I800982

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I800982 B**

(45)公告日： 中華民國 112 (2023) 年 05 月 01 日

(21)申請案號：110142549 (22)申請日： 中華民國 110 (2021) 年 11 月 16 日

(51)Int. Cl.： *G06F16/38 (2019.01)* *G06N3/08 (2006.01)*

(71)申請人：宏碁股份有限公司 (中華民國) ACER INCORPORATED (TW)
新北市汐止區新台五路一段 88 號 8 樓

(72)發明人：林意淳 LIN, YI-CHUN (TW) ；蔡岳洋 TSAI, YUEH-YARNG (TW) ；林品銓 LIN, PIN-CYUAN (TW) ；潘可涵 PAN, KE-HAN (TW) ；朱昇瑋 CHU, SHENG-WEI (TW)

(74)代理人：葉璟宗；卓俊傑

(56)參考文獻：

TW	201117024A	TW	201931170A
TW	202129533A	CN	110688491A
CN	111159416A		

審查人員：吳家豪

申請專利範圍項數：10 項 圖式數：4 共 24 頁

(54)名稱

文章標記資料的產生裝置及其產生方法

(57)摘要

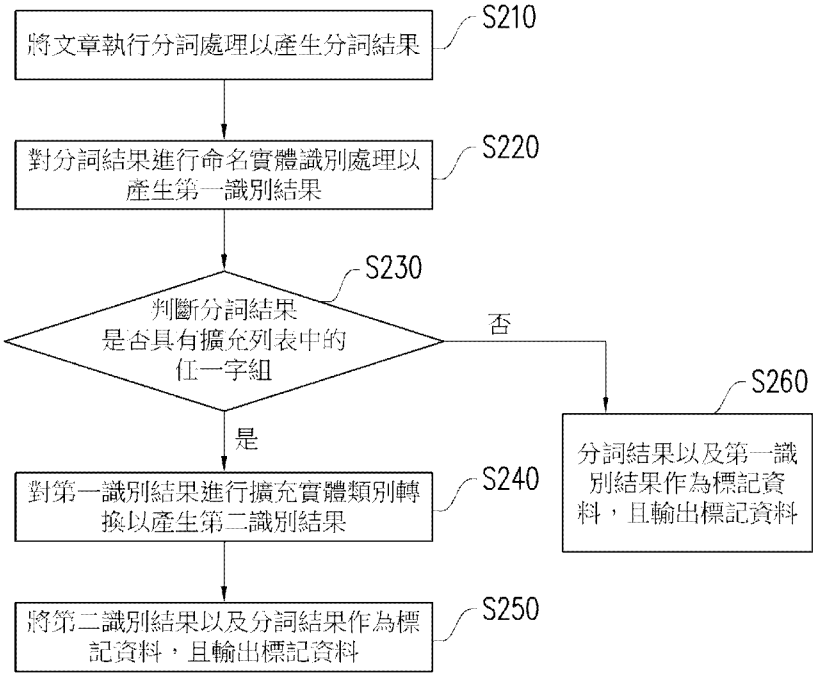
一種文章標記資料的產生裝置及其產生方法。文章標記資料產生方法包含：將文章執行分詞處理以產生分詞結果；對分詞結果進行命名實體識別處理以產生第一識別結果；判斷分詞結果是否包括有擴充列表中的任一字組；對第一識別結果進行擴充實體類別轉換以產生第二識別結果；將第二識別結果以及分詞結果作為標記資料。

A device and method for generating article markup information are provided. The method for generating article markup information includes: generating a segmentation result by executing a segmentation processing on an article; generating a first recognition result by executing name entity recognition on the segmentation result; identifying whether the segmentation result includes the same word as any word in an expansion list; executing an expanded entity classification conversion on the first recognition result to generate a second recognition result; being used the second recognition result and the segmentation result as a markup information.

指定代表圖：

符號簡單說明：

S210、S220、S230、S240、S250、S260:步驟



【圖3】

專利案號：110142549

公告本

I800982

【發明摘要】

【中文發明名稱】文章標記資料的產生裝置及其產生方法

【英文發明名稱】DEVICE AND METHOD FOR GENERATING ARTICLE MARKUP INFORMATION

【中文】一種文章標記資料的產生裝置及其產生方法。文章標記資料產生方法包含：將文章執行分詞處理以產生分詞結果；對分詞結果進行命名實體識別處理以產生第一識別結果；判斷分詞結果是否包括有擴充列表中的任一字組；對第一識別結果進行擴充實體類別轉換以產生第二識別結果；將第二識別結果以及分詞結果作為標記資料。

【英文】A device and method for generating article markup information are provided. The method for generating article markup information includes: generating a segmentation result by executing a segmentation processing on an article; generating a first recognition result by executing name entity recognition on the segmentation result; identifying whether the segmentation result includes the same word as any word in an expansion list; executing an expanded entity classification conversion on the first recognition result to generate a second recognition result; being used the second recognition result and the segmentation result as a markup information.

