Chunwijitra and Wutiwiwatchai *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:24
DOI 10.1186/s13636-017-0121-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Classification-based spoken text selection for LVCSR language modeling

Vataya Chunwijitra* and Chai Wutiwiwatchai

Abstract

Large vocabulary continuous speech recognition (LVCSR) has naturally been demanded for transcribing daily conversations, while developing spoken text data to train LVCSR is costly and time-consuming. In this paper, we propose a classification-based method to automatically select social media data for constructing a spoken-style language model in LVCSR. Three classification techniques, SVM, CRF, and LSTM, trained by words and parts-of-speech are comparatively experimented to identify the degree of spoken style in each social media sentence. Spoken-style utterances are chosen by incremental greedy selection based on the score of the SVM or the CRF classifier or the output classified as "spoken" by the LSTM classifier. With the proposed method, just 51.8, 91.6, and 79.9% of the utterances in a Twitter text collection are marked as spoken utterances by the SVM, CRF, and LSTM classifiers, respectively. A baseline language model is then improved by interpolating with the one trained by these selected utterances. The proposed model is evaluated on two Thai LVCSR tasks: social media conversations and a speech-to-speech translation application. Experimental results show that all the three classification-based data selection methods clearly help reducing the overall spoken test set perplexities. Regarding the LVCSR word error rate (WER), they achieve 3.38, 3.44, and 3.39% WER reduction, respectively, over the baseline language model, and 1.07, 0.23, and 0.38% WER reduction, respectively, over the conventional perplexity-based text selection approach.

Keywords: Spoken language model, LVCSR, Classification

1 Introduction

Large vocabulary continuous speech recognition (LVCSR) systems now play an increasingly significant role in daily life. Many commercial applications of LVCSR are widely employed, e.g., medical dictation, getting weather information, data entry, speech transcription, speech-to-speech translation, railway reservation, etc. However, in some systems, e.g., a speech-to-speech translation and interactive voice response (IVR) for customer service, speech input is highly conversational while it is more of a written style in medical dictation. A spoken language and a written language are different in several aspects including the word choice and the sentence structure. Hence, it is important to consider the language style for creating an efficient language model (LM) for a LVCSR system.

Typical speech recognition uses a LM to introduce linguistic restrictions that helps the recognizer figure out a word sequence. In general, a LM is built by using a text corpus, and its performance depends on the data size and text quality. For creating LVCSR systems in different domains and styles, it is necessary to find appropriate text sources well matched to the task domain as well as the style of speech input. A straightforward way to create a conversational text corpus is to transcribe recorded human conversations. However, transcribing is much costly and time-consuming, and thus it is quite difficult to get a large amount of conversational data to reliably train a LM. Much effort has been devoted to the unsupervised and semi-supervised acoustic model training [1–3] to exploit the untranscribed data. Most of these works focus on generating better quality hypothesis and on improving confidence measure for better data selection. In other direction, acquiring large text from the Internet is a popular way nowadays. Filtering text appropriate for the targeted LM is also an important step towards effective use of these data. Word perplexity was used as a similarity criterion to select text for a target domain [4, 5]. Another approach based on comparing

*Correspondence: vataya.chunwijitra@nectec.or.th
NECTEC, National Science and Technology Development Agency (NSTDA), 112 Pahonyothin Road, Pathumthani 12120, Thailand

the entropy between domain-specific and general domain LMs was proposed [6, 7]. The relative entropy-based criterion in [8] was also chosen for building topic-specific LMs. They showed that these techniques could produce a better domain-specific LM than that by random data selection.

This paper targets on building Thai LVCSR serving general daily conversations. To enlarge the LM training data suitable for this task, acquiring text, and filtering for spoken-style text are needed. Twitter, a well-known social media microblog, is attractive as the text length up to 140 characters tends to make the language more informal and sometimes produces incomplete sentences, which are one characteristic of the spoken language. Using the Twitter text, called tweets, to build LM was first introduced in [9], where the in-vocabulary hit rate was used as a criterion to select useful tweets. In addition to the classical entropy-based sentence selection methods previously proposed, in this paper, modern machine learning algorithms including support vector machines (SVM), conditional random fields (CRF), and long short-term memory neural network (LSTM) are comparatively investigated to improve the precision of spoken language sentence selection. With training data representing highly written language such as newspaper and highly spoken language such as telephone conversation, the trained model is expected to estimate the degree of spoken style of an input text. This model is suitable for selecting spoken sentences from mixed-style data such as tweets. The selected tweets are finally used to construct a spoken-style LM which is interpolated to a baseline LM to improve the overall Thai LVCSR performance. The resulted LM is comparatively evaluated with a LM trained by a set of exact spoken text, in terms of both the perplexity and the LVCSR word error rate (WER) on two different tasks.

This paper is organized as follows: we first briefly introduce the characteristics of spoken and written languages and also describe existing Thai large vocabulary speech corpora in Section 2. We present our process for collecting data from Twitter in Section 3. In Section 4, the proposed method of style-based data selection for constructing spoken LM is explained. In Section 5, we describe the experiments to evaluate the proposed style classifier in terms of the classification accuracy, the perplexity of the LM, and the LVCSR recognition performance. We finally conclude our work and discuss our future direction in Section 6.

