Journal of Engineering Science and Technology
Vol. 13, No. 1 (2018) 167 - 186
© School of Engineering, Taylor's University

AN EMPIRICAL MACHINE LEARNING APPROACH TO EXTRACT AND RANK MULTI-WORD PRODUCT NAMES USING iCIW APPROACH

SIVASHANKARI R.*, B. VALARMATHI

School of Information Technology and Engineering, VIT University, Vellore, India
*Corresponding Author: sivashankari.r@vit.ac.in

Abstract

Entity Extraction or product name extraction is a suitable premise for sentiment analysis. A sentiment discovers opinion of the customers on the product stated in the sentence. Extracting product names using the existing approaches from the customer's reviews are not exact most of the time. Almost many existing approaches mainly lack in addressing the product name having multiple words or sequence of words (multi-word). When compared with single word named products, extracting opinion on multi-word named products are non-trivial task as customer use either full name or part of the product name (sub-word) in product reviews. Therefore, it is the foremost challenging task in sentiment analysis to recognize appropriate complete name of a specific product using the sub-word names. Secondly, the multi-word named products or sub-word named products may occur anywhere in the review sentences, the position of the product names cannot be predicted in advance. It may occur in the beginning, middle or at the end and also with any number of times. So, identifying the position of these product names is another key issue. Therefore, this research attempt to design a novel automated improved Context Information Weight (iCIW) approach to resolve the exiting issues. The iCIW assimilates the concept of lexicon and statistical approach. The proposed iCIW model is more suitable for document analysis related to medical reports, product details repository and political documents. The experimental result reveals that the proposed method performs very efficient than existing approaches in multi-word named product extraction.

Keywords: Data mining, Entity extraction, Sentiment analysis, Text mining, Document level mining, Machine learning, POS tagging.

1. Introduction

As a recent trend, analyzing online social media data's is one of the widespread research activities in Machine learning domain. The machine learning approach

Nomenclatures

$f_c(mpn)$	Number of *mpn* occurrences in the corpus
l_{msn}	Number of multi-words in lexicon contains *msn*
$l_s(mw)$	Number of lexicon words
mpn	Multi-word product name
msn	Multi-subword product name

Greek Symbols

δ_{M}	Total number of Multi-Word Product Names, N
δ_{m}	Total number of extracted multi-subword product names from corpus, N
Φ	Customer Review Corpus
Ω	Word/Term

Abbreviations

ANNIE	A Nearly-New Information Retrieval Information System
GATE	General Architecture for Text Engineering
HAL	Hyperspace Analogue to Language
iCIW	improved Context Information Weight
IWAL	Importance-Weighted Active Learning Algorithm
NLP	Natural Language Processing
POS	Part of Speech

uses statistical and artificial intelligence techniques; the machine understands the training dataset and discovers the knowledge from that dataset. This discovered knowledge is then applied into the testing dataset to produce the desired result. Therefore, machine learning can significantly improve the accuracy and reduce the computation time in the process of knowledge discovery. Among the various research activities 'Entity Extraction' is one of the popular investigational areas in machine learning. The existing approaches involved in the entity extraction are extracting only the single word named entity. But nowadays, most of the product names are not a single word. Rather, it is a sequence of words or multi-words. The order of the sequence consists of organization name, product model name, version number and manufactured year. For example, a popular product name of a bike is TVS Victor 2016. Here, 'TVS' is a name of the motorcycle manufacturing company; 'Victor' is the product model name and '2016' is the manufactured year. Recognizing this multi-word named product in the review documents is not a trivial task. The reason behind in exploring the recognition of multi-word named products is it to resolve two main issues.

The first issue is the user does not mention the full name of the product to write the opinion of that product in their reviews. Secondly, the same sub-words may exist in two different product names. For example, 'Nokia 1110' and 'Nokia 1110 Classic' are two different mobile products. It is difficult for user to understand, to distinguish on which product the customer has written the review, if that review contains the words Nokia 1110. This chaotic situation occurs due to the multi-word product names. Hence, this research introduces an improved CIW method, to extract the multi-word named products, prioritize them and identifies dominate product from the review corpus.

Frantzi and Ananiadou [1] proposed a context based statistical approach that assigns the weight to the extracted entities. The experimented results of context based statistical approach is not suitable for extracting multi-word named products from customer reviews and also the behaviour is similar to the frequency based approach. This paper modifies the approach adopted by them to obtain and identify the dominant multi-word named product in the review sentences. Therefore, this paper attempts to resolve these shortcomings of the existing approaches and propose a novel approach called improved context based statistical approach known as iCIW. The proposed approach is very effective in extracting dominate product as well as ranking product names that drawn from the review corpus. The remaining section is as follows. Section 2 provides a detailed survey on existing approaches in entity extraction, Section 3 describes the proposed approach in detail, Section 4 explains the evaluation of proposed method, Section 5 shows the experimental results and Section 6 compares the proposed approach with an existing approach.

2. Literature Survey

Ayat et al. [2] have proposed a probabilistic method to find similarity in tuples (rows) in the tables. A tuple consists of a set of columns. In this approach, any two columns from two different tuples may have similar values with different representation. The resemblance can be recognized by using semantic association between words in the columns. The highest probability value of similarity association shows the similarity of two tuples. For this similarity pair identification, the context-free similarity function is used. Bellare et al. [3] investigated whether two entities are illustrating about the same object. To ascertain this, they have used Importance-Weighted Active Learning Algorithm (IWAL) black-box active learning approach. This IWAL active learner determines whether the entity pairs are referring the same object. In addition, they had used ConvexHull algorithm to maximize the recall of the existing active learning approaches.

Bhattacharya and Getoor [4] have proposed a novel clustering algorithm to detect the multiple references of the same entity in research papers. The proposed clustering algorithm, verifies whether the author(s) in two research papers has same author(s) or some of the authors are same or all the authors are different. They have suggested a graph based Naïve relational entity resolution approach to determine the co-occurrence of these authors in different papers. Derczynski et al. [5] have presented various existing approaches to find out named entity recognition.

