Wölfpak™: A Neural Network for Multilingual Text Analysis

Paper presented at the 25th annual meetings of the
International Society for Network Analysis
Los Angeles, CA, February 16-20, 2005

Joe K. Woelfel Raymond Hsieh
MERL University at Buffalo

Hao Chen Jennie Hwang
University at Buffalo University at Buffalo

Pauline Hope Cheong Devon Rosen
University at Buffalo University at Buffalo

Joseph Woelfel
University at Buffalo

Abstract

Text analysis systems based on linguistic or grammatical characteristics and heuristics are restricted to the language for which they have been developed. Neural Networks, on the other hand, which operate only by recognizing, storing and retrieving patterns, have great potential for working across a wide variety of languages. Catpac II™, for example, has been used successfully in several languages as different as English, German, Korean and many other languages, but is restricted to languages that can be expressed in the ASCII character set.

Wölfpak™, a new variation of Catpac, is a neural text analyzer which recognizes the UNICODE set, and may be used across a very wide array of languages. In this paper, we show the results of applying Wölfpak™ in English and Chinese.

Introduction: The Problem

All text analysis systems involve a search for recurrent patterns. Sometimes this search involves thoughtful human analysts scrutinizing texts for whatever patterns they may identify intuitively. Other systems involve a computer algorithm searching for predetermined combinations of words and phrases based on some linguistic or substantive theory. A third class of systems involves computer algorithms that search for any patterns whatsoever that happen to occur in the text. The most common forms of this third class are co-occurrence models and, less commonly, neural networks (Klein, http://www.textanalysis.info/).

The first two classes of text analysis algorithms are inherently language specific. Human analysts can only analyze languages they understand, and theories and heuristics based on dictionaries of words or search phrases apply only to the language for which they have been written. Both co-occurrence models and neural networks, since they make no reference to any linguistic or language-specific rule or heuristic, can, in general, be applied to any language.

Catpac™, (Woelfel, 1993) perhaps the first neural text analysis program, began as a co-occurrence model in the early 1980's, but added a neural network engine in 1988, and first sold commercially in 1990. Early versions allowed both co-occurrence and neural analysis as options, but later versions left out the co-occurrence model because the neural model provided consistently deeper and more informative results than the co-occurrence model. (Early versions including the co-occurrence option are still available from The Galileo Company.) Catpac™ has been used successfully in a wide variety of languages, including such diverse languages as English, German, and Korean, but it only reads the ASCII character set, and requires translation of non-ASCII languages into their ASCII equivalent.

Wölfpak™ is a new variation of the original Catpac™ program. It is a platform independent program based on UNICODE and thereby requires no translation. While existing descriptions of neural software often describe data as running words through scanning windows, in fact what both programs see is not words, but a bit stream. Consequently, it is of no consequence whether this bit stream represents words or characters, such as Kanji or Chinese characters. This is because the neural network will identify, store and retrieve recurrent patterns in the bit stream regardless of what they represent. All neural algorithms' performance depends on several non-linear and interactive parameters, and there is no fixed setting of these parameters that is optimal for all texts. For this and other reasons, each version of Catpac™ has been slightly different from its relatives, and Wölfpak™, while based on Catpac™, yields results very similar, but not identical, to Catpac II™.

Analysis of Multi-Lingual Texts

In this paper, Wölfpak™ analyses were run on both English and Chinese texts. This is preliminary analysis, which aims to test the capabilities of the new variation of the original Catpac™ program in a language other than English. For illustrative purposes, one test was run on an English web site, and another analysis was conducted on a Chinese newspaper article.

a) Wölfpak™ in English

Figure One shows the results of Wölfpak™ analysis of the first half of the first page of the Galileo Website. It does a good job of capturing the underlying clusters of ideas represented there, which includes an analysis of the philosophical differences among Aristotle, Plato and early Greek philosophers and the comparative scientific method of Galileo Galilei and his successors.

