MIKE: An Interactive Microblogging Keyword Extractor using Contextual Semantic Smoothing

Osama Ahmed Khan, Asim Karim
Department of Computer Science, SBASSE
Lahore University of Management Sciences (LUMS)
Lahore, Pakistan
oakhan@lums.edu.pk, akarim@lums.edu.pk

ABSTRACT
Social media, such as tweets on Twitter and Short Message Service (SMS) messages on cellular networks, are short-length textual documents (short texts or microblog posts) exchanged among users on the Web and/or their mobile devices. Automatic keyword extraction from short texts can be applied in online applications such as tag recommendation and contextual advertising. In this paper we present MIKE, a robust interactive system for keyword extraction from single microblog posts, which uses contextual semantic smoothing; a novel technique that considers term usage patterns in similar texts to improve term relevance information. We incorporate Phi coefficient in our technique, which is based on corpus-based term-to-term relatedness information and successfully handles the short-length challenge of short texts. Our experiments, conducted on multi-lingual SMS messages and English Twitter tweets, show that MIKE significantly improves keyword extraction performance beyond that achieved by Term Frequency, Inverse Document Frequency (TFIDF). MIKE also integrates a rule-based vocabulary standardizer for multi-lingual short texts which independently improves keyword extraction performance by 14%.

KEYWORDS: Keyword Extraction, Microblogs, Short texts, Semantic Smoothing, SMS, Romanized Urdu, MIKE.

Type	Text
Message	*aj friday hay is jaldi chutti hogayi, aur wassey mein mob lekare jata hun, tm ne kal program dekha tha kya?* [It is Friday, therefore I got off early, otherwise I take mobile with me. Did you see the program yesterday?]
Tweet	*Tweet 3x Lens Cap Keeper Holder with Elastic Band Loop Strap: US$6.93 End Date: Sunday Sep-05-2010 11:15:10 PDTBuy it N...http://bit.ly/cZXiSP*

Table 1: Examples of short texts

1 Introduction

Recently microblogs (e.g. Twitter) have become very popular for rapid information sharing and communication (Kwak et al., 2010). Typically, microblog posts and SMS messages are short-length documents written in an informal style. Two short-text examples are given in Table 1. The SMS message is a conversational text written primarily in Urdu (but in Latin script) mixed with some English words, while the tweet represents a short advertisement in English. Key challenges of short texts include noise (e.g. ‘βetter’, ‘:)’), multiple languages (e.g. ‘friday hay’ [it is Friday]), varied vocabulary (e.g. abbreviations like ‘Interpol’, contractions like ‘tc’ [take care], acronyms like ‘BBC’ [British Broadcasting Corporation], proper nouns like ‘Paris’), different styles (e.g. slangs like ‘lol’ [laughing out loud], colloquialism like ‘wanna’), poor quality (e.g. typos like ‘achieve’), and out-of-vocabulary words (e.g. ‘gr8’). It is therefore necessary to propose and evaluate new text processing techniques for this varied document type.

Extensive preprocessing can address some of these challenges, but established resources and tools are not yet available for short texts. Furthermore, due to multi-varied nature of short texts (e.g. multi-linguality), external resources like Wikipedia and WordNet are not applicable. Therefore, in addition to adapting standard preprocessing procedures, we adopt a specialized standardizer (Khan and Karim, 2012) for transforming varied usage of terms in multiple languages to their standard forms.

Despite extensive preprocessing, individual short texts contain limited information for term relevance determination. This is because such documents are short in length (about 13 terms on average), while vocabulary size is large. Moreover, typically a specific term appears no more than once in a short text document, thus providing no information for its relevance rank in the document. However, we can exploit term-to-term semantic relatedness information harvested from similar short texts to reduce sparsity and improve relevance ranking of terms; a technique labeled as contextual semantic smoothing. Semantic smoothing of document models has been utilized previously for clustering and classification (Zhang et al., 2006; Nasir et al., 2011). To the best of our knowledge this is the first time that semantic smoothing has been applied to short-text processing in general and to keyword extraction in particular. Also, the incorporation of Phi coefficient in our technique validates its usefulness for improved term association in sparse datasets (Tan et al., 2002).

