Thai Word Segmentation Web Service

Seksan Poltree (seksan.poltree@gmail.com)
Asst. Prof. Kanda Saikaew (krunapon@kku.ac.th)
Department of Computer Engineeering
Faculty of Engineering
Khon Kaen Univerity

Agenda

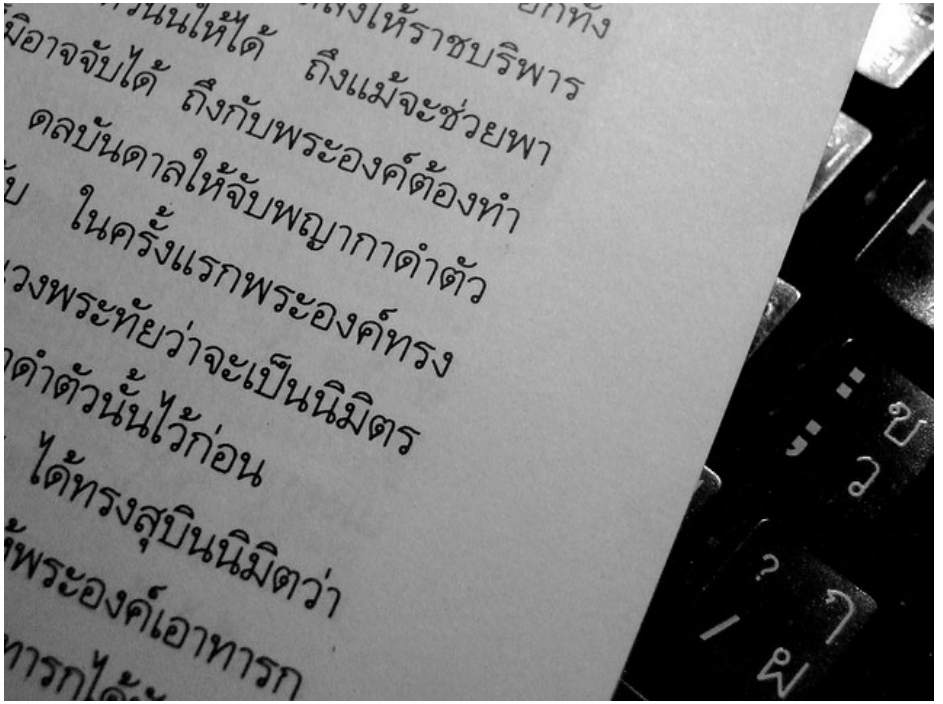
- Thai vs English text processing
- Current Thai Software and Service
- Why segmentation web service
- System Overview
- Web Application Example
- Provided Service Methods
- Comparing Service vs TLEX
- Conclusion and Future work

Current Thai Software and Service

Resource	description	Licensing
libthai	Segmentation software + word list corpus Maximal Matching	GNU LGPL
SWATH	Segmentation software + word list corpus Maximal matching/ longest matching	GNU GPL
ORCHID	Thai Part-Of-Speech tagged corpus	NECTEC (BSD-like)
BEST	Thai segmentation solution corpus	NECTEC (BSD-like)
TLeX Service	SOAP Web service Conditional Random Field technique	Free to use

Thai vs English in Text Processing

- Extract Thai Words?
  - no boundaries
  - no delimiters
- Word Segmentation is a classical issue
- Need word and sentences segmentation



