형태소 단위 자질을 이용한 콘텐츠 기반 한국어 SMS 스팸 필터링

손대능[0], 신중휘, 이정태, 이승욱, 임해창
고려대학교 컴퓨터.전파통신공학과
{danny, jhshin, jtlee, swlee, rim}@nlp.korea.ac.kr

Contents-Based Korean SMS Spam Filtering Using Morpheme Unit Features

Dae-Neung Sohn[0], Joong-Hwi Shin, Jung-Tae Lee, Seung-Wook Lee, Hae-Chang Rim
Dept. of Computer and Radio Communications Engineering, Korea University

요 약

본 논문에서는 형태소 분석을 이용한 확률 기반 한국어 SMS 스팸 필터링 기법을 제안한다. 기존 연구에서는 단어 및 문자 단위 어휘 정보를 자질로 이용한 영어 및 스페인어 SMS 스팸 필터링 방법들이 있다. 하지만 교착어인 한국어의 경우, 어근과 접사의 조합에 의해서 다양한 어절이 형성될 수 있다. 따라서 어절단위 어휘 정보를 자질로 사용할 경우, 미등록어(out of vocabulary) 문제가 발생한다. 특히, 매우 적은 수의 단어들로 구성된 SMS 메시지의 경우에는 이 문제가 매우 심각하다. 본 논문에서는 형태소 분석을 이용하여 이러한 문제점을 해결하고자 하였다. 실험 결과, 제안하는 방법은 기존 연구와 비교하여 10.6%의 스팸 분류 정확률 향상을 보였다. 또한 미등록어만을 포함하는 SMS 메시지의 수는 약 77% 감소하였다.

1. 서 론

휴대전화의 대중화와 더불어 사용자들 사이에서 SMS(단문 메시지 서비스)기능은 일상생활에서 커뮤니케이션의 하나의 도구로 비중 있게 자리 잡고 있다. 이를 악용하여 각종 성인광고, 대출광고, 게임 광고를 포함한 스팸 메시지가 불특정 다수에게 발송되고 있는 실정이다[1]. 이러한 SMS 스팸은 받는 사람으로 하여금 불쾌감을 유발하고 불필요한 정보를 제공함으로써 휴대전화 사용자들의 불편을 가중시킨다. 이로 인해 SMS 스팸이 사회적 이슈가 되고 있으며, 이를 방지하고자 각 이동통신사업자들은 2006년부터 휴대전화 번호 당 SMS 문자메시지 1일 발송량을 1,000통으로 제한하고, 스팸 트랩시스템을 두어 스팸 발송 번호를 차단하고 있다. 그럼에도 불구하고 2007년 상반기 1인당 일평균 0.54통의 SMS 스팸을 수신하고 있다고 통계 조사에서 밝혀졌다[1]. 이러한 방지책들은 주로 제도적 관점이기 때문에 SMS 콘텐츠 자체에 기반을 둔 SMS 스팸 필터링 기법이 필요한 실정이다.

스팸 필터링의 필요성은 전자우편 서비스 분야에서 먼저 대두되었으며, 콘텐츠 기반 문서분류(Text classification) [2][3] 관점에서 접근한 기법들이 주를 이룬다. 가장 잘 알려진 기법 중에 하나인 베이지안 분류기를 이용한 콘텐츠 기반 방법[4][5][6]은 전자우편이 담고 있는 단어 등의 어휘를 기본자질로 사용하였으며 좋은 성능을 보여주었다.

SMS 스팸 필터링 분야에서는 기존에 전자우편 스팸 필터링에서 사용하던 기술들을 영어와 스페인어 SMS 스팸 필터링에 적용하고자 한 연구가 있었다[7][8]. 이 연구들에선 SMS 메시지의 길이제한과 언어의 형태변형(Morphological variance)으로 인한 정보부족문제를 고려해 콘텐츠 상의 단어자질 뿐만 아니라 더 작은 단위에서 단어의 의미를 추출하기 위해 문자-bigram(Character-bigram)과 문자-trigram(Character-trigram)등을 추가로 사용하였으며 영어와 스페인어 SMS 스팸 필터링에서 이것이 유용하다는 것을 보였다.