【指定代表圖】圖3。

【代表圖之符號簡單說明】

S210、S220、S230、S240、S250、S260:步驟

【特徵化學式】

無

【發明說明書】

【中文發明名稱】文章標記資料的產生裝置及其產生方法

【英文發明名稱】DEVICE AND METHOD FOR GENERATING ARTICLE MARKUP INFORMATION

【技術領域】

【0001】 本發明是有關於一種文章標記資料的產生裝置及其產生方法，且特別是有關於一種可以自動產生標記資料的文章標記資料的產生裝置及其產生方法。

【先前技術】

【0002】 在人工智慧、機器學習模型及深度學習模型的建立中，訓練資料為重要的要件之一。其中，用於監督式學習的訓練資料，每筆資料都需要有相對應的答案標記。

【0003】 目前的技術是透過人工手動地進行逐筆資料的標記，導致耗費時間且容易發生標記錯誤的情況，進而造成後續模型訓練表現不佳或是訓練過程中發生錯誤。因此，現有產生用於訓練模型的標記資料仍有改善的空間。

【發明內容】

【0004】 本發明提供一種文章標記資料的產生裝置及其產生方法，可根據預設的字組以及實體類別產生標記文章中的字組，進而自動產生可用於訓練模型的標記資料。

【0005】 本發明的一種文章標記資料的產生裝置，包含處理器、以及收發器。處理器耦接收發器，且處理器用以：將文章執行分詞處理以產生分詞結果；依據命名實體識別模型對分詞結果進行命名實體識別處理以產生第一識別結果；依據擴充列表判斷分詞結果是否包括有擴充列表中的任一個字組；當分詞結果包括有擴充列表中的字組，依據擴充列表以及分詞結果對第一識別結果進行擴充實體類別轉換以產生第二識別結果；以及，將第二識別結果以及分詞結果作為標記資料且輸出標記資料。

【0006】 本發明的文章標記資料的產生方法包括：處理器將文章執行分詞處理以產生分詞結果；依據命名實體識別模型處理器對分詞結果進行命名實體識別處理以產生第一識別結果；依據擴充列表處理器判斷分詞結果是否包括有擴充列表中的任一個字組；當分詞結果包括有擴充列表中的字組，依據擴充列表以及分詞結果處理器對第一識別結果進行擴充實體類別轉換以產生第二識別結果；以及，處理器將第二識別結果以及分詞結果作為標記資料，且輸出標記資料。

【0007】 基於上述，本發明的文章標記的產生裝置可自動地產生具有關於擴充列表的實體類別的文章標記資料。並且，標記資料可用於做為命名實體識別模型的訓練資料。

【圖式簡單說明】

【0008】

圖 1 根據本發明的一實施例繪示一種文章標記資料的產生裝置的示意圖。

圖 2 根據本發明的一實施例繪示儲存媒體的示意圖。

圖 3 根據本發明的一實施例繪示一種文章標記資料的產生方法的流程圖。

圖 4 根據本發明的另一實施例繪示一種文章標記資料的產生方法的流程圖。

【實施方式】

【0009】 為了使本發明之內容可以被更容易明瞭，以下特舉實施例作為本發明確實能夠據以實施的範例。另外，凡可能之處，在圖式及實施方式中使用相同標號的元件/構件/步驟，係代表相同或類似部件。

【0010】 圖 1 根據本發明的一實施例繪示一種文章標記資料的產生裝置的示意圖。文章標記資料的產生裝置 1 可包含處理器 110、以及收發器 120。文章標記資料的產生裝置 1 可用於自動地產生標記資料，以用於擴充命名實體識別模型的訓練樣本，進而強化與擴充命名實體識別模型的識別範圍與功效。

【0011】 處理器 110 例如是中央處理單元(central processing unit，CPU)，或是其他可程式化之一般用途或特殊用途的微控制單元（ micro control unit，MCU)、微處理器（ microprocessor)、數位信號處理器（ digital signal processor，DSP)、可程式化控制器、特殊

應用積體電路（application specific integrated circuit，ASIC）、圖形處理器（graphics processing unit，GPU）、影像訊號處理器（image signal processor，ISP）、影像處理單元（image processing unit，IPU）、算數邏輯單元（arithmetic logic unit，ALU）、複雜可程式邏輯裝置（complex programmable logic device，CPLD）、現場可程式化邏輯閘陣列（field programmable gate array，FPGA）或其他類似元件或上述元件的組合。處理器 110 可耦接至收發器 120。

【0012】 收發器 120 以無線或有線的方式傳送及接收訊號。收發器 130 還可以執行例如低噪聲放大、阻抗匹配、混頻、向上或向下頻率轉換、濾波、放大以及類似的操作。

【0013】 於另一實施例中，產生裝置 1 更可包括儲存媒體 130，儲存媒體 130 耦接處理器 110。儲存媒體 130 例如是任何型態的固定式或可移動式的隨機存取記憶體（random access memory，RAM）、唯讀記憶體（read-only memory，ROM）、快閃記憶體（flash memory）、硬碟（hard disk drive，HDD）、固態硬碟（solid state drive，SSD）或類似元件或上述元件的組合，而用於儲存可由處理器 110 執行的多個模組或各種應用程式。如圖 2 在本實施例中，儲存媒體 130 可儲存包含爬蟲模組 131、分詞處理模型 132、命名實體識別（named entity recognition，NER）模型 133 以及訓練模組 134 等多個模組，其功能將於後續說明。處理器 110，並且存取和執行儲存於儲存媒體 130 中的多個模組和各種應用程式。

