The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 271-283

命名實體識別運用於產品同義詞擴增

Using Named Entity Recognition Increases the Synonym of Products

洪智力 Chihli Hung
中原大學資訊管理學系
Department of Information Management
Chung Yuan Christian University
chihli@cycu.edu.tw

黃政華 Jheng-Hua Huang
中原大學資訊管理學系
Department of Information Management
Chung Yuan Christian University
joyhung1993@gmail.com

鍾瑞嘉 Rui-Jia Zhong
中原大學資訊管理學系
Department of Information Management
Chung Yuan Christian University
barry67024444ab@gmail.com

陳良圃 Liang-Pu Chen
財團法人資訊工業策進會
Institute for Information Industry
eit@iii.org.tw

楊秉哲 Ping-Che Yang
財團法人資訊工業策進會
Institute for Information Industry
maciaclark@iii.org.tw

摘要

本研究提出產品名稱機率比對法，嘗試解決命名實體識別領域中，常被忽略的同義詞辨識問題。在意見探勘領域中，正確並使用相同的用詞用語描述一項產品或服務，能有效

提升意見探勘或情感分析的成果。然而，口碑或意見為使用者自行產生的內容(UGC; user generated content)，口碑文章缺乏一定的撰寫規則與統一的命名規定，不同作者甚至相同作者對於相同的產品，常會產生命名不一致的現象，導致口碑敘述的產品名稱和正式產品名稱不一致。此產品名稱不一致的現象，在搜尋或整理口碑文章時，則會產生資訊遺漏的問題。本研究提出產品名稱機率比對法，透過產品共現詞彙集、Word2vec語言模型，擴增產品的同義詞，並使用機率比對方式，嘗試找到適合的產品同義詞，以解決資訊遺漏的問題。根據本研究的初步實驗顯示，本研究所提出的方法對於同義詞辨識問題，有發展的潛力。

Abstract

This research proposes the probability mapping approach for a product name to attempt to resolve the synonym identification problem, which is usually ignored in the field of Named Entity Recognition. Using the same name to describe a product or service may effectively improve the results of opinion mining or sentiment analysis. However, as WOM is a user generated content (UGC), different names may be used by the same or different users. Besides, there is no unified naming rule when writing the WOM. Even though the authors are the same or different, they may use different names to describe the same products. In this case, searching or organizing the WOM article without the consideration of the naming issue may lead to the problem of information loss. Thus, we propose the probability mapping approach via the co-occurrence naming dataset and the Word2vect language model in order to reduce the naming issue. According to our initially experimental results, the probability mapping approach for a product name present its potential in the naming issue.

關鍵詞：意見探勘、命名實體辨識、深度學習、同義詞辨識

Keywords: Opinion Mining, Named Entity Recognition, Deep Learning, Synonyms Identification.

一、 前言

本研究提出產品名稱機率比對法結合語意概念模型和Word2vec的方法[1]，以機率方式擴增產品的非正規用詞為目標產品的同義詞，以改善口碑所描述的產品名稱和廠商產品名稱不一致的問題，降低目標產品口碑搜尋時，所產生的資訊遺漏。

網際網路的發展，造就了消費者對於表達意見的方式有巨大的改變。消費者從被動的接收訊息，轉變成主動對於產品的使用經驗，在各種網際網路的平台上，發表並分享相關的評論[2]。網際網路上所散佈的電子口碑具有巨大行銷價值，消費者可以透過網路口碑，做出購買與否的決定[3]；廠商可以由流傳網路的口碑，分析消費者對於所屬產品的意見，作為產品改進的重要參考[4]。口碑意見以非結構化的文字資料呈現，轉化這些非結構化的大量資料為廠商和網民可以直接利用的知識，一直為意見探勘或情感分析領域的一項重要挑戰[2]。產品名稱不一致的問題，加深此挑戰的困難度，也讓意見探勘成為相關學術界和實務界共同關注的一項重要議題。

