Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 1-16

蘊涵句型分析於改進中文文字蘊涵識別系統

Entailment Analysis for Improving Chinese Recognizing Textual Entailment System

楊善順*、吳世弘*、陳良圃+、邱宏昇+、楊仁達+

Shan-Shun Yang, Shih-Hung Wu, Liang-Pu Chen,

Hung-Sheng Chiu, and Ren-Dar Yang

摘要

文字蘊涵是自然語言處理最近興起的研究課題。文字蘊涵識別(Recognizing Textual Entailment, RTE)可以應用到其他許多自然語言處理的研究中。在本文中將介紹我們在觀察 NTCIR-10-RITE-2 資料集後發現過去系統的缺陷，進而提出如何改進中文文字蘊涵系統的方法。過去的系統處理文字蘊涵多使用機器學習分類文題的方法，所有輸入句子都用同樣的分類器處理，對於某些特別的問題往往會產生誤判。我們認為應該針對於特定類型的問題做處理，增加系統可以處理的問題類型。實驗結果顯示配合之前提出的機器學習方法，增加四種特殊類型分類對特殊類型句子進行個別處理，可以有效改進系統，實驗結果系統在識別簡體中文蘊涵兩類的正確率從原本 67.86%提昇到 72.92%。

關鍵詞：中文文字蘊涵識別、蘊涵分析

*朝陽科技大學資訊工程系 Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan (R.O.C)
E-mail: { s10027619; shwu}@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.
+財團法人資訊工業策進會 Institute for Information Industry, Taipei, Taiwan (R.O.C)
E-mail: {eit; bbchiu; rdyang}@iii.org.tw

Abstract

Recognizing Textual Entailment (RTE) is a new research issue in natural language processing (NLP) research area. RTE can be a useful component in many NLP applications. In this paper, we introduce our finding on the entailment analysis of the NTCIR-10 RITE-2 dataset, and use the observation to improve our system. In the previous works, all the input pairs are treated equally in a standard classification architecture. We find that is not suitable for some special cases. We believe that by isolating the special cases and building separated classifiers, a RTE system can perform better. After implementing modules for four special cases into our system, the result is significantly improved from 67.86% to 72.92% on the binary class classification task.

Keywords: Chinese Recognizing Textual Entailment, Entailment Analysis

1. 緒論

文字蘊涵(Textual Entailment, TE)(Dagan *et al*., 2006)是自然語言處理(Natural Language Processing, NLP)最近興起研究議題，文字蘊涵識別目標為給定一個句子對(T1,T2)系統能夠準確的推斷這兩句子之間的蘊涵關係。因此文字蘊涵可以應用在自然語言處理其他領域研究中，例如問答系統、資訊抽取、資訊檢索、機器翻譯(Dagan & Glickman, 2004 ; Ou & Yao, 2010)等等。文字蘊涵最基本的方法就藉由句子字面上的資訊例如語意、句法(Hua & Dinga, 2011)等等字面上的相似性進而推斷句子是否有著蘊涵關係。因此利用這個特性，文字蘊涵有助於問答系統找到資料庫中與輸入問句最相近的問句進而回應最適當回答。以資訊檢索來說檢索詞的好壞對資訊檢索有著很大影響，藉由文字蘊涵找到與檢索詞相關的字詞(例如同義詞)加入檢索條件這樣可以讓使用者更容易找到使用者所需要的資訊。

目前文字蘊涵的研究分成兩種層面，首先兩類(Binary Class, BC)的任務的目標是單純判別 T1 與 T2 之間是否有蘊涵關係，但句子之間蘊涵關係並不能單純以有或沒有這麼簡單就區分開，因此為了表示不同情況下的句子之間蘊涵關係，NTCIR RITE 另外定義多類(Multi Class, MC)這項任務將句子之間的蘊涵作更為明確的分類。假設這個句子對具有蘊涵關係，我們可以很合理認為這兩個句子所表達是相同的意思，但有可能兩個句子如表 1 中正向蘊涵的例句一樣兩個句子所包涵的資訊數量不同，造成我們可以從 T1 句子可以推論出 T2 句子的完整的意思，但是不能從 T2 推論出 T1 句子完整的意思，這樣情況我們就稱正向蘊涵。反之如表 1 中雙向蘊涵的例句一樣 T1 句子可以推論出 T2 句子的含意，T2 也可以推論出 T1 句子完整的意思，兩個句子之間可以相互推論這樣的情況我們就稱為雙向蘊涵關係。假設句子對之間沒有蘊涵關係，我們可以很合理認為兩個句子所表達的意思不相同，但這並不完全正確的想法。如同表 1 矛盾例句一樣可能兩個句子所包涵的資訊大致相同只是少部份資訊例如:是與不是或是時間點不同造成句子的意