그러나 한국어 SMS 스팸 필터링에 있어서 SMS 콘텐츠의 단어 자질을 그대로 사용하는 것은 적합하지 않다. 한국어는 교착어(膠着語)에 속하며, 문장 내에서 각 단어는 어근에 접사가 결합되는 형태로 표현된다. 이는 영어와는 달리 수많은 어휘조합 생성이 가능하다는 것을 의미한다. 더군다나 SMS 메시지의 길이제한과 통신어의 다양성을 고려해 볼 때 미등록어가 나타날 가능성이 높다. 또한 한국어는 단어보다 더 작은 단위에서 의미를 나타내는 형태소가 존재하기 때문에 영어와 스페인어 SMS 스팸 필터링에서 같은 역할로 쓰였던 문자-bigram이나 문자-trigram은 한국어에 적합하지 않을 수 있다.

이러한 한국어의 특징을 고려하여, 본 논문에서는 형태소 단위 자질을 이용한 SMS 스팸 필터링 기법을 제안한다. 검증을 위해 기존에 영어와 스페인어에서 쓰였던 자질들과 형태소 단위 자질을 자체 구축한 실제 한국어 SMS 메시지 말뭉치를 이용해 비교 실험하였다. 아울러 본 논문은 한국어 SMS 스팸 필터링에 콘텐츠를 이용한

확률 기반 분류모델을 적용한 최초의 논문으로 사료된다.

본 논문의 구성은 다음과 같다. 2장에선 기존 영어와 스페인어 SMS 스팸 필터링에 관한 연구를 중점적으로 살펴본다. 3장에선 본 연구에서 제안하는 형태소 단위 자질 기반 한국어 SMS 스팸 필터링에 대한 설명과 분류모델로 사용하는 최대 엔트로피 모델에 대해 소개한다. 4장에선 실험 데이터와 실험 과정, 실험 결과 및 분석에 대해 기술한다. 5장에선 결론 및 향후연구에 대해 언급한다.

2. 관련 연구

본 연구와 가장 관련이 있는 연구는 콘텐츠 기반 SMS 스팸 필터링 연구이다[7]. 영어와 스페인어를 대상으로 한 이 연구는 기존 전자우편 스팸 필터링 연구들에서 쓰인 콘텐츠기반 방법을 SMS 스팸 필터링에 적용하기 위해 어떤 요소들을 고려해야 하는지를 제안하였다. 기본적으로 전자우편 스팸 필터링과 SMS 스팸 필터링은 문서분류 관점에서 단어를 주요자질로 사용하는 공통점을 지닌다. 그러나 SMS 메시지의 경우 160바이트의 길이제한과 통신언어 표현의 다양성 때문에, 전자우편에 비해 스팸 필터링 시 필요한 정보가 부족하다고 볼 수 있다. 이를 해결하고자 SMS 메시지의 콘텐츠를 나타내는 토큰(Token)으로 단어뿐만 아니라 문자-bigram, 문자-trigram, 단어-bigram을 함께 사용하였다. 각각에 대한 설명은 아래와 같다.

- 단어: 영어 알파벳과 숫자를 포함한 문자열을 말하며 그 이외의 공백이나 특수기호 등은 단어의 경계역할을 하는 것으로 본다.
- 문자-bigram과 문자-trigram: 단어 중 연속된 2개 또는 3개의 문자로 이루어진 문자열을 말한다. 예를 들면, 영어문장 “the quick car”는 문자-bigram “th”, “he”, “e_”, “_q”, “qu”, “ui”, “ic”, “ck”, “k_”, “_c”, “ca”, “ar”를 포함하고 있다. 문자-bigram과 문자-trigram은 단어의 형태변형에 견고하게 만들기 위해 사용하는 것이다.
- 단어-bigram: 현재 단어를 포함하고 이 단어의 문장 내 위치를 기준으로, 다음에 오는 단어들 중 윈도우 사이즈 5안에 있는 다른 단어와의 조합을 의미한다. 단어-bigram의 예를 들면, 영어문장 “ the quick blue car”는 “the quick”, “the blue”, “the car”, “quick blue”, “quick car”, “blue car”를 포함하고 있다.

[7]에서는 위에서 언급한 토큰화(Tokenization) 방법이 SMS 스팸 필터링에서 매우 중요하며, 콘텐츠 기반 전자우편 스팸 필터링 기법이 영어와 스페인어를 대상으로 한 SMS 스팸 필터링에도 적용 될 수 있음을 보였다.