Ding and Jiang [6] suggested Conditional Random Field to extract the product name in the review sentences. For this extraction, domain specific ontology database is used and the product name is obtained based on its features. The features of the products are maintained in a database and compared with the review sentences. In this approach POS tagging method had used to classify the word structure to identify the sentiment of the sentences. Frantzi and Ananiadou [1] recommended contextual cues for extracting and ranking multi-word terms in the sentences automatically. Guo et al. [7] proposed a novel approach for entity extraction to extract the entities from the sensor data for the military intelligent system. For this method, they had used NLP, GATE and information retrieval system (ANNIE) to develop the rule-based approach for the extraction of entities from the sensor dataset.

Konkol et al. [8] applied machine learning approach to improve named entity recognition. Latent Dirichlet Allocation is an effective machine learning algorithm used to determine the semantic relationship between the neighbour words in the sentences. Hyperspace Analogue to Language (HAL) is another method used to identify the similarity between the co-occurrence of words. Korkontzelos et al. [9] have suggested several methods such as Dictionary based entity discovery, pattern generation with regular expressions and multinomial logistic regression methods to recognize the drug entities in biomedical dataset. The Multinomial logistic regression classifier is trained using samples to predict the drug entities. In the Dictionary based entity discovery approach, a dictionary is used to verify whether the predicted entity is present in that collection of entities. Regular expression is also used to recognize the drug entities that are undisclosed by the classifier and dictionary approaches.

Nothman et al. [10] proposed logistic regression classifier to extract the entities from Wikipedia corpus. This classifier exploited the Wikipedia data and Wikipedia entity page links to retrieve the entities. Si et al. [11] used mixture model of Dirichlet topic model to extract the topics of the tweets and predicts the opinion of that tweets.

Song et al. [12] adopted dictionary based approach to extract the biomedical entities from the dataset. Stanford CoreNLP and rule-based methods are applied to retrieve the target entities. Weninger et al. [13] proposed a novel approach to extract web pages, web pages with entities and mapped entities web pages from the database. A combined web mining approaches of Hybrid List Extraction algorithm (HyLiEn), parallel paths and link paths are used to perform this extraction. Xu et al. [14] have recommended Weakly Supervised Latent Dirichlet Allocation (WS-LDA), to extract book name, movie name and game title in the web data. Zafarani et al. [15] suggested a novel approach Link-based user identification approach to map the usernames that exist across the online data. Li et al. [16] used SVM based learning method to extract the relationship between the entities.

Zhao et al. [17] proposed a meta-learning system to discover the relationship between the entities. The entities considered for this approach are person, organization, location, vehicle, weapon and facility. Yan and Zhu [18] evaluated the entity extraction methods to find out the performance of those methods in scientific publications. For this evaluation, few popular approaches such as, Conditional Random Field (CRF), Wikipedia based dictionary and keyword based approaches are considered. Peled et al. [19] proposed a classifier to predict the matching entities of online social networks. The classifier is built using Weka tool. Agerri and Rigau [20] recommended robust named entity extraction system with cluster based approach to extract the entities from CoNLL dataset and other datasets. Callan and Mitamura [21] presented the learning rules on knowledge based approach to discover the entities. Fersini et al. [22] suggested CRF to discover entities from CoNLL dataset.

Hsu and Kao [23] proposed Co-occurrence Interaction Nexus model with named entity extraction approach to find the entities in toxic genomic database. The identified entities are chemical name and disease name. Tianlei et al. [24]

used KeEL to improve entity linking and automatic biography construction. Kang et al. [25] presented a web-scale entity ranking and machine learning approaches to discover the entities and also discover the pair-wise entities from the social network data.

3. Improved Contextual Information Weight Approach

The proposed iCIW is an assimilation of statistical and lexicon approach to extract the multi-word product names from customer reviews. In this approach, the review sentences written in the English language have been taken for evaluation. Before extracting product names, identification of product names is essential. However, the customer reviews are language dependent and unstructured. With this complex nature of the sentence, identification of product names is a non-trivial task. To handle this, the review sentence is first converted into a unique representation by using Stanford POS tagger. This POS tagger is used to generate tagged sentences for review sentences. From this tagged sentence, consecutive noun sequences are selected as product names. But all the extracted noun sequences would not be the product names i.e. not all noun phrase sequences are relevant to the entity names. For example, the following noun phrase sequences are irrelevant with respect to the product names.

- Performance NN speed NN
- Waste NN hours NNS
- customer NN service NN department NN
- amazon NN return NN period NN
- panel NN button NN layout NN
- January NNP 13 CD 2004 CD
- 80 CD % NN

So, these irrelevant noun phrases have to be identified and eliminated from the noun phrase sequence set. The proposed system architecture is shown in Fig. 1.

To eliminate these irrelevant noun phrase sequences, lexicon-based approach is used to recognize the actual multi-word product names. The lexicon is consisting of an initial seed of multi-word product names. Subsequently the user may not use the complete name of a product to write the review. Hence, this lexicon based approach in product name extraction is very essential. The customer may use the partial name or part of the product name to write the opinion about the product. This part of the product name is called as multi-subword product name. When a multi-subword exists in the review sentences, it indicates the aspects of its multi-word product name. So, identification of dominant product without considering these sub-word product names from the review corpus is worthless. This proposed system calculates improved context information weight for both multi-word named products and the multi-subword named products.

