基於單語言機器翻譯技術改進中文文字蘊涵

Improving Chinese Textural Entailment by Monolingual Machine Translation Technology

楊善順 Shan-Shun Yang, 吳世弘 Shih-Hung Wu*
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)
{s10027619, shwu}@cyut.edu.tw
*Contact author

陳良圃
財團法人資訊工業策進會
Institute for Information Industry, Taipei, Taiwan (R.O.C)
eit@iii.org.tw
謝文泰
曹承礎
國立台灣大學資訊管理學系
Department of IM, National Taiwan University, Taipei, Taiwan (R.O.C)
wentai@iii.org.tw
chou@im.ntu.edu.tw

摘要

在本文中敘述了我們如何透過單語言機器翻譯提高中文文字蘊涵識別系統效能。在之前我們的做法是基於標準的監督式機器學習分類方法。我們整合單語言機器翻譯系統與其他可用的計算中文的自然語言處理的應用建設為語言資源處理系統。我們觀察訓練語料，並列出了所有可用的特徵。這些特徵包括表面文字，語義和語法的資訊，如：詞性標註、同義詞替換和上下位關係。從訓練語料中被標出特徵被應用於中文文字蘊涵識別的訓練分類模型之上。實驗結果表明單語言機器翻譯技術，可以提高我們的系統效能。

關鍵詞：中文文字蘊涵、 語言特徵、分類

一、緒論

文字蘊涵是一個重要的自然語言處理(NLP)問題，它有著許多方面的應用，例如問答系統、資訊抽取、機器翻譯[1]。截至 2011 年為止在中文領域缺乏認識文字蘊涵(RITE)的相關理論，所以現在很難評估它的效能。在 2011 年由 NTCIR-9RITE-1 提供繁體及簡體中文的共同任務中文文本蘊涵的數據集。該數據集包含一個分兩類(BC)和分多類(MC)的測試集。BC 子任務是假設為一個給定的文本對(T1,T2)，測試 T1 句子是否(推論到)T2

句子，MC 子任務將句子分類成 5 大類的方式來檢測是否有(正向/反向/雙向)蘊涵關係或沒有(矛盾/獨立)蘊涵關係[2]，在表一中舉出是否蘊涵的例子。
假設我們可以從 T1 的資訊得到 T2 相關資訊那麼我們可以認為 T1 與 T2 有蘊涵關係。在數據集中一些蘊涵的例子我們可以視為釋義[3]。也就是說，T1 和 T2 是描述同一件事，並有許多共同的字詞，這是比較容易檢測是否有意譯的關係在較複雜的蘊涵關係之上方法。在本文中，我們的分析著重於文字蘊涵問題的意譯部分。我們測試單語言機器翻譯技術的方式是否可能識別於中文蘊涵的數據集的正向蘊涵的意譯。

表一、蘊涵關係例句

類別	例句
蘊涵	T1：日本時間 2011 年 3 月 11 日，日本宮城縣發生芮氏規模 9.0 強震，造死傷失蹤約 3 萬多人 (Japan time March 11, 2011, Miyagi Prefecture, Japan, a magnitude 9.0 earthquake occurred, causing casualties of about 30,000 people missing or dead.)
	T2：日本時間 2011 年 3 月 11 日，日本宮城縣發生芮氏規模 9.0 強震 (Japan time March 11, 2011, Miyagi Prefecture, Japan, a magnitude 9.0 earthquake occurred)
獨立	T1：黎姿與”残障富豪”馬廷強結婚(Gigi married with the “disability rich” Mating Jiang married)
	T2：馬廷強為香港”東方報業集團”創辦人之一馬惜如之子(Mating Jiang is the son of Ma Xi Ru, one of the founders of Hong Kong, "the Oriental Press Group")