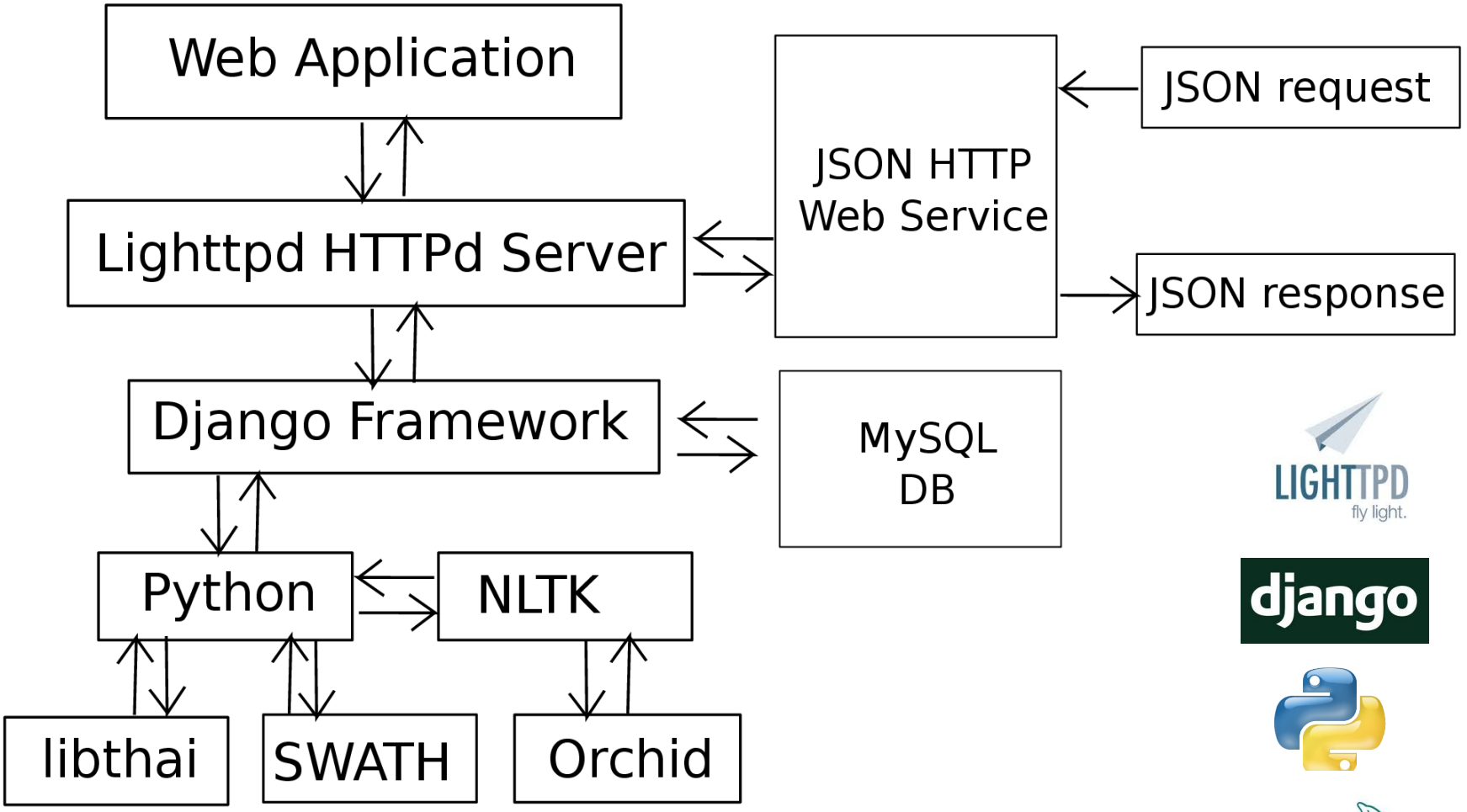
http://www.flickr.com/photos/geoff_b/5332735639/sizes/z/in/photostream/