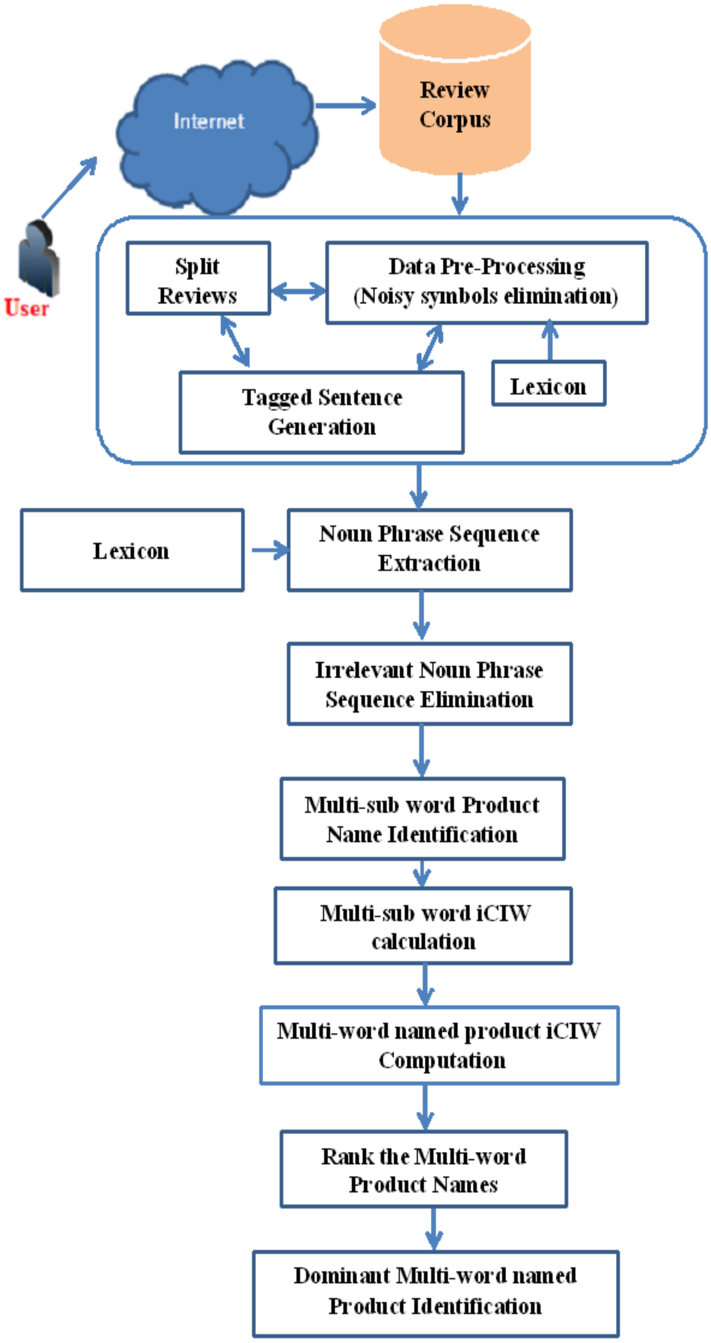
Internet
Review Corpus
User
Split Reviews
Data Pre-Processing (Noisy symbols elimination)
Tagged Sentence Generation
Lexicon
Lexicon
Noun Phrase Sequence Extraction
Irrelevant Noun Phrase Sequence Elimination
Multi-sub word Product Name Identification
Multi-sub word iCIW calculation
Multi-word named product iCIW Computation
Rank the Multi-word Product Names
Dominant Multi-word named Product Identification

Fig. 1. iCIW System architecture.

The calculated iCIW value of sub-word named products is used to calculate the cumulative iCIW of multi-word named products. From the calculated iCIW values, a multi-word named product which has scored the highest CI weight among all multi-word named products is ranked as a dominant product in the review corpus. The iCIW calculation steps are given below.

Let φ is a customer Review Corpus, $\varphi = \{\omega_1, \omega_2, \omega_3 \cdots \omega_d\}$ where ω_i is a term or word and $\omega_i \in \varphi$.

M is a multi-word named product, m is a multi-sub-word named product and m is a sub-sequence of M , $m \in M$.

$|m|$ is the length of m and $|M|$ is the length of M .

$M = \{t_1, t_2, t_3, \cdots, t_j, t_{j+1}, \cdots t_p, \cdots, t_{q-1}, t_q\}$, $m = \{t_j, t_{j+1}, t_{j+2} \cdots t_p\}$, where t_j is a word.

δ_m = Total number of selected multi-subword product names

δ_M = Total number of Multi-word named products

$$f_l^m = \sum_{m \in M}^{\delta_M} f_\varphi^M \tag{1}$$

Equation (1) computes the number of occurrences of lexicons in corpus φ which contains m .

f_φ^M = Number of occurrences of multi-word named product M in corpus φ

f_φ^m = Number of occurrences of sub-word m in corpus

f_M^m = Number of lexicons which contains multi-subword named product m .

$$B_\varphi^m = w_f * \left(\left(\frac{f_M^m}{\delta_M} \right) * \left(\frac{f_\varphi^m}{f_l^m} \right) \right) \tag{2}$$

Equation (2) calculates Contextual weight of multi-subword named product m (B_φ^m). w_f = Weight factor. Default value is 0.5.

$$iCIW(M) = \sum_{m \in M}^{\delta m} B_\varphi^m \tag{3}$$

Equation (3) shows the formula to compute iCIW. Unlike existing methods, iCIW does not fix the locations to extract the product names. Rather, the entire review sentence will be searched.

4. Evaluation of ICIW

The proposed iCIW technique considers a document consisting of review sentences as input for evaluation. The iCIW methodology first reads the review sentences one by one from the review corpus and converts it into a language independent pattern called phrases sentence or tagged sentence. A sample review sentence is given below.

"*in encoding a mpeg-2 film with pinnacle studio 7, the intel pentium 4/2200 is clearly faster than the and athlon xp 2000+."*

The above review sentence is converted into tagged sentence using Stanford POS Tagger.

"*in_IN encoding_VBG an_DT mpeg-2_NN film_NN with_IN pinnacle_NN studio_NN 7_CD, _, the_DT intel_NN pentium_NN 4/2200_CD is_VBZ clearly_RB faster_RBR than_IN the_DT amd_JJ athlon_NN xp_IN 2000_CD +_CC._..*"

The above tagged sentence has split into tagged words. A rule-based approach is employed to automate the process of extracting the noun phrase sequences. This proposed approach is fully targeted on automation of entire process. So, there is no manual work carried out from the beginning of the execution till the output generation stage. Consequently, the tagged words are further experimented with the rule-based approach to extract the noun-sequences alone from the tagged sentences. The following noun phrase sequences are extracted from the above tagged sentence. The noun sequence is generally a combination of noun phrases, conjunction, numeral, cardinal, preposition and list item marker.

mpeg-2:NN
film:NN
mpeg-2 film: NN NN
mpeg-2 film with: NN NN IN
mpeg-2 film with pinnacle NN NN IN NN
mpeg-2 film with pinnacle studio: NN NN IN NN NN
mpeg-2 film with pinnacle studio 7: NN NN IN NN NN CD
intel Pentium:NN NN
intel pentium 4/2200: NN NN CD
athlon xp :NN IN
athlon xp 2000: NN IN CD
athlon xp 2000+: NN IN CD CC

The review sentence mentioned in section 4 expresses the opinion on the product Athlon XP 2000+. It is a multiword product name and the customer has used its full name in the review. However, the customer(s) may not use this full name every time when he/she writes the review of that product. Because the user is interested to use commonly used sub-words (part of the full name) to write the review of that multi-word named product. The following review sentence deliberates about Athlon XP 2000+ by using its sub-word Athlon XP.

