 위한 것이다. 구조체 SMS_CHAR는 각 문자에 대해 입력 문자와 보정된 문자를 16비트 정수형으로 저장하여 입력 문자와 보정된 문자를 1:1로 매핑이 되도록 한다. 또한, 입력 문자 및 보정된 문자의 유형을 각각 tag1, tag2에 표시함으로써 핵심어를 추출하는데 활용하고자 한다.

다음 단계는 여러 가지 형태로 변형된 문자들을 정상적인 한글로 보정하는 단계이다. 변형된 문자들은 한글 자음 또는 모음과 유사한 2바이트 기호, 영문자, 숫자 등으로 표현된다. 예를 들어, 자음 'ㅇ'을 '@/o/O/0/(o)'으로 대치하거나 모음 'ㅐ'를 'h/H', 모음 'ㅣ'를 'i/I/l/1'로 대치하는 경우가 많다. 또한, 발음이 영문자와 유사한 음절 '비'는 'b/B', '지'는 'g/G'로 대치하기도 한다.

3.3 자동 띄어쓰기에 의한 키워드 추출

본 연구에서 구현한 문자 메시지 보정 시스템을 이용하여 단계별로 문구가 보정되는 과정을 구체적으로 살펴보면 다음과 같다. SMS 문자 메시지는 초성, 중성, 종성을 모두 분리하는 경우도 있지만 '코ㄹ'과 같이 종성만 분리하거나 또는 복합모음 'ㅘ'를 'ㅗㅏ'로 분해하는 언어 파괴 현상들이 발견되며 '왕'의 경우에 '와ㅇ' 또는 'ㅇㅗㅏ'으로 변형시키기도 한다. 본 논문에서는 다양한 형태의 문구 변형 및 한글을 왜곡한 유형들에 대하여 정상적인 한글 문장으로 복원을 한다. 문구가 보정되면 불필요한 기호들을 제거한 후에 자동 띄어쓰기, 복합명사 분해 과정을 거쳐 최종으로 키워드를 추출한다.

- SMS 입력 문장 예제

  (o)ㅑ(ㅁ)ㅏ토/ㄷㅐㄹI ㅋr드,서B스서β스보다저.렴한ㅇi율

- 문자 보정 결과

  야마토/대리카드,서비스서비스보다저.렴한이율

- 기호 제거 및 자동 띄어쓰기
  야마토대리 카드 서비스 서비스보다 저렴한 이율

SMS 문장을 살펴보면 한글 문장이 음절뿐만 아니라, 자음, 모음, 영문자, 숫자, 특수기호가 포함된 다양한 유형의 문자들로 구성되는 것을 알 수 있다. 자음과 모음이 분리되고 유사한 문자로 대치되어 있는 변형된 문장을 정규화하기 위하여 각 문자들에 대해 유형을 구분하는 태그를 정의하는데 정의된 태그 유형들은 문자 변환 및 음절 구성의 편의성을 목적으로 정의된다.

음절: K, 자음: L, 모음: M, 한자: N, 기호: O
영문: A, 숫자: B, 문장부호: C, 태그없음: _

이와 같이 입력 문자들에 대해 태그를 부착한 이유는 한글 음절과 자모가 분리된 문자들을 구분함으로써 정상적인 한글 음절인지, 자모 변환과 음절 조합 등 문구 보정이 필요한 문자인지 판단하기 위한 것이다. 입력 문장에 대한 태그열과 문구가 보정된 문자열에 대한 태그열을 구성해 보면 다음과 같다.

(o)ㅑ(ㅁ)ㅏ토/ㄷㅐㄹ1 ㅋr드,서B스 서β스보다저.렴한ㅇi율
OMOMKCLMLB LAKCKAK KOKKKKCKKLAK

야마토/대리 카드,서비스 서비스보다저.렴한이율
K_K_KCK_K_ K_KCKKK KKKKKKCKKK_K

휴대폰의 SMS 문자 메시지에서는 띄어쓰기를 무시하는 경우가 매우 빈번하게 발생한다. 그 이유는 초기에 문자 메시지의 요금 부과 기준이 80자 단위로 구분되었던 문자 메시지 사용 환경과도 밀접한 연관이 있다. 따라서 스팸 문자 인식 시스템은 한글 문장의 변형뿐만 아니라 띄어쓰기가 무시되는 현상이 복합되어 있다. 띄어쓰기를 하지 않은 문장들은 스팸 문자 차단 시스템의 성능을 저하시키게 되므로 이에 적합한 자동 띄어쓰기 모듈을 적용한다[10]. 그림 3은 문구 보정된 결과에 대해 자동 띄어쓰기를 위해 사용되는 자료구조이다.

