The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 116-118

探究不同領域文件之可讀性分析

Exploring Readability Analysis on Multi-Domain Texts

曾厚強　Hou-Chiang Tseng, 陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
ouartz99@gmail.com, berlin@ntnu.edu.tw

宋曜廷　Yao-Ting Sung
國立臺灣師範大學教育心理與輔導學系所
Department of Educational Psychology and Counseling
National Taiwan Normal University
sungtc@ntnu.edu.tw

摘要

可讀性(Readability)是指閱讀材料能夠被讀者所理解的程度[1],[2],[3],[4]，當讀者閱讀高可讀性的文件時，會產生較好的理解及學後保留效果[2],[3]。由於文件的可讀性在知識傳遞扮演極為重要的角色，因此西方的可讀性公式發展的非常早[5], [6]。然而這些傳統的可讀性研究大多使用較淺層的語言特徵來發展線性的可讀性公式，其實並不足以反映文件難度。Graesser、Singer 和 Trabasso 便指出，傳統語言特徵公式無法反映閱讀的真實歷程，文件的語意語法只是文件的淺層語言特徵，沒有考量文件的凝聚特性[7]。Collins-Thompson 亦指出傳統可讀性公式僅著重在文件的表淺資訊，而忽略文件重要的深層特徵。這也讓傳統可讀性公式在預測文本可讀性的結果常遭受到質疑[8]。直到今日，可讀性的研究仍持續不斷。研究人員為了克服傳統可讀性公式的缺點，嘗試利用更細緻的機器學習演算法來發展出非線性的可讀性模型，並納入更多元的可讀性指標來共同評量文本的可讀性，以提升可讀性模型的效能[9],[10],[11]。然而可惜的是，研究人員發現採用一般語言特徵的可讀性模型在應用於特定領域文時，一般語言特徵並無法判斷詞彙在不同領域文本時背後所代表的意義。因此開始有學者去針對特定領域文本的知識結構研發出專屬於該領域的特徵來取代一般語言特徵[12],[13]，使可讀性模型可以正確評估特定領域文本的可讀性。由上述的研究可知，不論是過去一般語言特徵或是針對特定領域文本的知識結構所設計的文件表示(Document Representation)技術，長久以來都需要仰

賴專家來研發，有著耗時費力等問題。近年來，有所謂表徵學習(Representation Learning)方法可以自動從原始資料中去擷取有用的特徵以建立文本的向量表示，能有助於分類模型的訓練和預測[14]。使得模型所需要的特徵可以逐漸不需仰賴專家，成功開啟了另一個研究的方向。因此，本研究基於近年來熱門的卷積神經網路(Convolutional Neural Network, CNN)[15]或快速文本(festText)[16]等不同的表示學習法來自動抽取文本特徵，訓練出一個能夠分析跨領域文件的可讀性模型。實驗結果顯示兩種模型皆有優異的效能；本研究亦發現兩種模型在預測錯誤的程度上是有所差異的。在未來的研究中，本研究也將探討如何整合不同類型的類神經網路模型的優點來促使可讀性模型在預測錯誤時，其誤差也能夠盡可能的往適讀年級集中。

關鍵詞：可讀性，詞向量，卷積神經網路，表示學習法，快速文本

參考文獻

[1]E. Dale and J. S. Chall, “The concept of readability,” *Elementary English*, vol. 26, pp. 19–26, 1949.

[2]G. R. Klare, “Measurement of Readability,” 1963.

[3]G. R. Klare, “The measurement of readability: useful information for communicators,” *ACM Journal of Computer Documentation (JCD)*, vol. 24, pp. 107-121, 2000.

[4]G. H. McLaughlin, "SMOG grading: A new readability formula," *Journal of reading*, vol. 12, pp. 639–646, 1969.

[5]B. A. Lively and S. L. Pressey, “A method for measuring the vocabulary burden of textbooks,” *Educational administration and supervision*, vol. 9, pp. 389–398, 1923.

[6]M. Vogel and C. Washburne, “An objective method of determining grade placement of children's reading material,” *The Elementary School Journal*, pp. 373–381, 1928.

[7]A. C. Graesser, M. Singer, and T. Trabasso, “Constructing inferences during narrative text comprehension,” *Psychological Review*, vol. 101, p. 371, 1994.

[8]K. Collins-Thompson, “Computational assessment of text readability: A survey of current and future research,” *International Journal of Applied Linguistics*, vol. 165, pp. 97–135, 2014.

[9]S. E. Petersen and M. Ostendorf, “A machine learning approach to reading level assessment,” *Computer Speech & Language*, vol. 23, pp. 89–106, 2009.

[10] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, “A comparison of features for automatic readability assessment,” in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 276–284.

[11] Y. T. Sung, J. L. Chen, J. H. Cha, H. C. Tseng, T. H. Chang, and K. E. Chang, “Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning,” *Behavior research methods,* vol. 47, pp. 340–354, 2014.

[12] X. Yan, D. Song, and X. Li, “Concept-based document readability in domain specific information retrieval,” in *Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, pp. 540–549.

[13] A. Borst, A. Gaudinat, C. Boyer, and N. Grabar, “Lexically based distinction of readability levels of health documents,” *Acta Informatica Medica,* vol. 16, pp. 72–75, 2008.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning (adaptive computation and machine learning series)*. MIT Press, 2016.

[15] Y. LeCun, "Generalization and network design strategies," *Connectionism in perspective,* pp. 143-155, 1989.

[16] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*.