

國立交通大學

資訊科學與工程研究所

碩士論文

中文篇章中時間關係的辨識研究

A Study of Temporal Relation Recognition in Chinese
Texts

研究生：黃慧庭

指導教授：梁 婷 教授

中華民國 九十七 年 七月

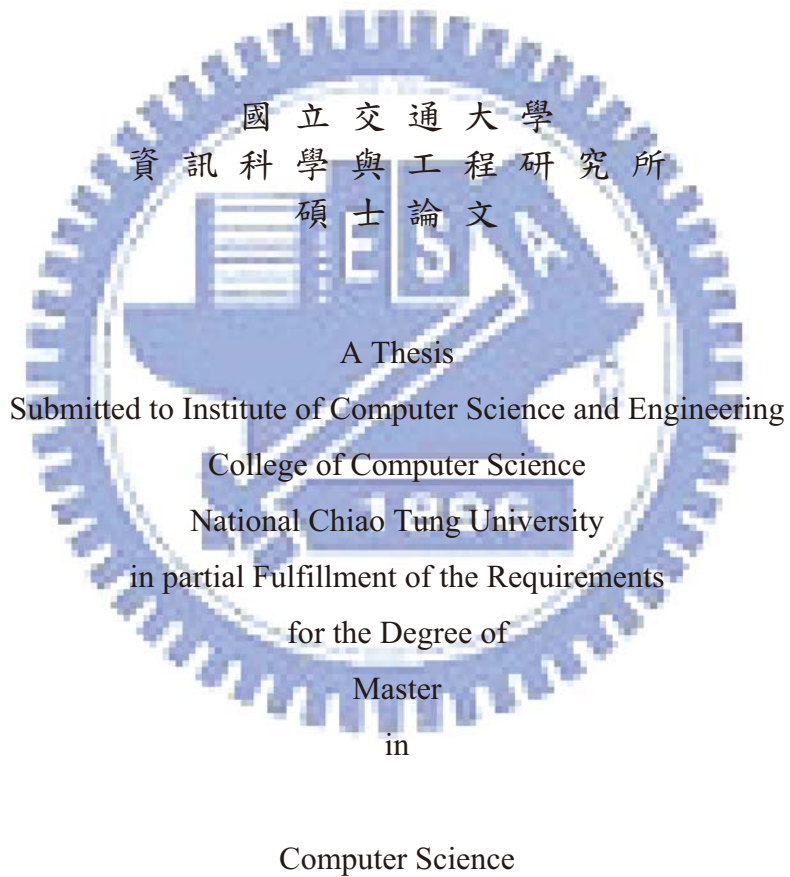
中文篇章中時間關係的辨識研究
A Study of Temporal Relation Recognition in Chinese Texts

研究生：黃慧庭

Student : Hui-Ting Huang

指導教授：梁 婷

Advisor : Tyne Liang



July 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年七月

中文篇章中時間關係的辨識研究

研究生：黃慧庭

指導教授：梁婷

國立交通大學資訊科學與工程研究所

摘要

一般文本式的篇章中，語段間常包含許多不同的語意銜接關係，諸如：轉折、並列、因果……等等。在此篇論文中，我們針對時間關係進行探討，時間關係存在於具備時間先後順序的雙語段間。本論文提出一個統計式的辨識中文語篇時間關係的方法，並實作一個中文語篇時間關係自動標記的系統。時間關係辨識可應用於其他的自然語言處理領域，例如可用於實作作文評分系統。我們使用所挑選出的特徵作為辨識依據，這些特徵包含位置特徵、人物主詞、主要動詞組、時間詞組、動詞前動詞、副詞、時態字根、主要動詞詞性組以及主題詞；然後，利用自動產生的訓練語料進行訓練，並建立出辨識邏輯。我們嘗試三種訓練方式，貝氏分類器與 C4.5 決策樹，當中正確率最高者達 78%。

關鍵詞：語篇關係、連貫關係、時間關係、自動標記

A Study of Temporal Relation Recognition in Chinese Texts

Student : Hui-Ting Huang

Advisor : Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung Univeristy

The logo of National Chiao Tung University is a circular emblem with a gear-like border. Inside the circle, there is a stylized building and the year '1896'. The word 'ABSTRACT' is overlaid on the logo.