【0014】 下文中，將搭配文章標記資料的產生裝置 1 中的各項裝

置、元件及/或模組說明本發明實施例所述之方法。本方法的各個流程可依照實施情形而隨之調整，且並不僅限於此。

【0015】 圖 3 根據本發明的一實施例繪示一種文章標記資料的產生方法的流程圖，其中這文章標記資料的產生方法可由如圖 1 所示的文章標記資料的產生裝置 1 實施。在本實施例中，處理器 110 將文章執行分詞處理以產生一分詞結果(步驟 S210)。於一實施例中，處理器 110 依據爬蟲模組，而透過爬蟲技術取得多篇文章（article）。舉例來說，爬蟲模組可通過收發器 130 存取新聞網站或醫療網站，並且利用爬蟲技術以從這新聞網站或醫療網站中取得多篇新聞與文章。在另一實施例中，爬蟲模組可根據預設週期重複地儲存文章至儲存媒體 130 之中。

【0016】 在一實施例中，處理器 110 透過分詞處理模型 132 對待標記文章(即，文章)進行分詞處理。舉例來說，本發明所述之分詞處理模型 132 可透過雙向編碼器表徵（bidirectional encoder representations from transformers，BERT）演算法的詞法分析器(Tokenizer)等執行，但本案不應以此為限。舉例來說，待標記文章為「John believes that only around 20% of the country's 126 million population has been fully vaccinated against Covid-19.」，經過處理器 110 對這待標記文章執行分詞處理以獲得對應標記文章的分詞結果。在這實施例中，這分詞結果為「John,believes,that,only,around,2,%,of,the,country,',s,126,million,population,has,been,fully,vaccinated,against,Covid,-,19,.」。由上述

可以得知，本實施例所使用的分詞處理為標點符號與字詞皆進行分詞的分詞處理，但本案不應以此為限。

【0017】 在一實施例中，在處理器 110 獲得分詞結果之後，處理器 110 依據命名實體識別模型 133 對分詞結果進行命名實體識別處理以產生第一識別結果(步驟 S220)。具體而言，處理器 110 透過命名實體識別模型 133 對分詞結果執行命名實體識別處理。在另一實施例中，步驟 S210 與步驟 S220 可整合於一個步驟之中，也就是說處理器 110 依據命名實體識別模型 133 將文章進行命名實體識別處理後，即可獲得分詞結果以及對應分詞結果的第一識別結果。

【0018】 舉例來說，命名實體識別模型 133 是基於包括 Transformer 架構的自然語言處理演算法的深度學習所訓練的。舉例來說，命名實體識別模型 133 可透過雙向編碼器表徵（bidirectional encoder representations from transformers，BERT）演算法、ELMo 演算法或 GPT-2 演算法所訓練。藉由命名實體識別模型 133，處理器 110 將分詞結果中與命名實體識別模型 133 中相同的字組標記為對應的實體類別。舉例來說，處理器 110 根據命名實體識別模型 133 對上述分詞結果執行命名實體識別處理後，處理器 110 可以獲得對應的第一識別結果。這第一識別結果為「B-PER,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O」。在本實施例中，B-PER 表示人名，O 表示非命名實體或其他，但本案不應以此為限。命名實體識別模型 133 可產生對應於字組的實體類別

（entity classification）。舉例來說，命名實體識別模型 126 可將字組分類為「人名」、「地名」、「機構名」、「時間」、「數字」、「其他實體」或「其他」等實體類別的其中之一。

【0019】 在處理器 110 獲得第一識別結果之後，在一實施例中，處理器 110 依據擴充列表判斷分詞結果是否具有擴充列表中的任一個字組(步驟 S230)。在一實施例中，這擴充列表中的多個字組是經分詞處理且/或經格式統一處理的字組，且上述格式統一處理可以為將每一字組中的文字統一轉換為大寫形式，或將每一字組中的文字統一轉換為小寫形式的文字。並且，擴充列表為使用者預先設定的字組列表。

【0020】 舉例來說，這擴充列表為使用者預先設定的傳染病字組列表，字組的實體類別均為 DIS，例如，表(1)：

表(1)

字組	同義字組	同義字組	同義字組	同義字組
Covid-19	Wuhan pneumonia	SARS-CoV-2	Corona virus 2019	Coronavirus pandemic
Dengue fever	dengue virus	dengue	DEN-1	NS1 rapid test
ZIKA	ZIKV	Zika virus	Zika virus infection	Microcephaly
novel influenza	avian flu	Novel Influenza A Virus Infections	Pandemic influenza	H5N1

由擴充列表的範例(即，表(1))可以得知，擴充列表包括字組、同義

字組。使用者可自行設定與擴充這擴充列表中的字組以及同義字組，舉例來說，使用者可新增關於書名的擴充列表、法律用語的擴充列表或其他專有名詞的擴充列表。並且，於步驟 S230 中，處理器 110 透過擴充列表所包括的字組以及同義字組以提高其判斷的精準度。舉例來說，當文章(即，待標記文章)中的 dengue 是以「DEN-1」、「Dengue fever」，或以其他方式描述 dengue 之時，或是文章中的 West Nile Fever 是以「West Nile virus」、「WNV」等其他方式描述 West Nile Fever 之時，處理器 110 皆能夠根據擴充列表中的字組及對應的同義字組以判斷分詞結果中是否包括擴充列表中任一字組或任一同義字組，以具有高精準度的效益。擴充列表中的同義字詞也可用於統一疾病名稱(或實體類別)使用，例如文章中 dengue fever 可被標示為 DIS 實體類別或是 Dengue 名稱，本發明不在此限。