2 Thai spoken and written languages

2.1 Distinction between difference text styles

One of the sociolinguists, Prasithrathsint [10], suggested that it is better to recognize spoken and written languages in terms of language styles rather than communicative methods. For instance, we can talk using a written-style language as in giving a formal speech; on the other hand, we can write using a spoken-style language as in a personal letter. A spoken language and a written language are different in several aspects including vocabulary choice and sentence structure. The spoken text uses words that would be inappropriate in a written text. The words employed in the conversation are simpler and shorter, and the speakers do not have much time in a conversation to choose their vocabulary carefully. Sentences are relatively unplanned, less structured and interactive. In contrast, the written text is planned, well organized and transactional. Some certain words used in the written text would seem unusual if they appeared in a conversation. The writers have more time to chose their vocabulary appropriate for their particular purpose. There are also more nouns and longer words utilized in the written text than in the spoken text. The characteristics of Thai spoken and written languages [11] in general are shown in Table 1.

2.2 Thai Corpora for analysis of spoken and written styles

Several Thai speech corpora have been developed for speech processing research. These corpora are different in terms of acoustic characteristics of speech signals, conditions of input channels, and application domains. Among these corpora, three of them, LOTUS [12], LOTUS-Cell 2.0 [13] , and LOTUS-SOC [14, 15], are considered large

Table 1 Characteristics of Thai spoken language vs. written language

Spoken language	Written language
(1) A sentence is incomplete or fragmented (missing a subject or a verb) [10, 34]. Connected phrases maybe found continuously [34].	(1) A sentence is complete.
(2) A sentence is less sophisticated: fewer subordinate clauses [34].	(2) A sentence is more sophisticated: more subordinate clauses [34].
(3) A sentence starts with a topic-comment structure [34].	(3) A sentence starts with a subject-predicate form [34].
(4) Repetition, word duplication or paraphrasing, often appears [35].	(4) A sentence contains less repetition [35].
(5) A filler, a word or expression which is filled up when a speaker is in the process of thinking, often appears [35].	(5) A filler does not appear [35].
(6) A final particle, e.g. /khâʔ/, /khráp/, /nîːaʔ/, and /c-âːʔ/, often appears [35].	(6) A sentence contains fewer final particles [35].
(7) Slang and foreign words are often used.	(7) Formal lexicon is used.

vocabulary speech corpora. The detail of each corpus is described below while their statistics are given in Table 2.

LOTUS is a Large vOcabulary Thai continUous Speech corpus specially designed for developing Thai LVCSR. Utterances in the LOTUS corpus were recorded in a reading style where the reading prompts were taken from a Thai text corpus, ORCHID [16]. This corpus contains articles from various sources, such as magazines, Thai encyclopedia, and journals, and thus can be considered as a general-domain written-style corpus. The LOTUS corpus contains 55 h of speech from 48 speakers.

LOTUS-Cell 2.0, or LOTUS-Cell for short in the context of this paper, is a large Thai telephone speech corpus. It contains three parts of speech data, answers to closed-ended questions, answers to open-ended questions and dialog speech. The questions and discussion topics were designed to elicit speech data which have their contents conformed to the domains of potential automatic speech recognition (ASR) applications such as transportation, tourism, and health care. The corpus contains recorded speech from 212 speakers with gender balance. The amount of recorded speech is 90 h.

LOTUS-SOC is a Large vOcabulary Thai ContinUous Speech SOCial media corpus which aims at reducing the effort in creating a conversational speech corpus so that a larger corpus could be collected. The LOTUS-SOC corpus is created by recording Thai native speakers uttering Twitter messages through a mobile application. By using this data collection method, we can avoid the process of segmenting the recorded speech into utterances and also the need to transcribe the data. The corpus contains recorded speech from 208 speakers and approximately 172 h of speech. The age range of speakers is 11–58 years old.

VoiceTra4U-M (VT) is a speech translation application in sport and travel domains developed under the Universal Speech Translation Advanced Research (U-STAR) consortium (http://www.ustarconsortium.com/qws/slot/u50227/index.html). This consortium, consisting of 26 research institutes from 23 countries, aims to enable people in the world to communicate without any language barriers. The application allows five people to chat simultaneously in 23 different languages including Thai. It is available on the iOS platform, but the consortium also plans to make it available on Android platform due to a rapid growth of Android smartphones in many countries. The corpus contains speech from various Thai speakers and obtained approximately 22 h of speech recorded on mobile devices in real environments. The speech data were manually transcribed.

Table 2 The amount of texts in Thai large vocabulary speech corpora

Corpus	Text style	Number of utterances	Number of word tokens	Vocabulary size
LOTUS	Written	4887	90,336	5112
LOTUS-CELL	Spoken	55,457	284,498	9595
LOTUS-SOC	Spoken/Written	78,264	1,601,230	13,739
VoiceTra4U-M	Spoken	9899	30,876	2141

To be used in this paper, the LOTUS corpus was divided into two sets: LOTUS-TRN and LOTUS-DEV. Similarly, the LOTUS-Cell corpus was divided into CELL-TRN and CELL-DEV. The VoiceTra4U-M corpus was also divided into VT-DEV and VT-TST. SOC was randomly selected from the LOTUS-SOC corpus. The detail of each set is presented in Table 3.

3 Twitter data collection

Twitter data are used as a data source for performing style-based data selection to build our spoken LM. In this study, we collected approximately 2 million Thai tweets during February to March 2013 via the available Twitter REST API. This API allows us to specify desired keywords when acquiring the data. We will refer to this text collection as a Thai Twitter text corpus throughout the rest of this paper. 150,000 tweets were randomly selected as initial data for examining in our experiments.