在台灣常使用的口碑資料可分為中文或英文，本研究鎖定中文口碑。中文口碑和英文口碑最大的不同處在於，中文並無自然存在的詞間斷詞符號，而英文的空白即為詞與詞之間的分隔依據。單一中文方塊字(Chinese character)的字義過於模糊，無法表達完整的概念，如輸贏的「輸」和運輸的「輸」，字型相同字義卻截然不同。所以對於中文的文字處理，需要再額外的處理中文的斷詞問題[5]。另一方面，口碑文章為使用者所產生的內容(UGC; user generated content)，撰寫風格與用詞用語不受規範，無法避免縮寫字、同音字、新創字、錯別字、別名、同義詞等非正規用詞的發生。當這些非正規用詞，發生於產品、服務或公司名稱時，導致採用關鍵字收集口碑時，因與正式名稱字型不同，會產生資訊遺漏的問題，例如欲尋找「義美豆奶」的口碑文章，則會遺漏誤打為「義美豆漿」的口碑文章。

定義產品、服務或公司名稱的比對的研究稱為命名實體識別(NER; name entity recognition)。常用的方法有二種，第一種為法則法，由人工建立正確的命名實體集，採取比對方式；第二種為機器學習法，使用機器學習演算法，從已標記的資料集建立命名實體識別模型[6]。文獻上，命名實體研究，以找出文章中事先定義的命名實體類別為主要的目的，例如找出人名、地名或是公司名的實例[7]，並忽略因縮寫字、同音字、新創字、錯別字、別名所產生的同義詞等非正規用詞的存在問題。亦即，傳統NER的研究，可以判斷命名實體A和B同屬於人名、地名或是公司名，但並不處理命名實體A和

命名實體B是否相同。因此，文獻上，仍然缺乏判斷命名實體間的同義詞關係的相關研究，本研究則專注於命名實體的同義詞研究，除了改善在命名實體中所產生的資訊遺漏的問題，也嘗試填補在此研究領域文獻上的不足。

二、 文獻探討

(一)、 命名實體識別

命名實體識別主要用於辨識名稱特徵的表達方式，這些特徵可以是人名、地名[8]。除了名稱單位之外，數字的表達辨識也是研究的範圍，如時間、日期、貨幣…等等[9]。命名實體識別是將資料進行資訊萃取，其主要的功能包括識別和分類某些種類的資訊元素名稱。因此，其結果可以作為語義標註、本體的建構等應用。同時命名實體識別也是意見探勘的基礎，在口碑意見中透過命名實體的技術可以改善意見的匹配結果[10]。

命名實體識別在自然語言處理中發展許久，命名實體識別主要分為以規則為主的辨識[11]、或是以機器學習法為主的辨識。常見使用於命名實體識別的機器學習法，如隱馬可夫模型(HMM; hidden Markov model)、決策樹(DT; decision tree)、最大熵支援向量機(MESVM; maximum entropy support vector machine)。命名實體識別所辨識的文字資訊種類也是相當多元的，目前主流的研究都是以英文為主[11]，也有以土耳其語為基礎的混合型命名實體識別，其透過規則的建立進行識別，此模型可以同時對於新聞、財經新聞、童話故事、歷史文字資料進行命名實體的識別，其所提出的模型有良好的識別精確率，缺點則為必須依賴繁瑣的規則[11]。

(二)、 文字向量表示

向量空間模型(VSM; vector space model)為主要的文字向量表示方法，廣泛運用於文字探勘、資訊檢索、自然語言處理等研究領域中。向量空間模型將字視為向量的純量或屬性，資料集中的所有不同字構成高維度向量空間，一篇文章使用一個文字向量來表示，兩篇概念類似的文章因使用類似的文字，預期映射到相近的向量空間。單純的文字亦可向量化，相似的文字將出現於概念類似的文章中，因此，概念類似的文字，預期能映射的相近的向量空間中。Baroni [12]將分析文字向量間的關係區分為兩種方法：詞頻

