Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6495–6505
Marseille, 20-25 June 2022

Survey on Thai NLP Language Resources and Tools

Ratchakrit Arreerard, Stephen Mander, Scott Piao
School of Computing and Communications
Lancaster University
Lancaster LA1 4WA
{r.arreerard,s.mander3,s.piao}@lancaster.ac.uk

Abstract

Over the past decades, Natural Language Processing (NLP) research has been expanding to cover more languages. Recently particularly, NLP community has paid increasing attention to under-resourced languages. However, there are still many languages for which NLP research is limited in terms of both language resources and software tools. Thai language is one of the under-resourced languages in the NLP domain, although it is spoken by nearly 70 million people globally. In this paper, we report on our survey on the past development of Thai NLP research to help understand its current state and future research directions. Our survey shows that, although Thai NLP community has achieved a significant achievement over the past three decades, particularly on NLP upstream tasks such as tokenisation, research on downstream tasks such as syntactic parsing and semantic analysis is still limited. But we foresee that Thai NLP research will advance rapidly as richer Thai language resources and more robust NLP techniques become available.
Keywords: Natural Language Processing, Thai NLP, Survey, Language Resource, NLP tools

1. Introduction

Over the past decades, Natural Language Processing (NLP) has advanced rapidly, having become a major research area with a wide range of applications in AI (Artificial Intelligence) and ICT (Information Communication Technology) systems. Today, NLP language resource and tools such as WordNet (Fellbaum, 1998), CoreNLP (Manning et al., 2014) and NLTK (Loper and Bird, 2002) cover numerous languages, and keep expanding to include more languages. However, there are still many languages for which more language resources and NLP tools need to be developed. In particular, unique features of some languages demand specifically tailored NLP techniques and approaches. Thai language is one of the under-resourced languages in the NLP domain, although it is spoken by approximately 70 million people globally[1].

In fact, Thai NLP research has a long history, started about three decades ago (Sornlertlamvanich, 2019). The earliest published Thai NLP work we could trace back includes an automatic Thai syllabus analysis of Poowarawan (1986) and Multi-Lingual Machine Translation (MMT) Project (Funaki, 1993). Since then, numerous NLP language resources and tools have been developed and published for Thai language, such as TLTK (Aroonmanakun, 2002) and AIforThai platform[2].

Despite recent significant development of NLP research in Thai NLP community, however, compared to some other major languages such as English, there is still much room for further development. In particular, Thai language has some unique features, for instance Thai writing system lacks explicit delimiter of words (similar with Chinese and Japanese languages), and Thai syntactic grammar allows all Subject-Verb-Object, Subject-Object-Verb and Object-Subject-Verb structures, as shown by Figure 1 of Thai example 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

Figure 1: example Thai sentences

Such unique features of Thai language require NLP techniques and tools to be tailored to provide an efficient performance.

In this paper, we survey on publicly available NLP language resources and tools developed for Thai language processing and examine the remaining gaps in this research area. As will be shown by our survey, although the amount and scale of Thai NLP resources and tools have been increasing, the development is not well balanced across main aspects of NLP tasks, such as automatic sub-word, morpho-syntactic and semantic analysis. For example, remarkable efforts of Thai NLP community have been focused on morpho-syntactic tools development, such as tokenisers and Part-of-Speech (POS) taggers, but development of semantic tools is still limited.

[1] https://en.wikipedia.org/wiki/Thai$_{l}anguage$

[2] See https://aiforthai.in.th/

The remainder of this paper is organised as follows. Section 2 surveys the development of Thai language resources, Section 3 discusses publicly available Thai NLP tools, Section 4 discusses the current situation of Thai NLP research, and Section 5 concludes our work.

2. Thai Language Resources

2.1. Lexical resources

In NLP context, a lexical resource is a language resource that contains various information about lexemes of one or more languages, such as general dictionaries and more specifically structured lexical database like WordNet and USAS Lexicon (Piao et al., 2016). In addition to providing word sense definitions, such lexical resources provide various information with which lexemes are inter-connected and grouped in a network. The lexical resources provide critical lexical knowledge base for many NLP systems. In this section, we survey Thai lexical resources that have been developed by the NLP community so far.

A major existing Thai lexical resource is Lexitron[3] which contains 42,221 Thai words. Based on it, Lexitron 2.0[4] and Yaitron (Satayamas, 2019b) were developed which are Thai-English bilingual dictionaries[5]. The Lexitron 2.0 contains 53,000 Thai/English word pairs and 83,000 English/Thai word pairs. Of them, the Yaitron was developed based on Lexitron 2.0 in XML format. These resources also provide information of synonyms, part of speech, and example sentence.

Another Thai lexicon is available at GitHub cite[6]. It contains various lexicon types, such as 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), and misspelled Thai words from Wikipedia (approx. 1,032).

Other similar resources include two Thai versions of the WordNet in which English words are grouped into synonyms, dubbed synsets. These Thai WordNets were built in 2008[7] by Leenoi (2008) and Akaraputthiporn (2008), which employ 493 and 491 concepts respectively. Later in 2009, another version of Thai WordNet was created by Thoongsup et al. (2009)[8] containing 82,504 Thai words in 73,350 synsets.