In this paper, we make the following key contributions. First, we develop a keyword extraction system for short texts, called MIKE that is based on contextual semantic smoothing of the TFIDF matrix using Phi coefficient. This methodology does not require an external knowledge source, is more efficient than iterative graph-based techniques, and caters for all of the above mentioned challenges of short texts. We also demonstrate robustness of MIKE, which is interactive in nature, for multi-varied and sparse short texts. Second, we evaluate the impact of various preprocessing procedures on keyword extraction performance.

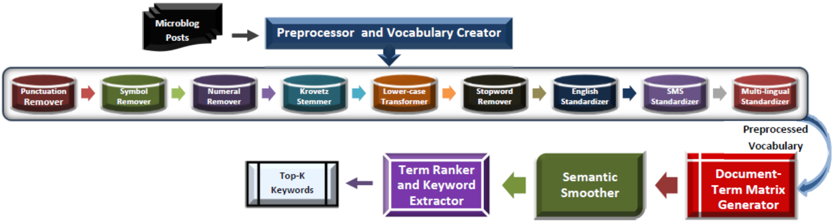

Figure 1: System Architecture of MIKE

In particular, we show that standardization of multi-lingual and multi-varied short texts through a rule-based standardizer can improve keyword extraction performance by 14%. Third, we perform our experiments on two short-text collections: a unique SMS collection from an SMS-based online social network (group messaging service) running in Pakistan, and a tweet collection from Twitter. The SMS collection contains significant proportions of messages typed in Romanized Urdu and other local languages, while the Twitter collection comprises English tweets only. This is the first time that an interactive keyword extraction system has been developed for short texts, and is evaluated on predominantly Romanized Urdu SMS messages.

The rest of the paper is organized as follows. We present the system architecture of MIKE in Section 2. Results from various experiments are provided in Sec 3. Our demonstration plan is laid out in Section 4.

2 System Architecture

In this section, we present the interactive system architecture of MIKE, which involves four processing modules: (1) preprocessor and vocabulary creator, (2) document-term matrix generator, (3) semantic smoother, and (4) term ranker and keyword extractor. These modules are discussed in detail ahead and in Figure 1. Given a collection of short texts or documents $\mathscr{D} = \{d_1, d_2, \ldots, d_N\}$ and domain-based information (stopword list and standardization lists), a keyword extraction technique outputs the top K most descriptive terms from document d_i as its keywords.

2.1 Preprocessor and Vocabulary Creator

The first module in MIKE preprocesses the collection of microblog posts and builds the vocabulary of terms out of it. Due to the presence of substantial variability and 'noise' in short-text documents when compared to conventional documents, several preprocessing procedures may be required. In this work, we implement the following procedures: (1) Punctuation removal; (2) Symbol removal; (3) Numeral removal; (4) Transformation to lower-case; (5) Stemming using Krovetz stemmer (Krovetz, 1995), which produces complete words as stems rather than truncated ones that can be produced by other stemmers. (6) Removal of stopwords using stopword list containing English articles and pronouns, obtained from AutoMap software[1]. (7) Application of English, SMS, and national/local language standardization lists.

[1] http://www.casos.cs.cmu.edu/projects/automap/

In order to tackle the issue of multi-varied composition of short texts, we apply three standardization lists. The English standardization list is obtained from AutoMap software[1]. This list maps British English and general English term variations to their respective American English forms. The SMS standardization list is built from two online sources[2] [3]. This list transforms frequently used terms in English microblog texts to their standard forms.

For our SMS collection, which contain messages written in Urdu (national language) and local languages using Latin script, we apply the specialized standardizer that we have developed earlier (Khan and Karim, 2012). This standardizer is based on a rule-based model for multi-lingual texts, and maps varied usage of terms to their unique standard forms.

2.2 Document-Term Matrix Generator

Once the vocabulary is built, the second module represents the documents in vector space of size M, where M is the vocabulary size. TFIDF is utilized here for weighting each term in a short text document since it is considered state-of-the-art among keyword extraction techniques (Wu et al., 2010). Now each document $d_i \in \mathscr{D}$ is mapped into the vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$, where $x_{ij} \geq 0$ is the weight of term t_j in document d_i. Notice that the computation of TFIDF requires the entire document collection; thus this method incorporates a corpus-based statistic of a term (its document frequency) alongwith local document information (its term frequency). After this transformation, the entire document collection $\mathscr{D}$ can be represented by the $N \times M$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$.