SMS 메시지의 길이에 비추어 볼 때 SMS 스팸 필터링은 단문 분류 문제(Short Text Classification Problem)와도 관련이 있다고 볼 수 있다. [9]에서는 단문 분류 시 정보 부족 문제를 해결하기 위해 트레이닝 데이터와 관련되어 있는 원시배경지식(Unlabeled Background Knowledge)을 사용하였다. 예를 들어, 자연과학관련 논문의 제목을 분류하는 작업이 있으면, 원시배경지식으로 자연과학논문들의 초록을 사용 수 있을 것이다. 그러나 이 방법은 분류 대상에 따라 그에 적합한 원시배경지식을 사용해야 한다는 단점이 있다. 또한 SMS 메시지처럼 특별한 주제와 관련 없이 일상의 대화가 주를 이루는 문장에 대해서는 원시배경지식으로 무엇을 사용해야 할 것인가에 대한 기준이 모호해지므로 이 방법은 SMS 스팸 메시지 필터링에 적합하지 않다.

3. 제안하는 방법

본 논문에서는 형태소 단위 자질을 이용한 콘텐츠 기반 한국어 SMS 스팸 필터링을 제안한다. SMS 스팸의 기준은 개인마다 다를 수 있지만, 본 논문에서 상업광고 성격이 짙은 SMS 메시지를 SMS 스팸으로 정의하며 자세한 기준은 4장에서 언급한다.

3.1 SMS 메시지 특징

관련 연구에서 언급하였듯이 SMS 메시지의 토큰화는 SMS 스팸 필터링에서 매우 중요한 부분을 차지한다[7]. 여기서 SMS 메시지의 토큰화란 메시지가 포함하고 있는 전체 콘텐츠를 의미적으로 더 작은 단위로 나누는 것을 말한다. 예를 들어 문장을 단어나 음절-bigram, 음절-trigram 단위로 나누는 것을 들 수 있다. 이러한 토큰들은 분류모델의 입력으로 모델을 구축하는데 사용되기 때문에 성능에 매우 큰 영향을 끼친다. 따라서 성능향상에 도움이 되는 토큰 추출을 위해 SMS 메시지의 특징을 살펴볼 필요가 있다.

- 문자메시지의 길이 제한: 표준 SMS 메시지의 길이는 최대 160바이트로 제한되어 있다. 이는 곧 SMS 메시지 분류 시 사용 가능한 정보가 전자우편이나 길이가 긴 문서들에 비해 적다는 것을 의미한다.
- 통신어적인 특성: SMS 메시지의 대부분이 모바일 기기에서 생성되는 것이어서 그 특성 상 띄어쓰기 오류 및 입력 편이를 위한 의도적 무시가 빈번하다. 또한 언어파괴, 변형 현상을 볼 수 있으며 이모티콘, 특수기호의 사용도 관찰 된다.

3.2 형태소 단위 자질

한국어의 경우 위와 같은 SMS 메시지의 일반적인 특징뿐만 아니라 교착어인 한국어의 특성이 맞물려 SMS 스팸 필터링을 더욱 어렵게 한다. 어근에 접사가 결합되어 단어조합이 가능한 한국어는 그 특성상 수많은 어휘생성이 가능하다. 또한, SMS의 길이 제한으로 인한 어휘정보부족으로 미등록어가 나타날 가능성이 높다. 기존 연구[7]에선 정보부족문제 해결을 위해 단어자질 뿐만 아니라 더 작은 단위에서 단어의 의미를 추출하기 위한 음절-bigram이나 음절-trigram을 함께 사용했다. 그러나 한국어에서는 단어보다도 더 작은 단위에서 의미를 나타내는 형태소가 존재하고, 문장이나 단어에서 형태소

를 추출하는 고도화된 툴[10]이 있기 때문에 본 연구에선 이를 활용한다. 형태소를 이용함으로써 기대되는 효과는 아래와 같다.