Why Segmentation Web Service

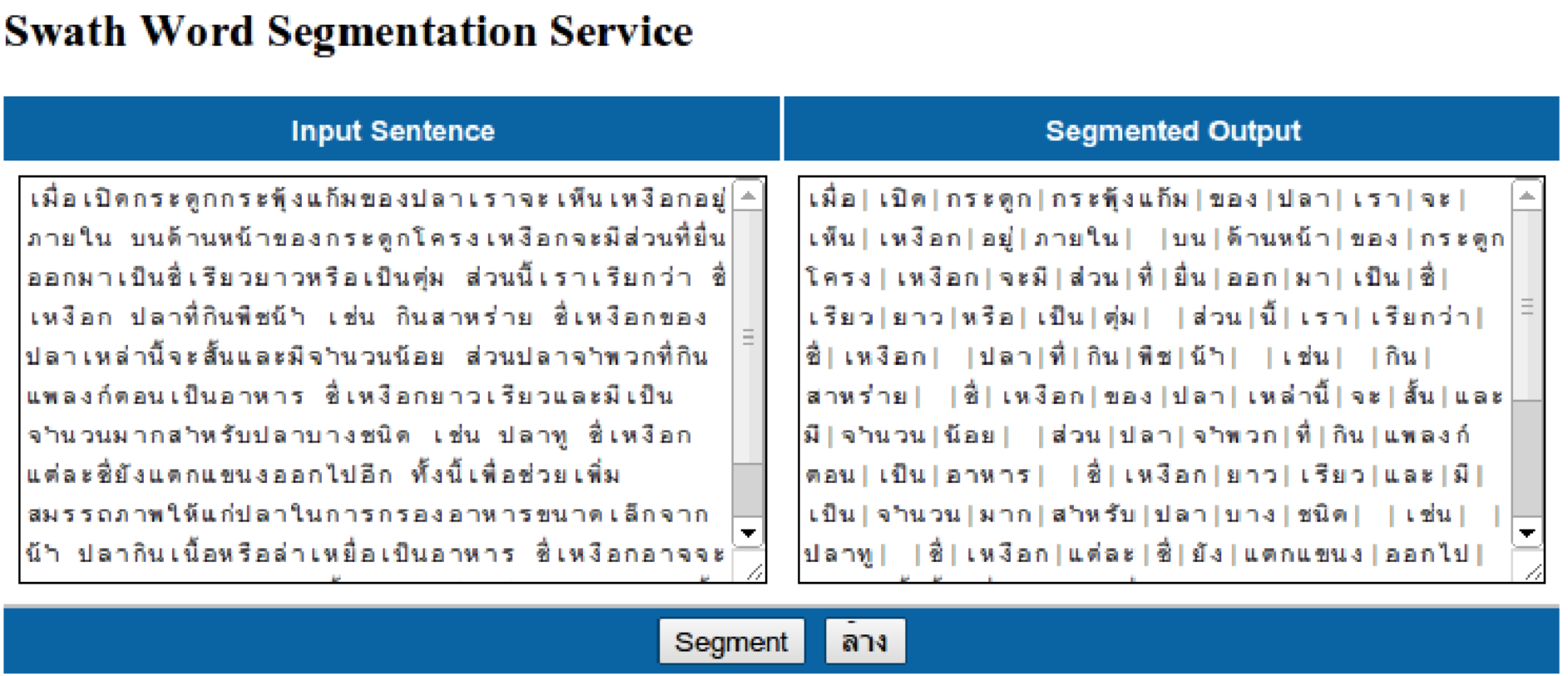
- Increasing of web application and services
- Reducing user learning time of segmentation algorithms
- Make use of existing Thai language resources