思互相衝突，這樣的情況我們就稱之為矛盾蘊涵。或是兩個句子本身包涵的資訊毫無關係這樣的情況我們就稱之為獨立蘊涵，藉由上述的四種蘊涵關係將句子之間的蘊涵關係細分，使得文字蘊涵系識別的研究更有其意義。

本篇重點在處理簡體中文與繁體中文方面的文字蘊涵，使用 NTCIR-10 RITE-2 所提供的訓練資料基於機器學習的方法 SVM 建立一個中文文字蘊涵系統。系統的一開始先將輸入的文句對資料進行預處理，由於處理是中文資料必須先進行斷詞以便接下來的工作，使用問題類型分類將可以個別處理的類型抽出特別處理，接著句子對經由特徵擷取的子系統取得各項特徵值，最後將取的特徵值使用 SVM 分類處理。

本篇接下來章節如下，在第二段介紹過去研究方法，第三段將介紹系統架構與預處理部分以及系統使用到的特徵值跟我們所觀察到特殊類型問題分析，第四段是實驗結果與討論最後是結論與未來工作。

表 1. 各種蘊涵關係例句

類型	例句
正向蘊涵 (F)	t1：異位性皮膚炎好發於具過敏體質的嬰幼兒、兒童及青少年，常見症狀是臉、頸、手肘窩、膝窩、或四肢背側等部位出現搔癢紅疹、皮膚變厚、變粗糙。
	t2：異位性皮膚炎產生的發紅皮疹好發於臉頰頸側、手肘窩或膝蓋等彎曲部位。
雙向蘊涵 (B)	t1：洋基球團保護選手的立意甚篤，要求「只投一場且不超過一百球」。
	t2：洋基球隊開出「只投一場且不超過一百球」的保護條件。
矛盾蘊涵 (C)	t1：印尼蘇門答臘西岸外海發生芮氏規模九的強震，主震與餘震所引發的海嘯席捲南亞諸國的海岸。
	t2：印尼蘇門答臘北部外海廿八日深夜發生芮氏規模八點七的強震。
獨立蘊涵 (I)	t1：中研院基因體中心正和何大一合作，研發新一代的禽流感基因疫苗。
	t2：中研院院士何大一最近正在研發禽流感基因疫苗。

2. 過去研究

之前的研究文獻中有許多不同的方法應用在英文文字蘊涵識別，例如定理證明或使用 WordNet 等等不同的詞意語料資源。在中文文字蘊涵方面(Wu *et al.*, 2011)等人參考其他語言的方法提出一個基礎機器學習利用機器翻譯效能評估的 BLEU 分數及句子長度做為特徵以及句子長度做為特徵來訓練分類器，建立基礎中文文字蘊涵識別系統，(Zhang *et al.*, 2011)等人提出加入語意相關資訊作為特徵處理，藉由上下位詞、同義詞與反義詞等資訊來進行的推論以及使用多個機器學習的方法，最後使用投票機制選出最合適蘊涵關係提高系統準確率。

隨後(Yang *et al.*, 2012)等人提出使用單語言機器翻譯的方法，藉由 GIZA++中字詞對齊的匹配值進行計算出句子之間相似度作為新特徵使用以有效提升正確率，但這個方法在處理兩類中效果較好再處理多類蘊涵關係時效果不如兩類，之後(Wang *et al.*, 2013)等人提出反向的矛盾字詞對齊，有效改善之前正向字詞對齊容易產生的誤判問題。

在文字蘊涵識別的推論時，各種語義資訊與上下文資訊是必要的處理。雖然 NTCIR-10 RITE-2 (Watanabe *et al.*, 2013) 任務目的在於評鑑各種語義/上下文處理系統，但這有一個問題注重具體語言現象的研究是不容易的達成。在 RITE-2 日文子任務的資料集中含包涵了識別 T1 和 T2 之前蘊涵關係必需的語言學現象，從兩類(BC)子任務資料集中擷取句子對作為樣本，建立具有語言學現象的句子，將這樣的句子加入資料集中作為單元測試使用。單元測試數據相當於多類(BC)任務資料集的一個子集。雖然這個資料集並不多，但您可以將它用於各種研究，包括分析 RITE 資料集中出現的語言學問題，評測每一個語言學現象識別正確度和為各種語言學現象的分類器訓練與測試資料。

3. 系統架構

我們的系統的系統流程圖如圖 1 所示，基本組成部分”預處理”、”斷詞”、”中文簡繁轉換”、”特殊類型分類”、個別” SVM 分類器”和最後”結果整合”。

3.1 預處理

3.1.1 表示格式正規化

在預處理的部份第一步就是句子的字詞格式正規化，我們系統預處理正規化模塊將括號中字詞視為括號前字詞的同義詞，例如”葉望輝(Stephen J. Yates)”，括號中的字詞”Stephen J. Yates”是另一個字詞”葉望輝”是相同意思。另外我們觀察訓練資料發現有部分句子之間對於時間表示格式不同，因此時間表示方法必須統一以便接下來的處理工作，如表 2 中所示的時間表示格式常見的例子，測試資料中數字的格式也相當不一致造成系統判讀上的困難，這個部份也是我們需要統一的部份。將句子正規化後的資料有著更高的相似度也更容易被識別，在機器翻譯方面也句子中的字詞更容易文字對齊進而提高兩個句子可翻譯機率。