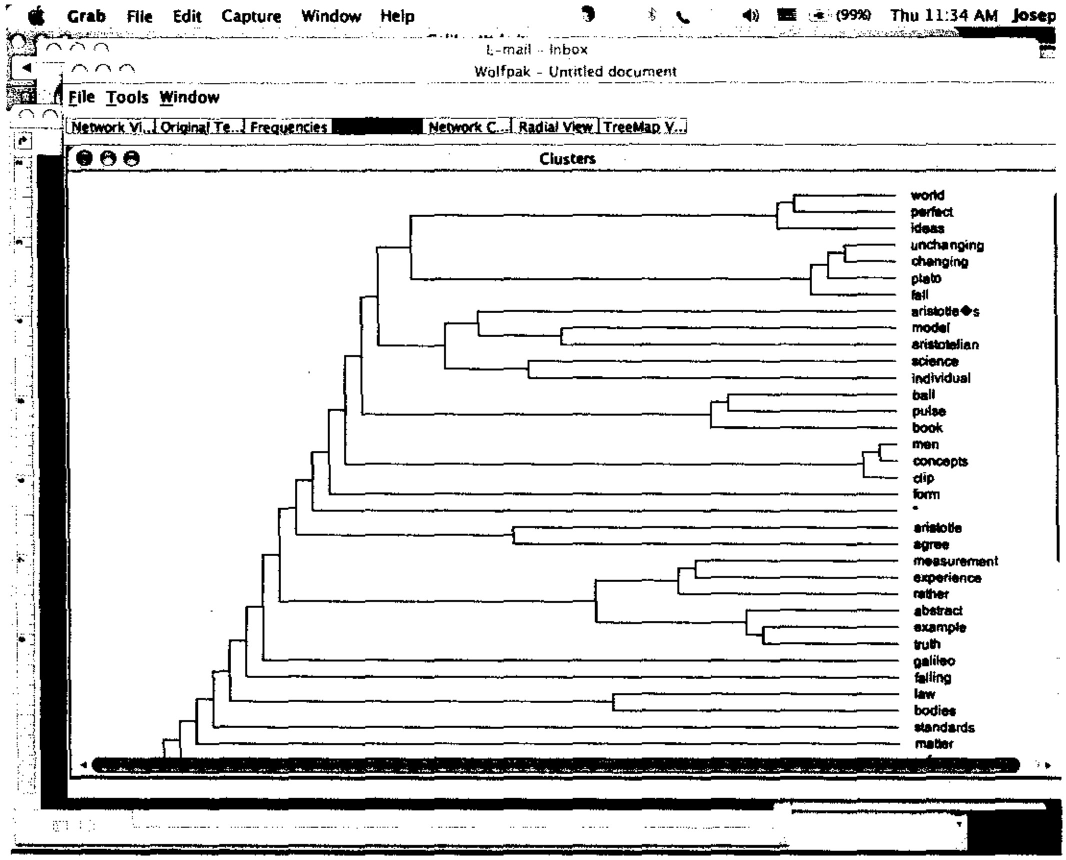

Figure 1: Clusters underlying The Galileo Website page "What is it?"

b) Wŏlfpak™ in Chinese

Figure two shows the results of Wŏlfpak™ analysis of a news article extracted from People's Daily Online (http://www.people.com.cn/). The article concerns foreign policy, in particular on the U.S. agreement with Indonesia to request foreign rescue missions to register at the Indonesian army. (See appendix for article).

The article explains Indonesia's rationale for insisting that foreign rescue groups register with the army. Reasons include a) The Muslim ethnic majority in Indonesia have an aversion towards foreign armies intruding on their land, b) The U.S. rescue mission is perceived to be self-serving its own strategic objectives, and c) The general unease towards the dispatchment of Japanese troops overseas. Results show that Wŏlfpak™ captures the underlying clusters of ideas represented in the news article well.

