Survey on Thai NLP Language Resources and Tools

Data Science | Lancaster University

Ratchakrit Arreerard Stephen Mander Scott Piao

Research Purposes

- To survey on the past development of Thai NLP resources and tools, including lexicon resources, corpora, and Thai NLP tools.
- To understand current state and future research direction of Thai NLP.

Special Features of Thai Language

- Thai language is one of the under-resourced languages in the NLP domain, although it is spoken by nearly 70 million people globally.
- Thai language has unique writing system which lacks explicit delimiter of words.
- Thai syntactic grammar allows all Subject-Verb-Object, Subject-Object-Verb and Object-Subject-Verb structures.

Sample Thai sentences

[Subject Verb Object]

Teacher hit me.

ครูตีฉัน

S = ครู Teacher; V = ตี hit; O = ฉัน me

[Object Subject Verb]

I am hit by teacher.

ฉันถูกครูตี

O = ฉัน I; S = ครู teacher; V = ถูก..ตี am hit ; *ถูก is auxiliary verb

[Subject Object Verb]

Woman and her bicycle that being ridden.

หญิงสาวกับจักรยานที่ถูกปั่น

S = หญิงสาว woman; O = จักรยาน bicycle; V= ถูก..ปั่น being ridden

Lexical Resources

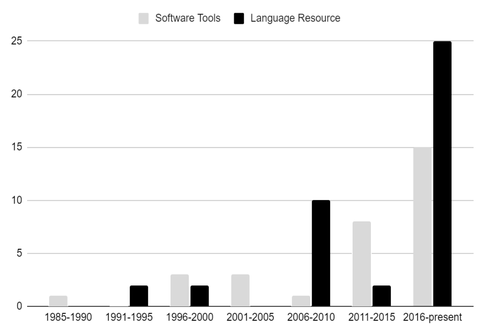
Table below lists existing Thai lexical resources available publicly including Thai lexicons, Thai-English dictionaries, Thai WordNets, semantic lexicons, and Thai word embedding.

Thai Lexicon		
Lexitron	42,221 words	
Thai-lm	Thai words (over 40,000) abbreviations (263) Thai name entities (6,061) Thai swear words (95) English-Thai transliteration (approx. 547) Thai words variants (approx. 286) misspelled Thai words (approx. 1,032)	
Thai-English bilingual dictionaries		
Lexitron 2.0	53,000 Thai/English word pairs 83,000 English/Thai word pairs	
Yaitron	A dictionary developed based on LEXiTRON in XML format	
WordNet		
Leenoi	493 concepts	
Akaraputthiporn	491 concepts	
Thoongsup et al.	82,504 Thai words in 73,350 synsets	
Semantic Lexicon		
Phatthiyaphaibun	633 words with positive and negative label	
NRC Emotion Lexicon	14,100 English words	Eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiment orientations (negative and positive)
Polyglot	1,279 Thai sentiment words	
Word embedding		
FastText	A large-scale pre-trained set of word embeddings using Continuous Bag of Words (CBOW) with position-weights	
Thai BERT	Built from scratch for the Thai language	
Thai2fit	Thai Universal Language Model Fine-tuned (ULMFiT) with 60,005 embeddings, trained on Thai Wikipedia Dump	

Thai NLP development Trend

The bar chart illustrates chronological view of publications on Thai NLP tools and language resources. It shows:

- Thai NLP research had experienced a slow development until 2006, since when the number of published Thai language resources has increased steadily.
- Since 2016 the number of Thai NLP publications, particularly on Thai language resources, has increased drastically.



Corpus Resources

Table below lists Thai corpus resources, including benchmark, treebank, name entity, sentiment, translation, and text classification and analysis corpora.