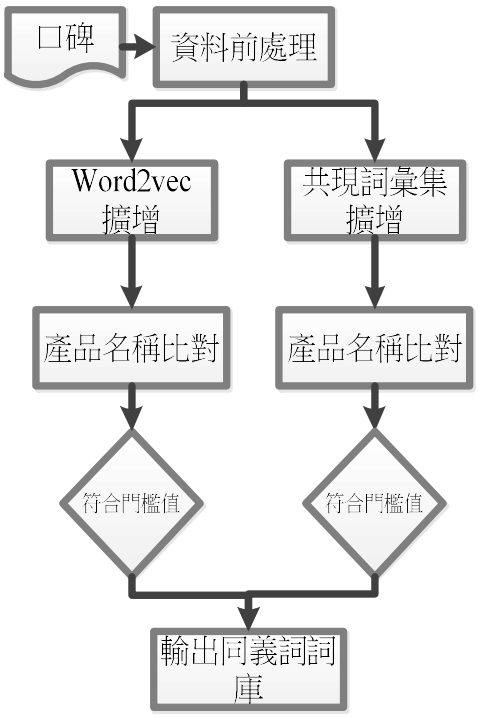
統計法(count-based method)，如潛在語意分析(LSA; latent semantic analysis)和預測法(predictive method)，如類神經機率語言模型(NPLM; neural probabilistic language model)。詞頻統計法從大量文字資料中統計目標字和鄰近字的共現關係，預測法藉由鄰近字預測目標字。

機器學習方法主要運用電腦模擬方式，採用演算法則，其中深度學習(DP; deep learning)在 AlphaGo 機器人程式連續挑戰韓國圍棋棋王成功後聲名大噪。深度學習目前已使用於計算機視覺的研究領域中，如影像分類、目標偵測、影像檢索，其學習模型可以分為四類，如限制波爾茲曼機器(restricted Boltzmann machine)、細胞式類神經網路(cellular neural network)、自編碼器(autoencoders)、稀疏性堆疊自編碼器(stacked sparse autoencoder)，深度學習的模型為類神經網路的一種衍生模型[13]，主要的論點在於採用大量的資料往往能讓簡單的演算模型的效果，高於只使用較少資料但卻設計複雜的演算模型[1]。深度學習除了可以運用於計算機視覺中，也可用於文字資料之中，如用於文字資料的摘要萃取[14]。

Word2vec 為 Google 於 2013 釋出之開源工具之一，根據 Mikolov et al. (2013) 發表的文章所開發。Word2vec 主要的功能為將辭彙轉化成為文字向量，透過向量空間模型可以計算語義間的相似度，屬於類神經機率語言模型的一種。其模型的預測方式分為兩種，分別是 CBOW (continuous bag-of-word)和 Skip-Gram [1]。CBOW 並未使用類神經網路常用的非線性隱藏層(non-linear hidden layer)，在輸入層的所有單詞皆共享隱藏層，其訓練目標是給定一個目標詞的上下文鄰近詞，以預測目標詞出現的機率，此方法適合較小的資料集。Skip-gram 與 CBOW 不同，使用一串文字中的一個目標詞，來預測鄰近詞發生的機率，此方法適合較大的資料集。Word2vec 之所以會受到關注是因為 Word2vec 的高效率和可用性，因為 Word2vec 不像類神經網路架構的方式，必須使用大量的訓練詞彙向量，即可預測不同目標詞的鄰近詞出現的機率[15]。Word2vec 適合做為文字向量特徵的運算工具，如 Zhang [16]透過 Word2vec 進行語意特徵的計算再利用 SVM 分類器進行文本的情感分類，經實驗得知，此方法能夠得到相當高的分類正確率。

三、 產品名稱機率比對法

本研究中提出的產品名稱機率比對法擴增同義詞詞庫模型，分別為手動擴增和自動化輸入。手動擴增則為以人工方式建立已知商品同義詞詞庫，自動化擴增如下圖一，透過共現詞彙集、Word2vec、產品名稱比對，分別敘述如下。



圖一、產品名稱機率比對法

(一)、 口碑收集

針對中文情感口碑網站，如 PTT 網路論壇(http://ptt.cc)蒐集口碑資料，並去除 HTML 標籤及非本文內容，提取口碑文本。本研究口碑文章的收集依賴關鍵字，因此所收集的口碑文章，均含有目標關鍵字，在初步的研究中，本研究採取模擬方式，將所收集的口碑文章添加目標關鍵字的同義詞，如複製所收集有關「義美豆奶」的口碑文章，但將其目標關鍵字，如「義美豆奶」修改為「義美豆漿」。

(二)、 資料前處理

依前步驟所收集的口碑雖已無 HTML 標籤、CSS 標籤、Java script 語法，但仍含有部分雜訊，如標點符號或是特殊的標籤，本研究去除非中文字元，採用 Jieba (Jieba