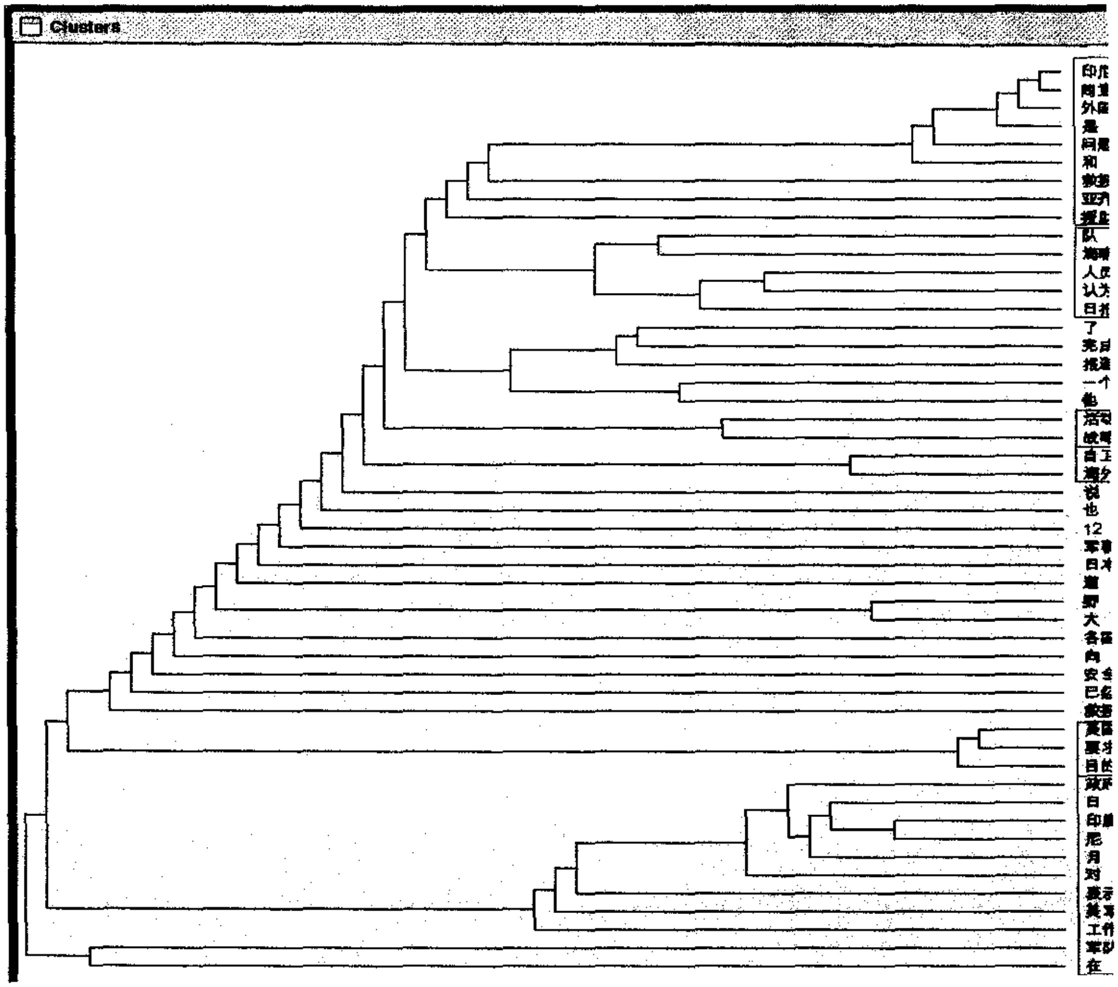

Figure 2: Clusters underlying the Chinese news article 'United States agree with Indonesia's request for the registration of foreign troops with the Indonesian army' 美国同意印尼要求外国军队限期撤出亚齐"

From the results shown above, the strongest clusters are:

- 印尼 – 同意 – 外国 – 是 – 问题 – 和 – 救援 – 亚齐 – 援助

- 队 – 海啸 – 人员 – 认为 – 日报

- 活动 – 战略

- 自卫 – 海外

- 美国 – 要求 – 目的

- 政府 – 日 – 印度 – 尼 – 月 – 对 – 表示 – 美军 – 工作 – 军队 – 在

The translations of the clusters above are:

- Cluster 1: Indonesia – Agree – Foreign Countries – Is – Problem – And – Rescue – Aceh – Aids

- Cluster 2: Group – Tsunami – Staff – Think – Daily

- Cluster 3: Activity – Strategic

- Cluster 4: Self-Defense – Oversea

- Cluster 5: U.S. – Request – Objective

- Cluster 6: Government – Day – India – (Part of the word of Indonesia) – Month – To – Express – U.S. Army – Work – Army – On

Appendix 1: Chinese News article

美国同意印尼要求外国军队限期撤出亚齐

新华社雅加达１月１４日电（记者吴强 余谦梁）美国驻印度尼西亚大使帕斯科１３日表示，美国同意印尼政府关于外国军队在３月２６日之前从印尼西部省亚齐撤出的要求，并认为印尼政府要求在亚齐的外国救援人员向印尼军队登记是合理的。

据印尼媒体１４日报道，帕斯科说，亚齐存在安全问题，而且短时间内大量外国人涌入肯定给印尼政府带来许多困难。因此，印尼政府要求外国人登记是合理的，这也是一个主权国家的权利。他表示，只要印尼政府认为美国等国的外国军队已经完成了救灾工作，美国军队马上就会从亚齐撤走。