二、研究方法

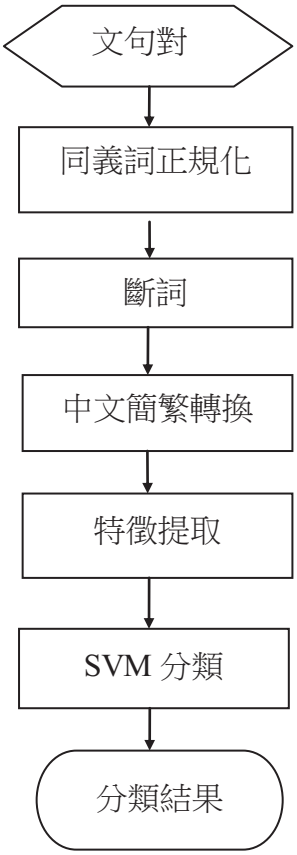
在以前的文獻之中有許多不同的方法被應用在中英文文字蘊涵識別之上，如定理證明或使用 WordNet 等等不同的詞意語料資源[4]。我們的研究方法，嘗試使用單語言機器翻譯作為一個標準的監督學習分類系統[5]的特徵。我們透過觀察訓練語料，使用可用的計算中文的自然語言處理的應用建設成語言資源處理系統。發展過程如下所述，首先我們觀察到的訓練語料，然後列出的各種可用特徵。這些特徵包括表面文字、語義和語法的信息，如：詞性標註、命名實體識別(NER)標註和單語言機器翻譯特徵。然後，從訓練集我們執行了子系統提取到各個特徵。最後我們建立一個分類系統，使用將訓練資料分成 10 等份 9 等分用於訓練 1 等分用於測試，這樣不斷交叉測試稱為”10 倍交叉驗證”方法對訓練數據的特徵測試，並發現哪些特徵是文字蘊涵識別更為有用。

三、系統架構

我們的系統的系統流程圖如圖 1 所示。的基本組成部分”同義詞正規化”、”斷詞” 、”中文簡繁轉換”、”特徵提取”和”SVM 的分類”。

(一)、同義詞正規化

在這裡我們將 T1 和 T2 句子中具有相同的含義的字詞統一取代成相同字詞，因此在後續句子匹配步驟更容易執行。



圖一、系統流程圖

1、 表示格式正規化

預處理的部份的第一個目標是句子中的符號正規化。在我們的系統，預處理模塊正規化中括號中的字可以視為一個在括號前字詞的一個代名詞。由於文件中括號通常表示在前面的括號一詞的音譯或翻譯。例如"車諾比核事故(切爾諾貝利核事故)"，括號中的字詞是翻譯另一個相同字詞"切爾諾貝利核事故"。時間表示方法也將正規化。如表 2 中所示的例子。在語料庫中有許多方式來表示語料庫中的日期或時間，正規化後的資料更容易被識別，在機器翻譯方面也更容易執行文字對齊。

表二、各種時間表示方式之例句

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
範圍	1999年延長至2001年

2、背景知識的替換

預處理的部份第二個目標是代替它們的同義詞。從維基百科收集同義詞的資訊。明確的時間和地點資訊也視為同義詞替換的問題。

另外有關的時間表示的還有另一個問題在於中國或日本歷史上不同朝代的同一年的表示。例如”乾隆”是西元 1735 年和”昭和”是西元 1925 年。時間表示式需要背景知識才能正規化。
另一個類似的問題是需要擴展到文字匹配前的充分表示地點的縮寫。例如”台、印、美”是指”台灣、印度、美國”，因此將需要正規化為”台灣、印度、美國”。

3、 斷詞與中文簡繁轉換

我們的系統中使用的斷詞工具是 ICTCLAS 的斷詞系統，這是由中國科學院計算技術研究所提供。該工具包的功能包括斷詞、詞性標註、NE 識別、新字詞識別，以及自訂字典。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，我們必須將繁體中文文句對轉換成簡體中文文句，然後我們使用的簡繁轉換工具為 google 的線上機器翻譯系統。

(二) 特徵提取

在我們的系統中使用到的特徵在表三列出以前的文字蘊涵識別工作[9]大多數可用的特徵。而本篇提出的單語言機器翻譯將在下一小節描述。在前三個特徵測量T1和T2中根據一般中國類似的字元。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三個特點[7]。Bleu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。

最後四個特徵是 T1 和 T2 的句子長度。我們的系統根據文字和字詞計算 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。

表三、我們使用到的特徵

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	Sub-tree mapping
12	Time mapping

1、 剖析子樹匹配

一個句子的語法資訊也是一個重要的問題。在一個句子的依賴已用於識別意譯的關係[8]。以前的一些文獻表明以不同的方式來衡量兩個解析樹，如樹的編輯距離之間的相似性。子樹的匹配是通過比較兩個句子解析樹的方式來計算兩個句子之間的相似性。在之前我們相信子樹匹配是對系統有所幫助，然而在我們之後的實驗結果其使用後系統效能會略有下降。

2、 時間匹配

當我們觀察到的訓練集資料時我們發現，有許多文句對之中含有時間表示式，然而一些時間表示式是句子重要組成一部分。如表四所示我們分析分為四種類型的匹配。在 T1 和 T2 的時間表達方式可以是：(1)完全匹配、(2)部分匹配、(3)部分匹配、(4)完全不匹配。時間匹配和不匹配的是有用的訓練數據，然而在我們的實驗結果，此特徵沒有提高測試集效能。