Corpus	Developer	Data Size
BEST	NECTEC	5 million words
HSE Thai	HSE School of Linguistics	50 million tokens
LST20	NECTEC	3 million words, 288,020 Named Entities
Nattadaporn	Nattadaporn Lertcheva	178,474 words with 2,463 Named Entities
Nutcha	Nutcha Tirasaroj	367,673 words with 16,179 Named Entities
Sasiwimon	Sasiwimon Kalunsima	80,513 words with 2,954 Named Entities
VISTEC-2021	VISTEC and CMU	3.39M words, 49,997 sentences
Wongnai-corpus	Thongthanomkul et al.	500,000 unique words, 39,999 reviews
Thai-Nest	NECTEC	45,000+ Named Entities
ClickBait	Wannaphong Phatthiyaphaibun	350 sentences
Mt-opus	VISTEC	5.4 million sentence pairs
Orchid	NECTEC	30,000 sentences
Prachathai-67k	Phatthiyaphaibun et al.	67,000 sentences
TALPCo	Nomoto et al.	1,372 sentences
Thai Universal Dependency	UD Thai PUD	1,000 sentences
Thai WIKI QA	NECTEC	17,000 sentences
Thai Literature Corpora	Jitkapat Sawatphol	756,478 lines, 47 documents
Toxic tweet	Sirihattasak et al.	3,300 tweets
Wisesight	Suriyawongkul et al.	26,737 messages
Thai QA	NECTEC	4,000 questions
Thai-joke-corpus	iApp Technology	449 jokes
Prime Minister 29	Phatthiyaphaibun et al.	6 documents, 338KB
Thai Plagiarism	NECTEC	1,050 plagiarism texts, 554MB Source docs
Scb-mt-en-th-2020	VISTEC and SCB	1 million Thai-English texts
Blackboard Treebank	NECTEC	130,561 Trees
Wikipedia dumps	Wikipedia	2.08GB

NLP Software Tools

Table below lists publicly available Thai NLP tools and their reported accuracies, including basic Thai NLP, syntactic, and semantic tools.

Basic Thai NLP tools			
Tool	**Developer**	**Functionality**	**Accuracy**
TCC	Theeramunkong et al.	Tokenisation	Unreported
ETCC	Jeeragone et al.	Tokenisation	Unreported
JTCC	Wittawat Jitkrittum	Tokenisation	Unreported
AIforThai	NECTEC	Tokenisation (Lexto+ 96.30%)	Unreported
AttaCut	Chormai et al.	Tokenisation	89.00-91.00%
CutKum	Pucktada Treeratpituk	Tokenisation	95.00%
Deepcut	Kittinaradorn et al.	Tokenisation	98.10%
Multi-Candidate	Lapjaturapit et al.	Tokenisation	97.00%
OpenNLP	Apache Software Foundation	Tokenisation, POS tagging	Unreported
OSKut	Limkonchotiwat et al.	Tokenisation	96.18-97.03%
PyThaiNLP	Phatthiyaphaibun et al.	Tokenisation, Spell checker, POS-tagging, NE recognition	Unreported
SERF Cut	Limkonchotiwat et al.	Tokenisation	83.90-92.50%
SWATH	Meknavin et al.	Tokenisation	53.52-99.77%
Spacy-thai	Koichi Yasuoka	Tokenisation, POS-tagging, dependency-parser	Unreported
SynThai	Wutthiphat Phuriphatwatthana	Tokenisation(97.59%), POS-tagging(91.85%)	Unreported
ThaiLMCut	Seeha et al.	Tokenisation	98.78%
TLex/TLex+	NECTEC	Tokenisation	93.90-97.50%
TLTK	Wirote Aroonmanakun	Tokenisation (96.76-97.97%), POS-tagging (91.68%), NE Detection (88%)	Unreported
WordCut	Satayamas & Pakkapon	Tokenisation	Unreported
RDRPOSTagger	Nguyen et al.	POS-tagging	94.15-94.21%
Syntactic and Semantic Analysis Tools			
CF Parser	Seenual et al.	Parser	73.11-83.89%
GrammarAnalyser	Thodsaporn Chay-intr	Grammar extraction	Unreported
Lalita	AIBuilders	Chinese/Thai machine translation	15.53, 8.42 Bleu
Machinetranslator	VISTEC-depa	English/Thai Machine Translation	29, 17.77 Bleu
Polyglot-NER	Al-Rfou et al.	NE recognition	Unreported
SentimentAnalyser	Raymond	Sentiment analysis	Unreported

Contact: {r.arreerard, s.mander3, s.piao}@lancaster.ac.uk