Before using the Twitter text for building a LM, it is necessary to perform data cleaning and text normalization. In the cleaning process, we have removed Twitter symbols, such as "RT" (re-tweet), mention markers (@username) and hashtag (#hashtag), and also URLs from the text. We perform duplicated sentence removal to avoid spams and re-tweets. Known abbreviations, numbers and special characters in Twitter messages were normalized into more suitable forms with a set of rules shown in Fig. 1. For the nonstandard words in tweets, extensive normalization is required to transform these words into their normalized forms which mean words in a dictionary. In this work, we used a similarity-based text normalization method to handle frequent nonstandard types in Thai tweets including homophonic, spelling error, and insertion [11]. The similarity between a nonstandard word and a normalized word in terms of spelling and pronunciation are measured by an edit distance while the similarity in terms of context is measured by a Kullback-Leibler (KL) distance. With this cleaning and normalization techniques, a 3-g LM trained by preprocessed tweets has been improved regarding both the perplexity and WER performance [11].

Like Chinese, and Japanese, Thai is an unsegmented script language, i.e., there is no boundary marker between words while boundary markers on phrase and sentence levels are often ambiguous. Furthermore, there is no capital letter to indicate the beginning of a new sentence or a proper noun. Therefore, after the text is cleaned and normalized, we need to identify word boundaries. In this work, we use TLex [17], a Thai word segmentation tool

Table 3 The detail of each speech corpora set used in this study

Corpus	Data set	Usage	Number of utterances
LOTUS	LOTUS-TRN	Training the classification model	4330
	LOTUS-DEV	Evaluation the classifier performance	557
LOTUS-CELL	CELL-TRN	Training the classification model	40,000
	CELL-DEV	Evaluation the classifier performance	15,475
VoiceTra4U-M	VT-DEV	Optimization the selection of confidence groups	7982
	VT-TST	Evaluation the recognition performance	1917
LOTUS-SOC	SOC	Evaluation the recognition performance	4000

based on CRF, to automatically identify word boundaries in tweet texts.

4 Style-based data selection

The goal of this work is to retrieve spoken-style text from social media data, Twitter, for building an appropriate LM in spoken-style LVCSR. Since Twitter data may also contain written-style text such as formal news tweets from news agencies, we first explore the use of text categorization to classify text into two categories: spoken and written. After that, the selected spoken-style sentences and the existing spoken text, CELL-TRN, are used to train the spoken LM. Finally, an interpolated LM between prepared baseline and spoken LMs is used as the final LM for LVCSR. An overall diagram, baseline LM construction, classification features, methods, and data selection are discussed in Sections 4.1, 4.2, 4.3, 4.4, and 4.5, respectively.

4.1 Overall diagram

The system overall diagram is shown in Fig. 2. The social media data, Twitter, is used as the text resource. It contains 150K utterances of Thai tweets collected from February to March 2013 via the Twitter REST API. The original data are preprocessed by removing unnecessary data such as extraneous tags and duplicated sentences, and then normalizing to suitable forms for reading as described in Section 3. Next, these data are segmented into word sequences by CRF-based word segmentation, and afterwards the stylistic text classifications are applied. In this work, we compare three classification approaches, SVM [18], CRF [19], and LSTM [20]. The SVM or CRF classifier gives for each sentence an output score indicating the degree of being a spoken style, i.e., a large score for "spoken" and a small score for "written." In the LSTM case, each sentence is directly classified into "spoken" or "written" with no score.

In the case of SVM or CRF, sentences are separated into groups according to classification scores.

The group of sentences with the lowest perplexity scores on VT-DEV set are selected. In the LSTM case, only sentences classified as "spoken" are selected. After the above selection process, a spoken LM is built by using these selected sentences combined with CELL-TRN sentences. The spoken LM is then used to interpolate with a baseline LM to produce a final LM for the LVCSR system. A linear interpolation algorithm is employed as follows.

$$LM = \lambda \cdot LM_{\text{baseline}} + (1 - \lambda) \cdot LM_{\text{spoken}} \quad (1)$$

where LM_{baseline} is the baseline LM trained by the existing data as described in Section 4.2, LM_{spoken} is the LM trained by the selected spoken-style sentences combined with the CELL-TRN, and λ is the weighting factor for tuning the final model.

1) Expand known abbreviations into their full forms according to an abbreviation dictionary, e.g., ธ.ค. --> ธันวาคม (December).
2) Normalize special characters (/ : , ; + - * @) into words, e.g., ``+'' --> บวก (plus).
3) Convert Thai numerals to Arabic numberals, e.g., ๑ --> 1.
4) Convert numbers to words, e.g., 5 --> ห้า (five).
5) Remove all other special characters.
6) Replace repeated characters which occur more than 3 times with a single character using a regular expression, e.g., มากกกกก --> มาก (many).

Fig. 1 Normalization rules for Twitter messages

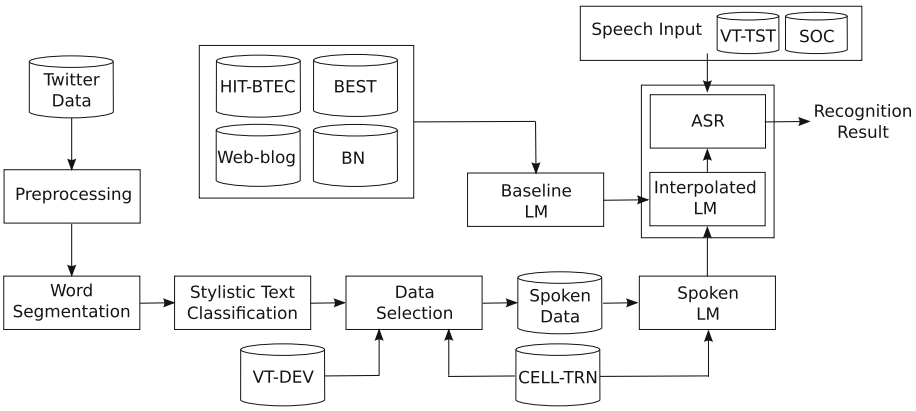

Fig. 2 Overall diagram

4.2 Baseline LM

In this study, four large text corpora, HIT-BTEC, BEST, Web-blog, and LOTUS-BN, are used to build a baseline LM. To handle OOV words, a hybrid 3-gram LM technique [21] was adopted. As these corpora cover a variety of domains, they are useful resources for open-vocabulary LVCSR. Table 4 summarizes the amount of text from each training corpus.