１２日，印尼政府要求在亚齐的外国军队在３月２６日之前完成工作后撤离，并表示灾区目前急需的是医疗卫生和工程援助人员。印尼政府还要求在亚齐的外国

救援人员把救援工作集中在班达亚齐、米拉务和亚齐贝萨尔三个城市，以便协调工作和保证安全。

印尼要求美国等外国军队尽快撤离灾区

印度洋海啸发生后，美国、日本、澳大利亚等国家派遣了军舰、飞机和士兵前去援助。这种援助隐含的政治外交考量引起受灾国的疑虑。据外国媒体报道，印度尼西亚政府12日要求参与海啸灾难救援的外国军队尽快撤走。

路透社12日自雅加达报道说，印尼副总统优素福·卡拉表示，外国军队在该国逗留的时间不能超过三个月，他们一完成救援工作就要撤离亚齐特区。他强调，亚齐在不久的将来需要的是工程师和外国医务人员，而不是军事援助，“我们不再需要外国军队”。

报道认为，外国军队进入印尼是个敏感问题，穆斯林占人口绝大多数的印尼一直不愿让外国军队，特别是美澳军队踏上它的领土。针对美军在印尼的救援行动，印尼政府反对他们携带武器，也不同意其在印尼沿海地区建立大本营。

美国航母离开印尼班达亚齐附近的水域

据法新社12日报道，印尼对苏门答腊岛海啸灾区的外国援助行动提出了限制，这包括为灾民运送救援物资的美国军队。由于印尼的反对，作为美国救援行动主要基地的“亚伯拉罕·林肯”号航空母舰已经离开班达亚齐附近的水域。而且，由于印尼对安全的考虑，向幸存者提供帮助的美国海军陆战队也削减了陆地援救人员的数量，一些陆战队员已经回到舰上。

美军海啸救援具有战略目的

日本《东京新闻》13日报道说，在针对苏门答腊地震和海啸的救灾活动中，美军派出了1．4万人、多架飞机和船只进行空前规模的紧急援助活动。面对伊拉克问题、阿富汗问题和朝鲜问题，美军应该已是应接不暇，为何还在人道救援上投入如此巨大的力量呢？美军开展如此迅速的大规模救援活动目的何在呢？军事评论家藤井治夫分析说：“美军此举目的是显示自身力量，同时要推进针对‘不稳定地区’战事的计划，打算在此次救援活动中模拟演习美军主导的联合作战。”这种观点主要

着眼于反恐、马六甲海峡运输线防卫和天然气开发等。军事评论家神浦元彰则认为：“由于伊拉克战争，世界各国尤其是伊斯兰国家对美国的印象恶化。美军希望通过对最大的伊斯兰教国家印度尼西亚提供人道援助，来改变穆斯林对美国的印象。”军事评论家稻垣治就此指出：“此次援助活动也有明确的战略目的，即以作出大家能够目睹的贡献来获得加分。‘9·11’事件后，布什政府一面实施先发制人理论，一面也奉行‘战争之外的作战’这一灵活战略。”

亚洲各国警惕日本海外派兵

日本《朝日新闻》13日报道报道说，亚洲各国警惕日本自卫队扩大海外活动。日本防卫厅长官大野功统结束对新加坡和马来西亚等东南亚三国的访问，于12日进入最后的一个访问国————韩国。大野这次历访是为了向各国说明时隔9年后首次修改的新防卫计划大纲，与此同时也有强调自卫队的海外活动是自卫队“本来任务”的意义，海外活动是自卫队的本来任务已写入新的防卫计划大纲。但是，在防卫厅长官大野所到之处，人们所关注的不是对自卫队扩大作用的期待感，而是警戒感。印尼国防部长尤沃诺9日同大野会谈时，对包括日本自卫队在内的外国军队的救援工作表示了谢意，但是他又叮嘱说：“驻扎外国军队是一个很微妙的问题。”印尼国防部长尤沃诺却接二连三地列举了自卫队应“遵守的规则”：“当前的活动期限是三个月”，“如果延长时间，则必须得到印尼政府的同意胁。”

来源：新华社
2005年01月14日16:40
(http://world.people.com.cn/GB/1029/42355/3120770.html)

References

(Klein, http://www.textanalysis.info/).

Woelfel, J., "Artificial neural networks in policy research: A current assessment," Journal of Communication, 43(1), 63-80. 1993.