"*Finally, AMD made a rather desperate move and introduced Athlon XP along with a new model rating, which gives the processor a model number*"

All the extracted sub-words of multi-word named products are displayed in Table 1. For example, the extracted multi-subword named products of Athlon XP 2000+ from the review corpus are Athlon XP and Athlon XP 2000. Using Dependency relationship among the words, the entity which is pointed by the pronouns such as It, This, etc. can be recognized. The dependency relations such as ref, def, combined dependency relations such as (nsubj & dobj) and (case & nmod) are useful to recognize the entities pointed by the pronouns. After identification of all the subword of a multi-word named product, its iCIW value is calculated. In location based approach, the positions of the entities are fixed and predefined in the sentences. Such positions are,

- Words at the first position of the phrase
- Words written with the first letter in uppercase.
- Words written totally in uppercase.
- Words in quotes

The above rules are applicable for the structured sentences when those sentences are framed with grammatical rules. The location based approach is not suitable for all the cases. For example, “Sheela, Mala rushing to the hospital madhan also joined with them”. In this sentence the word madhan is not positioned at the beginning or at the end. In general, the location is not fixed or predicted in advance for the unstructured sentences like review sentences. In our approach, we used customer review dataset which consists of unstructured sentences. In this dataset, if we use fixed locations to find the product names then we would have not been identified these many product names. In general, the entity can occur anywhere in the sentence. So, location based entity identification will be well suitable for structured dataset but not for the review sentences.

5. Results and Discussion

The performance of the proposed system is evaluated using Jindal and Liu [26] dataset. This dataset contains only customer reviews about various products such as Scan DVD player, Nomad Jukebox Zen Xtra, etc. This dataset contains more than 8500 product reviews. A review is a sequence of sentences. The review has segmented into set of sentences using line breaker and end of the line characters. From each segmented sentence, some of the characters such as extra spaces and noisy characters (☺,☹, etc.) are removed.

These noisy characters include images and icons. These characters are not useful in product name extraction and dominant product name identification. So, these characters are removed from the segmented sentences. When a special character ‘&’ is encountered between noun phrases, left side noun phrase will be considered as one product name and the right-side noun phrase is treated as another product name. Edit Distance method is used to identify the misspelled sub-words names. In this proposed system the minimum distance value is fixed as 1. In the tested dataset, most of the sub-words are spelled correctly by the reviewer. So, using this edit distance method very few misspelled sub-words are identified and corrected. Those are, Nomad is misspelled as Nomand and Coolpix is misspelled as Colpix.

All the predicted noun sequences of segmented sentences are not actual product names. A few of them are irrelevant to the name of the products. Those irrelevant noun sequences are eliminated by comparing the lexicon words. Table 1 shows the number of occurrences of multi-word product names in the review corpus and their iCIW value. Figure 2 demonstrates that Pentium 4 has highest frequency among other products.

Table 1. Multi-word product name frequencies and its iCIW.

Multi-word product name	Frequency	iCIW
Apex AD-2600	**11**	**0.272**
Pentium 4	**78**	**0.742**
Pentium III Coppermine	**5**	**4.1**
Scan DVD Player	**1**	**41**
Nokia 6610	**10**	**0.3**
Athlon XP 2000+	**15**	**0.999**
Nikon Coolpix 4300	**6**	**0.583**
Nomad Jukebox Zen Xtra	**5**	**3**
Canon Powershot G3	**14**	**3.5**

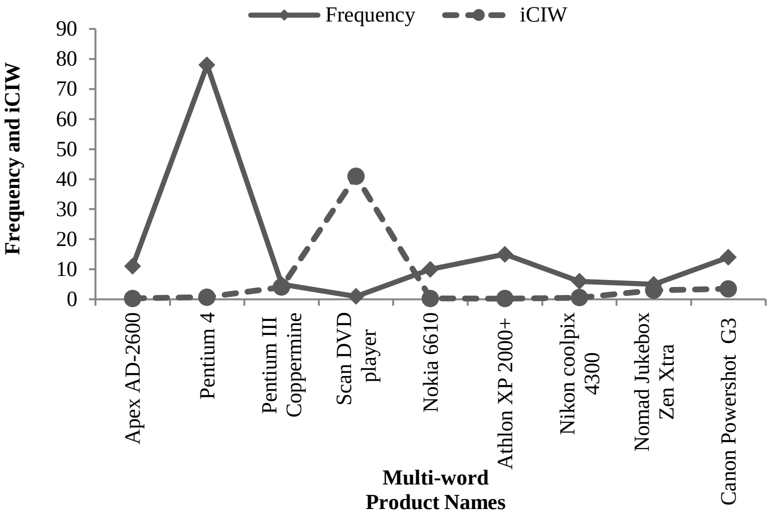

Fig. 2. Frequency and iCIW comparison.

It implies that the Pentium 4 has discussed mostly by the reviewers when compared to other multi-word named products. Therefore, in case of frequency based calculation the Pentium 4 has mostly conferred multi-word product. On other hand, Figure 2 shows that Scan DVD player is reviewed by less number of reviewers. The total occurrences of Scan DVD player in the review corpus is one. But iCIW score of Scan DVD player has highest score among others. Because, the iCIW value calculation for the multi-word named product Scan DVD player considers full name as well the sub-words of the Scan DVD player. The DVD Player is a sub-word named product of Scan DVD player and in the review corpus it has been reviewed more than 140 times. Thus, in iCIW score, the Scan DVD player is mostly conferred multi-word product name among all the other products. The important factor in the calculation of iCIW is the sub-word named products. This iCIW calculation details are explained in section 6.

In frequency based approach, the sub-word product names are not considered for the calculation of multi-word named products. So, frequency based approach is suitable for single word named entities. If this frequency based approach is applied to multi-word named products, it can produce inconsistent results. The reason is shown with one example. Nomad Jukebox Zen Xtra is a multi-word named entity. For example, in a data set, the customer reviews on the product **Nomad Jukebox Zen Xtra** is expressed with its multi-subword name **Zen Xtra** as 10 times and another multi-subword **Nomad Jukebox** as 12 times. Then the frequency based approach will be computed as follows.