HIT-BTEC is a Thai translated version of the multilingual Olympic-domain corpus developed by Harbin institute of technology (HIT) [22] and the Basic Travel Expression Corpus (BTEC) [23]. This corpus was initially created for speech-to-speech translation research. The HIT corpus includes utterances in five domains related to Olympic games, namely, traveling, dining, sports, traffic, and business. The Thai version of BTEC and HIT was constructed under the Universal Speech Translation Advanced Research (USTAR) consortium. The total amount of data in the HIT-BTEC corpus is nearly 160,000 utterances.

BEST [24] is a Thai text corpus developed under the BEST (Benchmark for Enhancing the Standard of Thai language processing) project. Articles in BEST were collected from eight different genres: academic article, encyclopedia, novel, news, Buddhism, law, lecture, and Wikipedia. These articles were manually segmented into words by linguists. The total amount of available data is approximately 7 million words.

Web-blog is a large collection of Thai web text from chatting blog and discussion forum on a famous website in Thailand. The corpus consists of articles from eight different genres: mobile, travel, camera, films, residence, news, automobile, and woman's life. This corpus was collected from June 2011 to March 2012. Tlex [17] was used to automatically identify word boundaries. The total amount of data in this corpus is nearly 140,000 utterances.

LOTUS-BN [25] is a Thai television broadcast news corpus which includes audio recordings of hour-long news programs and their transcriptions. Each news program is segmented into small parts, i.e., sections and utterances. There are 18 news topics in the corpus, for instance, politics, sport, and weather. The corpus contains approximately 156 h of speech from 43 female speakers and 38 male speakers. Data in LOTUS-BN are divided into 3 sets: a training set (TR), a development test set (DT), and an evaluation test set (ET). There is no overlapping speaker among the TR, DT and ET sets. Only the TR set (around 50,000 utterances) was used to train the LM.

Table 4 Text resources for language model training

Corpus	Number of utterance	Number of token	Vocabulary size
HIT-BTEC	159,718	1,745,680	20,307
BEST	410,648	7,818,410	110,334
Web-blog	1,380,932	58,698,866	449,743
LOTUS-BN (TR)	50,187	929,810	35,327
ALL	2,001,485	69,192,766	615,711

4.3 Classification features

According to the difference between spoken and written languages summarized in Table 1, we consider using words in the input text and their parts of speech (POSs) as features for style classification. The word feature represents different word choices between the spoken and written language. For instance, informal words are more likely to occur in a spoken-style utterance while formal words are more likely to occur in a written-style utterance. Similarly, some POSs, e.g., particle and personal pronoun, are more likely to occur in spoken language while some POSs, e.g., conjunction which indicates a complex sentence, are more likely to occur in the written language. To automatically tag the POS of each word, we constructed

a POS tagger using CRF. We trained the CRF model with manually tagged data which contain 3 million words taken from the BEST corpus [24]. Articles in BEST were manually segmented into words and then POS tagged by linguists. The tagset consists of 35 POSs [26] which was modified from the 47-POS tagset used in the ORCHID corpus [16]. We use the word and its contexts, i.e. the previous word and the following word, as features for predicting the POS of each word. Our CRF-based POS tagger has 97.6% accuracy.

4.4 Stylistic text classification methods

In this work, three machine learning approaches, SVM, CRF, and LSTM are compared for style classification. SVM is one of the most effective machine learning algorithms for many complex binary classification problems. One remarkable property of SVM is its ability to learn regardless of the dimensionality of the feature space. Given a category, each example belongs to one of two classes referred to as the positive and negative class.

For the kernal function in this study, we investigate two basic kernels, Linear and Radial Basis Function (RBF).

An input vector of the SVM classifier contains 40,964 elements representing all lexical words and all POSs. Each element value is the frequency of the word or POS in the input sentence. We train the SVM to predict two classes, a positive class (+) for a spoken text utterance and a negative class (−) for a written one. Given an input utterance, the SVM outputs a real value of the decision function, the larger value the output of SVM is, the more spoken the utterance becomes. It is noted that the SVM output score used in this study is the predicted value that can be used to order the test utterance for ranking the spoken-like degree.

CRF is undirected graphical model for segmenting and labeling structured data [19]. The CRF model represents a conditional distribution $p(y|x)$ and dependencies over the observation sequence x. We assume that $X = (x_1, x_1, \ldots, x_T)$ is a input sequence and $Y = (y_1, y_2, \ldots, y_T)$ is a set of label sequence.

We train a CRF by maximizing the log-likelihood of a label sequence in the training data. Our CRF classifier works at word level.

Each word is labeled as "spoken" in spoken training utterances or "written" in written ones. We also use words and POSs as classification features. The current word, previous word and the following word along with their POSs are used to predict a classification label, "spoken" or "written," for each word. The output of CRF for each input sentence is a conditional probability; a high probability reflects "spoken" and a low reflects "written."