【0021】 換句話說，在步驟 S230 中，處理器 110 可判斷擴充列表中的字組與分詞結果(即，經分詞處理後的文章)中任一字組是否匹配。若擴充列表中的字組與分詞文章中的字組匹配，則進入步驟 S240。若擴充列表中的字組與分詞結果中的字組不匹配，則進入步驟 S260。舉例來說，若分詞結果中包括「(dengue, fever)」字組，則處理器 110 可依據擴充列表(如表(1))中字組的分詞結果包括「(dengue, fever)」分詞字組，而判斷分詞結果與擴充文章的字組匹配。若分詞結果中不包含擴充列表中的任一字組的分詞，處理器 110 則判斷分詞結果與擴充列表的字組不匹配。

【0022】 在一實施例中，當分詞結果(即，經分詞處理的文章)包括有擴充列表中的任一個字組/任一字組的分詞結果，處理器 110 依據擴充列表以及分詞結果對第一識別結果進行擴充實體類別轉換以產生一第二識別結果(步驟 S240)。在擴充實體類別轉換中，處理器 110 根據分詞結果中的文字(即，單詞或字組)與擴充列表的字組/字組的分詞結果相匹配(即，相同)，而將第一識別結果中對應的實體類別轉換成擴充列表中的對應的實體類別。舉例來說，分詞結果為「 John,believes,that,only,around,2,%,of,the,country,',s,126,million,population,has,been,fully,vaccinated,against,Covid,-,19,.」，且第一識別結果為「 B-PER,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O」，而擴充列表(如表(1))中字組的分詞結果包括「Covid,-,19」。處理器 110 根據分詞結果中的「Covid,-,19」把第一識別結果中對應於「Covid,-,19」的原實體類別「O,O,O」的轉換為「B-DIS,I-DIS,I-DIS」，以獲得第二識別結果。在這實施例中，第二識別結果為「 B-PER,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O, B-DIS,I-DIS,I-DIS,O」。需補充說明的是，經過分詞處理之後的文章，這文章之中的標點符號也會被逐一斷開。

【0023】 在一實施例中，處理器 110 獲得第二識別結果之後，處理器 110 將第二識別結果以及分詞結果作為標記資料，且輸出標記資料(步驟 S250)。

【0024】 在一實施例中，當分詞結果中不包括擴充列表中的任一字組/任一字組的分詞結果，處理器 110 將分詞結果以及第一識別結果作為標記資料，且處理器 110 輸出標記資料(步驟 S260)。在本發明之中，處理器 110 將標記資料作為訓練資料以及驗證資料，進而用以訓練命名實體識別模型。在另一實施例中，處理器 110 將標記資料根據不同命名實體識別模型所對應的標記資料格式及檔案類型寫成對應的檔案類型(例如，csv、xml、json 或 txt)。如此一來，透過本發明的產生裝置以及產生方法，可以正確無誤地進行大量資料的自動標記，進而自動地產生可用於訓練模型的標記資料、節省人力成本以及提升模型的效能。值得說明的是，透過本發明的文章標記資料的產生裝置 1 與產生方法也可應用於其他需要重新標記文件的情況，本案不應僅以用於訓練模型為限。

【0025】 在另一實施例中，在步驟 S230 中，處理器 110 更依據擴充列表中的每一字組的分詞結果以及對應字組的分詞結果的多個窗口長度，對分詞結果進行搜尋處理以提高識別文字時的準確度以及降低錯誤發生率，來判斷分詞結果是否包括任一字組。具體而言，運算模組 122 可判斷分詞結果中的字組是否為擴充列表中多個字組的分詞結果的其中之一，並且判斷過程中包括比對擴充列表中字組的分詞結果的窗（window）尺寸(即，窗口長度)。每一字組的分詞結果皆具有對應的窗尺寸，舉例來說，字組「Covid-19」的分詞結果為(Covid,-,19)，窗尺寸為 3(即，搜尋長度為 3)；字組「Dengue」的分詞結果為(Dengue)，窗尺寸為 1(即，搜尋長度為

1) ；字組「SARS-CoV2」的分詞結果為(SARS,-,CoV2)，窗尺寸為3(即，搜尋長度為 3)。以字組「Covid-19」為例，處理器 110 在步驟 S230 中以搜尋長度為 3 且字組的分詞結果為(Covid,-,19)對文章的分詞結果進行逐字的搜尋處理。

【0026】 圖 4 根據本發明的另一實施例繪示一種文章標記資料的產生方法的流程圖。如圖 4 所示，步驟 S240 中更包括：依據擴充列表與第一識別結果，處理器 110 將分詞結果與擴充列表中字組的分詞結果進行比對以獲得至少一字組於分詞結果中對應的多個位置(index)(步驟 S241)；以及依據上述位置及擴充列表的預設實體類別，處理器 110 將第一識別結果中的對應於這些位置的實體類別轉換為擴充列表的預設實體類別，以產生第二識別結果(步驟 S242)。舉例來說，分詞結果為「John,believes,that,only,around,2,%,of,the,country,',s,126,million,population,has,been,fully,vaccinated,against,Covid,-,19,.」，第一識別結果為「B-PER,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O」，且擴充列表如表(1)所示，經過比對處理，處理器 110 獲得擴充列表中的字組「Covid-19」對應於分詞結果中第 21 個、第 22 個以及第 23 個的字組。接著，處理器 110 將第一識別結果中第 21 個、第 22 個以及第 23 個的實體類別「O,O,O」轉換為「B-DIS, I-DIS, I-DIS」的實體類別。據此，第二識別結果為「B-PER,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O, B-DIS,I-DIS,I-