- Frequency of Zen is 10
- Frequency of Xtra is 10
- Frequency of Nomad is 12 times
- Frequency of Jukebox is 12 times

The summation of all the terms in **Nomad Jukebox Zen Xtra is: 12+12+10+10=44.** The result is incorrect, since the Zen and Xtra frequencies are considered separately but it actually occurred together in the review sentences. This existing problem is eliminated using the proposed iCIW approach. The cumulative iCIW of a multi-word product name is the summation of all of its sub-words iCIW values. The sub-word occurrences and its iCIW value of Scan DVD player is higher than other multi-word product names, hence, the Scan DVD player becomes the leading product or mostly discussed product among all other products of the review document. So, the dominating multi-word named product in the review document is Scan DVD player.

In entity extraction, the number of occurrences of a product name alone is not sufficient to find the dominating product of a document. The sub-word names and multi-word product names occurrences are also to be considered for the evaluation. The iCIW calculation provides a measurement for the number of occurrences of a product in account of its sub words. But in frequency based calculation, the number of occurrences of the product full name alone was considered. This calculation is appreciated when the product name is a single word. But when the product name is more than one word, then the frequency calculation is not accurate. So, using iCIW, the number of occurrences of all the subwords is uniquely computed.

Another popular approach in multi-word product name extraction is C-value [27] score method. This method is also a language and frequency independent approach. This method can be used to extract the multi-word names. Using this approach, the extracted noun phrases are experimented with lexicon words and then C-value of multi-word product name and multi-subword named products are calculated separately. The formula for the calculation of C-value score for individual multi-word is given in Eq. (4) [27]. Similarly, Eq. (5) [27] indicates the C-value score function of multi-word named product.

$$C-Value(mw)=\log_{10}(|mw|)\left(f(mw)-\frac{1}{f(R)}\sum_{r\in R}f(r)\right) \tag{4}$$

mw= Multi word or multi sub-word product names

$|mw|$ = Number of words in mw

$f(mw)$ = Number of occurrences of mw in the review corpus

$f(R)$ = Number of multi-word terms that contains R

$f(r)$ = Frequency of multi-word term r in the corpus

$$C-value(mpn) = C-value(mpn) + \left(\sum_{msn \in mpn} C-value(msn) \right) \tag{5}$$

Figure 3 depicts the comparative chart of C-value and iCIW. The computed C-value score function also reveals Scan DVD player as the dominant product in the review document. This existing C-value score function is also implemented with python programming language of 230 lines for the sake of comparison. Although, C-value performance is similar to iCIW in terms of dominating product name identification, it has different views with respect to iCIW.

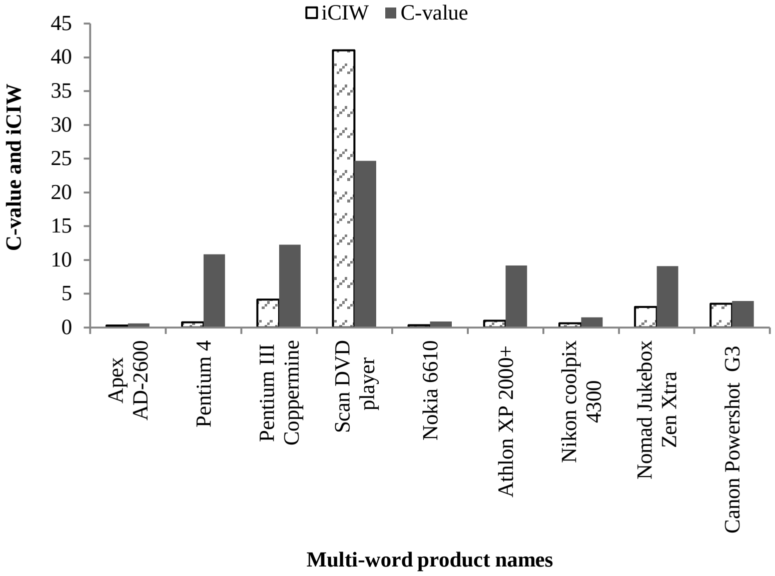

Fig. 3. C-value and iCIW comparison.

$$wei(msn) = w_f * \left(\frac{l(msn)}{l_s(mw)} + \frac{f_c(mpn)}{f_l(msn)} \right) \tag{6}$$

$wei(msn)$ = Weight factor of multi-subword product msn.

$l(msn)$ = Number of multi-words in lexicon contains msn.

$l_s(mw)$ = Number of lexicon words.

$f_c(mpn)$ = Number of occurrences of mpn appears in the corpus.

$f_l(msn)$ = Number of occurrences of msn appears in lexicon words.

$$CIW(mw) = \sum_{y \in mw} \sum_{msn \in y} wei(msn) + 1 \tag{7}$$

$CIW(mw)\, or\, weight(mw)$ = weight of multi word product name

$$C-value\,based\,CIW((mpn)) = \sum_{msn \in mpn}^{ls} (C-value(msn) * CIW(mw)) \tag{8}$$

The other widespread approach in automatic multi-term extraction is C-value based CIW technique. The CIW calculation is given in Eq. (6) [1]. Equation (7) [1] indicates the weight factor calculation of multi-subword named products. Equation (8) [27] shows the C-value based CIW calculation of multi-word named products. The C-value based CIW of (mpn) is the product of C-value (mpn) and CIW (mpn).

Figure 4 shows the comparison analysis graph of proposed method and C-value based CIW. Compared to C-value, the C-value based CIW is more co-incidents with proposed method. Even though C-value, C-value based CIW and iCIW approaches prove that the Scan DVD player is the dominant multi-word named product for the same review corpus. Figures. 3 and 4 depict the similarity results with respect to iCIW.