Chinese text segmentation, https://github.com/fxsjy/jieba) 斷詞。Jieba 斷詞系統有三種斷詞方式，分別為精確模式、全模式、搜尋引擎模式。由於本研究需要進行詞彙的詞性篩選，因此採用 Jieba 精確斷詞模式。經多次實驗，本研究最後保留的詞性為名詞、動詞和形容詞，因為在中文口碑中此三類詞彙和命名實體關係最為密切。Jieba 斷詞結果仍有過於零碎的問題，對於目標詞彙，則採取完整保留策略，亦即不斷詞，如目標詞彙為「義美豆奶」，則確保不會將其斷開，而產生「義美」、「豆奶」。

(三)、 共現詞彙集建立

目標詞彙的共現詞彙和其同義詞的共現詞彙應具備某種程度的相似性，例如「義美豆奶」的共現詞彙和「義美豆漿」的共現詞彙應該類似。因此，我們從所 PTT 網路論壇收集的口碑文章中，經過資料前處理步驟，依照字詞出現次數，建立共現詞彙集，如表一。產品詞彙組合的相似度會透過公式一，計算餘弦相似度，在此的詞彙相似度組合僅以產品詞彙與口碑詞彙進行運算，不進行口碑詞彙間的運算，如運算「義美豆奶」與「陽光」、「義美豆奶」與「穀物」等的餘弦相似度，不計算「陽光」與「好喝」等口碑詞彙間的餘弦相似度。因此，我們可以得到所有口碑詞彙和產品詞彙的共現值。

表一、共現詞彙集

	陽光	義美豆奶	好喝	無糖
陽光	-	1	1	1
義美豆奶	1	-	1	2
好喝	1	1	-	1
無糖	1	2	1	-

$$\text{Cos}(a,b)=\frac{\sum_i a_i \times b_i}{\sqrt{\sum a_i^2}\times\sqrt{\sum b_i^2}} \qquad (1)$$

其中 a 表口碑文章中的產品詞彙，b 表口碑文章中的口碑詞彙。

(四)、 相似度比對

當輸入未知產品的口碑文章後，如該文章包含「義美豆漿」但並未包含「義美豆奶」，對於未知產品的口碑文章，同樣經過資料前處理步驟後，其口碑詞彙高於共現詞彙集一

定門檻值(如 60%)，則此口碑文章和目標產品詞彙具備高度關係。如某口碑文章斷詞後的口碑詞彙(全家 有 無糖 義美豆漿 好喝)，在共現詞彙集中發現，「無糖」、「好喝」、「全家」均和「義美豆奶」有高度共現關係，則此口碑文章所描繪的產品可推定為「義美豆奶」，此時會將口碑中的詞彙(全家、有、無糖、義美豆漿、好喝)視為可能的同義詞讓後續的產品名稱比對法進行比對。

然而，此方法並未考慮產品詞彙和口碑詞彙間的字序關係，因此，互相競爭的產品，如統一和義美的豆奶，其所使用的口碑詞彙有可能完全相同，因而在此階段，只能找出類別相同的產品。因此，我們進一步採用深度學習中的 Word2vec 方法，以弭補共現詞的缺點。

(五)、 同義詞擴增

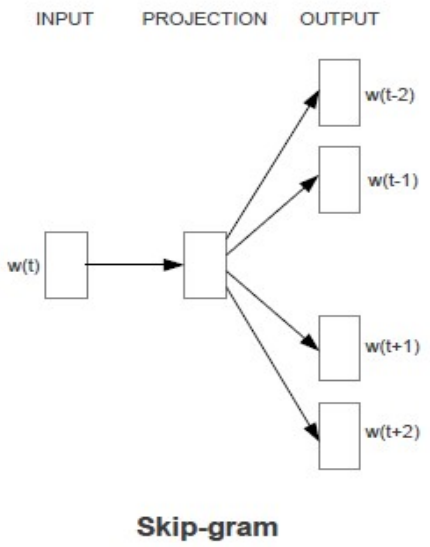
Word2vec 模型屬於深度學習的一種應用，訓練詞彙時可以選擇兩種模式進行詞彙的訓練：CBOW 和 skip-gram，由於 skip-gram 適合用來處理大量資料，本研究的初期實驗採用的方式為 skip-gram，其類神經網路示意圖，如圖二，能夠根據目標詞彙和其前後詞彙的關係建立模型，當輸入目標詞彙時，產生和目標詞彙最具關係的前後詞彙。此方法和本研究提出，利用共現詞彙集比對的方式不同，會考慮產品詞彙和口碑詞彙間的字序關係。當「義美豆奶」和「義美豆漿」所產生的共現詞具備高度的重複性時，則可推斷為同義詞。