There are also some Thai semantic lexical resources. One of them was compiled by Phatthiyaphaibun (2017c) and contains 633 verbs and adjectives that are labelled as negative and positive. Another semantic lexicon is the multilingual NRC Emotion Lexicon (Mohammad and Turney, 2010; Mohammad and Turney, 2013) which includes Thai language. This is a large lexical resource containing over 14,100 English words and their translation equivalents in some languages. Many of the multilingual words are associated with one of eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiment orientations (negative and positive) through their English equivalent words. The annotations were manually carried out through Amazon's Mechanical Turk. In this dataset, the Thai words were translated from English words using google translate in November 2017, and approximately 10,000 Thai words are contained in this dataset. Polyglot[9] is another Thai sentiment lexical resource (Chen and Skiena, 2014) that contains 1,279 Thai sentiment words with a 0.51 ratio of positive words.

Word embedding is an approach to represent a word as a vector based on contextual information, generally context words. The vector contains real-number values obtained by using language model and feature learning techniques and indicates the word's semantic meaning. Some Thai word embedding models have been developed, too. For example, a Thai word embedding model is available in a large-scale pre-trained set of word embeddings of FastText (Bojanowski et al., 2017), which provides pre-trained word vectors for 157 languages. FastText itself is a library for efficient learning of word representations that train models using Continuous Bag of Words (CBOW) with position-weights.

BERT (Devlin et al., 2019) is another major pre-trained unsupervised natural language processing model, prepared for fine-tuning to perform NLP downstream tasks significantly. Chay-intr (2020) introduced a Thai BERT that was built from scratch for the Thai language. Thai2fit (Polpanumas and Phatthiyaphaibun, 2021) is Thai Universal Language Model Fine-tuned (ULMFiT) with 60,005 embeddings, trained on Thai Wikipedia Dump.

2.2. Corpus Resources

Over the past years, a number of corpora have been constructed for research or commercial purposes. Each of them has unique feature of data and usage, such as the knowledge corpus of Wikipedia website [10] that consists of articles from various topics.

Some Thai corpora were built as benchmark or contribution for improving the tokenization process for Thai NLP, including Orchid[5] (Sornlertlamvanich et al., 1999) and BEST[4]. These corpora were compiled using encyclopedia, news and novels, and were annotation with word boundaries or sentence boundaries. In the Thai Literature Corpora, there are 2 datasets: TNHC and TLC[11]. The TNHC data is from Thai National Historical Corpus while the TLC data is from Vajirayana Digital Library[12]. Of them, the TNHC dataset was

[3] http://www.sansarn.com/lexto/download-lexitron.php

[4] https://aiforthai.in.th/corpus.php

[5] https://www.nectec.or.th/corpus/index.php?league=pm

[6] https://github.com/Knight-H/thai-lm

[7] http://pioneer.chula.ac.th/ãwirote/resources/corpora–data.html

[8] http://compling.hss.ntu.edu.sg/omw/

[9] https://polyglot.readthedocs.io/en/latest/Installation.html

[10] https://dumps.wikimedia.org/thwiki/

[11] https://attapol.github.io/tlc.html

[12] https://vajirayana.org/

manually tokenised.

Some Thai corpora contain richer annotations. For example, Thai-Nest[4] (Theeramunkong et al., 2010), Prime Minister 29 (Phatthiyaphaibun, 2017b), Nattadaporn (Lertcheva, 2010), Nutcha (Tirasaroj, 2010), Sasiwimon (Kalunsima, 2010) and LST20[4] (Boonkwan et al., 2020) are annotated with named entity information. In addition, LST20 is more informative, tagged with sentence boundaries, word boundaries, part of speech and clause boundaries.

The largest available Thai social media corpus is VISTEC-2021[13] (Limkonchotiwat et al., 2021), which contains 49,997 sentences with 3.39M words. This corpus is manually annotated with word segmentation, misspelling correction and named-entity boundaries.

Blackboard Treebank[4] is a Thai syntactic dependency corpus whose annotation follows the LST20 (Language and Semantic Technology Lab) annotation guideline (Boonkwan et al., 2020). It features dependency structures, syntactic constituency structures, word boundaries, named entities, clause boundaries and sentence boundaries. This corpus is available in the CoNLL-U format for universal compatibility. Likewise, Thai Universal Dependency[14] is a part of the Parallel Universal Dependencies (PUD) Treebanks. The sentences were translated to the Thai language from English, and then the data was annotated morphologically and syntactically by a Google team according to Google universal annotation guidelines.

Parallel corpora containing Thai language include TALPCo[15] (Nomoto, 2019) which consists of Japanese sentences and their translations into eight languages, with Thai being one of them. Each sentence was assigned with interpersonal meaning annotation of the speaker, addressee, and lexical. HSE Thai Corpus[16] is a collection of modern texts written in Thai language. Each token was tagged with its English translation and part of speech. Some other grammatical tagging was also assigned where applicable.

English-Thai Machine Translation Dataset (scb-mt-en-th-2020)[17] is a collection of online news, conversations, Wikipedia and official documents. This dataset contains more than a million Thai and English text pairs and the translation was carried out manually or by a sentence alignment algorithm. Mt-opus[18] is an English-Thai machine translation corpus, with OPUS[19] data containing a total 5.4 million sentence pairs, 68.8 million English tokens and 53.1 million Thai tokens. Another similar corpus is the Transliteration Corpus[5] which contains 31,801 Thai words transliterated from English.