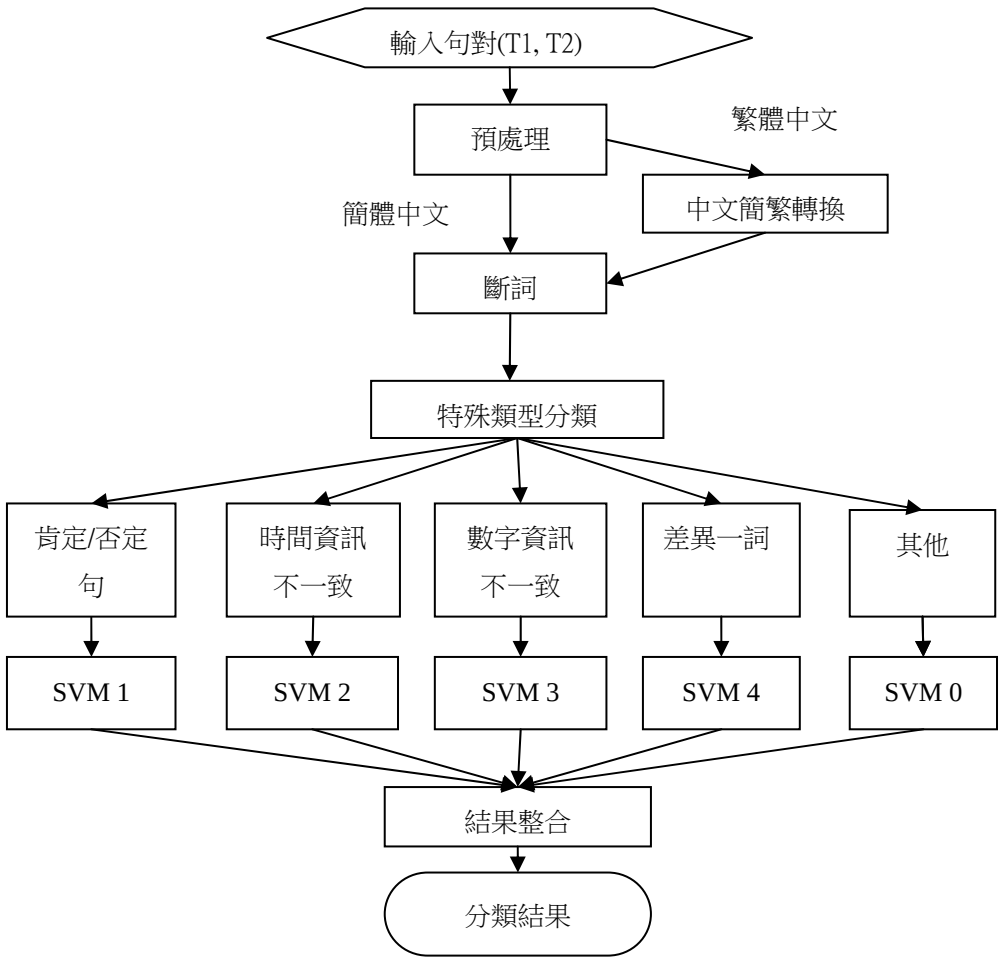

圖 1. 系統流程圖

表 2. 各種時間表示格式之例句

時間型態	時間表示方式
中文	一九九七年二月廿三日
數字全形	１９９７年２月２３日
數字半形	1997年2月23日
數字以「-」隔開	1999-05-07
數字以「/」隔開	1985/12/20
範圍	1999年延長至2001年

3.1.2 背景知識的替換

預處理的部份第二步是代替它們的同義詞，同義詞的資訊可以從”維基百科”、”HowNet (Liu & Li, 2002)中取得，其中從維基百科中可以取得句子必要的背景知識，包括明確的時間和地點的資訊作為同義詞替換補充所需的資訊。另外有關時間表示格式還有另一個難題就是歷史上不同朝代如何轉換成同一個時間表示，這是需要相當的背景知識才能正確的轉換。例如”乾隆”是西元 1735 年和”昭和”是西元 1925 年或是例如”北京奧運”擁有背景知識的人可以很輕易就知道”北京奧運”發生於西元 2008 年。所以時間表示式問題需要相當背景知識才能進行正規化，因此我們從維基百科取得各朝代資訊建立一個朝代資料庫，作預處理時偵測句子中是否含有朝代字詞，若有朝代字詞將其替換成對應年份後加上朝代字詞後數字減一，如”乾隆 3 年”將替換成”西元 1737 年”。另一個類似的問題是地名的表示問題有時候可能因作者的不同對於相同的地方有不同的表示方式，所以遇到這樣的情況我們必須進行地名正規化的步驟，例如縮寫”台、澎、金、馬”是指”台灣、澎湖、金門、馬祖”或是代稱，觀察過去的語料資料建立地名資料庫，偵測句子中是否含有地名字詞，若有地名字詞將其替換成統一字詞。

