รหัสโครงการ 11p34c339

คูวส์ (CUWS)

BEST 2009 – **การแบ่งคำไทย** (Thai Word Segmentation)

รายงานฉบับสมบูรณ์
เสนอต่อ
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ
กระทรวงวิทยาศาสตร์และเทคโนโลยี

ได้รับทุนอุดหนุนโครงการวิจัย พัฒนาและวิศวกรรม
โครงการแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์แห่งประเทศไทย ครั้งที่ 11
ประจำปีงบประมาณ 2551

โดย
ผู้พัฒนา
นายวิชญ์ เนียรนาทตระกูล
นายไพโรจน์ ลีลาภัทรกิจ
นายจิรัฏฐ์ ศรีสวัสดิ์

บทคัดย่อ

การแบ่งคำภาษาไทยนั้นเป็นหัวใจหลักในการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ซึ่งสามารถนำไปประยุกต์ใช้ได้ในหลายด้านโดยเฉพาะด้านธุรกิจสิ่งพิมพ์ และด้านการประมวลผลภาษาธรรมชาติแบบงานอื่นๆ เช่น การแปลข้อความอัตโนมัติด้วยเครื่อง แต่เนื่องด้วยการวิจัยทางด้านการแบ่งคำไทยมีน้อยผนวกกับคลังข้อความที่ใช้ในการประมวลผลนั้นยังไม่เพียงพอ ผู้พัฒนาได้เล็งเห็นประโยชน์ถึงการจัดการแข่งขันเพื่อหาอัลกอริธึมที่ดีที่สุดสำหรับการแบ่งคำภาษาไทย คูวส์ (CUWS) เป็นโปรแกรมแบ่งคำไทยเอนกประสงค์ที่รับข้อความเข้ามาเป็นข้อมูลเข้า และส่งข้อความที่ถูกแบ่งคำแล้วเป็นข้อมูลออก โดยใช้ความรู้ทางด้านภาษาศาสตร์และความรู้จากการเรียนรู้ด้วยเครื่องมาช่วยในการประมวลผล เพื่อให้สะท้อนถึงประสิทธิภาพของคูวส์ คูวส์ได้ถูกวัดโดยใช้วิธีการแบ่งส่วน k (k-fold cross validation)

Abstract

Thai word segmentation is the main function in Natural Language Processing (NLP). It can be applied in various domains, especially in publishing business and other NLP tasks, e.g., machine translation. However, this research field still lack of deep studying and still need additional corpus. Developers see the contribution of this challenge to find the best algorithm for word segmentation. CUWS, the general propose segmentation program, efficiently receives text as an input and returns a set of segmented words as an output. To make CUWS work, the Thai language model and machine learning technique are utilized in processing, and to evaluate our CUWS, k-fold cross validation technique is used in experimental evaluation.

บทนำ

แนวคิด ความสำคัญ และความเป็นมาของโครงการ

เนื่องจากมีการแข่งขันการแบ่งคำไทยจากหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ผู้พัฒนาได้เล็งเห็นประโยชน์ถึงการจัดการแข่งขันเพื่อให้ได้อัลกอริธึมที่ดีที่สุดสำหรับการแบ่งคำภาษาไทย ซึ่งการแบ่งคำภาษาไทยถือว่าเป็นหัวใจหลักและสำคัญมากสำหรับการวิจัยและพัฒนาทางด้านการประมวลผลภาษาธรรมชาติ ทั้งนี้ผู้พัฒนาหวังว่าความรู้และประสบการณ์การวิจัยและการทำงานของผู้พัฒนาจะเป็นส่วนหนึ่งของการแก้ปัญหานี้ไม่มากก็น้อย คูวส์ (CUWS) เป็นโปรแกรมแบ่งคำไทยเอนกประสงค์ที่รับข้อความเข้ามาเป็นข้อมูลเข้า และส่งข้อความที่ถูกแบ่งคำแล้วเป็นข้อมูลออก โดยใช้ความรู้ทางด้านภาษาศาสตร์และความรู้จากการเรียนรู้ด้วยเครื่องมาช่วยในการประมวลผล

สารบัญ

วัตถุประสงค์และเป้าหมาย

- เพื่อเสนออัลกอริธึมที่สามารถแบ่งคำไทยได้ถูกต้องมากที่สุดตามแนวทางการแบ่งคำไทยที่คณะกรรมการได้กำหนดขึ้น โดยใช้เวลาประมวลผลอย่างเหมาะสม

รายละเอียดของการพัฒนา

ทฤษฎีที่เกี่ยวข้อง

เทคนิคการแบ่งคำด้วยวิธี Longest matching, Maximal matching, N-gram model, และ HMM

ทฤษฎีหลักการและเทคนิคหรือเทคโนโลยีที่ใช้

ใช้เทคนิคการแบ่งคำที่มีอยู่แล้ว (Longest matching, Maximal matching, และ N-gram model) และเทคนิคที่ได้วิจัยและพัฒนาขึ้นมาเองสำหรับการแข่งขัน (จะอธิบายในรายงานฉบับสมบูรณ์)