2.3 Semantic Smoother

A key contribution of this work is the evaluation of a contextual or corpus-based term-to-term semantic relatedness measure (Phi coefficient) on semantic smoothing and keyword extraction from short texts. The document-term matrix **X**, defined in the previous subsection, captures the relevance of each term in a document via its frequency in the document and the corpus (TFIDF); this matrix does not incorporate the semantic relatedness of terms in the collection $\mathscr{D}$. Therefore, we define an $M \times M$ term-term matrix **R** whose (i, j) element, identified as r_{ij}, quantifies the semantic relatedness of terms t_i and t_j. This matrix serves as a smoothing or scaling matrix for the original document-term matrix **X** to yield the modified document-term matrix (Nasir et al., 2011): $\bar{\mathbf{X}} = \mathbf{XR}$.

Matrix $\bar{\mathbf{X}}$ now incorporates local document information, global corpus information, as well as semantic relatedness between terms. For example, if a document contains three terms (t_i, t_j, t_k) with equal frequency, and the semantic relatedness of t_i is high with both t_j and t_k in the collection (while t_j and t_k are semantically related to t_i only), then the smoothed weight of t_i will become larger than that for both t_j and t_k. We state our hypothesis as: "A term in a document should be ranked high for keywordness in the document, if the term possesses high local document structure value as well as high global (but contextually relevant since it is based on short texts from the same microblog collection) semantic relatedness value with the other terms present in the document". It is worth noting that this module generates the semantically smoothed document-term matrix via a single multiplication of the document-term and term-term matrices, as opposed to the multiple iterations required in graph-based techniques (Mihalcea and Tarau, 2004).

[2] http://www.sms-text-guide.com/sms-text-language.html
[3] http://www.smsdictionary.co.uk/abbreviations

2.3.1 Semantic Relatedness measure

We select Phi coefficient as the term-to-term semantic relatedness measure for constructing matrix **R**:

$$PhiCoefficient = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}} \tag{1}$$

The pairwise term co-occurrence distributions are formed by the term co-occurrence contingency table (Table 2). Each cell in this 2 × 2 table shows the number of documents exhibiting a particular term co-occurrence behavior (e.g. b is the number of documents in the collection in which t_i occurs and t_j does not [identified with $\bar{t}_j$]). Due to the short length of microblog posts, we assume the co-occurrence window size to be the document length. This choice is supported by our preliminary results which demonstrated that this setting outperformed all other settings where window size was assigned a value less than the document length.

	t_j	$\bar{t}_j$
t_i	a	b
$\bar{t}_i$	c	d

Table 2: Term co-occurrence contingency table

The Phi coefficient is the Pearson's correlation coefficient for binary variables and its value lies in the interval $[-1, +1]$. It is a statistically sound measure of correlation which additionally possesses the following two significant characteristics (Tan et al., 2002). First, it is anti-symmetric under row and column permutations of the contingency table, i.e., it effectively distinguishes between positive and negative correlations between items. Second, it is a symmetric inversion invariant measure which does not get affected when the contingency table is inverted. Phi coefficient is the only measure among co-occurrence based measures that possesses these strong properties.

2.4 Term Ranker and Keyword Extractor

The final module outputs the top K terms from each document as its keywords. Given the original or smoothed document-term matrix (**X** or $\bar{\mathbf{X}}$), the top K keywords for document d_i are the top K terms in the document (in the ith row of the document-term matrix) with the highest term weights. In the next section we evaluate performances of two keyword extraction techniques: the original document-term matrix (TFIDF), and the smoothed document-term matrix (TFIDF × **R**) on multiple short-text collections.

3 Experimental Results

Tables 3, 4 and 5, alongwith Figure 2 highlight significant results generated by our experiments.