In addition, there are corpora for text analysis and classification as follows;

- Thai Plagiarism[4] contains source documents from Wikipedia and public websites with 1,050 plagiarism texts in 4 different aspects (i.e., copy-based change, lexicon-based change, structure-based change and semantic-based change).
- Thai QA and Thai WIKI QA[4] are question-answer (QA) pairs from various Wikipedia topics. Thai WIKI QA questions are categorized as factoid questions and yes/no questions while Thai QA dataset is based on standard question words (e.g. what, when, where).
- Wisesight (Suriyawongkul et al., 2019), a sentiment corpus, contains social media messages labelled as negative, positive, neutral, or question.
- ClickBait (Phatthiyaphaibun, 2017a) holds 350 sentences labelled as clickbait collected from website headlines.
- Toxic tweet data (Sirihattasak et al., 2018) contains 2,027 toxic tweets and 1,273 non-toxic and toxic keywords. The tweets were manually labelled based on a majority decision.
- Prachathai-67k (Phatthiyaphaibun and Polpanumas, 2018), a dataset collected from Prachathai news between August 2004 to November 2018, contains 67,000 articles each of which has at least 500 words.
- Wongnai-corpus (Thongthanomkul et al., 2019) provides customer reviews on food or respective restaurant ratings from 1 to 5 and a sample of search queries from customers with word boundaries segmented.
- Thai-joke-corpus (Viriyayudhakorn, 2019), Thai jokes scraped from four Thai jokes Facebook pages collected by iApp Technology Co, Ltd.

Table 1 is the summary of the corpus resources discussed previously. These corpora were reported using different measures of data size, such as number of words, sentences, documents etc. As the result, their sizes are not directly comparable. To mitigate this issue, in the table, we clustered those corpus resources together which have same or similar measures of data size.

3. NLP Software Tools

In this section we survey Thai NLP software tools that have been developed and published over the past years. The NLP research community has designed and developed a variety of software libraries and tools, such as

[13] https://github.com/mrpeerat/OSKut/tree/main/VISTEC-TP-TH-2021

[14] https://universaldependencies.org/

[15] https://github.com/matbahasa/TALPCo

[16] http://web-corpora.net/ThaiCorpus/search/

[17] https://airesearch.in.th/releases/machine-translation-datasets/

[18] https://github.com/vistec-AI/mt-opus

[19] https://opus.nlpl.eu/

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 (TNHC set)	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

Table 1: Thai corpus resources.

OpenNLP (Hockenmaier et al., 2004; Apache Software Foundation, 2014), CoreNLP, NLTK and Stanza (Qi et al., 2020), which provide models for the Thai NLP tools. For example, OpenNLP has been extended to provide a Thai tokenizer, a Thai part-of-speech tagging and a Thai sentence detector. Stanza is the Stanford NLP Group's official Python NLP library, which also supports tokenisation of Thai language now, trained using Orchid.

3.1. Sub-word Analysis

In NLP, sub-word analysis tools are important for basic level NLP tasks, such as automatically identifying syllables and morphemes. Similarly, for Thai language processing, it is an important task to automatically identify sub-word units.

The Thai language consists of 44 characters, 21 vowels and 4 tone marks. The vowels can be placed in 4 different positions around the character: front, behind, upper, and below. However, the tone marks can only be placed above the character or the upper vowel. According to the Thai spelling grammar, the Thai characters can be grouped into an inseparable unit called Thai Character Clusters (TCCs). Automatic identification of these units is beneficial to various NLP processes and the following tools have been developed to deal with this issue.

JTCC (Jitkrittum, 2017) is a Java library to tokenise Thai text into a list of TCCs. The rules used to determine TCCs' boundaries are implemented as a grammar using ANTLR[20] (ANother Tool for Language Recognition).

TCC (Theeramunkong et al., 2000) and Enchanted Thai Character Clusters (ETCC) (Jeeragone et al., 2001) are built-in functions in PythaiNLP, a Python library. TCC is a rule-based algorithm. Therefore, context-free grammar can represent the rule for segmenting Thai text into TCCs. The ETCC is an improved version of TCC by adding a new set of rules that are capable of detecting larger cluster groups than TCC.

3.2. Word Tokenisation

As mentioned earlier, the Thai language is written without an explicit word boundary. Thus, word segmentation is a crucial step before analyzing the text further. Tools for extracting words from texts have been developed over the past decades. In the early stage, the word segmentation process, SWATH (Meknavin et al., 1997), relied on dictionary-based algorithms such as the longest string matching or maximal string match-

[20]https://www.antlr.org/

ing. Later learning algorithms, RIPPER and Winnow, were used to extract features from training corpus. In 2005, WordCut was introduced and has been under continuous development to the present date. WordCut supports Python (Satayamas, 2019a), Node.JS (Satayamas, 2021) and Coffee-script (Phongthawee, 2018). To segment words, a word graph is created based on the dictionary, where nodes represent the position of a character in the input sentence, and edges represent the existing word formed by characters between the starting node and the ending node. The word segmentation is determined by the shortest path of the word graph.