• 형태소 단위 자질은 스팸 SMS 메시지와 햄(정상) SMS 메시지를 구분하는데 중요한 자질역할을 하는 어휘의 정확한 빈도 및 분포 추정에 도움을 준다. 형태소 대신 단어를 그대로 사용하면 <표 1>에서의 예시처럼 "연락" 이란 하나의 어근에 여러 종류의 접사가 붙어 각각이 새로운 어휘가 된다. 이로 인해 실질적 의미를 지니는 "연락" 이라는 어휘가 스팸과 햄 SMS 메시지에서 어떻게 분포되어 있는지를 정확히 추정하기 어려워진다. 하지만 형태소만을 이용하면 "연락" 뒤에 어떤 접사가 오더라도 "연락" 이란 어휘 정보를 추출해 낼 수 있어 이러한 문제점이 줄어든다. 정확한 어휘 분포를 추정함으로써, 분류 모델의 성능향상에도 도움을 줄 것이다.

<표 76>어근-접사 조합과 형태소 정보의 예

어근-접사	형태소
연락주세요	**연락**+주+시+어요
연락줄래	**연락**+주+ㄹ래
연락주실거죠	**연락**+주+시+ㄹ+거+이+죠
연락줘	**연락**+주+어

• 미등록어를 줄이는 데에 도움을 준다. 그 이유로 크게 두 가지로 볼 수 있는데 첫 째로 의미를 갖는 최소 단위인 형태소는 단어보다 더 작은 단위이므로 단어의 활용형 매칭에 대해 견고하기 때문이다. 두 번째는 토큰화 과정에서 생성되는 토큰의 수가 기존연구[7]의 단어, 문자-bigram, 문자-trigram, 단어-bigram을 포함한 토큰 수보다 월등히 작기 때문이다. 이는 <표 2>[1]에서 보는 바와 같이 전체 SMS 메시지를 표현하는데 필요한 어휘의 개수가 줄어든다는 것을 의미한다.

<표 77> SMS 메시지에서의 자질 구성 별 생성된 토큰 수

자질 구성	생성된 총 토큰 수
기존연구[7]	365,069
제안하는 방법	13,622

3.3 SMS 스팸 필터링 모델

본 연구에서는 최대 엔트로피모델[11] 기반으로 한국어 SMS 스팸 필터링 모델을 구축하였다.[2] 주어진 SMS 메시지를 x라고 하고, 이 메시지의 분류를 $y=\{spam, ham\}$, 즉 스팸과 햄으로 분류 될 수 있다고 하자. 본 SMS 스팸 필터링 모델의 목적은 SMS 메시지 x가 주어졌을 때 스팸 분류에 속할 확률인 조건부 확률 $p(y=spam|x)$를 구하는 것이다. 최대엔트로피 모델[3]을 이용하면 이 조건부 확률 $p(y|x)$는 다음 수식과 같이 계산 될 수 있다.

$$p(y|x) = \frac{1}{Z} exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right]$$

$f_i(x,y)$: i 번째 자질
λ_i : i 번째 자질의 가중치
Z : 정규화 인수

최대 엔트로피 모델은 조건부 확률을 추정할 시에 자질로 사용되는 정보들을 통합하여 사용가능한 장점이 있다. 전자우편 스팸 필터링 분야에서 이미 베이지안 분류기보다 최대 엔트로피 모델이 좋은 성능을 보여주었던 관련 연구가 있으며[12], 문서 분류 작업 시에도 어느 정도 견고한 성능을 낸다고 알려져 있다[13].

4. 실험

4.1. 실험 데이터

본 연구에서는 실제 SMS 메시지를 수집하여 실험에 사용하였다. 전화번호나 계좌번호 같은 개인 정보는 모두 숫자 '0' 으로 대체하거나 삭제하였다. 스팸과 햄 여부는 각 SMS 메시지가 포함하고 있는 콘텐츠를 보고 상업광고메시지[4]인지의 여부에 따라 수작업으로 분류하였다. 실험을 위해 총 20,000개의 SMS 메시지를 이용하여 이 중 18,000개를 학습 집합으로, 2,000개를 실험 집합으로 사용하였다. 학습 집합과 실험 집합에서의 햄과 스팸의 구성은 9:1의 비율로 이루어져 있다.

4.2 SMS 메시지 토큰화

학습과 실험을 위해 각각의 SMS 메시지에 대한 전처리 및 토큰화가 이루어져야 한다. 전처리 단계에서는 한글 SMS 메시지에서 한글과 영어 알파벳, 숫자를 제외한 모든 문자들을 공백으로 대체하였다. 이렇게 처리된 한글 SMS 메시지를 철자오류교정모델[14], 띄어쓰기교정모델[15]을 이용해 오류 교정 과정을 거친다. 교정을 거친 SMS 메시지는 이후 본 연구가 제안하는 형태소 단위 자질 기반 SMS 스팸 필터링과 기존연구[7]간의 성능비교를 위해 다음에 언급한 토큰들로 표현된다.