DIS,O」。據此，本發明將對應字組的第一個字的實體類別標記為「B-實體類別名稱」，且將對應字組的其餘字標記為「I-實體類別名稱」，實體類別名稱為 DIS，但本發明不在此限。 在另一實施例中，若對應字組為單詞，舉例來說「dengue」，第一識別結果為「O」，第二識別結果為「B-DIS」。換句話說，I-實體類別名稱只會出現在複數對應字詞中。當文章中包含多個疾病名時，透過本發明的文章標記資料的產生方法，進而將文章中所有疾病名的實體類別進行修正。換句話說，本發明會對待標記文章掃過一遍擴充列表中的字組及其同義詞，如果待標記文章有多個疾病(即，與擴充列表的字組相對應)，例如疾病 A 、疾病 B 、疾病 C ，且疾病 A、疾病 B、疾病 C 可為相同疾病名或互為同義詞或為不同疾病名，本發明的產生裝置 1 及產生方法會依序記住疾病 A 、疾病 B 、疾病 C 於待標記文章及對應的分詞結果中的位置。據此，本發明會將識別結果中對應於疾病 A、疾病 B、疾病 C 位置的實體類別轉換為擴充列表中的實體類別(例如，DIS)。舉例來說，待標記文章為「Brazil study:

People who have had dengue in the past are twice as likely to develop symptoms of COVID-19 if infected.」，則這待標記文章的分詞結果為

「 Brazil,study,:,People,who,have,had,dengue,in,the,past,are,twice,as,likely,to,develop,symptoms,of,COVID,-,19,if,infected,.」。並且，這分詞結果的第一識別結果為「B-LOC,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O」，而

第二識別結果為「B-LOC,O,O,O,O,O,O,B-DIS,O,O,O,O,O,O,O,O,O,O,O,O,B-DIS,I-DIS,I-DIS,O,O,O」。如此一來，本發明的產生裝置 1 標記的文章，可以確保對應的每一字組都有被標記，而不會發生產生修正了疾病名 A 的命名實體類別卻漏掉修正疾病名 B 的命名實體類別的情況，進而提高文章標記的準確度。換句話說，透過本發明於步驟 S241 中，處理器 110 紀錄文章中與擴充列表的字組相對應的文字，以及紀錄這些文字於分詞結果中的位置(index)。再者，處理器 110 根據這些位置將第一識別結果中的實體類別轉換為擴充列表中對應的實體類別，進而提高轉換實體類別時的正確率。

【0027】 在本發明中，透過將複數個的字組中的第一個字組的實體類別標記為 B-DIS 或 B-BOOK，且將其餘的字組的實體類別標記為 I-DIS 或 I-BOOK，以增加實體類別之間的明確度，進而提升後續使用(例如，將識別結果作為訓練資料及驗證資料用於訓練模型)的便利性。值得說明的是，擴充列表的預設實體類別可以是「車子品牌」、「疾病名稱」等實體類別，及包括對應的複數字組。

【0028】 在一實施例中，於步驟 S210 之中，處理器 110 將文章將執行分詞處理以及形式轉換以產生分詞結果。形式轉換為將文章中的每一個為大寫形式的文字轉換為對應的小寫形式的文字。易於理解的，處理器也可透過形式轉換將文章中的每一個文字轉換為大寫形式的文字，本案不應以此為限。處理器 110 透過將文章中的文字形式轉換為同一形式(統一為大寫或小寫)以提高處理器

110 在識別文字時的準確度與正確率。具體來說，在本實施例中，處理器 110 將文章執行分詞處理以產生未經形式轉換的分詞結果，且處理器 110 將文章執行分詞處理以及形式轉換以產生分詞結果。接著，於步驟 S220、步驟 S230 與步驟 S240 中，處理器 110 所使用的分詞結果皆是經形式轉換與分詞處理的。值得說明的是，於步驟 S250 與步驟 S260 中，作為標記資料的分詞結果為未經形式轉換的。也就是說，作為標記資料的分詞結果中的文字形式與未經處理的文章中的文字形式(例如，大寫形式、小寫形式)一致，進而提高本發明輸出資料(即，標記資料)的相容性。

【0029】 圖 4 根據本發明的另一實施例繪示一種文章標記資料的產生方法的流程圖。在一實施例中，處理器 110 取得多篇文章(步驟 S410)。在本發明中，處理器 110 可透過爬蟲模組 131 以利用爬蟲技術取得多篇文章（article）。舉例來說，爬蟲模組 131 可通過收發器 120 存取新聞網站，並且利用爬蟲技術從新聞網站中取得多篇新聞文章。在一實施例中，爬蟲模組 121 可根據預設週期以及設定值重複地執行步驟 S410。

【0030】 在一實施例中，在取得多篇文章後，處理器 110 從多篇文章中每次提取一篇文章(步驟 S420)。另一方面，在步驟 S260 以及步驟 S250 之後，處理器 110 判斷這文章是否為多篇文章中的最後一篇文章(步驟 S430)。若文章為最後一篇文章則結束流程，若文章不是最後一篇文章則回到步驟 S420。