Word2vec 訓練時，必須輸入所要考慮的字序長度，當考慮的字序愈長時，所需訓練的時間愈久，若設定的字序長度為 2，亦即考慮目標詞(如義美豆奶)前 1~2 個字和後 1~2 個字的關係。舉例如下：

(1)全家 有 義美豆奶 無糖 好喝

(2)全家 有 義美豆漿 無糖 好喝

此範例中兩句中只有「義美豆奶」與「義美豆漿」不同，整句話所描述的主角其實都是「義美豆奶」這一個產品。預期 Word2vec 能夠藉著字序關係找出較具可能的同義詞彙。



圖二、Skip-gram 模型[1]

(六)、 產品名稱比對

本研究所提出的產品名稱機率比對法為，輸入一產品名稱後，進入到比對的過程，過程中會將已斷詞的口碑文章與產品名稱進行比對。透過此比對方法會找尋相同或相似的名稱，如果找到相同的產品名稱，就會輸出口碑文章並結束比對流程。如果找到相似的產品名稱，計算相似值，當符合門檻值時，以相似度最高者為其產品名稱，並將相似的產品名稱列入產品語料庫中，相似度的計算方式採取口碑文章中產品單一中文字(Chinese character)佔所比對的口碑語料庫產品名稱總字數百分比。相似度公式如公式(2)所示：

$$\text{最大相似度} = \text{Max}\left(\frac{w_i \cap p_j}{w_i} \times \frac{w_i \cap p_j}{p_j}\right) \qquad (2)$$

其中 w_i 表口碑文章中 i 產品的總字數，p_j 表產品詞庫中 j 產品的總字數，$w_i \cap p_j$ 表相同的字數。

四、 實驗

(一)、 實驗設計

本研究的實驗尚處於測試階段，目前資料數量為 2141 筆口碑資料，資料集所包含之口碑的產品名稱為義美豆奶、王子麵、可口可樂，資料集來源為透過資策會的 API(http://api.ser.ideas.iii.org.tw/)擷取 PTT 的口碑資料。並由人為方式產生同樣筆數但將「義美豆奶」產品詞彙代換為「義美豆漿」、「王子麵」代換為「小王子麵」、「可口可樂」代換為「可樂」，此人工資料做為測試資料集，建立目的是本研究是否能夠找到目標的同義詞彙，在實驗中 Word2vec 模型參數如表二，此參數依照 Word2vec 官方文件進行參數設定 (https://code.google.com/archive/p/word2vec/)。

表二、Word2vec 參數表

參數值	參數模式
cbow =0	使用 Skip-gram 模型
Size= 400	輸出詞向量的維度
Window= 5	訓練時包含前後文的長度
Hs= 1	使用 Hierarchical Softmax 最佳化
iter =10	迭代訓練回數

(二)、 實驗結果

透過 2141 筆口碑資料所訓練的義美豆奶、王子麵、可口可樂的共現詞彙集，與人工建立之測試口碑資料經過比對、比對門檻為 50%，與義美豆奶可能為同義詞的詞彙為，統一豆漿、全聯、燕麥、義美豆漿…等等。透過共現詞比對所產生的詞彙眾多在此不一一列出。根據 Word2vec 模型參數所計算出與「義美豆奶」相似的詞彙如表三所示，研究將訓練完成模型，透過 Python 的 Gensim 套件計算詞彙的相似度，其顯示最相似詞排序分別為義美豆漿、義美、食品。在此會將可能的同義詞進行後續的產品名稱比對。