表四、時間匹配度例子

匹配程度	例子
時間為完全匹配	t1：據他所知，這是查爾斯首度參加雪梨-荷芭特帆船賽，而查爾斯一向是注重安全、非常謹慎的人，他更想參加2000年雪梨奧運帆船賽。
	t2：2000年奧運在雪梨舉辦
部分時間匹配(1)	t1:若望保祿二世一九七八年十月十六日被選為教宗
	t2：若望保祿二世於1978年當上教宗
部分時間匹配(2)	t1：蘇哈托 1921年6月8日出生
	t2：蘇哈托（Suharto，民間常用「Soeharto」，1921年6月8日－2008年1月27日）
時間完全不匹配	t1：張藝謀1987年以「紅高粱」拿下柏林影展金熊獎
	t2：柏林電影節應該是張藝謀的福地。1988年，他執導的《紅高粱》贏得了最佳影片金熊獎，成為中國電影的首個金熊獎

3、 單語言機器翻譯

我們認為在系統中增加單語機器翻譯作為一項新的可用特徵是有意義的，這項新方法不同於以往的文字蘊涵識別系統。

在我們的實驗中，我們使用 GIZA++作為我們的單語機器翻譯的特徵工具。藉由 GIZA++對訓練集建立一個翻譯模型並計算測試集中文句對集對齊的機率。文字對齊是統計機器翻譯系統訓練很重要的程序。GIZA++[10]這是用於這樣工作的經典工具，GIZA++執行 IBM1-5 模型以及其延伸的的 HMM 模型和更複雜模型 6。產生的這些所有模型是不對

稱的，也就說由選定的翻譯方向，讓他們多對一的進行對齊，但不是一對多的路線。通常進行訓練相反的兩種翻譯方向和對稱，產生字詞對齊提高字詞對齊的品質。兩個對齊模式訓練完全相互獨立，在 GIZA++使用到的 HMM 的對齊模型計算公式如下:

$$p_\alpha(t_1 \mid t_2) = \varepsilon(m \mid l)\sum_a \prod_{j=1}^{m} (t_\alpha(t_{1_j} \mid t_{2_j}) a_\alpha(a_j \mid a_{j-1}, l) \qquad (1)$$

我們將 GIZA++計算出來的對齊機率應用於的分兩類的文字蘊涵任務中作為我們的第 13 個特徵。在我們的系統所使用的計算公式如下：

$$p = \frac{\log\left\{\prod_{i,j=0}^{i,j=\max} p(t1_i \mid t2_j)\right\}}{n} \qquad (2)$$

在下面舉了一個例子說明文句對對齊的機率的應用：

T1:外交部長　胡志強　坦言　以　告　國人，台灣　外交即將　面臨　暴風雨　。

T2:台灣　外交部長　是　胡志強　。

如表五所示為上面例子執行 GIZA++後計算出來的 T1 與 T2 對齊機率，經由公式(1)計算後就即是我們將增加系統的第 13 個特徵。

表五、GIZA++對齊機率例子

T1 單詞	T2 單詞	機率
外交部長	外交部長	0.9951
胡志强	胡志强	0.9512
坦言	台灣	0.2014
台湾	台灣	0.9812
台灣	是	0.0151

P=log(0.9951*0.9512*0.9812*0.0151)/4≒-0.46381736

四、實驗與討論

在這個章節中，我們報告的訓練集與測試集上進行的幾個不同的實驗設定的實驗結果。我們的系統在給定的 407 對訓練集和測試集做 10 倍交叉驗證訓練，並使用相同的系統處理另一個 407 對文句開放測試集。表六中列出的四個設定的實驗結果。在我們實驗結果中第二的設定得到最好的效能，其正確率為 0.69。

表六、實驗結果總表

	10 倍交叉驗證訓練	公開測試集

1~10 特徵 [9]	0.6560	0.6830
1~10 特徵與機器翻譯	**0.6658**	**0.6904**
1~12 特徵 [9]	0.6461	0.5577
1~12 特徵與機器翻譯	0.6560	0.5749

(一)實驗結果

如表七所示在第一個實驗中我們使用到表三 1 到 10 個特徵進行 10 倍交叉驗證實驗，接著我們加入單語言機器翻譯特徵如表八所示可以提昇其效能。

表七、使用 10 特徵 10 倍交叉驗證實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	70	42	112
N	98	197	295
Total	168	239	407

表八、使用 10 特徵加入機器翻譯特徵 10 倍交叉驗證實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	72	40	112
N	96	199	295
Total	168	239	407