•단어: 한글과 영어 알파벳, 숫자를 포함한 문자열을

1) 표에 표기된 총 토큰 수는 본 연구를 위해 구축한 실험 데이터를 이용해 추출한 것으로 토큰화 과정은 4.2절에서 기술한다.

2) 본 연구는 형태소 단위 자질이 기존 영어와 스페인어 SMS 스팸 필터링에서 사용했던 자질들 보다 한국어에 효과적임을 보이기 위한 것임으로, 분류 모델로 무엇을 사용하느냐는 중요하지 않다.

3) 본 연구는 필터링 모델 구축 시 Zhang Le의 최대 엔트로피 모델(http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html#intro) 툴킷을 사용하였다.

4) 성인광고, 대리운전광고, 게임광고 등 이윤을 목적으로 상품이나 서비스를 소개하는 SMS 메시지.

지칭하며 공백은 단어의 경계역할을 하는 것으로 본다.

•음절-bigram: 영어의 문자-bigram과 동일한 역할을 하는 것으로써 연속된 두 음절이 해당된다. 예를 들면, 문장 "휴대폰 메시지"는 음절-bigram "휴대", "대폰", "폰_", "_메", "메시", "시지"을 포함하고 있다.

•음절-trigram: 영어의 문자-trigram과 동일한 역할을 하는 것으로써 연속된 세 음절이 해당된다. 예를 들면, 문장 "휴대폰 메시지"는 음절-trigram "휴대폰", "대폰_", "폰_메", "_메시", "메시지"를 포함하고 있다.

•단어-bigram: 특정 크기의 윈도우 내에 출현한 단어들의 모든 가능한 두 단어의 조합을 말한다. 예를 들면, "나는 배가 너무 고프다" 문장의 경우 "나는 배가", "나는 너무", "나는 고프다", "배가 너무", "배가 고프다", "너무 고프다"를 포함하고 있다.

•형태소: 전처리를 거친 SMS 문자 메시지에서 형태소 분석기[10]를 사용해 추출한 형태소를 지칭한다.

4.3 자질 구성 및 선택

본 연구에서는 기존연구[7]와 제안하는 형태소 단위 자질 기반의 SMS 스팸 필터링 방법을 비교하기 위해 두 가지의 자질 구성을 사용하여 실험을 진행하였으며 각각은 <표3>에 나타나 있다.

<표 78> 자질 구성

기존연구[7]	단어, 음절-bigram, 음절-trigram, 단어-bigram
제안하는 방법	형태소

각각의 자질 구성은 많은 수의 토큰들을 포함하고 있으므로, 실험에는 이들 중 정보이득(Information gain)이 높은 것부터 SMS 스팸 필터링 시 도움이 되는 것으로 간주하고 사용하였다[16][17]. 정보이득에 의한 자질 선택은 토큰 수를 줄여주면서 분류 정확도에 큰 해를 끼치지 않는 것으로 알려져 있다.

4.4 실험 평가 방법

각 자질 구성 별 정보이득 값이 높은 토큰을 최대 2,000개까지 사용해 학습 및 실험을 진행하였다.

성능 평가에는 <표 4>를 참고해, 전자우편 스팸 필터링 분야에서 주로 이용하는 <표 5>에 기술한 방법들을 사용한다.

<표 79> 평가 시 사용되는 지표

이름	의미
TP (True Positive)	스팸을 스팸으로 분류한 횟수
TN (True Negative)	햄을 햄으로 분류한 횟수
FP (False Postivie)	햄을 스팸으로 분류한 횟수
FN (False Negative)	스팸을 햄으로 분류한 횟수

<표 80> 평가 척도

Precision	*TP* / (*TP* + *FP*)
Recall	*TP* / (*TP* + *FN*)
F_1-*measure*	2 x *Precision* x *Recall* / (*Precision* + *Recall*)
False-positive rate	*FP* / (*FP* + *TN*)
Unseen rate	(실험 집합에서 미등록어로만 구성된 SMS 메시지의 총 개수) / (실험 집합의 SMS 메시지 총 개수)

•*Precision* : 스팸을 스팸으로 정확히 분류한 것의 비율이다.