ABSTRACT

It's called temporal relation when there is a temporal sequence between a segment and another. In this thesis, we proposed a Chinese discourse temporal relation extraction method. We implemented a Chinese discourse temporal relation extraction system using features. Temporal relation can be widely applied to other natural language research. We selected some features among lexical, Part-Of-Speech, and other surface features. Three kind of training methods, Naive Bayes Classifier and C4.5 decision tree, are used. We use the automatically generated training corpus to do the training. Then we use the resulting extraction logic to construct the automatic labeling system.

Keywords: Discourse Relation, Rhetorical Relation, Coherence Relation, Temporal Relation, automatic labeling.

致 謝

能夠完成這篇論文，首先要感謝我的指導教授梁婷教授，感謝老師對於碩士生的要求，讓我這兩年沒有白過，有一些確實的收穫。另外也感謝我的口試委員們，張教授俊盛、劉教授美君、胡教授毓志，認真的給我論文上的建議。同時也感謝實驗室的學長們，感謝吳典松學長、龔自良學長、鄭建富學長、鄭守益學長在我剛來實驗室的時候給我的建議；感謝朱俊榮學長、陳淳齡學姐、楊哲青學長、劉正義學長不只常常答覆我的問題，也在閒暇之餘陪我聊天打屁；感謝潘善均學長、曾盛興學長這一兩年的陪伴。最後也感謝我的家人，特別是媽媽跟妹妹在我忙碌的時候為我打氣；還有室友昱婷在我忙碌的時候幫我買飯還有處理一些雜事；也感謝以前的同學佳霖、惠君在我最辛苦的時候給我安慰；然後還有我的兩隻小兔子，估狗跟兔比在我壓力大的時候給我無聲的撫慰。非常感謝大家，我能順利完成這篇論文，每一分的幫助都是重要的。

目 錄

中文篇章中時間關係的辨識研究i

A Study of Temporal Relation Recognition in Chinese Texts..... ii

致 謝 i

目 錄 ii

表格目錄 iv

圖片目錄 vi

第一章 緒論	1
1.1 時間關係探討.....	1
1.2 中文文本探討.....	2
1.3 相關文獻探討.....	4
1.4 研究方法概觀.....	6
第二章 特徵探勘	7
2.1 語料庫介紹.....	7
2.2 訓練與測試語料的建置.....	8
2.2.1 正例.....	8
2.2.2 負例.....	11
2.2.3 中文語料性質處理.....	12
2.3 特徵擷取.....	13
2.3.1 位置特徵.....	14
2.3.2 主詞為人物.....	14
2.3.3 主要動詞組.....	15
2.3.4 時間詞組.....	16
2.3.5 主要動詞前動詞.....	17
2.3.6 主要動詞前後副詞.....	18
2.3.7 時態字根.....	20

2.3.8	主要動詞詞性組.....	21
2.3.9	小句重心一致性.....	22
第三章	機器學習法.....	24
3.1	訓練方法.....	24
3.1.1	貝氏分類器.....	24
3.1.2	C4.5 決策樹.....	24
3.2	實驗介紹.....	25
3.3	實驗二：訓練實驗.....	26
3.4	實驗二：特徵組合實驗.....	28
第四章	標記系統	30
4.1	標記方法.....	30
4.1.1	切分文章.....	30
4.1.2	辨識句內時間關係.....	30
4.1.3	辨識句間時間關係.....	31
4.1.4	合併辨識結果.....	31
4.2	實驗四：標記結果實驗.....	33
第五章	以長句為單位進行句內時間關係辨識	39
5.1	測試資料.....	39
5.2	特徵擷取.....	40
5.2.1	主詞為人物.....	40
5.2.2	主要動詞組.....	41
5.2.3	時間詞組.....	41
5.2.4	主要動詞前動詞.....	41
5.2.5	主要動詞前後副詞.....	41
5.2.6	時態字根.....	42
5.2.7	主要動詞詞性組.....	42
5.2.8	小句重心一致性.....	42
5.3	實驗三：句內特徵擷取實驗.....	42
第六章	結論	44