LSTM is an alternative architecture for recurrent neural network inspired by the human memory systems. The LSTM is a model that allows us to input a sequence of inputs. At every step, the model will update its internal representation of the input so far, giving it a form of memory.

Using the LSTM classifier in this work, both words and their POSs are also employed as classification features at an utterance level. We implemented the LSTM with a single hidden layer, 0.01 of learning rate for weight updates, and a stochastic gradient-based optimization algorithm [27] for weight optimization. Each training sentence is labeled as "1" for a spoken class and "2" for a written class. Since we are attempting to classify the whole sentence, not an individual word, we only consider the last output of the network as the actual classification result. The LSTM output contains class elements, a class "1" for a spoken-style and a class "2" for a written-style.

4.5 Data selection methods for building spoken LM

After classifying, the utterances are organized into groups based on the scores from SVM or CRF, or the output classes from LSTM. From SVM and CRF classification process, the output score is a real value.

We can then, by observing, cluster the utterances into 10 groups as shown in Table 5.

The top N groups having the highest spoken-style scores are selected for spoken LM training. To find an optimal N, we prepare accumulated groups of C0 to CN, denoting as C0, C0–C1, C0–C2, and so on. Each accumulated group is used to train a LM. The constructed LM is then evaluated by the perplexity of the VT-DEV which is a known spoken-style data set. The accumulated group giving the lowest perplexity is considered the optimal one. Figures 3 and 4 show the perplexity changes and the number of picked sentences with the accumulations of groups for SVM- and CRF-based scoring, respectively. As shown in the figures, at the point of C0–C4 for SVM and C0-C2

Table 5 The organized groups of SVM- and CRF-based scoring calculated from Twitter text data

Group	SVM-based scoring	CRF-based scoring
C0	CELL-TRN data	CELL-TRN data
C1	Score $\geq$ 4.0	Score $\geq$ 0.9
C2	4.0 > score $\geq$ 3.0	0.9 > score $\geq$ 0.8
C3	3.0 > score $\geq$ 2.0	0.8 > score $\geq$ 0.7
C4	2.0 > score $\geq$ 1.0	0.7 > score $\geq$ 0.6
C5	1.0 > score $\geq$ 0.0	0.6 > score $\geq$ 0.5
C6	0.0 > score > −1.0	0.5 > score $\geq$ 0.4
C7	−1.0 $\geq$ score > −2.0	0.4 > score $\geq$ 0.3
C8	−2.0 $\geq$ score > −3.0	0.3 > score $\geq$ 0.2
C9	−3.0 $\geq$ score > −4.0	0.2 > score $\geq$ 0.1
C10	Score $\leq$ −4.0	Score < 0.1

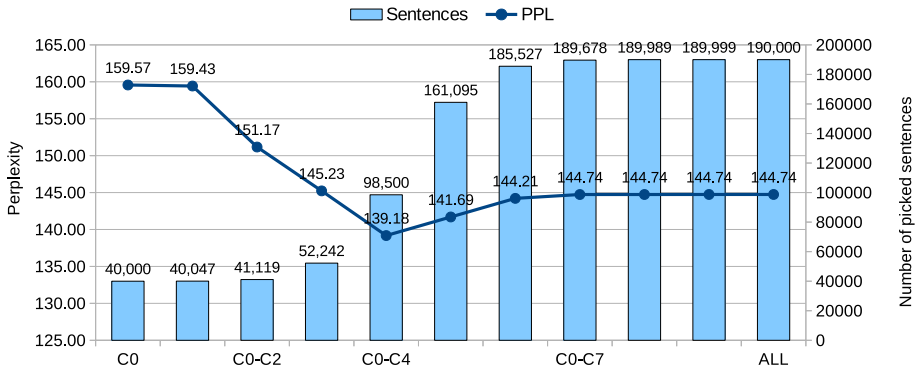

Fig. 3 Perplexities with accumulation of groups for SVM-based scoring

for CRF, the perplexity is the minimum. These accumulated groups, containing totally 98K sentences for SVM and 174K sentences for CRF, are regarded as the optimal clusters.

Using the LSTM classifier, the output contains class elements, the class "1" for a spoken text utterance and the class "2" for a written one. Consequently, in this work, the spoken LM are trained from the union of the existing spoken-style corpus (CELL-TRN) and the selected sentences with the class "1." With this technique, the total number of utterances for building spoken LM are 151K.

5 Experiments

We evaluate the proposed style classifier in terms of the classification accuracy, the perplexity of the result LM and the recognition performance. Experimental results are discussed in the following sub-sections.

5.1 Experimental conditions

Acoustic model training data of our LVCSR composes of 773 h of speech from LOTUS [12], LOTUS-BN [25], LOTUS-SOC [14], VoiceTra4U-M, and other unpublished sources. VoiceTra4U-M is a speech translation application in sport and travel domains developed under the Universal Speech Translation Advanced Research (U-STAR) project (http://www.ustar-consortium.com/). Twenty-two h of speech were recorded on mobile devices in the real environment. We used the Kaldi Speech Recognition Toolkit [28] to first train a conventional GMM-based acoustic model. We then applied the Maximum Mutual Information (MMI) discriminative training technique described in [29]. Each frame of speech data was converted into a sequence of 39 dimensional feature vectors of 12 MFCCs appended with a log energy, and their first and second derivatives. We used a 25-ms frame length with 10-ms window shift. Features from a context window of 3 frames to the left and right were also included. A Linear Discriminate Analysis (LDA) was also applied to the feature space to reduce feature dimensions to 40.