•F_1-*measure* : 분류 모델의 *Precision*과 *Recall*을 1:1 비율로 고려해 종합한 평가방법이다.

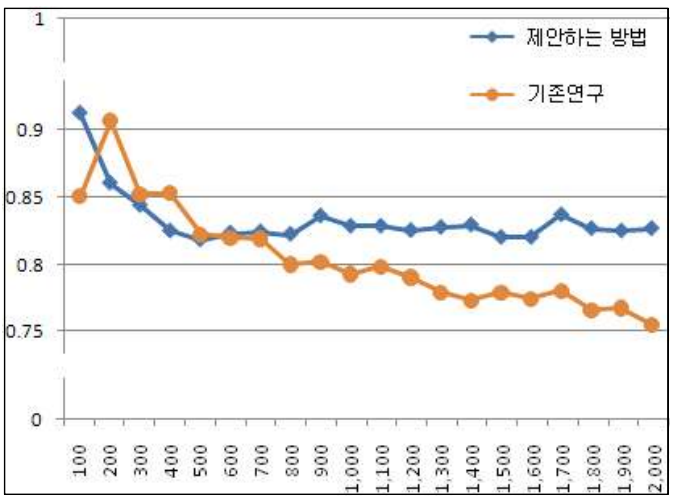
•*False-positive rate* : 햄을 스팸으로 잘못 분류한 것의 비율이다[7].

•*Unseen rate* : 실험 집합에서 미등록어로만 구성된 SMS 메시지의 비율을 뜻한다. 형태소 단위 자질을 이용하는 제안하는 방법이 기존연구[7]에 비해 미등록어를 줄이는데 얼마나 기여했는지 평가하기 위함이다.

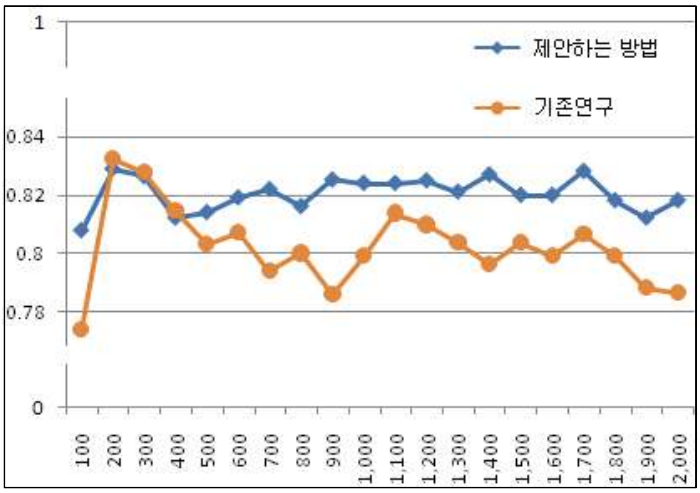
4.5 실험 결과 및 분석

앞서 언급한 바와 같이 기존연구의 방법[7]과 제안하는 방법을 비교하기 위해 두 가지의 자질 구성을 사용하여 실험을 진행하였다. 각각의 그림에는 4.3절의 <표 3>의 자질 구성을 사용한 실험 결과가 나타나 있다. 실험결과는 스팸 분류 모델의 성능을 평가하는 부분과 *Unseen rate*의 결과를 분석하는 부분으로 나누어 설명한다.

• 스팸 분류 모델의 성능

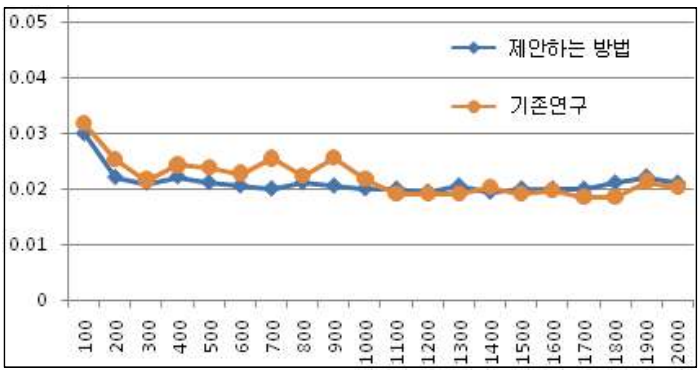


<그림 203> 토큰 개수 변화에 따른 *Precision*



<그림 204> 토큰 개수 변화에 따른 F_1-measure

<그림 1> 과 <그림 2>는 토큰 개수별 *Precision* 과 F_1-measure를 나타내고 있다. 대체적으로 제안하는 방법의 성능이 높게 나옴을 볼 수 있다. 이는 한국어 SMS 스팸 필터링의 경우 기존연구[7]에서 사용하였던 자질 구성보다 형태소를 사용하는 것이 더 효과적임을 입증하는 결과이다.