【0031】 綜上所述，本發明可突破現有的命名實體識別模型的限

制而自動地擴充與產生訓練資料，且訓練資料可用於訓練命名實體模型。如此一來，本發明的產生裝置所產生的文章標記資料可以用於擴充命名實體識別模型的識別範圍。其中，在轉換實體類別的過程中，透過記錄對應的字組於分詞結果的位置，用以逐一轉換對應字組的實體類別，進而提高轉換實體類別的正確率。另一方面，透過將待標記文章及擴充列表中的文字轉換為同一形式(統一為大寫或小寫)以提高識別文字時的準確度以及降低錯誤發生率。

【符號說明】

【0032】

1:文章標記裝置

110:處理器

120:收發器

130:儲存媒體

131:爬蟲模組

132:分詞處理模型

133:命名實體識別模型

134:訓練模組

S210、S220、S230、S240、S241、S242、S250、S260、S410、S420、S430:步驟

【發明申請專利範圍】

【請求項1】 一種文章標記資料的產生裝置，包括：

一收發器；以及

一處理器，耦接該收發器，用以：

將一文章執行分詞處理以產生一分詞結果；

依據命名實體識別模型對該分詞結果進行命名實體識別處理以產生一第一識別結果；

依據一擴充列表判斷該分詞結果是否具有該擴充列表中的多個字組的任一個字組；

當該分詞結果包括有該擴充列表中的任一個該些字組，依據該擴充列表以及該分詞結果對該第一識別結果進行擴充實體類別轉換以產生一第二識別結果；以及

將該第二識別結果以及該分詞結果作為一標記資料，且輸出該標記資料；

其中將該文章執行分詞處理以產生該分詞結果的步驟中，該處理器更用以：

將該文章執行分詞處理以及形式轉換以產生該分詞結果，其中形式轉換為將該文章中的每一個為大寫形式的文字轉換為對應的小寫形式的文字；

其中，作為該標記資料的該分詞結果為未經形式轉換。

【請求項2】 如請求項1所述的文章標記資料的產生裝置，其中當該分詞結果不包括該擴充列表中的任一個該些字組，該處理器以該分詞結果以及該第一識別結果作為該標記資料，且該處理器輸出該標記資料。

【請求項3】 如請求項1所述的文章標記資料的產生裝置，其中該分詞處理是透過一分詞處理模型所執行，且該分詞處理模型與該命名實體識別模型分別是基於一深度學習所訓練的，且該深度學習包括基於Transformer架構的自然語言處理演算法。

【請求項4】 如請求項1所述的文章標記資料的產生裝置，其中該依據該擴充列表對該分詞結果判斷是否包括有該擴充列表中的任一該些字組的步驟之中，該處理器更用以：

依據該擴充列表中的每一該些字組以及對應該些字組的多個窗口長度，對該分詞結果進行搜尋處理以判斷該分詞結果是否包括任一該些字組。

【請求項5】 如請求項1所述的文章標記資料的產生裝置，其中對該第一識別結果進行擴充實體類別轉換以產生該第二識別結果的步驟之中，該處理器更用以：

依據該擴充列表與該第一識別結果將該分詞結果與該擴充列表進行比對以獲得該些字組於該分詞結果中對應的多個位置；

依據該些位置以及該擴充列表將該第一識別結果中的對應於該些位置的實體類別轉換為該擴充列表的擴充實體類別，以產生該第二識別結果。

【請求項6】 如請求項1所述的文章標記資料的產生裝置，其中該命名實體模型是基於一深度學習，該深度學習包括基於Transformer架構的自然語言處理演算法以及一預設字組列表所訓練。

【請求項7】 如請求項6所述的文章標記資料的產生裝置，其中該處理器更用以：

依據該標記資料訓練該命名實體模型，以產生一擴充命名實體模型。

【請求項8】 如請求項1所述的文章標記資料的產生裝置，其中該擴充列表中的該些字組為經分詞處理且/或經格式統一處理的字組，其中該格式統一處理為將每一字組中的文字統一轉換為大寫形式的文字或小寫形式的文字。

【請求項9】 一種文章標記資料的產生方法，包括：

一處理器將一文章執行分詞處理以產生一分詞結果；

該處理器依據命名實體識別模型對該分詞結果進行命名實體識別處理以產生一第一識別結果；

該處理器依據一擴充列表判斷該分詞結果是否包括有該擴充列表中的多個字組的任一個字組；

當該分詞結果包括有該擴充列表中的任一個該些字組，該處理器依據該擴充列表以及該分詞結果對該第一識別結果進行擴充實體類別轉換以產生一第二識別結果；以及

該處理器將該第二識別結果以及該分詞結果作為一標記資料，

且輸出該標記資料；

其中將該文章執行分詞處理以產生該分詞結果的步驟中，更包括：

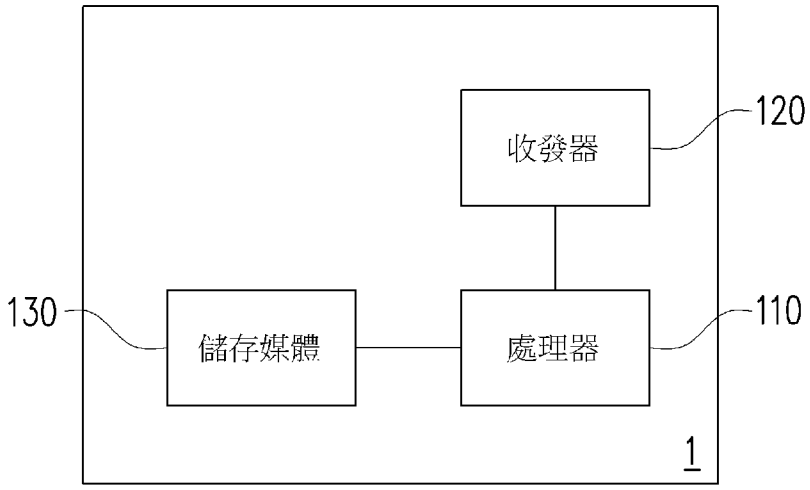
該處理器將該文章執行分詞處理以及形式轉換以產生該分詞結果，其中形式轉換為將該文章中的每一個為大寫形式的文字轉換為對應的小寫形式的文字，其中作為該標記資料的該分詞結果為未經形式轉換。