3.1.3 斷詞與中文簡繁轉換

由於我們使用的斯坦福剖析器只能處理簡體中文以及英文，因此我們的系統中使用的斷詞工具是 ICTCLAS 斷詞系統，這是由中國科學院計算技術研究所提供。該工具功能包括斷詞、詞性標註、NE 識別、新字詞識別，以及自訂字典。為了配合斷詞系統我們處理繁體中文時必須將繁體中文文句對轉換成簡體中文文句，一開始我們使用 Google 翻譯之後改使用我們自行開發的機器翻譯系統(Li *et al*., 2010)。

3.2 特徵提取

在我們的系統中使用到的特徵在表 3 列出以前的文字蘊涵識別工作(Yang *et al*., 2012)大多數可用的特徵。在前三個特徵測量 T1 和 T2 中根據一般中文類似的字元。Unigram recall、unigram precision、unigram F-measure 可以視為在 T1，T2 的字元比例和幾何平均，我們的系統使用 BLEU (Papineni *et al*., 2002)三個特點。Bleu (Zhou *et al*., 2006)當初是被設計來測量機器翻譯(machine translation)的品質。一個良好的機器翻譯需要包含適當、準確以及流暢的翻譯，我們的系統會將其翻譯為原來的文字 T1 和 T2 得到 log Bleu recall、log Bleu precision 和 log Bleu F measure values。

第七到第十這四個特徵是 T1 和 T2 的句子長度。我們的系統根據字元和字詞計算 T1 和 T2 的句子中長度的差異，並使用了這兩個特徵的絕對值在我們的系統中。

最後特徵是由 GIZA++(Och & Ney, 2003)字詞對齊分數，這是 T1 句子以單一語言機器翻譯到 T2 句子的機率。機器翻譯可該功能有助於 RTE (Quang *et al*., 2012)，我們的系統採用的單一語言機器翻譯作為一個特徵。在我們的系統中，我們使用 GIZA ++做為單一語言機器翻譯工具，GIZA++根據 IBM 模型製作而成，我們經由 GIZA++ 計算出兩個

輸入句子之間的匹配得分，在使用下列公式計算最後應用在 SVM 上的特徵值：

$$p_n = \frac{\log\left\{\prod_{i,j=0}^{i,j=\max}\left[p(t1_i|t2_j)\right]\right\}}{n} \quad (1)$$

公式中 $t1_i$ 與 $t2_j$ 分別代表 T1 與 T2 句子中各字詞， $p(t1_i \mid t2_j)$ 代表 $t1_i$ 與 $t2_j$ 字詞對齊的機率，使用 GIZA++計算句子中字詞對齊機率後連乘取 log 在除以連乘的次數 n 後就是使用在 SVM 的特徵值 p_n 。

表 3. 分類器使用的特徵

編號	特徵
1	Unigram recall
2	Unigram precision
3	Unigram F-measure
4	Log bleu recall
5	Log bleu precision
6	Log bleu F-measure
7	difference in sentence length (character)
8	absolute difference in sentence length (character)
9	difference in sentence length (term)
10	absolute difference in sentence length (term)
11	GIZA++

3.3 特殊類型處理

我們觀察 NTCIR10 RITE-2 所提供的訓練資料，發現有許多的句子是過去的系統所無法正確處理容易產生誤判。如表 4 所示我們檢視 NTCIR10 RITE-2 我的們系統正式評測結果整理過去系統容易誤判問題類型。

在表 4 中 Caes1 是一個獨立關係被系統誤判成雙向蘊涵的例子，我們發現兩個句子字面上非常相似，只是因為粗體部份不同使兩個句子變成獨立關係，假如對句子中單詞進行比對即可解決這個誤判可能。Caes2 是一個矛盾關係誤判成雙向蘊涵的例子，從字面上可以發現兩個句子非常相似，只是因為句子中數字資訊不相同產生互相矛盾關係的情況，在此可以知道我們系統對於數字資訊比對不足這是未來改進的方向，在觀察系統誤判題目時發現有許多如 Caes3 這種句子中包含字詞完全相同就誤判成正向的情形的存在，Caes4 的例子由於我們的系統不具有相關背景知識可以推論句子中的”倫敦”就指是”英國”的意思，因此將正向誤判成獨立關係，除了上面提到的同義詞問題，Caes5 中可以發現反義詞也是未來需要解決的問題之一。