자동 띄어쓰기를 적용하더라도 대부분의 문자 메시지들이 조사, 어미가 생략된 채 명사들만 나열하는 경우가 일반적이다. 자동 띄어쓰기는 띄붙 오류보다 붙띄 오류가 유해성이 높기 때문에 붙띄 오류를 최소화하기 위하여 가급적 공백 삽입을 보수적으로 하는 것이 바람직하다.[1] 붙띄 오류에

```
typedef struct sms_token {
    char str_nobl[SMS_SIZE];
          // input string with no white space
    int n_ch; // number of characters: ignore punc. marks
    char tag_sp[SMS_SIZE]; // 'c': char., '2': 2nd char. of
DBCS, '.': punc. symbols
    char tag_bl[SMS_SIZE]; // blank location marks by
automatic word spacing
    char wlen[SMS_SIZE]; // byte length of a word in each
position in str_nobl

    int n_word; // no. of words in a sentence
    char *wp[MAX_WORDS]; // start position of word: pointer
to str_nobl
    int wl[MAX_WORDS]; // byte-length: word

    int n_token; // number of tokens
    char *tp[MAX_WORDS]; // token beginning postition:
pointer to str_nobl
    int tl[MAX_WORDS]; // byte-length: token
    char pos[MAX_WORDS]; // POS of token
} SMS_TOKEN;
```

Fig. 3. Data structure for auto-spacing and tokens

의해 한 어절이 두 어절로 분리되는 부작용을 최소화하기 위한 전략으로 인하여 많은 복합명사들이 단위명사로 분리되지 않게 된다. 따라서 자동 띄어쓰기를 적용한 후에 다시 복합명사 분해 모듈을 통해 복합명사를 분해하여 최종적으로 문장을 복원하는 과정을 마무리한다[11]. 그 결과로부터 키워드들을 추출하고 추출된 키워드 중에서 스팸 어휘집에 등록된 키워드가 탐지되면 스팸 문자로 간주하게 된다.

이때, 스팸 어휘집에 등록된 어휘가 포함된 문자들을 모두 스팸으로 간주하는 것은 아니다. 예를 들어, “김대리, 오늘 회식은 회사 신용카드로 결제하세요”라는 문자 메시지에서 ‘대리’, ‘카드’ 때문에 스팸으로 간주하면 안 된다. 일반적으로, 스팸 차단 시스템은 대량의 문자 발송 등 스팸이 의심스러운 문자 메시지에 대해 스팸 차단 기능을 적용한다. 또한, white 리스트와 black 리스트를 반영하고 있으며, 특히 자신의 주소록에 저장된 번호로 발신된 문자 메시지는 스팸으로 처리하지 않도록 하는 방법을 적용하여 false-positive 오류(false alarm)를 최소화하는 방법을 사용한다[2].

4. 실험 및 결과 분석

변형된 SMS 문자열의 정규화 효용성을 실험하기 위하여 휴대전화를 통해 전송되는 문자 메시지에서 스팸 문자로 차

1) ‘띄붙 오류’는 ‘가고있다’와 같이 띄어써야 하는데 붙여쓴 오류이며, ‘붙띄 오류’는 ‘검색시 스템’의 ‘시’와 ‘스’ 사이와 같이 붙여야 하는데 띄어쓴 오류이다. 띄붙 오류는 하나의 문자열에 2개의 어절이 연결되어 그 의미를 파악하는데 어려움이 없으나, ‘붙띄 오류’는 한 어절이 2개의 무의미한 어절로 분리되어 그 의미가 훼손되는 문제점이 발생한다.

단되는 문자 유형을 분석하였다. SMS 문자의 스팸 문자를 차단 방법은 크게 등록된 발신 번호와 회신 번호를 차단하는 '번호 차단' 기법과 등록된 문구를 차단하는 '문구 차단' 방식으로 구분된다. '번호 차단' 방식과 '문구 차단' 방식은 다시 모든 사용자들에게 공통적으로 적용되는 것과 각 개인이 등록한 것만 차단하는 것으로 구분된다.