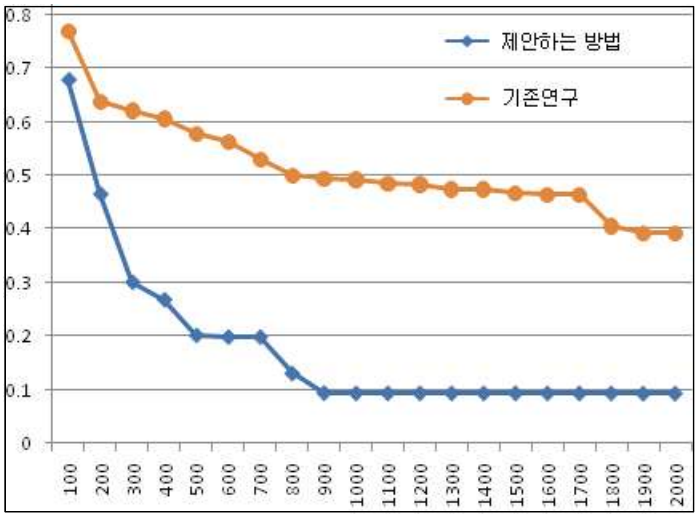


<그림 205> 토큰 개수 변화에 따른 *False-positive rate*

<그림 3>은 토큰 개수 별 *False-positive rate*을 나타내고 있다. 이 결과에 의하면 두 자질 구성 간 뚜렷한 차이는 보이지 않는다. 햄을 스팸으로 잘못 분류한 경우, 모바일 기기 사용자들에게 큰 불편을 초래할 수 있으므로, 향후 연구에서 개선되어야 할 부분으로 판단된다.

- *Unseen rate* 비교

<그림 4>의 결과에서 볼 수 있듯이 스팸 분류에 제안하는 방법인 형태소를 이용할 경우 미등록어 감소하여, 기존연구[7]의 자질 구성보다 분류가 불가능한 SMS 메시지가 현저히 줄어들었다. 토큰 개수를 2,000개 사용했을 때 제안하는 방법의 *Unseen rate*는 0.09(9%)이다. 같은 상황에서 기존연구[7] 자질 구성의 *Unseen rate*는 0.39(39%)인데 이는 총 1,800개의 실험 집합을 구성하는 SMS 메시지 중 702개를 제대로 표현하지 못했다는 것을 의미한다. 이러한 결과는 기존연구[7]에서 사용했던 토큰들이 한국어 SMS 메시지를 표현하는데 있어서 형태소에 비해 적합하지 않다는 것을 보여준다.



<그림 206> 토큰 개수 변화에 따른 *Unseen rate*

각 자질 구성별 토큰 개수가 2,000개일 때의 각 평가방법들의 결과수치를 종합하면 <표 5>와 같다.

<표 81> 각 자질 구성 별
토큰 개수가 2,000개일 때의 성능

평가방법	기존연구[7]	제안하는 방법
Precision	0.75	0.83(+10.6%)
F_1-*Measure*	0.78	0.82(+5.1%)
False-Positive rate	0.02	0.02(+0.0%)
Unseen rate	0.39	0.09(-77.0%)

5. 결론

본 논문에서는 한국어의 특징을 고려하여 형태소 단위 자질을 이용한 콘텐츠 기반 한국어 SMS 스팸 필터링 기법을 제안하였다. 실제 한국어 SMS 메시지를 수집하여 실험 데이터를 구축하였고, 기존의 영어와 스페인어에서 쓰였던 콘텐츠 기반 SMS 스팸 필터링 기법과 비교실험을 수행하여 한국어 SMS 스팸 필터링에는 형태소를 이용하는 것이 효과적임을 보였다. 한국어 SMS 스팸 필터링 시 형태소로 구성된 토큰을 사용하여 교착어의 특성으로 인해 생기는 문제점과 미등록어를 줄이면서 실제 스팸 분류 시 기존연구[7]의 방법에 비해 성능향상을 보인 것은 상당히 의미 있는 결과라고 판단된다. 아울러 본 논문은 한국어 SMS 스팸 필터링에 콘텐츠를 이용한 확률 기반 분류모델을 적용한 최초의 논문으로 사료되며 그 자체로도 의미를 지니고 있다.