http://www.flickr.com/photos/pipeapple/3280609082/

System Overview

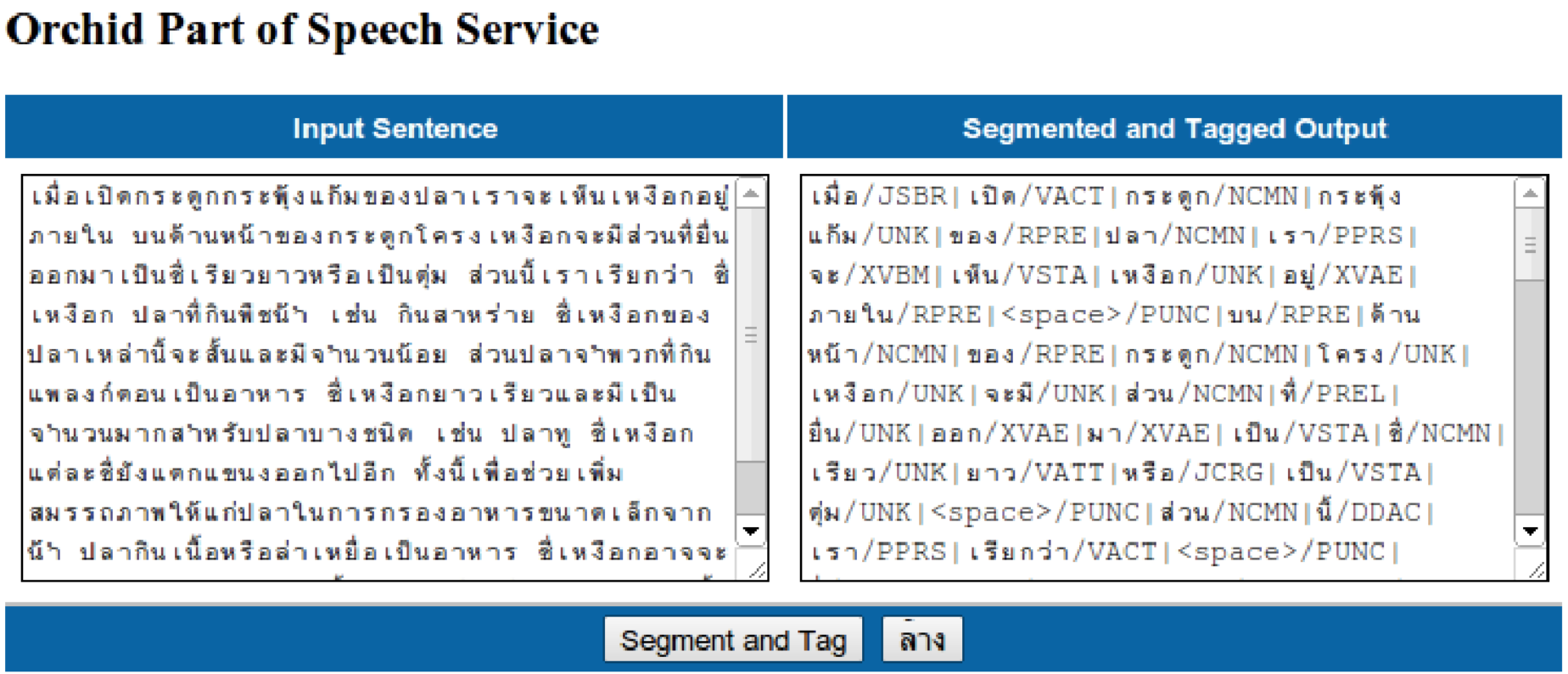


Web Application : SWATH



http://www.thaisemantics.org/service/swath/index

Web Application : ORCHID



http://www.thaisemantics.org/service/orchid/index

Current Provided Service Methods

	Request Format	Response Format
SWATH	api_key': 'YOUR API KEY', 'method': 'ORCHID', 'params': [['list','PoS'],['OF','PoS'],['list','PoS'], ['list','PoS']], }	{"status": 0, "result": ['list','of', 'segmented', 'words'], }
ORCHID	{'api_key': 'YOUR API KEY', 'method': 'ORCHID', 'params': [['list','PoS'],['OF','PoS'],['list','PoS'], ['list','PoS']], }	{"status": 0, "result": [list of tagged', 'words'], }
Wrong KEY	{ 'api_key': '', 'method': 'ORCHID', 'params': ['unicode strings'], }	{"status": 1, "result": ["Wrong API key."]}
Wrong JSON	{unknown or malform json format}	{"status": -1, "result": ["Unkown request"]}

Register to get Free API Key

- Using Facebook account instead of legacy registration

- Re-generated your API Key on demand

Why REST, not SOAP Service?