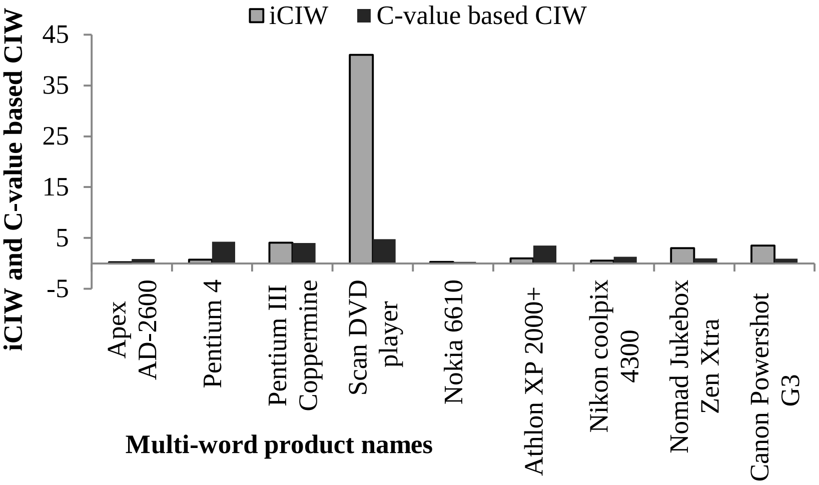

Fig. 4. C-value based CIW and iCIW comparison.

Table 2 displays the computed results of product name extraction with respect to baseline approaches, proposed method iCIW and frequency based approach. The comparison results depict that, there exists high variation between frequency and other methods.

Table 2. Frequency, C-value, iCIW and C-value based CIW values of multi-word named products.

No.	Multi-word product name(MPN)	Multi sub-word names (MSN)	C-value (MSN)	iCIW (MSN)	Frequency	C-value (MPN)	C-value based CIW (MPN)	iCIW (MPN)
1	Apex AD-2600	Apex ad-2600	0.602	0.227	11	0.602	0.882	0.272
		Apex ad	0.000	0.045				
2	Pentium 4	Pentium4	8.429	0.314	78	10.837	4.28	0.742
		Pentium IV	2.408	0.428				
3	Pentium III Coppermine	Pentium III Coppermine	1.431	0.4	5	12.268	4.00	4.1
		Pentium III	10.837	3.7				
4	Scan DVD player	DVD player	24.684	41	1	24.684	4.79	41
5	Nokia 6610	Nokia 6610	0.903	0.3	10	0.903	0.319	0.3
6	Athlon XP 2000+	Athlon XP	3.913	0.533	15	9.160	3.54	0.999
		Athlon XP 2000	5.248	0.466				
7	Nikon Coolpix 4300	Coolpix 4300	1.505	0.5	6	1.505	1.31	0.583
		Nikon Coolpix	0.000	0.083				
8	Nomad Jukebox Zen Xtra	Zen Xtra	6.924	2.4	5	9.082	.996	3
		Jukebox Zen Xtra	0.954	0.1				
		Nomad Jukebox	0.602	0.3				
		Jukebox Zen	0.602	0.2				
9	Canon Powershot G3	Canon Powershot	0.602	1	14	3.913	0.922	3.5
		Powershot G3	1.505	2.5				
		Canon G3	1.806	0				

The result shows that, the frequency based approach is not suitable for multi-word product names. As the sub-word frequencies and product name frequencies are considered as important factors for baseline approaches and proposed approach iCIW, the better accuracy is achieved in those methods compared to frequency based approach. Table 3 illustrates the ranking of multi-word named products with respect to frequency based approach and iCIW.

As per the frequency approach, the Pentium 4 is considered as the dominant product. According to the proposed iCIW model, Scan DVD player is the dominant product. This approach is suitable for document level sentiment analysis because it identifies dominant product of the document. The documents are categorized by the identified dominant products. The review sentence which contains the dominant products will be given as the input to the dependency parser. This parser finds the relationship among the words in the review sentences.

From the result, it can be recognized as whether a word is acting as an adjective or adverb or verb or noun phrase to that sentence. Then the sentiment of that sentence will be recognized using the identified adjective or adverb.

Table 3. Ranking of multi-word named products using frequency, C-value, iCIW and C-value based CIW methods.

Multi-word product name(MPN)	Frequency	C-value	C-value based CIW	iCIW
Apex AD-2600	4	9	8	9
Pentium 4	1	3	2	6
Pentium III Coppermine	7	2	3	2
Scan DVD player	8	1	1	1
Nokia 6610	5	8	9	8
Athlon XP 2000+	2	4	4	5
Nikon Coolpix 4300	6	7	5	7
Nomad Jukebox Zen Xtra	7	5	6	4
Canon Powershot G3	3	6	7	3

6. Comparison with Existing Methods

The section 5 demonstrates the comparison of various existing techniques with respect to iCIW. The proposed iCIW model is outperformed in terms of multi-word named product identification from customer reviews. This section compares various existing methods on ranking of multi-word named products. In a review corpus, the most discussed multi-word named product is a remarkable product. The end users may have an interest on finding the products that are highly remarkable product. In this paper, the proposed iCIW method assigns rank to the multi-word named products from highest to lowest, based on the calculated improved context information weight.

As discussed in section 5, the number of occurrences of a multi-word named product purely depends on itself and its sub-word product names. The iCIW weight calculation plays a role to discover the dominant multi-word named product and assigns rank to the multi-word named products. While considering iCIW value, the highest weight score of multi-word named product is Scan DVD player. Pentium III Coppermine and Canon Powershot G3 are having the second and third highest score. So, this document has been categorized as Scan DVD player review document and the user will pay more attention to choose this document if he wants to know about the Scan DVD player product. That is, this review document is one of the analysis reports for Scan DVD player product. This kind of analysis report is very valuable nowadays in the area of sentiment analysis and Business Intelligence. Consequently, following the Scan DVD player, the Pentium III Coppermine has the second highest iCIW value. So, the user wants to know the merits and demerits of the Pentium III Coppermine product, this document can be positioned as desirable to refer it. This ranking is very much essential to categorize the documents according to the entities or term(s) discussed in it. This sort of categorization will enhance the end user to select a particular document from the collection of documents, which is addressing a particular product, service or political issue with which he or she is specifically interested. This document categorization will help the marketing people to find opinion of the customers on their product by selecting appropriate document from the collection of documents. So, ranking of product names is very essential.

Figure 5 depicts iCIW ranking with respect to the frequency approach. The product ranking based on frequency approach for the multi-word named products is inaccurate and it is experimentally reported in section 5. Because, in frequency based approach, the dominating multi-word named product is Pentium 4. Whereas in iCIW based approach, the Scan DVD player is the dominant multi-word named product. When it comes to comparison of frequency and proposed approach, Figure 5 evidently shows that the ranking between frequency and iCIW totally deviates from each other and the discrepancy range is 8. For example, for the multi-word product name Apex AD-2000 rank is 9 with respect to iCIW and 4 with respect to frequency approach.