Moreover, words in the Thai language can be compounded into a new word, creating a different meaning from that of original words. To cope with such ambiguity, Thai Language Toolkit[21] (TLTK) (Aroonmanakun, 2002) implements a word segmentation method using syllable segmentation and syllable merging. There are various forms of syllables. For instance some syllables use more than one character for vowel forms, or some have two initial consonants. Typically, a syllable is composed of vowel forms, initial consonants, and final consonants. In Aroonmanakun (2002) study, about 200 patterns were defined, then a tri-gram model is trained. Finally, the syllables were grouped using the co-occurrence affinity metric of syllables calculated from the training corpus. It was concluded that the algorithm highly depends on the exhaustiveness dictionary-lookup to achieve the best performance. As the result, the unknown words have a significant impact on the performance of the tool.

Here the unknown words can be coined words or those containing intentional spelling errors. These errors were addressed by Haruechaiyasak and Kongthon (2013). They categorized the errors into four categories: insertion, transformation, transliteration and onomatopoeia. The repeated ending characters cause the insertion; for example, the word ”love—eeeeee” has ”eeeeee” as an unknown word. Transformation happens when words are altered, such as ”love” and ”luv”. They proposed LextoPlus[22] which could handle the insertion problems using a statistical model, conditional random fields (CRFs), as a tokeniser and the longest string matching with a dictionary to eliminate the errors. The elimination process will look for repeated character tokens, and remove them if they are not in the dictionary. However, inaccurate word segmentation can be caused by the out-of-vocabulary problem.

Other than the unknown words, the Thai language has long expressions comprising more than two words, whereas commonly compound words comprise only two words. To solve the problem, Kongyoung et al. (2015) proposed TLex+[22], a hybrid method developed based on Tlex (Haruechaiyasak and Kongyoung, 2009). Tlex is a word segmentation method based on CRF. The model predicts two classes for each character: word-beginning and intra-word characters. Tlex gives the best performance when trained with both characters and character types as features. Similar to TLex, TLex+ uses a list of long expression terms to identify the terms before continuing the segmentation process. The result showed that this approach improves the performance of the former method.

In recent years, to overcome the out-of-vocabulary problem, several tools have been developed based on deep learning. For example, CutKum (Treeratpituk, 2017) was developed for Thai Word-Segmentation by employing Recurrent Neural Network (RNN) based on Tensorflow in Python. It was trained on BEST2010 corpus and yielded an F-Measure of 97.1% at character level and 95.0% at word level.

Deepcut (Kittinaradorn et al., 2019), a method developed based on Convolutional Neural Networks (CNN) with character embedding and character type embedding as features. The model was trained on the BEST corpus to predict whether or not a character is the beginning of a word. The test result showed that the model produced 97.8% precision, 98.5% recall, and 98.1% F1.

AttaCut (Chormai et al., 2019) was designed to accelerate the segmentation speed, inspired by Deepcut. AttaCut experimented by disabling neurons in Deepcut's layer, resulting in a six times improvement in speed. The test result on the domain data set revealed 89%-91% F1 , while in the out-domain data set, the performance was 63%-81% F1.

SynThai (Phuriphatwatthana, 2016) and Multi-Candidate-Word-Segmentation (Lapjaturapit et al., 2018) are word segmentation methods based on Bi-directional LSTM (BiLSTM). SynThai is a word segmentation and part of speech tagging tool trained on the BEST2010 corpus. The input data is encoded strings of character (e.g. 'space'=0, 'a' = 1, etc.). If the output of the certain character is more than 2, then it is the last character of the word. In addition, the value of the output determines the part-of-speech tag of the word, such as 2='NN', 3='NR' etc. Multi-Candidate-Word-Segmentation can segment sentences and provide more than one segmentation solution determined by threshold. Each threshold indicates the confidence level of the model. Thus the first threshold only generates one segment symbol. The model uses TCC embedding and character embedding as features and was trained on InterBEST 2009 and 2010 datasets. It achieved F1 score of 97% and 98.95% for word-level and boundary-level respectively.

ThaiLMCut (Seeha et al., 2020) is a semi-supervised approach utilizing a bi-directional character language model (LM). The LM was trained on unlabeled corpora, and then the weights were transferred to a supervised word segmentation model to continue fine-tuning them on a word segmentation task. As a result, ThaiLMCut outperformed other open-source state-of-

[21] https://pypi.org/project/tltk/

[22] https://aiforthai.in.th/service_bn.php

the-art models, achieving an F1 Score of 98.78% on the standard benchmark InterBEST2009.

SEFR CUT[23] (Limkonchotiwat et al., 2020) is a Thai word segmentation model which adopts stacked ensemble. Here the general-decision maker or black-box model is the pre-train model from Deepcut and Attacut. The black box produces an output; then, the output is filtered using softmax entropy. The high entropy value indicates that the output needs further consideration from Domain-Specific, while the low value is ignored. Finally, classical learning methods as Domain-Specific are employed to determine the final output.