如表九所示在第二個實驗中我們使用到表三 1 到 12 個特徵進行 10 倍交叉驗證實驗，接著我們加入單語言機器翻譯特徵如表十所示可以提昇其效能。

表九、使用 12 特徵 10 倍交叉驗證實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	68	44	112
N	100	195	295
Total	168	239	407

表十、使用 12 特徵加入機器翻譯特徵 10 倍交叉驗證實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	70	42	112

N	98	197	295
Total	168	239	407

如表十一所示在第二個實驗中我們使用到表三 1 到 10 個特徵與公開測試集進行實驗，接著我們加入單語言機器翻譯特徵如表十二所示可以提昇其效能。

表十一、使用 10 特徵公開測試集實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	172	38	210
N	91	106	197
Total	263	144	407

表十二、使用 10 特徵加入機器翻譯特徵公開測試集實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	174	36	210
N	89	108	197
Total	263	144	407

如表十三所示在第二個實驗中我們使用到表三 1 到 10 個特徵與公開測試集進行實驗，接著我們加入單語言機器翻譯特徵如表十四所示可以提昇其效能。

表十三、使用 12 特徵公開測試集實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	126	43	169
N	137	101	238
Total	263	144	407

表十四、使用 10 特徵加入機器翻譯特徵公開測試集實驗結果

Predicted	**Actual**		**Total**
	Y	**N**	
Y	129	40	169
N	134	104	238
Total	263	144	407

(二)實驗討論

我們的系統的額外的機器翻譯新特徵提高了正確率，無論是的交叉驗證或是公開測試集的情況，實驗結果也證實第 11 和第 12 的特徵沒有改進我們的效能。從混淆矩陣我們可以發現，該系統是在數據分佈方面的強勁。yes/no 在訓練和開放測試集的分佈有很大的不同。該系統可以大多數識別正確。

五、結論

本篇報告改進我們參加 RITE1 的系統，我們增加了一個新的機器翻譯特徵到我們的系統中並取得了較好的效能，我們用機器翻譯方法來翻譯同一種語言作為一種方法來識別語言中的意譯。我們的系統是專門處理中文部份，然而同樣的想法以我們系統的基礎也可能應用在不同的語言。

從資料觀察得知，處理這個問題背景知識是最必要的條件，像中國或是日本人的朝代名稱，在時間匹配之前必須轉換成相同表示，地理知識也是必要的。這些需求是超出任何正常大小的訓練集和語言知識的內容。從 Web 挖掘需要的必要知識可能是一個有用的資源來源來。

在本篇我們提出了單語言機器翻譯來改進我們的系統，在實驗結果中發現這個新的方法的確可以改進我們的系統，但是改進的幅度並不高因為機器翻譯這個方法需要相當大量的訓練資料才能讓 GIZA++文字對齊效果更準確，所以在未來希望可以使用更大量的資料集來提昇文字對齊的精確度。

致謝

本研究依經濟部補助財團法人資訊工業策進會「100 年度數位匯流服務開放平台技術研發計畫」辦理。感謝國科會贊助部分研究經費，計畫編號 NSC- 100-2221-E-324 -025-MY2。

參考文獻

[1] Ido Dagan and Oren Glickman, 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.

[2] NTCIR 9, Recognizing Inference in TExt task, http://artigas.lti.cs.cmu.edu/rite/Main_Page.

[3] Ion Androutsopoulos and Prodromos Malakasiotis, “A survey of paraphrasing and textual entailment methods”, Journal of Artificial Intelligence Research, Volume 38, pages 135-187, 2010.

[4] Christiane Fellbaum, “WordNet: An Electronic Lexical Database”, The MIT Press, 1998.

[5] Prodromos Malakasiotis, Ion Androutsopoulos, “Learning textual entailment using SVMs and string similarity measures”, In Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 42–47, Prague, Czech Republic, 2007.

[6] Wan, S., Dras, M., Dale, R., & Paris, C., "Using dependency-based features to take the "para-farce" out of paraphrase", In Proceedings of the Australasian Language Technology Workshop, pages 131–138, Sydney, Australia, 2006.

[7] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation", In Proceedings of the 40th Annual Meeting on ACL, pages 311–318, Philadelphia, PA, 2002.

[8] Prodromos Malakasiotis, "Paraphrase recognition using machine learning to combine similarity measures", In Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2009.

[9] Shih-Hung Wu, Wan-Chi Huang, Liang-Pu Chen and Tsun Ku. 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.

[10] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models." *Computational Linguistics*, vol. 29, no. 1, pp. 19–51,2003.