또한, 본 연구에서 제안한 방법은 짧은 길이와 언어변형현상이 심한 콘텐츠에 기반을 두고 있기 때문에 인터넷 포탈사이트나 뉴스기사의 댓글 자동분류에도 사용 할 수 있다. 향후 연구로는 콘텐츠상의 특수기호나 전화번호같은 비텍스트 정보를 스팸 필터링에 이용하는 것을

고려해 볼 수 있다. 햄을 스팸으로 잘못 분류하는 오류를 최소하고 분류 정확도를 높이기 위해 자동으로 햄과 스팸을 판단하는 규칙을 추출해 스팸 필터링에 이용하는 방법을 고안하고자 한다.

감사의 글

이 논문은 2008년도 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2006-000-11162-0).

참고문헌

[1] 정보통신부, "이메일 스팸 계속 감소 추세", 뉴스와 이어, http://www.eenewsfeed.com/redirect/feed-item/281125, 2007.

[2] D. Lewis. "(spam vs.) forty years of machine learning for text classification", In Proceedings of the Spam Conference, http://www.daviddlewis.com/publications/slides/lewis-2003-0117-spamconf.html, 2003.

[3] F. Sebastiani, "Machine learning in automated text categorization", ACM Computing Surveys, Volume 34, Issue 1, pp. 1 - 47, 2002.

[4] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk email", In Proceedings of AAAI-98 Workshop on Learning for Text Categorization, pp.55-62, 1998.

[5] G. Paul, "Better Bayesian Filtering", Proceedings of the 2003 Spam Conference, http://www.paulgraham.com/spam.html, 2003.

[6] Yerazunis, W.S., "The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past It", MIT Spam Conference, http://crm114.sourceforge.net/docs/Plateau_Paper.html, 2004.

[7] Gómez, J.M., Cajigas, G., Puertas Sanz E., and Carrero García,F. "Content Based SMS Spam Filtering", Proceedings of the 2006 ACM Symposium on Document Engineering, pp. 107-114, 2006.

[8] Gordon V. Cormack , José María Gómez Hidalgo , and Enrique Puertas Sánz, Feature engineering for mobile (SMS) spam filtering, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 871-872, 2007.

[9] Zelikovitz, S., and Hirsh, H., "Improving short-text classification using unlabeled background knowledge to assess document similarity", In Proceedings of the Seventeenth International Conference on Machine Learning (ICML), pp. 1183-1190, 2000

[10] 이상주, 류원호, 김진동, 임해창, "품사태깅을 위한 어휘문맥 의존규칙의 말뭉치기반 중의성주도 학습", 한국정보과학회 논문지(B), 제26권, 제1호, pp.178-189, 1999.

[11] A. L. Bergerm, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing", Computational Linguistics, Vol.22, No.1, pp. 39-71, 1996.

[12] L. Zhang, and T. Yao. "Filtering junk mail with a maximum entropy model". In Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03), pp. 446-453, 2003.

[13] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification", In IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67, 1999.

[14] J.H. Byun, S.Y. Park, and H.C. Rim, "Automatic Spelling Correction Rule Extraction and Application for Spoken-style Korean Text", Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), pp. 195-199, 2008

[15] D.G. Lee, H.C. Rim, and D.S. Yook, "Automatic Word Spacing Using Probabilistic Models Based on Character n-grams", IEEE Intelligent Systems, Vol. 22, No. 1, pp. 28-35 2007.

[16] Yang Y., "An evaluation of statistical approaches to text categorization", Information Retrieval, No. 1(1/2), pp. 69-90, 1999.

[17] Yang Y., and J.O. Pedersen, "A comparative study on feature selection in text categorization", En Proceedings of the 14th International Conference on Machine Learning, pp. 412-420, 1997.