실험 대상 문자는 61,934,003개이고 그 중에서 937,683개가 스팸 문자로 분류되었다. 스팸으로 분류된 문자들을 '번호 차단' 방식과 '문구 차단' 방식, 그리고 '공통 차단'과 '개인 차단'으로 분류하면 표 1과 같다. 일반 문구에서 차단되지 못한 스팸 문자 중에서 본 논문에서 제안한 SMS 문장의 정규화 기법을 적용함으로써 스팸으로 차단된 것은 137,221개로 전체 스팸 문자의 14.6%이다. 이는 SMS 문자열 정규화를 하지 않았을 때 차단되는 스팸 문자 개수가 800,442개인데 비해 문자열 정규화를 통해 937,683개로 증가하여 스팸 문자 차단 효과가 17.1% 향상되는 효과가 있었다.

Table 1. SMS spam filtering result

	Number Filtering	Normal Filtering	Transformed filtering
Common	7.7% (71,890)	43.1% (404,307)	12.5% (117,271)
Personal	19.0% (178,160)	15.6% (146,105)	2.1% (19,950)
Total	26.7% (250,040)	58.7% (550,412)	14.6% (137,221)

5. 결 론

광고성 문자들은 문자 메시지가 자동으로 차단되는 것을 회피하기 위해 한글 문장을 다양한 형태로 변형하거나 왜곡시키고 있다. 이러한 문자 메시지를 자동으로 차단하려면 변형되거나 왜곡된 문장들을 정상적인 한글 문장으로 정규화해야 한다. 본 논문에서는 광고성 문자를 자동으로 인식하여 차단하기 위해 변형된 문자열을 정상적인 문자열로 정규화하는 방법을 제안하였다. 변형되거나 왜곡된 광고성 문자 메시지를 정상적인 한글 문장으로 변환하여 정규화하고 정규화된 문장으로부터 자동 띄어쓰기 및 복합명사 분해 과정을 거쳐 키워드를 추출하였으며, 이 방법을 적용하여 변형된 문자열을 정규화함으로써 스팸 문자 차단 효과를 17.1% 향상시키는 효과가 있었다.

Reference

[1] B. Y. Kim, *A Study on the Morphological Characteristics of Communicative Languages by the Statistical Frequency*, Master Thesis, Kookmin University, 2002.

[2] S. J. Lee and D. J. Choi, "Personalized mobile junk message filtering system," *Journal of the Korea Contents Association*, pp.122-135, 2011.

[3] S. S. Kang, "Junk-mail filtering by mail address validation and title-content weighting," *Journal of the Korea Multimedia Society*, Vol.9, No.2, pp.255-263, 2006.

[4] K. Tretyakov, "Machine learning techniques in spam filtering," *Data Mining Problem-oriented Seminar*, MTAT. 03. 177, pp.60-79, 2004.

[5] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques," *ACM Transactions on Asian Language Information Processing(TALIP)*, Vol.3, No.4, pp.243-269, 2004.

[6] C. Brutlag and J. Meek, "Challenges of the email domain for text classification," Proceedings of the 17th International Conference on Machine Learning, 2000.

[7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk E-mail," *Proceedings of the AAAI Workshop*, pp.55-62, 1998.

[8] M. Salib, "MeatSlicer: Spam classification with Naive Bayes and smart heuristics," *Proceedings of the Spam Conference*, MA, Jan., 2003.

[9] K. Schneider, "A comparison of event models for Naive Bayes anti-spam E-mail filtering," *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL 2003)*, pp.307-314, 2003.

[10] S. S. Kang and K. B. Hwang, "A language independent n-gram model for word segmentation," *Proceedings of AI'2006*, pp.557-565, 2006.

[11] S. S. Kang, "A decomposition algorithm of Korean compound nouns," *Journal of KIISE(B)*, Vol.25, No.1, pp.172-182, 1998.

강 승 식

e-mail : sskang@kookmin.ac.kr

1986년 서울대학교 정보컴퓨터공학부 (학사)

1988년 서울대학교 컴퓨터공학과 (석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년~2001년 한성대학교 정보전산학부 교수

2001년~현 재 국민대학교 컴퓨터공학부 교수

관심분야 : 한국어정보처리, 자연어처리, 정보검색, 기계학습