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 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 Liang-Pu Chen, 邱宏昇 Hung-Sheng Chiu, 楊仁達 Ren-Dar Yang
財團法人資訊工業策進會
Institute for Information Industry, Taipei, Taiwan (R.O.C)
{eit, bbchiu, rdyang } @iii.org.tw

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)的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。文字蘊涵識別最基本的方法是藉由句子字面上的資訊例如語意、句法[2]等等進而推斷句子是否有著蘊涵關係,因此文字蘊涵識別可以應用到其他自然語言處理的研究中，如問答系統、資訊抽取、資訊檢索、機器翻譯[3][4]等等。

我們所參與公開評測 NTCIR10 RITE-2[5]將文字蘊涵的研究分成兩種層面，首先是分兩類(Binary Class, BC) ，任務的目標是單純判別 T1 與 T2 之間是否具有蘊涵關係。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，NTCIR RITE 另外定義多類(Multi Class, MC)這項任務，將句子之間的蘊涵分類為正向、雙向、矛盾、與獨立四種關係。假設這個句子對具有蘊涵關係，但有可能兩個句子所包涵的資訊數量不同，造成我們只能從其中一個句子推論出另一個句子的完整的意思,這樣的情況我們稱為兩個句子間的蘊涵關係為正向蘊涵。反之兩個句子可以互相推論出另一個句子的含意，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。可能兩個句子所包涵的資訊大致相同只是少部份資訊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。

在本文中將介紹我們的觀察 NTCIR-10-RITE-2 資料集以及正式評測結果後發現過去系統[6]的缺陷，進而提出如何改進中文文字蘊涵系統。過去處理文字蘊涵大多使用機器學習的方法，這種一視同仁方法處理，對於比較特別的問題往往在處理時會產生誤判。我們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。與過去系統[6]最大的不同在於加入特殊類型問題處理的子系統,在系統處理完預處理後將可以特殊類型處理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結果，作整合得到最後的結果。目前我們已經實做了”肯定/否定句”、”時間資訊不一致”、”數字資訊不一致”、 ”主/受詞資訊不一致”四個特殊類型問題處理子系統，

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完成。

實驗結果顯示配合之前提出的機器學習方法，增加特殊類型分類對特殊類型句子進行個別處理，這樣的過程可以有效改進系統，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%。另外過去系統在繁體中文上的處理結果不佳，因此改使用我們自行開發的機器翻譯系統[7]，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，在兩類(BC)任務提正確率高 6.02%以及多類(MC)任務則是提高正確率 9.49%。

關鍵詞：中文文字蘊涵識別、蘊涵分析

參考文獻

[1] Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.) Machine Learning Challenges. Lecture Notes in Computer Science , Vol. 3944, pp. 177-190, Springer, 2006.

[2] Dong-Bin Hua, Jun Dinga," Study on Similar Engineering Decision Problem Identification Based on Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS", Systems Engineering Procedia Volume 1, 2011, Pages 406–413

[3] 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.

[4] Yongping Ou, Changqing Yao, "Recognize Textual Entailment by the Lexical and Semantic Matching", Computer Application and System Modeling, 2010 International Conference on V2-500 -504

[5] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng -Wel Lee, Chuan-Jie Lin , Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, Kohichi Takeda," 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.

[6] Shan-Shun Yang, Shih-Hung Wu, Liang-Pu Chen, Wen-Tai Hsieh, and Seng-cho T. Chou, Improving Binary-class Chinese Textural Entailment by Monolingual Machine Translation Technology, in Proceedings of the IEEE IRI 2012, Las Vegas, USA,8 Aug, 2012.

[7] Min-Hsiang Li, Shih-Hung Wu, Yi-Ching Zeng, Ping-che Yang, and Tsun Ku, Chinese Characters Conversion System based on Lookup Table and Language Model, Computational Linguistics and Chinese Language Processing, Vol. 15, No. 1, March 2010, pp. 19-36.