เครื่องมือที่ใช้ในการพัฒนา

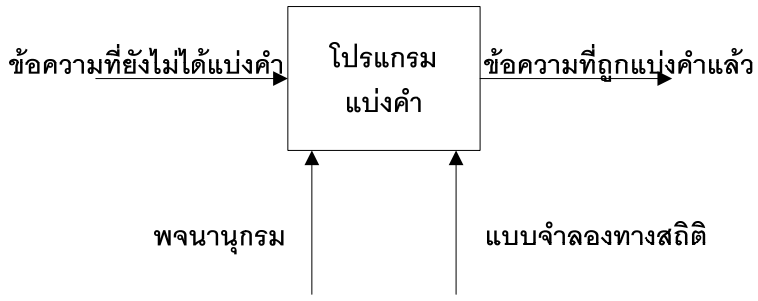
- Java 6 SDK เป็นชุดพัฒนาภาษาโปรแกรมจาวา 6
- Eclipse เป็น IDE สำหรับใช้พัฒนาซอฟท์แวร์ภาษา Java
- Subversion และ Tortoise SVN เป็นระบบ Version Control สำหรับการพัฒนาซอฟท์แวร์

รายละเอียดโปรแกรมที่ได้พัฒนาในเชิงเทคนิค (Software Specification)

- Input/Output Specification ข้อมูลเข้าคือข้อความที่ยังไม่ได้ถูกการแบ่งคำ ข้อมูลออกคือข้อความที่แบ่งคำเรียบร้อยแล้ว

- Functional Specification แบ่งคำภาษาไทยได้ด้วยความเร็วไม่น้อยกว่า 700 คำต่อวินาที

- **โครงสร้างของซอร์ฟแวร์** โปรแกรมแบ่งคำจะรับข้อความที่ยังไม่ได้แบ่งคำเข้ามาเป็นข้อมูลเข้า ระหว่างการประมวลผลโปรแกรมแบ่งคำจะอาศัยพจนานุกรมและแบบจำลองทางสถิติ และจะส่งข้อความที่ถูกแบ่งคำแล้วเป็นข้อมูลออก



ขอบเขตและข้อจำกัดของโปรแกรมที่พัฒนา

ไม่มี

กลุ่มผู้ใช้โปรแกรม

- นักพัฒนาโปรแกรมและนักวิจัยที่ต้องการใช้การแบ่งคำไทยเป็นส่วนหนึ่งของงาน

ผลของการทดสอบโปรแกรม

คูวส์ได้ทดสอบประสิทธิภาพด้วยวิธี k-fold cross validation ซึ่งในเบื้องต้นได้ทดสอบโดยแบ่งข้อมูลออกเป็น 2 ส่วน หรือ k = 2 และได้วัดค่า F-Measure เฉลี่ยเกินกว่า 99% คาดว่าในอนาคตประสิทธิภาพจะดีมากกว่านี้

หมายเหตุ ประเภทลิขสิทธิ์ของโปรแกรมนั้น ผู้เข้าแข่งขันเลือกที่จะสงวนสิทธิ์ทั้งหมดในซอฟต์แวร์ โดยจะส่งรายงานที่เปิดเผยถึงเทคนิคและอัลกอริธึมที่ใช้เป็นการทดแทนหลังจากการแข่งขันสิ้นสุดลง โดยจะไม่มีการส่งมอบโปรแกรมหรือซอร์สโค้ดให้กับทางเนคเทค

ปัญหาและอุปสรรค

- ข้อมูลฝึกสอนมีข้อผิดพลาดและไม่เพียงพอ

แนวทางในการพัฒนาและประยุกต์ใช้ร่วมกับงานอื่น ๆ ในขั้นต่อไป

- โปรแกรมนี้สามารถใช้ร่วมกับการพัฒนาโปรแกรมประยุกต์อื่นๆ เช่น โปรแกรม Word processing หรือการทำงาน NLP อื่นๆ

ข้อสรุปและข้อเสนอแนะ

- โปรแกรมนี้ได้พัฒนาเพื่อให้ผู้ใช้สามารถนำข้อความที่ยังไม่ได้ถูกแบ่งคำ นำมาผ่านโปรแกรมและสามารถนำคำที่ถูกแบ่งไปนั้นไปใช้ประโยชน์ต่อได้

เอกสารอ้างอิง (Reference)

• Aroonmanakun Wirote 2002. ภาษาศาสตร์คลังข้อมูล (Corpus Linguistics). [in Thai] Faculty of Arts, Chulalongkorn University, Bangkok.

• Aroonmanakun Wirote 2007. Thoughts on Word and Sentence Segmentation in Thai. Proceedings of the Seventh International Symposium on Natural Language Processing, 13th-15th September 2007, Pattaya, Thailand, pp. 85-90.