The baseline LM (*Base*) training data contain 9.4M words from three corpora, BEST [24], LOTUS-BN [25], and HIT-BTEC [23]. As these corpora cover a variety of domains, e.g. law, news, and travel, and the vocabulary size is as large as 121K, they are excellent resources for training a hybrid LM for open-vocabulary LVCSR system. For comparison, five sentence selection methods were applied on the same Twitter data set.

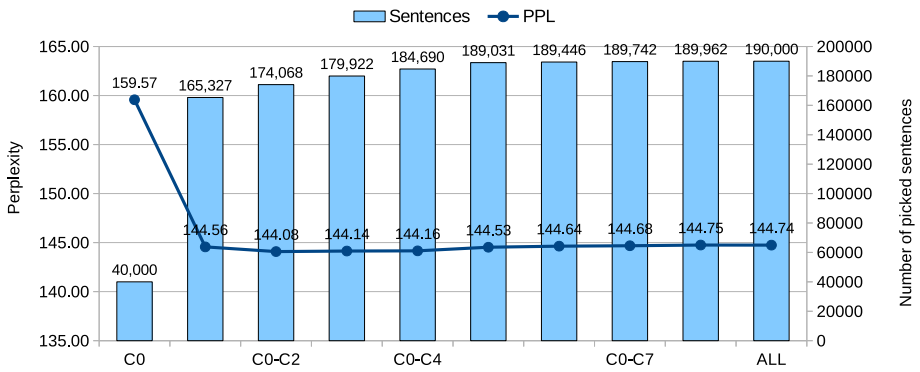

Fig. 4 Perplexities with accumulation of groups for CRF-based scoring

Table 6 The classification performance of style classifiers evaluated on LOTUS-DEV and Cell-DEV sets

Classifier	Precision (%)	Recall (%)	*F* score (%)	Accuracy(%)
SVM-linear	99.17	99.45	99.31	98.67
SVM-RBF	99.25	99.52	99.39	98.81
CRF	99.44	99.31	99.38	98.79
LSTM	98.99	99.75	99.37	98.66

(1) *ALL*: All sentences containing both "written" and "spoken."
(2) *Random*: A limited number of sentences randomly selected.
(3) *PPL*: Sentences selected by the perplexity-based method [4].
(4) *SVM*: Sentences selected by the SVM classifier. With this technique, 98K sentences are selected as spoken-style utterances.
(5) *CRF*: Sentences selected by the CRF classifier. In this case, 174K sentences are selected.
(6) *LSTM*: Sentences selected by the LSTM classifier. 151K sentences are chosen as spoken-style text.

As each selection techniques produces different numbers of utterances, for fair comparison, the number of sentences from *Random* and *PPL* cases are varied at 98K, 151K, and 174K as in *SVM*, *CRF*, and *LSTM* techniques, respectively.

For building each spoken LM, the CELL-TRN and the selected sentences from each method were adopted as described in Section 4.1. In the recognition step, the final LM is interpolated by these spoken LM to the baseline LM.

We evaluated our approaches in two different recognition tasks: Twitter posting (SOC) and VoiceTra4U-M speech-to-speech translation (VT). The evaluation data set is classified as either spoken or written utterances. It contained 4000 utterances from 5 speakers in the office environment taken from the LOTUS-SOC and 1916 utterances from the VT-TST of the VoiceTra4U-M mobile application.

5.2 Classification accuracy

To train a style classifier, we used the LOTUS corpus as a representation of written-style utterances and the LOTUS-Cell as a representation of spoken-style utterances. Every utterance in the LOTUS corpus was labeled as "written" while every utterance in the LOTUS-Cell corpus was marked as "spoken." Four thousand three hundred thirty utterances in the LOTUS-TRN set and 40,000 utterances in the Cell-TRN set, described in Section 2.2, were used to train classification models while 557 utterances in the LOTUS-DEV set and 15,475 utterances in the Cell-DEV set were used to evaluate the classifier performance.

For the LOTUS-DEV set, the classification accuracy is calculated from the number of utterances classified as "written." On the other hand, the accuracy of the Cell-DEV set is computed from the number of utterances classified as "spoken." The average classification performance in terms of precision, recall, *F* score and accuracy of both the LOTUS-DEV set and the Cell-DEV set are shown in Table 6.

You can see that *LSTM* classifier has the lowest precision value but its recall is the highest. The high value of recall indicates that *LSTM* classified the spoken utterance almost completely. However, many written sentences were also falsely classified as we can see from the low value of precision. On the other hand, we achieve the lowest recall and highest precision when using the *CRF* classifier. This indicates that a few written sentences were mixed selected, but not all the spoken utterances were selected. The results also demonstrated that both of the classification *F* scores and accuracies, the results of *SVM*, *CRF*, and *LSTM* classifiers are comparable. Therefore, in this work, we chose to compare multiple classification models, since we are interested to see how each classifier does in predicting the spoken degrees that are assigned to the utterances. For *SVM*, the RBF kernel has a slightly

Table 7 Perplexities of language models trained from mixed-style and spoken-style utterances evaluated on VT-TST and SOC sets

LM	PPL				
Base	146.71				
ALL	144.74				
LM (174K)	PPL	LM (151K)	PPL	LM (98K)	PPL
Random.174K	148.36	*Random.151K*	148.45	*Random.98K*	148.47
PPL.174K	144.91	*PPL.151K*	144.38	*PPL.98K*	139.56
CRF	144.08	*LSTM*	140.45	*SVM*	139.18