- **REST** : REpresentational State Transfer
  - Simple, Lightweight
  - But Lack of Standard
- **SOAP** : Simple Object Access Protocol
  - XML based, Schema, Standard
  - Need more bandwidth, Higher round trip time Latency
- No complex schema description need for segmentation
- REST is more suitable!

http://www.flickr.com/photos/tranchis/3378324051/sizes/z/in/photostream/

Why JSON not XML

- **XML** : eXtensible Markup Language
  - Self Descriptive language
  - Mark up overhead
- **JSON** : JavaScript Object Notation
  - Use simple brackets and notations
  - Suitable for simple transfer data
- No complex schema description need for segmentation , JSON is more suitable!

Comparing Service with TLeX

Original BEST data
เนื่อง|จาก|ทั้ง|ประเด็น|เรื่อง|"|ความ|ไม่|เป็นธรรม|"|และ|เรื่อง|"
คู่|ต่าง|เป็น|ประเด็น|ที่|ใหญ่|และ|ซับซ้อน|จึง|ส่ง|ผล|ให้การ|สะ
แต่|เพียง|บางส่วน|เท่า|นั้น| |อีก|ทั้ง|ประเด็น|ดัง|กล่าว|มี|ลักษณ
นั้น| |ข้อ|เสนอ|ต่าง ๆ|ที่|เสนอ|ใน|บท|นำ|ดัง|กล่าว|นี้| |จึง|เป็น|ก
ที่|ต้องการ|ให้|ท่าน|ผู้|อ่าน|ได้|นำ|ไป|ขบคิด|ถกเถียง|กัน|ต่อ|ไป

SWATH Output
เนื่องจาก|ทั้ง|ประเด็น|เรื่อง|"|ความ|ไม่|เป็นธรรม|"|และ|เรื่อง|"
คู่|ต่าง|เป็น|ประเด็น|ที่|ใหญ่|และ|ซับซ้อน|จึง|ส่งผล|ให้การ|สะ
ได้แต่|เพียง|บางส่วน|เท่านั้น| |อีกทั้ง|ประเด็น|ดังกล่าว|มี|ลักษ
ดังนั้น| |ข้อเสนอ|ต่าง ๆ|ที่|เสนอ|ใน|บทนำ|ดังกล่าวนี้| |จึง|เป็น
ที่|ต้องการ|ให้ท่า|น|ผู้อ่าน|ได้|นำ|ไป|ขบคิด|ถกเถียง|กัน|ต่อไป

TLexs Output
เนื่อง|จาก|ทั้ง|ประเด็น|เรื่อง|"|ความ|ไม่|เป็นธรรม|"|และ|เรื่อง
คู่|ต่าง|เป็น|ประเด็น|ที่|ใหญ่|และ|ซับซ้อน|จึง|ส่ง|ผล|ให้|การ|ส
ได้|แต่|เพียง|บาง|ส่วน|เท่า|นั้น| |อีก|ทั้ง|ประเด็น|ดัง|กล่าว|มี|ลั
ดัง|นั้น| |ข้อ|เสนอ|ต่าง ๆ|ที่|เสนอ|ใน|บท|นำ|ดัง|กล่าว|นี้| |จึง|เ
เจตนา|ที่|ต้องการ|ให้|ท่าน|ผู้|อ่าน|ได้|นำ|ไป|ขบคิด|ถกเถียง|กั

- Using BEST corpora as test data
- Create simple script and call each service
- TLEX and SWATH use difference method and implementation
- Just prove of concept

Evaluation Result

AVERAGE SERVICE CALLING USAGE TIME

Calling Service	real	user	sys
SWATH	1m39.293s	0m5.172s	0m1.176s
ORCHID	4m50.377s	0m11.077s	0m1.212s
TLexs	3m17.045s	0m6.632s	0m1.048s

EVALUATION RESULTS

Calling Service	Precision	Recall	F-Score
SWATH	85.93	77.62	81.57
TLexs	95.65	97.45	96.54

Conclusion and Future work

- Create Segmentation and POS-Tagger application and services
- Create Free JSON REST Web Service
- **http://www.thaisemantics.org**
- Comparing with existing TLeX SOAP web service to prove of concept
- Include more method and corpus in the future
- Using facebook account instead of registration

http://www.flickr.com/photos/nofrills/10895361/

References

[1] T. Karoonboonyanan, C. Silpa-Anan, P. Kiatisevi, P.Veerathanabutr and V. Ampornaramveth, "libthai Library". Available at: http://linux.thai.net/projects/libthai.