OSKut (Limkonchotiwat et al., 2021) uses domain adaptation method. OSKut employs three features; character n-gram embedding, character type n-gram, and probability and entropy values from the domain-generic model (Deepcut and Attacut). These features are concatenated and fed to a Bi-LSTM layer, followed by the attention (Vaswani et al., 2017) and fully connected network that ends with a single sigmoid output. OSKut achieved 94.57%-99.01% and 86.24%-97.33% F1 scores at character and word levels respectively. OSKut also offers a data augmentation method to increase the amount of training data. The method groups words together based on the output from the domain-generic model. Then the grouped words are replaced by words generated using Masked Language Model, WangchanBERTa (Lowphansirikul et al., 2021). With the data augmentation method, the OSKut performance reached 98.48%-98.67% and 96.18%-97.03% F1 scores at character and word levels on Wisesight corpus.

Some of the above methods are part of the PyThaiNLP (Phatthiyaphaibun et al., 2016), while the others are accessible via a GitHub page. PyThaiNLP is a Python library offering tools for the Thai language processing. It also has its own built-in methods for the word segmentation. For example, Newmm is the default word tokenization engine using dictionary-based maximal matching word segmentation, constrained with Thai Character Cluster (TCC) boundaries; Nercut is a dictionary-based maximal matching word segmentation, constrained with Thai Character Cluster (TCC) boundaries, combining tokens that are parts of the same named-entity.

After the segmentation process, the stems of tokens still need to be processed. For example, the segmentation result of sentence "Ilvoeeeeyouu" might be "I", "lvoe", "eee", "you", and "u". Noticeably, there are some unrecognised tokens from the segmentation result. Moreover, some token is misspelled, such as "lvoe". The unrecognised tokens ("eee" or "u") can be removed by checking with a lexical resource. However, it needs a suitable algorithm to correct word spelling.

A spell checker module is included in PythaiNLP. It uses Peter Norvig's algorithm (Norvig, 2007) to choose the most likely spelling correction given the word by searching for corrected candidate words based on edit distance. Then, it selects the candidate with the highest word occurrence probability. AiforThai[24] also provides a Word Approximation module to suggest the similar spelling words.

Thai2transformers (Lowphansirikul et al., 2021) provides customized scripts to train transformer-based masked language models on Thai texts with a variety of approaches for tokenisation. These approaches include: a subword-level token from SentencePiece library, a dictionary-based Thai word level tokenizer with maximal matching from PyThaiNLP, a similar dictionary-based tokenizer at a syllable level with maximal matching, and an ML-based Thai word level tokeniser SERF (Limkonchotiwat et al., 2020).

3.3. Part-of-Speech (POS) Tagging

The POS tagging is a major morpho-syntactic analysis task in NLP. According to the Orchid corpus, its POS tagset consists of 47 tags (word classes), whereas LST20 tagset consists of 16 tags. As mentioned in the previous section, SynThai and OpenNLP provide functionalities for this task. However, alternative tools are available.

RDRPOSTagger[25] (Nguyen et al., 2016) is a fast and accurate tool for Thai POS tagging. It constructs a SCRDR tree, where each node represents a set of rules, and the edge is called *if-not* or *expect* edge. Each word passes through the tree, then a tag is selected based on the condition of the stopping node. The experiment was conducted on 13 languages, and the model's accuracy for Thai language is 94.15%-94.21%.

spaCy-Thai[26] is a Thai Tokenizer, POS-tagger, and dependency-parser implemented in spaCy. The modal was trained on Universal Dependencies. In addition, PyThaiNLP offers POS tagging tools trained by 2 engines: averaged structured perceptron algorithm and unigrams. Similary, TLTK provides a POS tagging tool implemented based on NLTK perceptron with TNC data, which has achieved an accuracy of 91.68% on POS tagging.

3.4. Syntactic and Semantic Analysis Tools

Syntactic analysis tools, often called syntactic parser, analyses syntactic constituents and their relations, such as subject, predicate, object, etc. They also analyse the dependency relations between words within a sentence. Currently there are two Thai parser systems publicly available. One of them is the CF Parser (Seenual et al., 2018), which is a toolkit for Semi-automatic Thai treebank construction. This toolkit relies on WordCut for word segmentation, RDRPOSTagger for POS tagging, and NLTK for syntactic analysis and parsing. Seenual et al. carried out an experiment using three different methods, resulting in an F1 score ranging between

[23] https://github.com/mrpeerat/SEFR_CUT

[24] https://aiforthai.in.th

[25] https://github.com/datquocnguyen/RDRPOSTagger

[26] https://pypi.org/project/spacy-thai/

73.11% and 83.89%. Another parser is the Grammar Processing[27] which is a Python tool for transforming label brackets to Context-free grammar (CFG) and calculating the probability of all CFG.

Semantic analysis tools extract abstract level concepts and meanings from unstructured natural language data, such as named entities (NE), relations between them, sentiment and emotion etc.

Currently, most semantic tools for Thai language are available from PythaiNLP, such as ThaiNER, a Thai Named Entity Recognition (NER) tool. ThaiNER was trained with CRF algorithm using 5 features: *is_stop word*, *is_Thai word*, *is_space*, *POS tag*, and *is_digit*. It was trained with a data set which contains 6,456 sentences separated into 80% training data and 20% testing data.

TLTK also offers an NER module based on the CRF model adapted from sklearn[28], a Python library. The module was trained using AIforThai data, Sasimimon's and Nutcha's data. It was reported that the accuracy was below 88%[29].