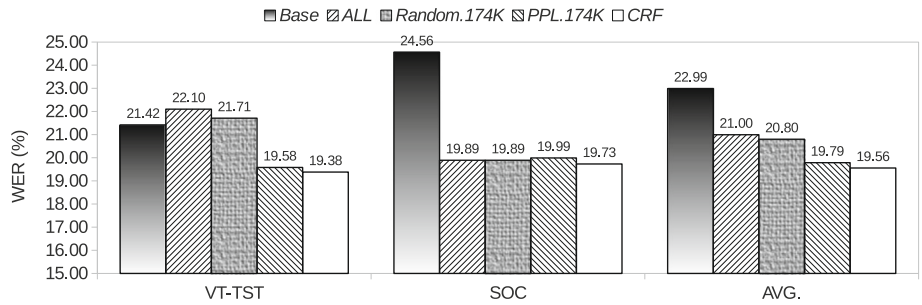

Fig. 5 Recognition performance (WER) using 174K selected sentences

better performance than the linear one. Therefore, the RBF kernelwas used to select spoken utterances for LM training in the latter experiment.

5.3 Language model perplexity

In this experiment, the performance of the style classifier was evaluated in terms of LM perplexity with respect to a known spoken-utterance test set (VT-DEV). 51.84, 91.61, and 79.93% of the utterances in the corpus were classified as spoken utterances by *SVM*, *CRF*, and *LSTM* classifiers respectively. We can see that the number of picked utterances of each classifier seem to be a difference due to our organized groups, shown in Table 5, which affect on the degree of spoken.

A trigram LM was trained by the SRILM toolkit [30] with modified Kneser-Ney discounting. Three LMs trained from spoken utterances selected by *CRF*, *LSTM*, and *SVM* classifiers respectively were interpolated with baseline LM (*Base*) and then evaluated on the VT-DEV set of spoken data. For comparison, we also investigated the LMs made with other selection methods, *ALL*, *Random*, and *PPL*, as described in Section 5.1. In cases of *Random* and *PPL*, we randomly selected 174K, 151K, and 98K utterances to train LMs to make the size of the training data comparable to each proposed method, *CRF*, *LSTM*, and SVM, respectively. These LMs were also interpolated with *Base* and then evaluated on the same VT-DEV set.

The perplexities are reported in Table 7. From this table, we can see that three interpolated LMs, *CRF*, *LSTM*, and *SVM*, trained from spoken utterances classified by *CRF*, *LSTM*, and *SVM* have lower perplexities than the *Base* and *ALL*. Moreover, comparing with *Random* and *PPL* on each different size of training data, the proposed selections also have lowest perplexity in all cases.

5.4 Recognition performance

In this section, we observed the recognition performance in terms of speech recognition word error rate (WER). The same evaluation sets, SOC and VT, were used in this experiment. For fair comparison, we also selected the utterance sizes of the *Random* and *PPL* cases at the same scale as each proposed methods. Figures 5, 6, and 7 show the recognition results of different data sizes.

From the results, we can see that all proposed classification-based selection methods effectively decreased the WER from those by *Base*, *ALL*, *Random*, and *PPL* in all test cases. In case of the VT test set, it is obvious that the *ALL* and *Random* selection methods deteriorate the performance of LM. However, in the SOC test set, the recognition results of *ALL* and *Random* selection methods are quite better than that of the *PPL*.