[2] P. Charoenpornsawat, "SWATH (Smart Word Analysis for THai)". Available at: http://www.cs.cmu.edu/~paisarn/software.html.

[3] T. Karoonboonyanan, "swath 0.4.1 Released". Available at: http://linux.thai.net/svn/software/swath.

[4] The Royal Institute of Thailand 2525, "Thai dictionary words from the royal institute of Thailand 2525". Available at: http://thailang.nectec.or.th/.

[5] V. Sornlertlamvanich, T.Charoenporn and H.Isahara, "ORCHID: Thai Part-Of-Speech Tagged Corpus". National Electronics and Computer Technology Center. Technical Report: TR-NECTEC-1997-001, 1997.

[6] National Electronics and Computer Technology Center (NECTEC), "BEST Corpus". Available at: http://thailang.nectec.or.th/best/.

[7] C. Haruechaiyasak and S. Kongyoung, "TLex: Thai Lexeme Analyser Based on the Conditional Random Fields", *Proc. 8th International Symposium on Natural Language Processing*, 2009.

[8] National Electronics and Computer Technology Center (NECTEC), "TLex". Available at: http://sansarn.com/tlex/.

[9] National Electronics and Computer Technology Center (NECTEC), "TLexs". Available at: http://www.sansarn.com/WSeg/wsdl/BnSeg.wsdl.

[10] K. Toutanova and C. D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger". *Proc. the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70, 2000.

[11] D. Roth and D. Zelenko, "Part of Speech Tagging Using a Network of Linear Separators", *The 17th International Conference on Computational Linguistics (1998)*, pp. 1136–1142, 1998.

[12] The World Wide Web Consortium. "Extensible Markup Language (XML) 1.0 (Fifth Edition)". Available at : http://www.w3.org/TR/REC-xml/.

[13] The World Wide Web Consortium. "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)". Available at : http://www.w3.org/TR/soap12-part1/.

[14] R. T. Fielding. "Representational State Transfer (REST)". Available at : http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm.

[15] Ecma International, "Introducing JSON". Available at: http://www.json.org/

[16] Ecma International, "Standard ECMA-262 5.1 Edition / June 2011 : ECMAScript Language Specification". Available at: http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf

[17] Wikipedia, "Comparison of web browsers". Available at: http://en.wikipedia.org/wiki/Comparison_of_web_browsers

[18] C. Haruechaiyasak, S. Kongyoung and M. N. Dailey. "A Comparative Study on Thai Word Segmentation Approaches", *Proc. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (2008)*, pp.125–128, 2008.

[19] G. van Rossum, "Python tutorial, Technical Report CS-R9526", *Centrum voor Wiskunde en Informatica (CWI)*, Amsterdam, May, 1995.

[20] S. Bird and E. Loper, "NLTK: The Natural Language Toolkit", *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pp. 62-69, 2002.

[21] Django Software Foundation, "The Django framework". Available at : https://www.djangoproject.com/.

[22] "LigHTTPD fly light". Available at : http://www.lighttpd.net/.

[23] C. D. Manning, P. Raghavan and H. Schutze. "An Introduction to Information Retrieval", *Cambridge University Press, Cambridge, England*, pp.155, 2009.

[24] G. Mulligan and D. Gračanin. "A Comparison of SOAP And REST Implementation of a Service Based Interaction Independence Middleware Framework", *Proceedings of the Winter Simulation Conference*, pp.1423-1432, 2009.

[25] G. Wang. "Improving Data Transmission in Web Applications via the Translation between XML and JSON", *Third International Conference on Communications and Mobile Computing*, pp.182-185, 2011.

http://www.flickr.com/photos/oberazzi/318947873/sizes/l/in/photostream/