POLYGLOT-NER (Al-Rfou et al., 2015) is a module in Polyglot built for named entity extraction for 40 languages, including Thai. It treats named entity extraction as a word classification task using information about phrases centered around key words, employing word embedding trained on Wikipedia data.

AIforThai[24] also provides API-based services to access their semantic analysis tools, including a) basic NLP processes (i.e., words similarity, NER, Cyberbully Expression Detector), b) tag suggestion for text, c) machine translation, d) sentiment analysis, e) chatbot, f) Question Answering, g) spam detection, h) intent classification, and i) feeling classification. The API can be accessed via HTTP request, which is free for academic and educational users. A sentiment analysis tool is also available to download at GitHub: `https://github.com/JagerV3/sentiment_analysis_thai`.

Machine translation is also an important aspect of semantic processing of language. There are two published major machine translation systems that involve Thai language. One of them is Thai/English machine translation (depa Thailand Artificial Intelligence Research Institute, 2020) and another is Thai/Chinese machine translation (Deelert, 2021). Both of them are based on a transformer model. The training datasets for Thai/English were *mt-opus* and *scb-mt-en-th-2020*. In an evaluation, they reported a BLEU score of 29 for Thai-to-English translation and 17.77 for English-to-Thai. The Thai-Chinese machine translation system was trained on OPUS Data Set (Open Subtitles v2018 and TED2020 v1), and they achieved BLEU scores of 15.53 and 8.42 for Thai-to-Chinese and Chinese-to-Thai translations respectively.

Table 2 shows the list of available tools discussed previously.

4. Discussion

Our survey shows that recently Thai NLP research has been advancing quickly and receiving an increasing attention from the NLP community. A number of corpus resources have been built that contain annotations of main layers of language information, such as Tokenisation, POS classes, syntactic dependency, named entities etc. Furthermore, some Thai lexicon resources have been built modelling on Major English counterparts like WordNet. Finally, a number of NLP software tools have been built to facilitate automatic analysis of language structures and the extraction of semantic information from Thai textual data.

However, current Thai NLP research has yet to develop a wider range of language resources and software tools to cater for the need of ever increasingly complex NLP tasks. For example, the most currently available Thai corpus resources are compiled from formal documents, news articles, etc., which fail to represent the full range of real-world language usage in today's communication media platforms and regional dialects of Thai language. For example, Thai language used in Twitter and Meta (Facebook) have distinct features from formal written documents. Such features can not be captured from the most current Thai corpus resources that are mostly reflect formal language use cases. For such purposes, more Thai corpus resources like VISTEC-2021 will need to be constructed.

With regards to Thai lexical resources, most existing major Thai lexical resources are in the form of Bilingual Dictionary or Thai version of existing English lexical resources, such as WordNet and the NRC emotion lexicon. There are very few dedicated large-scale Thai lexical resources, although some small-scale experimental Thai lexicons have been reported. It is a similar case for word-embedding based language models. As far as we know, only two major Thai word embedding models are available in the pre-trained BART and FastText models. With the increasing availability of Thai language data, particularly various online data, more larger-scale Thai language word embedding models can be generated.

Along with the Thai corpus and lexical resources, a number of Thai NLP software tools have been developed and reported. Our survey shows that remarkable efforts have been focused on the development of tools for automatic syllable analysis and tokenisation, achieving impressive 98%-99% top accuracies. This is reasonable, because of some unique linguistic features of Thai language such as the lack of word boundary delimiters, such tools are indispensable for any downstream NLP tasks. Such research has been fruitful in leading to the development of some downstream Thai

[27] https://github.com/tchayintr/simple-pcfgrammar/tree/master/grammar_processing

[28] https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

[29] For further details of the evaluation, see https://pypi.org/project/tltk/

Tool	Developer	Functionality	Accuracy
AIforThai	National Electronics and Computer Technology Center (NECTEC), Thailand	Tokenisation (Lexto+ 96.30%), Cyber bully detection, Chat bot, Feeling classification, Intent classification, Machine translation, NE recognition, Question answering, Sentiment analysis, Spam detection, Text tag suggestion, Word approximation, Word similarity	Unreported
AttaCut	Chormai et al.	Tokenisation	89.00-91.00%
CF Parser	Seenual et al.	Parser	73.11-83.89%
CutKum	Pucktada Treeratpituk	Tokenisation	95.00%
Deepcut	Kittinaradorn et al.	Tokenisation	98.10%
ETCC	Jeeragone et al.	Tokenisation	Unreported
GrammarAnalyser	Thodsaporn Chay-intr	Grammar extraction	Unreported
JTCC	Wittawat Jitkrittum	Tokenisation	Unreported
Lalita	AIBuilders	Chinese/Thai machine translation	15.53, 8.42 Bleu
Machinetranslator	VISTEC-depa	English/Thai Machine Translation	29, 17.77 Bleu
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%
Polyglot-NER	Al-Rfou et al.	NE recognition	Unreported
PyThaiNLP	Phatthiyaphaibun et al.	Tokenisation, Spell checker, POS-tagging, NE recognition	Unreported
RDRPOSTagger	Nguyen et al.	POS-tagging	94.15-94.21%
SentimentAnalyser	Raymond	Sentiment analysis	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
TCC	Theeramunkong et al.	Tokenisation	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

Table 2: Publicly available Thai NLP tools.