表 4. 容易誤判問題類型例子

編號	例句
Case1	T1: **流感**病毒可在人体外存活**三到六**小时
	T2: **冠状**病毒通常可在人体外存活**二到三**小时
Case2	T1: 大陆已有**四百万**人感染爱滋病
	T2: 大陆有**八十五万**人感染艾滋病
Case3	T1: 美国奉行一中政策和遵守三公报的立场并未改变；切尼则进一步表示不支持台湾独立
	T2: 美国不支持台湾走向独立
Case4	T1: 申奥成功的伦敦当局，在爆炸案后立即宣布取消庆祝活动。
	T2: 英国已停止所有庆祝申奥成功的活动。
Case5	T1: 我国生物技术可以与美国等先进国家相提并论，**解决**“异种核转殖”的问题并不难。
	T2: 我国生物技术可能**造成**异种核转殖等问题。

4. 蘊涵句型分析

為了解決這些容易誤判的問題，我們認為應將句子依照問題類型進行分類再各自使用最適合的方法進行處理。在系統處理完預處理後，將可以特殊類型的句子挑選出來，使用我們開發的子系統做處理，處理後的結果在與過去使用的機器學習方法作整合，得到最後的結果，以下是我們以實做出來的特殊類型，表 5 是對應的例句。

4.1 肯定/否定句

我們觀察訓練集資料發現有些句子對，句子包涵的資訊幾乎完全相同，只因為否定詞就使得兩個句子具有是否關係形成矛盾或獨立蘊涵關係，針對這種類型的句子對我們可以使用自行開發的系統進行句子對是否具有是否關係字詞，假設句子對只有一句具有是否關係字詞，這樣可以認為句子對沒有蘊涵關係。

4.2 時間資訊不一致

訓練資料之中發現有些句子中含有時間的資訊，然而這些時間的資訊對句子的含意有很重要的意義，所以我們認為應該將這些句子挑選出來針對句子中的時間部份比較，針對這種類型的句子在之前系統前處理部分會對句子中的時間格式做正規化，再對句子的時間進行比對可以處理特定時間點的句子。

4.3 數字資訊不一致

訓練資料中有些句子中含有數字的資訊,然而這些數字資訊對句子的含意有很大的影響,因此我們認為應該將這些句子挑選出來針對句子中的數字部份比較,對這種類型的句子在之前系統前處理部分會對句子中的數字格式做正規化,針對數字部分進行比對就可以處理這類大部分的句子。

4.4 差異一詞

我們觀察訓練資料時發現有些句子字面上相當相似,例如主詞或受詞或是部份字詞被替換就會導致句子的意思都改變了,針對這類型的句子可以對句子字詞進行比對處理。

當然特殊類型的問題不止上述的幾種,我們也歸納出更多特殊類型有待未來完成。

4.5 同義詞

我們觀察訓練資料時發現有些句子對部份的字詞有著同義詞的關係,對於這類的句子對應該對句子先進行處理,使用 E-hownet 等語料庫將同義詞都替換成相同字詞就可以使兩個句子變得更相似,這樣判斷蘊涵關係可以更容易。

4.6 背景知識

我們觀察訓練資料時發現有些句子提供的資訊即使是人類,在沒有一定的背景知識情況下也無法正確的判斷是否有蘊涵關係,例如新德里是印度的首都沒有這樣的背景知識很容易將新德里與印度認為成不同的地方,對於這類型的句子可以使用外部資源例如維基百科等獲得對應的背景知識作處理。

4.7 句法調換

我們觀察訓練資料時發現有些句子,只是將句子的順序進行調換就導致句子的含意有所改變,當句子中的主詞與受詞有所改變會導致整個句子意思整個改變,針對這類型的句子可以使用如史丹佛剖析器將句子進行剖析得到語意角色關係進而偵測句子主詞與受詞關係以及連接詞也進行偵測是否為前後調換語意不變的詞如與、一起等等。

4.8 一詞多義與命名實體

我們觀察訓練資料時發現有些句子雖然句子的字詞很相似,但其實含意大大不相同例如”沒完沒了”本身是一個成語但是這個詞在其他地方有著其他意思如歌名、電影名,正因如此這樣字詞造成句子整個含意大大不同其蘊涵關係也變成獨立關係了。