• Charnyapornpong Surin 1983. A Thai Syllable Separation Algorithm. Master Thesis of Engineering. Asian Institute of Technology, Pathumthani.

• Charoenpornsawat Paisarn 1998. การตัดคำภาษาไทยโดยใช้คุณลักษณะ (Feature-based Thai Word Segmentation). [in Thai] Master Thesis of Engineering. Department of Computer Engineering, Chulalongkorn University, Bangkok.

• Charoenpornsawat Paisarn, Sornlertlamvanich Virach 2001. Automatic Sentence Break Disambiguation for Thai. Proceeding of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL2001), May 2001. Seoul, pp. 231-235.

• Kooptiwoot Chompunuch 1999. การตัดคำกำกวมในข้อความภาษาไทยด้วยการโปรแกรมตรรกะเชิงอุปนัย (Segmentation of Ambiguous Thai Words by Inductive Logic rogramming). [in Thai] Master Thesis of Science. Department of Computer Engineering, Chulalongkorn University, Bangkok.

• Longchupole Sungkornsarun 1995. Thai Syntactical Analysis system by method of splitting sentences from paragraph for machine translation. [in Thai] Master Thesis of Engineering. King Mongkut's Institute of Technology Ladkrabang, Bangkok.

• Meknavin Surapant, Charoenpornsawat Paisarn, Kijsirikul Boonserm 1997. Feature-based Thai Word Segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97), 2nd-4th December 1997. Phuket, pp. 41-46.

• Poovarawan Yuen, Imarrom Wiwat 1986. การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี (Thai Syllable Separator by Dictionary). [in Thai] Proceedings of the 9th Annual Meeting on Electrical Engineering of the Thai Universities, 3rd-4th December 1986. Khonkaen.

• Promchan Pisit, Teng-Amnuay Yunyong 1998. Performance Comparison of Thai Word Separation Algorithms. Proceedings of the National Computer Science and Engineering Conference 1998 (NCSEC'98), 19th-21st October 1998. Bangkok.

• Raruenrom Samphan 1991. การแบ่งคำไทยด้วยพจนานุกรม (Dictionary-based Thai Word Separation). [in Thai] Senior Project Report. Department of Computer Engineering, Chulalongkorn University, Bangkok.

• Sawamipak Duangkaew 1990. การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์ (Construction of Thai Syntax Analysing Software under UNIX). [in Thai] Thammasart University Press, Bangkok.

• Sornlertlamvanich Virach 1993. การตัดคำไทยในระบบแปลภาษา (Word Segmentation for Thai in Machine Translation System). [in Thai] Machine Translation, pp. 50-56. NECTEC, Bangkok.

• Sornlertlamvanich Virach, Potipiti Tanapong, Charoenporn Thatsanee 2000. Automatic Corpus-Based Thai Word Extraction with the C4.5 Learning Algorithm. Proceedings of the 18th International Conference on Computational Linguistics, Vol. 2, Jul 2000, pp. 802-807.

• Sproat Richard, Shih Chilin, Gale William, Chang Nancy 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. Computational Linguistics, Vol. 22, No. 3, pp. 377-404.

• Thairatananond Yupin 1981. Towards the design of a Thai text syllable analyzer. Master Thesis of Science. Asian Institute of Technology, Pathumthani.

• Theeramunkong Thanaruk, Okumura Manabu 1996. Towards Automatic Grammar Acquisition from a Bracketed Corpus. Proceedings of the 4th International Workshop on Very Large Corpora (WVLC-4), Denmark, pp. 168-177.

• Theeramunkong Thanaruk, Tanhermhong Thanasan, Phatharakittikul Duangrumol, Sangvareethip Arunthep 2002. Non-Dictionary-Based Word Segmentation Using Local Context Statistics. Proceedings of the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop, May 2002, Hua Hin, Thailand, pp. 81-88.

• Theeramunkong Thanaruk, Usanavasin Sasiporn 2001. Non-Dictionary-Based Thai Word Segmentation Using Decision Trees. Proceedings of the First International Conference on Human Language Technology Research, 18th-21st Mar 2001, San Diego, California, pp. 251-256.

• Theeramunkong Thanaruk, Usanavasin Sasiporn, Machomsomboon Tanin, Opasanont Borisuth 2000. Thai Word Segmentation Without A Dictionary by Using Decision Trees.

Proceedings of the Fourth Symposium on Natural Language Processing, May 2000, Chiang Mai, Thailand, pp. 165-175.

• Varakulsiripunth Ruttikorn, Ngamwiwit Jongkol, Junwun Somsak, Chiwattayakul Suthatip, Thipchaksurat Sakchai 1989. การตัดคำจากประโยคในภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด (Word Segmentation in Thai Sentence by Longest Word Mapping). [in Thai], pp. 279-290. In: Sornlertlamvanich Virach (ed.) 1995. Papers on Natural Language Processing : Multilingual Machine Translation and Related Topics (1987-1994). NECTEC, Bangkok.