【請求項10】　如請求項9所述的文章標記資料的產生方法，包括：

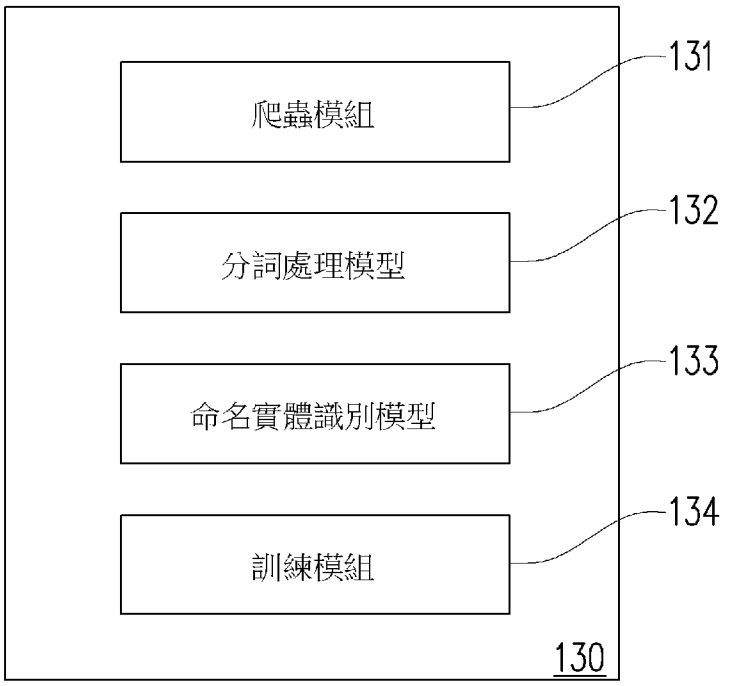
該處理器依據該擴充列表將該分詞結果與該擴充列表進行比對以獲得該些字組於該分詞結果中對應的多個位置；