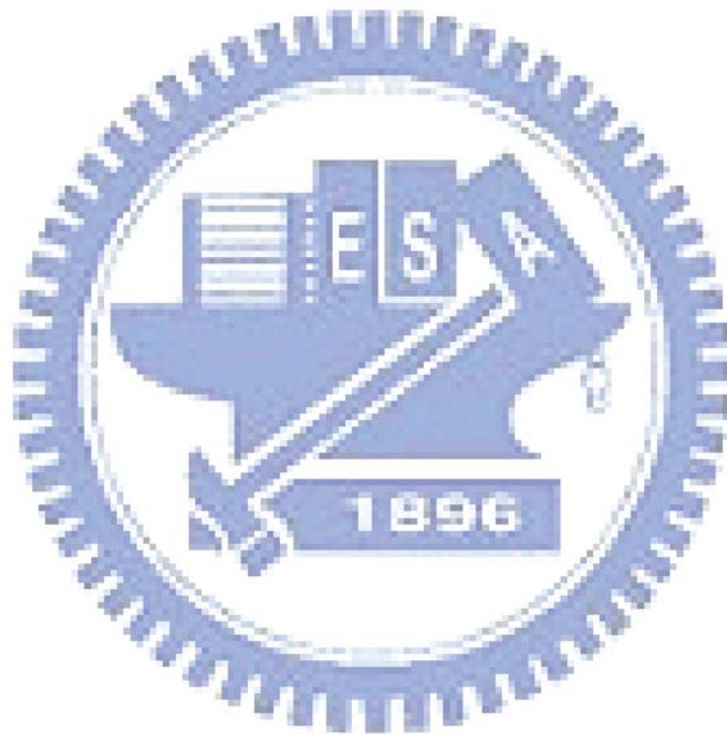
參考資料 46

附錄 47

表格目錄

表 1：語料庫資料.....	7
表 2：正例種子詞.....	9
表 3：正例例數.....	10
表 4：同一長句小句數量表.....	10
表 5：同一篇文章中小句數量表.....	10
表 6：負例 2 例數.....	11
表 7：句內特徵列表.....	13
表 8：句內特徵列表.....	13
表 9：關鍵詞出現位置統計（三分法）.....	14
表 10 時態字根列表.....	20
表 11：實驗結果情況列表.....	25
表 12：測資表.....	26
表 13：句間實驗結果.....	26
表 14：句內實驗結果.....	27
表 15：特徵組合句內實驗.....	28
表 16：標點符號列表.....	30
表 17：雙語段標記結果統計資料.....	34
表 18：句內標記結果統計資料.....	34
表 19：句間標記結果統計資料.....	34
表 20：正例種子詞.....	39
表 21：正例例數.....	40
表 22：句內特徵列表.....	40
表 23：句內實驗結果.....	42
表 24：特徵組合句內實驗.....	43
附表 25 正例產生關鍵詞與句型.....	47
附表 26 負例 2 產生關鍵詞與句型.....	49
附表 27 正例範例.....	51
附表 28 負例範例一.....	53
附表 29 負例範例二.....	55
附表 30 代名詞出現狀況統計資料.....	57
附表 31 動詞組出現狀況統計資料（前後 20 名）.....	58
附表 32 動詞出現狀況統計資料（前 20 名）.....	60
附表 33 副詞出現狀況統計資料（前 20 名）.....	61

附表 34 動詞詞性組出現狀況統計資料（前 10 名）	62
附表 35 實驗三動詞組出現狀況統計資料（前後 20 名）	63
附表 36 實驗三動詞出現狀況統計資料（前 20 名）	65
附表 37 實驗三副詞出現狀況統計資料（前 15 名）	66
附表 38 實驗三動詞詞性組出現狀況統計資料（前 10 名）	67
附表 39 字根出現狀況統計資料.....	68



圖片目錄

圖 1：合併演算法.....	12
圖 2：句間動詞組統計數值.....	16
圖 3：句內動詞組統計數值.....	16
圖 4：句間動詞統計數.....	18
圖 5：句內動詞統計數值.....	18
圖 6：句間動詞前副詞統計數值.....	19
圖 7：句間動詞後副詞統計數值.....	19
圖 8：句內動詞前副詞統計數值.....	19
圖 9：句內動詞後副詞統計數值.....	20
圖 10：句間動詞詞性組統計數值.....	22
圖 11：句內動詞詞性組統計數值.....	22
圖 12：句內合併演算法.....	32
圖 13：句間合併演算法.....	33
圖 14：標記結果一.....	35
圖 15：標記結果二.....	36