NLP tools for higher level morpho-syntactic analysis and semantic information extraction, such s POS taggers, syntactic parsers and sentiment analysis tools. Yet, Thai NLP tool development for downstream NLP tasks is still limited, and there is an urgent need for further development in this area.

5. Conclusion

In this paper, we reported our survey on the development of Thai NLP language resources and tools. As shown by our survey, Thai NLP research has achieved a significant achievement over the past three decades, and recently it has been advancing with increasing

pace. Nonetheless, while a remarkable achievement has been achieved in some aspects, particularly for upstream tasks such as tokenisation, research on downstream tasks such as syntactic parsing and semantic analysis is still limited. We envisage that, with the availability of richer Thai language resources and advance in NLP methodology and techniques, Thai NLP research will achieve rapid development in near future.

6. Acknowledgements

We would like to thank the Ministry of Higher Education, Science, Research and Innovation of Thailand for funding the first author's PhD research in The School of Computing and Communications of Lancaster University, UK, on which this paper is based., on which this paper is based.

7. References

Akaraputthiporn, P. (2008). The construction of thai wordnet of 2nd order entity common base concepts using a bi-directional translation method A study of the diversity of meanings affecting translational accuracy. Master's thesis, Department of Linguistics.

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594, British Columbia, Canada. SIAM.

Apache Software Foundation. (2014). opennlp natural language processing library. http://opennlp.apache.org/.

Aroonmanakun, W. (2002). Collocation and thai word segmentation. In *In Proceedings of the Fifth Symposium on Natural Language Processing The Fifth Oriental COCOSDA Workshop*, pages 68–75, Hua Hin, Prachuapkirikhan, Thailand.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information.

Boonkwan, P., Luantangsrisuk, V., Phaholphinyo, S., Kriengket, K., Leenoi, D., Phrombut, C., Boriboon, M., Kosawat, K., and Supnithi, T. (2020). The annotation guideline of LST20 corpus.

Chay-intr, T. (2020). Thaibert. https://github.com/tchayintr/thbert.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland, June. Association for Computational Linguistics.

Chormai, P., Prasertsom, P., and Rutherford, A. (2019). Attacut: A fast and accurate neural thai word segmenter.

Deelert, L. (2021). Lalita machine translation chinese-thai. Published 29 June 2021 by Lalita Deelert; last accessed on 05 January 2022.

depa Thailand Artificial Intelligence Research Institute, V. (2020). English-thai machine translation models. Published 23 June 2020 by VISTECxdepa; last accessed on 05 January 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Funaki, S. (1993). Multi-lingual machine translation (mmt) project. In *Proceedings of Machine Translation Summit IV*, pages 73–78, Kobe, Japan.

Haruechaiyasak, C. and Kongthon, A. (2013). LexToPlus: A Thai lexeme tokenization and normalization tool. In *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing*, pages 9–16, Nagoya Congress Center, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Haruechaiyasak, C. and Kongyoung, S. (2009). Tlex: Thai lexeme analyser based on the conditional random fields. In *Proceedings of 8th International Symposium on Natural Language Processing*.

Hockenmaier, J., Bierner, G., and Baldridge, J. (2004). Extending the coverage of a ccg system. *Research on Language and Computation*, 2(2):165–208.

Jeeragone, I., Yuanghirun, P., Paludkong, S., Nitsuwat, S., and Limmaneepraserth, P. (2001). Thai word segmentation using combination of forward and backward longest matching techniques. In *2001 In International Symposium on Communications and Information Technology (ISCIT)*, pages 37–40, Chiang Mai, Thailand.

Jitkrittum, W. (2017). Jtcc. https://github.com/wittawatj/jtcc.

Kalunsima, S. (2010). Thai named entity recognition: A study of person location and organization names. Master's thesis, Department of Linguistics.

Kittinaradorn, R., Achakulvisut, T., Chaovavanich, K., Srithaworn, K., Chormai, P., Kaewkasi, C., Ruangrong, T., and Oparad, K. (2019). DeepCut: A Thai word tokenization library using Deep Neural Network, September.

Kongyoung, S., Rugchatjaroen, A., and Kosawat, K. (2015). Tlex+: a hybrid method using conditional random fields and dictionaries for thai word segmentation. In *International Conference on Knowledge, Information, and Creativity Support Systems*, pages 112–125, Phuket, Thailand. Springer.

Lapjaturapit, T., Viriyayudhakom, K., and Theeramunkong, T. (2018). Multi-candidate word segmentation using bi-directional lstm neural networks. In *2018 International Conference on Embedded Systems and Intelligent Technology International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*, pages 1–6, Khon Kaen, Thailand.

Leenoi, D. (2008). The construction of thai wordnet

of 1st order entity common base concepts using a bi-directional translation method and with dictionaries of different compilational approaches. Master's thesis, Department of Linguistics.

Lertcheva, N. (2010). Thai named entity recognition A study of product names in economic news. Master's thesis, Department of Linguistics.

Limkonchotiwat, P., Phatthiyaphaibun, W., Sarwar, R., Chuangsuwanich, E., and Nutanong, S. (2020). Domain adaptation of Thai word segmentation models using stacked ensemble. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3841–3847, Online, November. Association for Computational Linguistics.