表三、義美豆奶相似詞彙表

產品名稱	相似值
義美豆漿	0.999682784081
義美	0.999646663666
食品	0.999582290649

本研究除了以「義美豆奶」做為實驗的產品詞彙，也針對王子麵、麥香紅茶、可口可樂三種產品進行同義詞擴充實驗，其可能的同義詞如表四所示。

表四、相似詞彙表

產品名稱	相似值	產品名稱	相似值	產品名稱	相似值
小王子麵	0.9982483	麥紅	0.999598503	可樂	0.997811
冬粉	0.9765478	服務	0.995698630	太古	0.985727
烏龍	0.9749084	主場	0.994993865	百事可樂	0.985078

經過前面兩個步驟所產生的可能同義詞，與其所對應的產品名稱比對後的結果如表五，其篩選相似值門檻為 0.5，透過本方法自動化篩選過後之同義詞，將提供人工進行後續的同義詞詞庫建置。

表五、產品名稱比對相似詞彙

產品名稱	同義詞	相似值
義美豆奶	義美豆漿	0.5625
王子麵	小王子麵	0.75
麥香紅茶	麥香紅	0.75
麥香紅茶	麥香奶茶	0.5625
可口可樂	可樂	0.5

五、 結論與未來展望

本研究所提出的產品機率比對法，主要利用產品詞彙與口碑詞彙的共現關係，找出和目標產品具備高度關係的同義詞，接著使用 Word2vec 模型進行同義詞擴增。根據本研究的初步實作結果，發現本研究所提出的方法，具備高度的應用潛力。文獻上，在命名實體識別(NER)的研究中，很少被應用於同義詞的辨識，本研究的提出，除了能運用於口碑搜尋中，減少資訊遺漏的問題外，也對於 NER 研究，擴充同義詞研究的方向。

未來的發展可以更進一步採用機率模型如交互資訊(MI; mutual information)模型，改善共現詞彙集，另外也可運用主題模型(topic model)，找出更具代表性的詞彙，讓向量空間模型更加緊密。

六、 致謝

本研究依經濟部補助財團法人資訊工業策進會「105 年度虛實整合智慧商務關鍵技術與平台研發計畫(3/4)」辦理。

參考文獻

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *ICLR Workshop*, Jan. 2013.

[2] F. H. Khan, U. Qamar, and S. Bashir, “SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection,” *Appl. Soft Comput.*, vol. 39, pp. 140–153, Feb. 2016.

[3] J. Berger, “Word of mouth and interpersonal communication: A review and directions for future research,” *J. Consum. Psychol.*, vol. 24, no. 4, pp. 586–607, Oct. 2014.

[4] B. Liu, M. Hu, and J. Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web,” in *Proceedings of the 14th International Conference on World Wide Web*, New York, NY, USA, 2005, pp. 342–351.

[5] N. Xue, “Chinese Word Segmentation as Character Tagging,” *Comput. Linguist. Chin. Lang. Process.*, vol. 8, no. 1, pp. 29–48, Feb. 2003.

[6] M. Konkol, T. Brychcín, and M. Konopík, “Latent semantics in Named Entity Recognition,” *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3470–3479, May 2015.

[7] 劉昭宏 and 吳宗憲, “使用前後文篩選之快速具名實體擷取技術,” *電腦與通訊*, no. 154, pp. 41–47, Dec. 2013.

[8] R. Agerri and G. Rigau, “Robust multilingual Named Entity Recognition with shallow semi-supervised features,” *Artif. Intell.*, vol. 238, pp. 63–82, Sep. 2016.

[9] C. V. Sundermann, M. A. Domingues, M. da S. Conrado, and S. O. Rezende, “Privileged contextual information for context-aware recommender systems,” *Expert Syst. Appl.*, vol. 57, pp. 139–158, Sep. 2016.

[10] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, “Named Entity Recognition: Fallacies, challenges and opportunities,” *Comput. Stand. Interfaces*, vol. 35, no. 5, pp. 482–489, Sep. 2013.

[11] D. Küçük and A. Yazıcı, “A hybrid named entity recognizer for Turkish,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2733–2742, Feb. 2012.

[12] M. Baroni and G. Dinu, “Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors,” in *In ACL*, 2014.

[13] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

[14] S. Zhong, Y. Liu, B. Li, and J. Long, “Query-oriented unsupervised multi-document

summarization via deep learning model," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8146–8155, Nov. 2015.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *NIPS*, 2013.

[16] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, Mar. 2015.