4 Demonstration Plan

In the demonstration we will show the system interfaces for user interaction during the process of feeding in real-time microblog posts and visualizing keywords automatically generated by MIKE in result. In addition we will present users option to explore different combinations of preprocessing procedures as desired for various types of microblog posts. Also, users can provide different values of K as input; the desired number of top keywords

	$K = 1$			$K = 3$			$K = 5$		
	PR	*RE*	*FM*	*PR*	*RE*	*FM*	*PR*	*RE*	*FM*
SMS TFIDF	**58.4**	**60.4**	**59.4**	47.1	50.0	48.5	39.7	61.9	48.4
SMS TFIDF+Phi	57.2	59.1	58.1	**48.9**	**52.0**	**50.4**	**41.7**	**65.0**	**50.8**
Twitter TFIDF	**49.4**	**50.0**	**49.7**	42.7	45.0	43.8	39.5	67.1	49.8
Twitter TFIDF+Phi	46.5	47.1	46.8	**44.0**	**46.3**	**45.1**	**40.9**	**69.5**	**51.5**

Table 3: Keyword Extraction Results for SMS and Twitter collections (*PR* = Precision, *RE* = Recall, *FM* = F-measure)

Size	SMS		Twitter	
	TFIDF	TFIDF+Phi	TFIDF	TFIDF+Phi
2,000	55.8	57.4	45.5	46.5
5,000	50.7	53.2	44.1	45.9
10,000	50.7	52.9	44.2	45.8
20,000	48.5	50.4	43.8	45.1

Table 4: Robustness of MIKE for SMS and Twitter collections in F-measure for $K = 3$

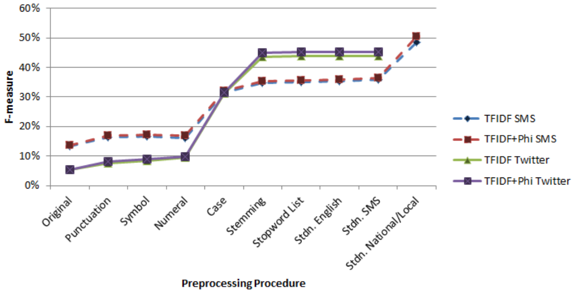

Figure 2: Significance of our rule-based multi-lingual standardizer among various preprocessing procedures in F-measure for $K = 3$

	$K = 1$	$K = 3$	$K = 5$
Message TFIDF	friday	friday jaldi [early] chhutti [holiday]	friday jaldi chhutti mobile program
Message TFIDF+Phi	friday	friday chhutti mobile	friday chhutti mobile program daikha [see]
Tweet TFIDF	lens	lens cap keeper	lens cap keeper band sunday
Tweet TFIDF+Phi	lens	lens cap holder	lens cap holder end sunday

Table 5: Keywords generated from microblog posts (see Table 1) for TFIDF and TFIDF+Phi techniques

for a document. The keywords generated by MIKE can be considered as personalized tags or advertising keywords, which are recommended for users for their respective collection of microblog posts in their selected social networks.

We will describe the system components briefly, alongwith their requirements, functionality and inter-connectivity. We will demonstrate the working of multi-lingual standardizer that we have built up earlier, and now is incorporated in MIKE. We will summarize the implementation technique on which MIKE is based and outline our key contributions in the system development. We will also narrate the practical development lessons learned through this work, our original research findings, and our first-hand experiences with this research prototype system.

References

Khan, O. A. and Karim, A. (2012). A rule-based model for normalization of SMS text. In *Proceedings of the 24th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI ’12. IEEE.

Krovetz, R. (1995). *Word sense disambiguation for large text databases*. Phd thesis, University of Massachusetts, Amherst, USA.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW ’10, pages 591–600, New York, NY, USA. ACM.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP’04, New York, NY, USA. ACL.

Nasir, J., Karim, A., Tsatsaronis, G., and Varlamis, I. (2011). A knowledge-based semantic kernel for text classification. In Grossi, R., Sebastiani, F., and Silvestri, F., editors, *String Processing and Information Retrieval*, volume 7024 of *Lecture Notes in Computer Science*, pages 261–266. Springer Berlin / Heidelberg.

Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’02, pages 32–41, New York, NY, USA. ACM.

Wu, W., Zhang, B., and Ostendorf, M. (2010). Automatic generation of personalized annotation tags for Twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL ’10, pages 689–692, Stroudsburg, PA, USA. ACL.

Zhang, X., Zhou, X., and Hu, X. (2006). Semantic smoothing for model-based document clustering. In *Proceedings of the 6th International Conference on Data Mining*, ICDM ’06, pages 1193–1198. IEEE.