Limkonchotiwat, P., Phatthiyaphaibun, W., Sarwar, R., Chuangsuwanich, E., and Nutanong, S. (2021). Handling cross- and out-of-domain samples in Thai word segmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1003–1016, Online, August. Association for Computational Linguistics.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, USA.

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., and Nutanong, S. (2021). Wangchanberta: Pretraining transformer-based thai language models.

Manning, C. D., Surdeanu, M., Bauer, J. B., Finkel, J. F., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA.

Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. (1997). Feature-based thai word segmentation. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, volume 97, pages 41–46, Phuket, Thailand. Citeseer.

Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, USA. Association for Computational Linguistics.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. (2016). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI communications*, 29(3):409–422.

Nomoto, H. (2019). Interpersonal meaning annotation for asian language corpora: The case of tufs asian language parallel corpus (talpco). In *Proceedings of the Twenty-Fifth Annual Meeting of the Association for Natural Language Processing*, Nagoya University, Furo-cho, Chikusa-ku, Nagoya.

Norvig, P. (2007). How to write a spelling corrector. Published February 2007; last accessed on 05 January 2022.

Phatthiyaphaibun, W. and Polpanumas, C. (2018). prachathai-67k. https://github.com/PyThaiNLP/prachathai-67k.

Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P. (2016). PyThaiNLP: Thai Natural Language Processing in Python, June.

Phatthiyaphaibun, W. (2017a). Clickbait. https://github.com/PyThaiNLP/lexicon-thai/tree/master/clickbait.

Phatthiyaphaibun, W. (2017b). Prime minister 29. https://github.com/PyThaiNLP/lexicon-thai/tree/master/thai-corpus/Prime%20Minister%2029.

Phatthiyaphaibun, W. (2017c). Thai sentiment lexicon. https://github.com/PyThaiNLP/lexicon-thai/tree/master/sentiment.

Phongthawee, P. (2018). Cutthai. https://github.com/pureexe/cutthai.

Phuriphatwatthana, W. (2016). Thai word segmentation and part-of-speech tagging with deep learning. Dissertation of Bachelor of Science in Computer Science, Chiang Mai University.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jimenez, R.-M., Knight, D., Kren, M., Lofberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., and Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of The Tenth International Conference on Language Resources and Evaluation (LREC2016)*, pages 290–297.

Polpanumas, C. and Phatthiyaphaibun, W. (2021). thai2fit: Thai language implementation of ulmfit, jan.

Poowarawan, Y. (1986). Dictionary-based thai syllable separation. In *The 9th Electronics Engineering Conference*, pages 409–418.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Satayamas, V. (2019a). Wordcutpy. https://github.com/veer66/wordcutpy.

Satayamas, V. (2019b). Yaitron. https://github.com/veer66/Yaitron.

Satayamas, V. (2021). Wordcut. https://github.com/veer66/wordcut.

Seeha, S., Bilan, I., Mamani Sanchez, L., Huber, J., Matuschek, M., and SchÃ¼tze, H. (2020). Thailmcut: Unsupervised pretraining for thai word segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6947–6957, Marseille, France, May. European Language Resources Association.

Seenual, P., Chay-Intr, T., and Theeramunkong, T. (2018). Cf planter: A toolset for semi-automatic thai treebank construction. In *2018 International Conference on Embedded Systems and Intelligent Technology International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*, pages 1–6, Khon Kaen, Thailand.

Sirihattasak, S., Komachi, M., and Ishikawa, H. (2018). Annotation and classification of toxicity for thai twitter. In *Proceedings of LREC 2018 Workshop and the 2nd Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS'18)*, Miyazaki, Japan.

Sornlertlamvanich, V., Takahashi, N., and Isahara, H. (1999). Building a thai part-of-speech tagged corpus (orchid). *Journal of the Acoustical Society of Japan (E)*, 20(3):189–198.

Sornlertlamvanich, V. (2019). Natural language processing research in thai context - a 29-year journey of thai nlp. https://www.slideshare.net/virach/nlp-historythaivirach20191025, October.

Suriyawongkul, A., Chuangsuwanich, E., Chormai, P., and Polpanumas, C. (2019). Pythainlp/wisesight-sentiment: First release, September.

Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong, T., and Chinnan, W. (2000). Character cluster based thai information retrieval. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, IRAL '00, page 75–80, New York, NY, USA. Association for Computing Machinery.

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. (2010). Thai-nest: A framework for thai named entity tagging specification and tools. In *Proceedings of the 2nd International Conference on Corpus Linguistics, 2010*, Universidade da Coruña, Spain, 05.

Thongthanomkul, E., Chuesathuchon, Y., Nearunchorn, T., and Polpanumas, C. (2019). Wongnai. https://github.com/wongnai/wongnai-corpus.

Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokarat, C., Sornlertlamvanich, V., and Isahara, H. (2009). Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),*, Suntec, Singapore.

Tirasaroj, N. (2010). Thai named entity recognition: The application of conditional random fields models. Master's thesis, Department of Linguistics.

Treeratpituk, P. (2017). Cutkum: Thai word-segmentation with lstm in tensorflow.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, California.

Viriyayudhakorn, K. (2019). thai-joke-corpus. https://github.com/iapp-technology/thai-joke-corpus.