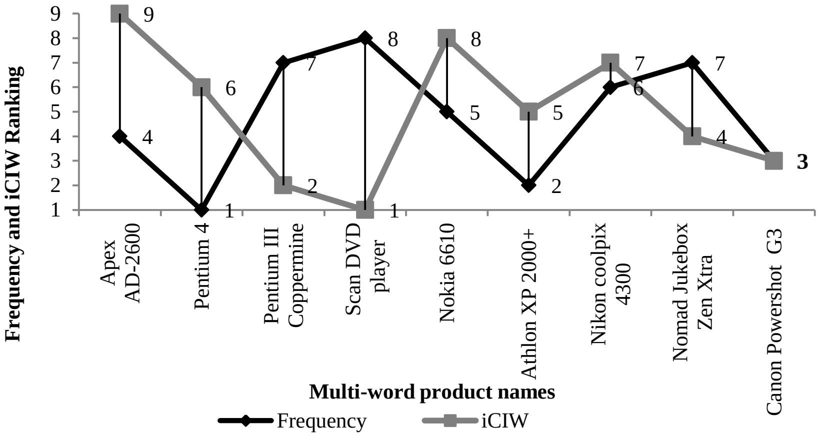

Fig. 5. Frequency and iCIW multi-word named products ranking chart.

Figure 6 displays the comparison between C-value and C-value based CIW ranking. The baseline approaches such as C-value and C-value based CIW performance is similar. So, these methods are compared with the proposed iCIW approach in Figs. 7 and 8 respectively.

In Fig. 7, the difference in the number of deviations on the ranking of the products between iCIW and C-value is 4. The deviation of the ranks for the products are Pentium 4, Athlon XP 2000+, Nomad Jukebox Zen Xtra and Canon Powershot G3. This deviation occurs, if the number of multi-subwords is high for a particular multi-word named product, then the C-value score function value becomes less and the deviation is slightly high. Whereas in Fig. 8, the deviation range is high between C-value based CIW and iCIW approaches and the number of deviations is 8. The C-value based CIW calculation is a time consuming process because its computation involves the calculation of both CIW and C-value approaches. If the number of multi-subwords are high for a particular multi-word named product or the length of multi-word named product is long, then the C-value based CIW approach requires lots of computation. In the proposed iCIW model, the frequency of individual multi-subwords are considered for the computation. However, C-value based CIW model ignores this factor. Hence, the proposed system is a truthful metric to identify the multi-word named products and ranking of the multi-word named products. Since the proposed approach is used to extract dominant multi-word named product and assigning ranks to the multi-word named products, thus iCIW approach is superior over other methods.

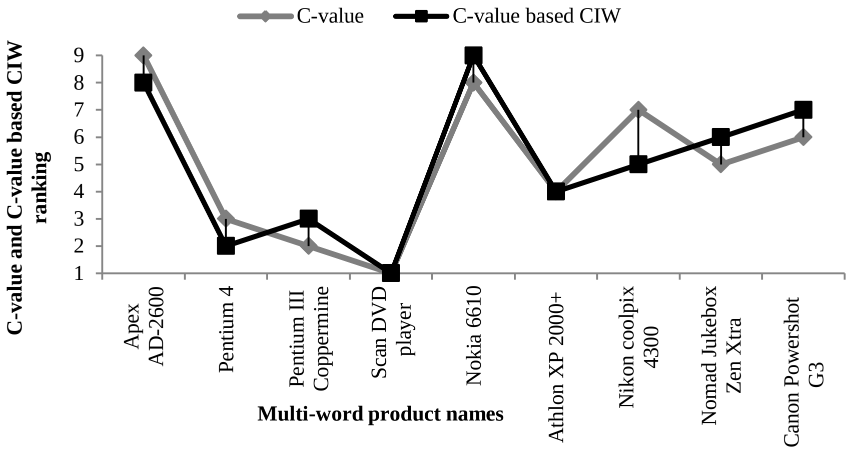

Fig. 6. C-value and C-value based CIW ranking chart.

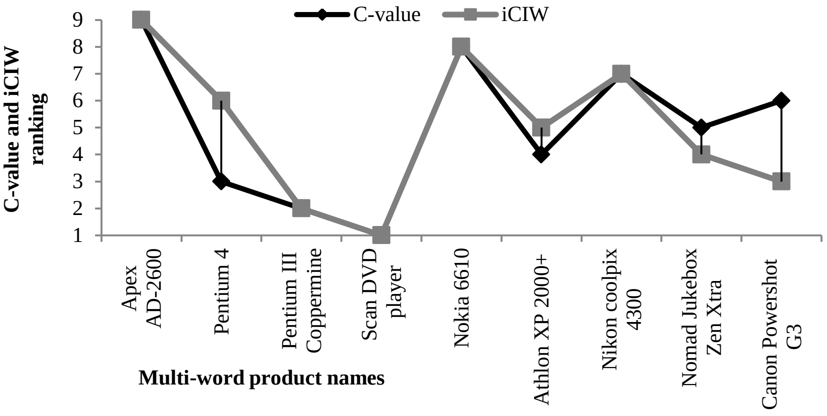

Fig. 7. C-value and iCIW multi-word named products ranking chart.

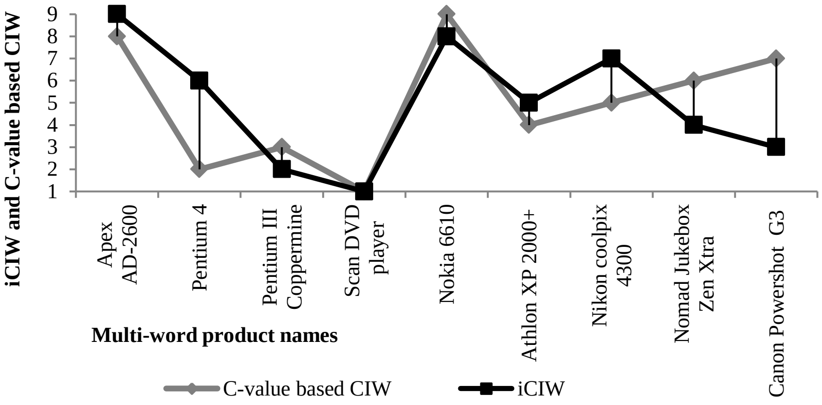

Fig. 8. iCIW and C-value based CIW multi-word named products ranking chart.