4.9 句子縮減

我們觀察訓練資料時發現有些句子將部份資訊去除掉就形成另一個句子,造成某個句子包涵著另一個句子所有的資訊,這樣兩個句子就有著方向性的蘊涵關係,對句子進行剖

析取得句子的剖析樹應用編輯距離的方法來取得句子重要資訊是否有所改變或相同來認定句子是否為縮減關係。

4.10 邏輯推理

我們觀察訓練資料時發現有些句子可以從句子的資訊可以合理推理出與另一個句子的含意，這樣的句子的我們初步將它分類在句子推理這一類。

表 5. 特殊類型例子

類型	例句
肯定/否定句	T1 319 槍擊案中，陳總統受槍擊時沒忘了穿防彈衣。
	T2 319 槍擊案中，陳總統受槍擊時沒穿防彈衣。
	兩句內容大都相同但因為 T1 句子含有反義詞造成文句對有著矛盾關係。
時間資訊不一致	T1 茱莉安德魯絲 1930 年出生
	T2 1935 年出生於英國的茱莉・安德魯絲。
	兩句由於時間有所不同造成句子表達的意思互相矛盾。
數字資訊不一致	T1 娜拉提諾娃一共獲得 18 座大滿貫金杯。
	T2 娜拉提洛娃一共取得了 58 個大滿貫的金杯。
	兩句由於數字有所不同造成句子訊息互相矛盾。
差異一詞	T1 一九七一年，印度協助東巴基斯坦脫離巴基斯坦成為孟加拉，結果印巴戰事再起。
	T2 一九七一年，印度協助西巴基斯坦脫離巴基斯坦成為孟加拉，結果印巴戰事再起。
	兩個句子可能只差異一個字詞就導致兩個句子的蘊涵關係改變，如例子兩個句子的主/受詞不同使得兩個句子的蘊涵變成互相獨立。
同義詞	T1 金泳三在一九九二年當選韓國第十四屆大統領。
	T2 金泳三在一九九二年當選韓國第十四屆總統。
	兩句中字詞雖然字面上有所不同但在語意上是相同的，因此這是雙向蘊涵關係。
背景知識	T1 2005 年全球恐怖攻擊活動不斷是第三世界向歐美霸權宣戰。
	T2 2005 年全球恐怖攻擊活動不斷是回教世界向歐美霸權宣戰。
	兩句中字詞雖然字面上有所不同但在有相當背景知識的人看來兩句表達的意思是相同，因此這是雙向蘊涵關係。
句法調換	T1 松花江汙染事件導致俄羅斯對大陸民眾反感日增
	T2 松花江汙染事件導致大陸對民眾俄羅斯反感日增

	兩句由於句子調換造成主詞與受詞關係改變表達的意思也有所改變。
一詞多義與命名實體	T1 馮小剛與和陳凱歌，吵的沒完了沒了
	T2《沒完沒了》是一部由馮小剛導演
	T1 與 T2 中都出現”沒完沒了”這個詞，然而所具有的涵義大大不同。
句子縮減	T1 在 1964 年 10 月中共成功試爆第一顆原子彈後，加快了中國本身的核子工業發展
	T2 1964 年中共第一次試爆原子彈
	T1 的句子部份資訊被刪掉但是不影響 T1 主要想表達的含意依然與 T2 含意相同，因此這是正向蘊涵關係。
邏輯推理	T1 張學良 1900 年 6 月 3 日出生，1920 年官拜少將
	T2 張學良廿歲官拜少將
	從 T1 句子可以得到關鍵字進而可以推理出 T2 句子的內容，因此這是一個正向蘊涵關係。

這些新的特殊類型目前需要使用人工的方式挑選出來，在未來將開發自動分類系統，然而有部分特殊類型並不像之前已完成的類型可以輕易程式化處理，例如:邏輯推理、背景知識、同義詞等特殊類型必須需要依靠語料資料才能夠正確的區分出來。

如表6所示RITE-2訓練與測試集資料中我們所分類所有的特殊類型句子中的數量，然而統計結果並非完全準確，這是因為一個句子對可能同時擁有多個特殊類型特性。

表 6. 特殊類型在訓練集與測試集涵蓋數量

特殊類型	**訓練集**	**測試集**
肯定/否定句	15(1.82%)	42(5.37%)
時間資訊不一致	43(5.28%)	60(7.68%)
數字資訊不一致	42(5.15%)	83(10.62%)
差異一詞	73(8.9%)	82(10.4%)
同義詞	53(6.5%)	43(5.5%)
背景知識	5(0.6%)	11(1.4%)
句法調換	67(8.2%)	52(6.6%)
一詞多義與命名實體	115(14.1%)	91(11.6%)
句子縮減	290(35.6%)	159(20.3%)
邏輯推理	111(13.6%)	86(11%)
粗略涵蓋	99.75%	90.47%

5. 實驗與討論

在這個章節中，我們將介紹 NTCIR10 RITE-2 測試集進行的幾個實驗設定的實驗結果。

5.1 實驗資料

本篇我們使用到的測試資料取自日本第十屆 NTCIR 研討會中 RITE-2(Recognizing Inference in Text-2)比賽子項目的開發資料(Development Data)以及公開測試集(open text)，在開發資料中 814 個文句對，另一個公開測試資料部份有 781 個文句對。