該處理器依據該些位置以及該擴充列表將該第一識別結果中的對應於該些位置的實體類別轉換為該擴充列表對應的擴充實體類別，以產生該第二識別結果。

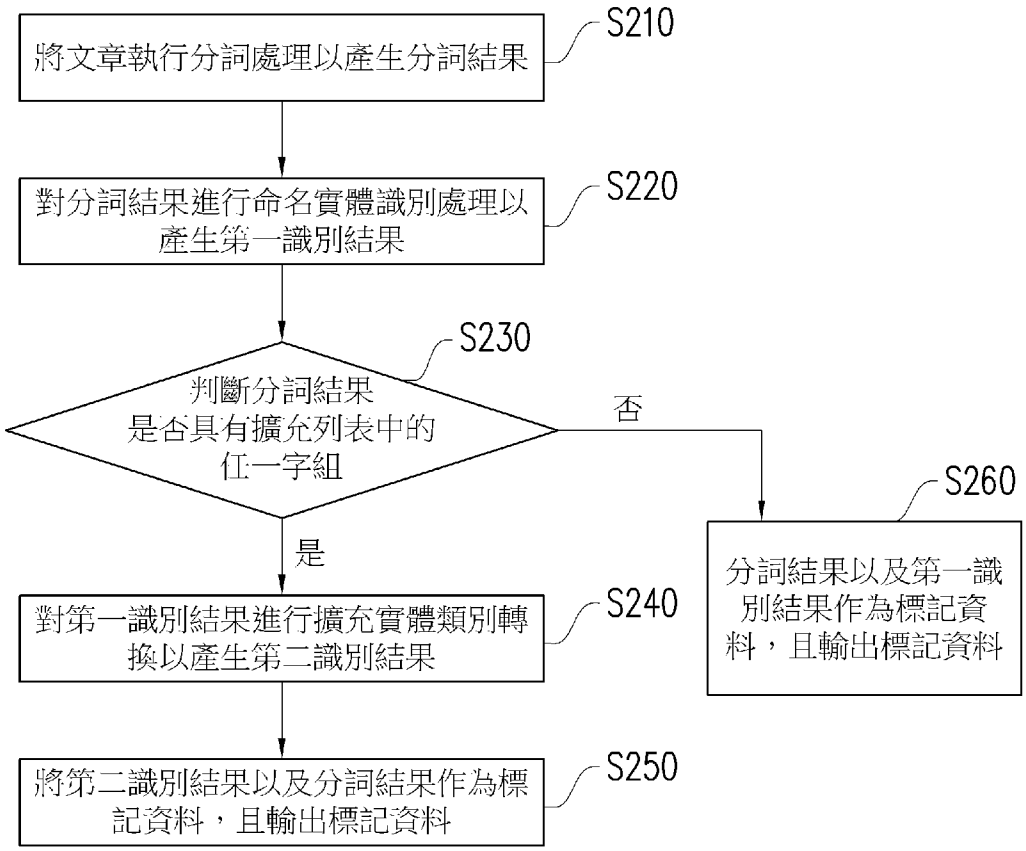
【發明圖式】


收發器
120
130
儲存媒體
處理器
110
1

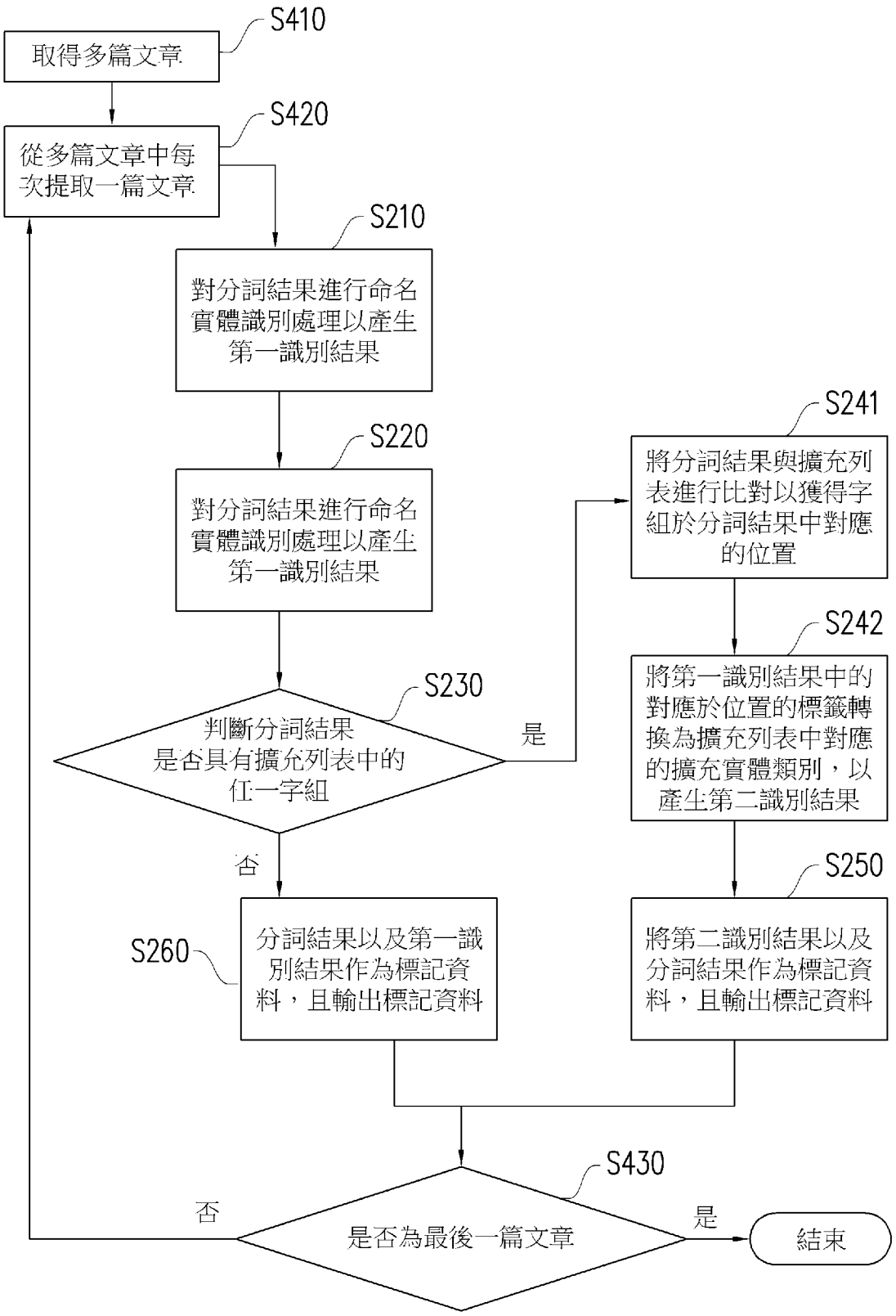
【圖1】


131
爬蟲模組
132
分詞處理模型
133
命名實體識別模型
134
訓練模組
130

【圖2】


將文章執行分詞處理以產生分詞結果
S210
對分詞結果進行命名實體識別處理以產生第一識別結果
S220
判斷分詞結果是否具有擴充列表中的任一字組
S230
否
是
分詞結果以及第一識別結果作為標記資料，且輸出標記資料
S260
對第一識別結果進行擴充實體類別轉換以產生第二識別結果
S240
將第二識別結果以及分詞結果作為標記資料，且輸出標記資料
S250

【圖3】


S410
取得多篇文章
S420
從多篇文章中每次提取一篇文章
S210
對分詞結果進行命名實體識別處理以產生第一識別結果
S220
對分詞結果進行命名實體識別處理以產生第一識別結果
S230
判斷分詞結果是否具有擴充列表中的任一字組
是
否
S241
將分詞結果與擴充列表進行比對以獲得字組於分詞結果中對應的位置
S242
將第一識別結果中的對應於位置的標籤轉換為擴充列表中對應的擴充實體類別，以產生第二識別結果
S250
將第二識別結果以及分詞結果作為標記資料，且輸出標記資料
S260
分詞結果以及第一識別結果作為標記資料，且輸出標記資料
S430
是否為最後一篇文章
否
是
結束

【圖4】