7. Conclusion

In this paper, the improved Context Information Weight (iCIW) is presented to train the machine to recognize multi-word product names and to identify the dominant multi-word named product in the customer reviews corpus. This proposed iCIW is a combination of lexicon based and statistical method. The existing approaches in entity extraction are capable to recognize single word named entities in the dataset. The proposed iCIW method is a novel approach to recognize the multi-word named entities from the review corpus. The sub-words of the multi-word named entities play a major role to categorize the documents. The proposed iCIW methodology is experimented with the existing approaches such as frequency based, C-value and C-value based CIW approaches. The summary of the proposed iCIW is as follows.

- Using iCIW approach, the multi-word product names are ranked according to its respective ICIW value.
- The iCIW method calculates the weights of each sub-word in the multiword product name.
- Using iCIW, the dominant product name of the document has been revealed. This dominating multi-word named product is the most discussed product by the customers in the review corpus. This dominant product extraction is very worthwhile in document categorization.
- Once the review documents are categorized based on the dominating products, a person who is interested on a particular entity (product, service and political member.) can select only that related documents from the categorized documents.
- The selected documents contain more relevant data of the customer's interest on a particular product compared to other documents.
- Therefore, this iCIW approach is well suitable for medical data analysis, business intelligence and decision support systems.

The proposed iCIW produces comparable results when its performance is compared with the existing approaches like frequency approach, C-value and C-value based CIW. In future, the proposed iCIW approach is enhanced to the document level sentiment analysis to identify the sentiment of the multi-word named product.

References

1. Frantzi, K.T.; and Ananiadou, S. (1997). Automatic term recognition using contextual cues. *Proceedings of 3rd DELOS Workshop*, Zurich, Switzerland.
2. Ayat, N.; Akbarinia, R..; Afsarmanesh, H.; and Valduriez, P. (2014). Entity resolution for probabilistic data. *Information Sciences*, 277, 492-511.
3. Bellare, K.; Iyengar, S.; Parameswaran, A.G.; and Rastogi, V. (2013). Active sampling for entity matching with guarantees. *ACM Transactions on Knowledge Discovery from Data (TKDD) - Special Issue on ACM SIGKDD 2012*, 7(3), Article No.: 12.

4. Bhattacharya, I.; and Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 1(1), Article No.: 5.
5. Derczynski, L.; Maynard, D.; Rizzo, G.; van Erp, M.; Gorrell, G.; Troncy, R..; Petrak, J.; and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51(2), 32-49.
6. Ding, S.; and Jiang, T. (2010). Comment target extraction based on conditional random field & domain ontology. *International Conference on Asian Language Processing* (*IALP*), IEEE Harbin, Heilongjiang, China, 28-30 December 2010, 189-192.
7. Guo, J.K.; Brackle, D.V.; LoFaso, N.; and Hofmann, M.O. (2015). Extracting meaningful entities from human-generated tactical reports. *Proceedings of the Complex Adaptive Systems*, San Jose, CA, USA, 72-79.
8. Konkol, M.; Brychcin, T.; and Konopík, M. (2015). Latent semantics in named entity recognition. *Expert Systems with Applications ACM*, 42(7), 3470-3479.
9. Korkontzelos, I.; Piliouras, D.; Dowsey, A.W.; and Ananiadou, S. (2015). Boosting drug named entity recognition using an aggregate classifier. *Artificial Intelligence in Medicine*, 65(2), 145-153.
10. Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; and Curran, J.R. (2013). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194, 151-175.
11. Si, J.; Mukherjee, A.; Liu, B.; Li, Q.; Li, H.; and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL*), Sofia, Bulgaria, 24-29.
12. Song, M.; Kim, W.C.; Lee, D.; Heo, G.E.; and Kang, K.Y. (2015). PKDE4J: entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, 57, 320-332.
13. Weninger, T.; Johnston, T.J.; and Han, J. (2013). The parallel path framework for entity discovery on the web. *ACM Transactions on the Web* (*TWEB*), 7(3), Article No.: 16.
14. Xu, G.; Yang, S.; and Li, H. (2009). Named entity mining from click-through data using weakly supervised latent dirichlet allocation. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 1365-1374.
15. Zafarani, R.; Tang, L.; and Liu, H. (2015). User identification across social media. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 10(2), Article No.: 16.
16. Li, L.; Zhang, J.; Jin, L.; Guo, R.; and Huang, D. (2015). A distributed meta-learning system for Chinese entity relation extraction. *Neurocomputing*, 149, 1135-1142.
17. Zhao, G.; Wu, J.; Wang, D.; and Li, T. (2016). Entity disambiguation to Wikipedia using collective ranking. *Information Processing and Management ACM*, 52(6), 1247-1257.

18. Yan, E.; and Zhu, Y. (2015). Identifying entities from scientific publications: A comparison of vocabulary-and model-based methods. *Journal of Informetrics,* 9(3), 455-465.
19. Peled, O.; Fire, M.; Rokach, L.; and Elovici, Y. (2016). Matching Entities across online social networks. *Neurocomputing*, ACM, 210, 91-106.
20. Agerri, R.; and Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238, 63-82.
21. Callan, J.; and Mitamura, T. (2002). Knowledge-based extraction of named entities. *Proceedings of the Eleventh International Conference on Information and Knowledge Management,* ACM, 532-537.
22. Fersini, E.; Messina, E.; Felici, G.; and Roth, D. (2014). Soft-constrained inference for named entity recognition. *Information Processing and Managemen*t *ACM*, 50(5), 807-819.
23. Hsu, Y.Y.; and Kao, H.Y. (2015). Curatable named-entity recognition using semantic relations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4), 785-792.
24. Tianlei, Z.; Xinyu, Z.; and Mu, G. (2015). KeEL: knowledge enhanced entity linking in automatic biography construction. *The Journal of China Universities of Posts and Telecommunications*, 22(1), 57-71.
25. Kang, C.; Yin, D.; Zhang, R.; Torzec, N.; He, J.; and Chang, Y. (2015). Learning to rank related entities in Web search. *Neurocomputing*, ACM, 166, 309-318.
26. Jindal, N.; and Liu, B. (2006). Identifying comparative sentences in text documents. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 244-251.
27. Frantzi, K.; Ananiadou, S.; and Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, Springer, 3(2), 115-130.