RITE-2 這個效能評鑑任務(task)，其目的在評鑑系統自動偵測句子之間「推論關係」的效能。句子之間的關係有很多種，比如說蘊涵(entailment)、意譯(paraphrase)以及矛盾(contradiction) 等。本次參與 RITE-2 評鑑的語言包含日文、簡體中文以及繁體中文，每個語言都包含數個「子任務」(subtask)。參與評鑑的系統需辨識給定的兩個文本(text)，輸出二選一(蘊涵與非蘊涵)或四選一(正向、雙向、獨立、矛盾)的蘊涵關係標記(label)。

5.2 實驗結果

我們在 NTCIR10 RITE-2 的正式評測簡體中文(CS)與繁體中文(CT)結果如表 7 表 8 所示，本次 NTCIR10 RITE-2 我們一共嘗試三種不同實驗設定，首先 CYUT-01 我們使用之前系統實驗結果效果最好實驗設定使用了上一個章節提到的前 10 個做測試，實驗設定 CYUT-02 延續 CYUT-01 的實驗設定加入之前提出的單一語言機器翻譯的方法作為的十一個特徵使用，實驗設定 CYUT-03 使用上述提到的 11 個特徵在加入一個是非句規則判斷作為的 12 個特徵使用，我們簡體中文以及繁體中文的實驗都是使用相同的實驗設定，我們在 NTCIR10 RITE-2 簡體中文這個項目取得第三名不錯成績，但在繁體中文這個項目表現比較不亮眼，相較於簡體中文繁體中文只得到第 8 名的成果。

表 7. NTCIR10 RITE-2 簡體中文正式評測結果

	BC (%)	**MC (%)**
CYUT-01	61.17	40.37
CYUT-02	63.11	42.52
CYUT-03	**67.86**	**40.37**

表 8. NTCIR10 RITE-2 繁體中文正式評測結果

	BC (%)	**MC (%)**
CYUT-01	**55.16**	**25.60**
CYUT-02	52.64	26.26
CYUT-03	51.58	23.51

5.2.1 改進實驗一

我們系統簡體中文與繁體中文使用相同的實驗流程但是如表 7 和表 8 所示，實驗結果有著顯著的不同，雖然簡體中文與繁體中文使用的測試資料不同，可能會造成兩個實驗結果有所不同，但我們認為不只這個因素可以造成如此大的差異。比較兩種中文的實驗流程，我們的系統處理繁體中文相對於簡體中文多了將繁體中文翻譯成簡體中文的步驟，深入研究翻譯過後繁體中文測試資料可以發現翻譯效果不佳造成之後抽取特徵時得到錯誤的數值造成之後 SVM 的誤判，因此改使用我們自行開發的機器翻譯系統，解決之前翻譯錯誤產生的空格與術語錯誤的問題提高系統效能，替換後的實驗結果如表 9 所示，在兩類(BC)任務提高 6.02 正確率以及多類(MC)任務則是提高 9.49 正確率。

表 9. 替換簡繁轉換系統實驗結果

簡繁轉換系統	BC (%)	MC (%)
Google 翻譯	53.64	26.26
CYUT 開發系統	**59.54(+6.02)**	**35.75(+9.49)**

5.2.2 改進實驗二

在前面章節提到特殊類型問題是過去系統所無法適當處理的問題，因此針對特殊類型問題進行開發個別專屬系統處理是有所必要，本次將兩類(BC)任務中特殊類型問題句子對抽出來進行個別實驗，實驗結果如表 10 所示，我們可以發現特殊類型使用特別開發的子系統作處理可以提高其正確率。

表 10. 個別特殊類型子系統做兩類(BC)實驗結果

	Case1 肯定/否定句	Case2 時間資訊不一致	Case3 數字資訊不一致	Case4 差異一詞
正式評測	52.38%	58.33%	52.25%	47.59%
個別特殊處理	71.42%	70%	71.25%	60.97%

在表 10 的實驗結果顯示，特殊類型子系統有助於提升兩類(BC)系統效果，因此我們認為處理多類(MC)時也可以提高系統效能，接著我們將特殊類型處理的方法加入系統之中做實驗，我們選擇在正式評測中效果最好的 CYUT-03 做測試，實驗結果如表 11 所示。

表 11. 加入特殊類型方法簡體中文實驗結果

項目	**BC (%)**	**MC (%)**
Cyut	67.86	40.37
Cyut+case1	68.88	42.08
Cyut+case1+case2	69.78	43.45
Cyut+case1+case2+case3	71.63	45.13
Cyut+case1+case2+case3+case4	**72.92**	**45.92**

5.3 實驗討論

在其他語言的文字蘊涵處理研究中也有與我們相似的類型處理例如在日文中(Shibata *et al.*, 2013)等人使用到的同義詞、上下位詞與我們所沒有的反義詞處理，(Makino *et al.*, 2013)等人使用了命名實體名稱作處理，在英文方面有許多文字蘊涵相關的研究也使用到 WordNet 提取同義詞與上下位詞作為擴增資訊，我們參考其他語言使用到的方法來個別開發子系統作處理，在 NTCIR10 RITE-2 日文測試資料中參雜了部份特殊語言現象，如同義詞、倒裝句、代詞等等更貼近現實人類使用的語言，因此特殊類型的處理是未來研究方向。

從實驗結果可以發現我們所提出的特殊類型問題需個別處理的方法有助於提升系統效果，然而我們可以 case4 的差異一詞改進的幅度不如其他子系統，仔細比較前後兩個系統結果後，發現差異一詞子系統雖然可以將原本錯誤的問題處理正確但也會將原本正確問題誤判，這是因為句子本身不只擁有一種特殊類型以及兩個句子之間差異詞對句子含義重要性不一致，子系統過度針對不重要的詞造成誤判。

6. 結論與未來工作

本文報告了我們的系統針對 NTCIR10 RITE-2 正式評測的結果所做的改進，在最後的實驗結果可以看到我們與之前系統改進幅度，改進簡體中文兩類系統，其成果從原本的 67.86%提升到 72.92%。改進簡體中文多類系統，其成果從原本的 40.37%提升到 45.92%。在繁體中文任務的結果發現一個好的簡繁轉換系統有助於我們系統改進，然而簡體中文與繁體中文之間還是有著些微差異有待克服。

在分析過去系統實驗結果後發現許多是以前的方法無法處理的問題，在其他語言的文字蘊涵研究中提出許多方法來增加系統可處理的句子數量，因此我們參考其他語言增加處理的方法提出特殊類型問題分類處理的構想，在個別特殊類型實驗結果證明我們提出的特殊類型問題因個別處理的想法是正確的，藉由個別問題使用個別子系統進行處理以提升系統整體效能，然而我們目前提出的特殊類型分類並非完美，還需要規畫更貼近問題本身的分類類型與建立更多特殊類型子系統來減少因為多重特殊類型造成系統的誤判，針對一些類型問題還需要收集更多語料資料建立資料集。

致謝

研究依經濟部補助財團法人資訊工業策進會「102 年度 數位匯流服務開放平台技術研發計畫 (4/4) 」辦理。本研究感謝國科會贊助部分經費，計畫編號為 NSC 100-2221-E-324-25-MY2，謹此致謝。

參考文獻

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL recognizing textual entailment challenge.

Dagan, I., & Glickman, O. (2004). Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France, 2004.

Hua, D. B., & Dinga, J. (2011). Study on Similar Engineering Decision Problem Identification Based on Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS. *Systems Engineering Procedia*, *1*, 406-413.

Liu, Q., & Li, S. J. (2002). Word Similarity Computing Based on How-net. *International Journal of Computational Linguistics and Chinese Language Processing*, *7*(2), 59-76.

Li, M. H., Wu, S. H., Zeng, Y. C., Yang, P. C., & Ku, T. (2010). Chinese Characters Conversion System based on Lookup Table and Language Model. *International Journal of Computational Linguistics and Chinese Language Processing*, *15*(1), 19-36.

Makino, T., Okajima, S., & Iwakura, T. (2013). FLL: Local Alignments based Approach for NTCIR-10 RITE-2. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Ou, Y. C. (2010). Recognize Textual Entailment by the Lexical and Semantic Matching. *Computer Application and System Modeling, 2010 International Conference* on V2-500 -504.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, 311-318, Philadelphia, PA.

Quang, M., Pham, N., Nguyen, L. M., & Shimazu, A. (2012). Using Machine Translation for Recognizing Textual Entailment in Vietnamese Language. *2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*.

Shibata, T., Kurohashi, S., Kohama, S., & Yamamoto, A. (2013). Predicate-argument Structure based Textual Entailment Recognition System of KYOTO Team for NTCIR-10 RITE-2. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013

Wu, S. H., Huang, W. C., Chen, L. P., & Ku, T. (2011). Binary-class and Multi-class Chinese Textural Entailment System Description in NTCIR-9 RITE. In *Proceedings of the NTCIR-9 workshop*, Tokyo, Japan, 6-10 Dec., 2011

Wang, X. L., Zhao, H., & Lu, B. L. (2013). BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RITE-2 Task. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Watanabe, Y., *et al*. (2013). Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Yang, S. S., Wu, S. H., Chen, L. P., Hsieh, W. T., & Chou, S. T. (2012). Improving Binary-class Chinese Textural Entailment by Monolingual Machine Translation Technology. In *Proceedings of the IEEE IRI 2012*, Las Vegas, USA, 8 Aug, 2012.

Zhang, Y., *et al*. (2011). ICRC_HITSZ at RITE: Leveraging Multiple Classifiers Voting for Textual Entailment Recognition. In *NTCIR-9 RITE, in Proceedings of the NTCIR-9 workshop*, Tokyo, Japan, 6-10 Dec., 2011.

Zhou, L., Lin, C. Y., & Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the Conference on EMNLP*, 77-84, Sydney, Australia, 2006.