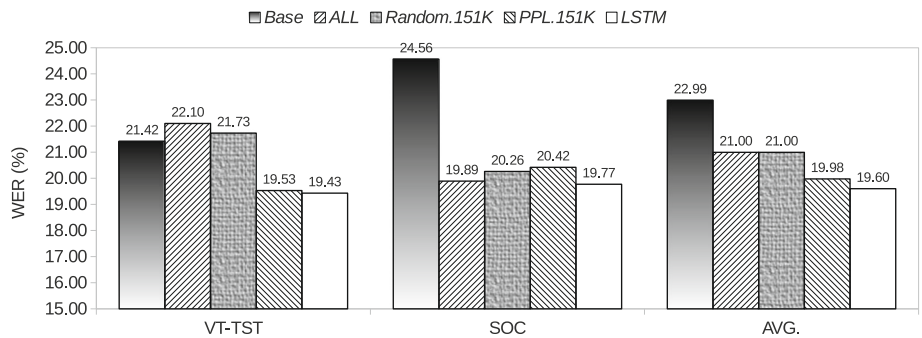

Fig. 6 Recognition performance (WER) using 151K selected sentences

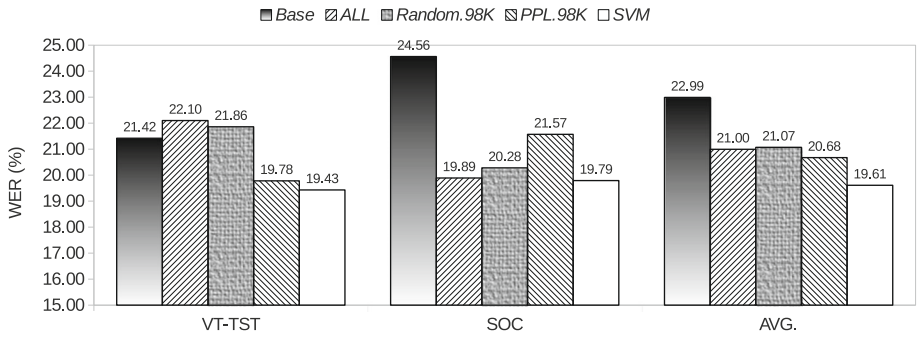

Fig. 7 Recognition performance (WER) using 98K selected sentences

The recognition results of our proposed techniques are also slightly improved compared to others. This might be due to the fact that we use a development set from the VoiceTra4U-M data (VT-DEV) to tune obtain an interpolation weight. However, the evaluation results on the SOC set demonstrate that the VT-DEV can be used even if in the open dataset. Compared with the *Base*, when no social media sentences were used, the average improvement with the proposed *CRF*, *LSTM*, and *SVM* were 3.44, 3.39, and 3.38%, respectively. With the increase of all social media data (*ALL*), which contains both "written" and "spoken" utterances, the average WER improvement of the proposed *CRF*, *LSTM*, and *SVM* became 1.44, 1.40, and 1.39%, respectively. This shows the fact that using a large amount of social media data from the Internet, without style-based classification, gives no benefit. Moreover, the *CRF*, *LSTM*, and *SVM* approaches achieved a reduction of 0.23, 0.38, and 1.07% in average WER over a conventional perplexity-based approach, respectively. It can conclude that the proposed techniques can obviously improve the selection of spoken-style data and still achieve slightly better recognition accuracies.

6 Conclusions

In this paper, we explored the possibility of using data from social media such as Twitter to augment the lack of large text corpora for LVCSR language modeling. The problem of mixed-style text, written- and spoken-like, in tweets was handled through our data selection approaches to determine spoken-like sentences for building LM in LVCSR. Three particular classification techniques were investigated to identify spoken-style sentences in a Twitter corpus; SVM, CRF, and LSTM. We trained each style classifier using both words and parts-of-speech as input features. With style classification, we were able to classify the spoken sentences based on output scores, of SVM or CRF.

For LSTM, spoken sentences were directly determined by the classifier.

The selected spoken-style text were used to construct a spoken LM, which was then interpolated with the baseline LM to build a final LM for the LVCSR system.

Our experiments showed that the LM constructed by our proposed techniques was efficient for conversational LVCSR as shown by the reduced LM perplexity and WER. Compared with the use of all data in the Twitter corpus, trigram language models trained from tweets selected by CRF, LSTM, and SVM methods achieved up to 0.66, 4.29, and 5.56% absolute perplexity improvement and 1.44, 1.40, and 1.39% absolute WER improvement, respectively.

In summary, it was confirmed that the proposed approach efficiently improved the selection of spoken-like sentences, and improved the LVCSR performance on spoken-style tasks.

In the future, we plan to use more features such as syntactic features to improve style classification. Moreover, more advanced classification methods will be investigated. Active learning can also be conducted to refine classification labels assigned to each sentence in the corpus. Approaches focusing on transforming written to cope with speaking disfluencies such as inserting filled pause, repetition, repair, and false start, have been proposed [31–33]. This idea is attractive and could be added after the sentence selection process in order to increase the degree of spoken-style of the corpus.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 22 May 2017 Accepted: 5 October 2017
Published online: 17 October 2017

References

1. E Egorova, JL Serrano, Semi-supervised training of language model on spanish conversational telephone speech data. Procedia Comput. Sci. **81**, 114–120 (2016)

2. S Novotney, R Schwartz, J Ma, in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference On*. Unsupervised acoustic and language model training with small amounts of labelled data (IEEE, 2009), pp. 4297–4300. https://scholar.google.co.th/scholar?hl=th&as_sdt=0%2C5&q=Unsupervised+acoustic+and+language+model+training+with+small+amounts+of+labelled+data&btnG=
3. K Yu, M Gales, L Wang, PC Woodland, Unsupervised training and directed manual transcription for lvcsr. Speech Commun. **52**(7), 652–663 (2010)
4. J Gao, J Goodman, M Li, K-F Lee, Toward a unified approach to statistical language modeling for chinese. **1**(1), 3–33 (2002). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Toward+a+unified+approach+to+statistical+language+modeling+for+chinese&btnG=
5. T Misu, in *Interspeech*. Kawahara: A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web text, (2006), pp. 9–13. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+bootstrapping+approach+for+developing+language+model+of+new+spoken+dialogue+systems+by+selecting+web+text&btnG=
6. RC Moore, W Lewis, in *Proceedings of the ACL 2010 Conference Short Papers*. Intelligent selection of language model training data, (2010), pp. 220–224. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Intelligent+selection+of+language+model+training+data&btnG=
7. A Axelrod, X He, J Gao, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*. Domain adaptation via pseudo in-domain data selection, (2011), pp. 355–362. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Domain+adaptation+via+pseudo+in-domain+data+selection&btnG=
8. A Sethy, P Georgiou, SS Narayanan, in *Proceedings of the Human Language Technologies (HLT) Conference*. Selecting relevant text subsets from web-data for building topic specific language models, (New York City, 2006), pp. 145–148. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Selecting+relevant+text+subsets+from+webdata+for+building+topic+specific+language+models&btnG=
9. A Jaech, M Ostendorf, Leveraging twitter for low-resource conversational speech language modeling. arXiv preprint arXiv:1504.02490 (2015). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Leveraging+twitter+for+lowresource+conversational+speech+language+modeling&btnG=
10. A Prasithrathsint, *Sociolinguistic Research on Thailand Languages. Language Sciences*, (1998). (https://www.sciencedirect.com/science/article/pii/0388000188900174)
11. A Chotimongkol, K Thangthai, C Wutiwiwatchai, in *Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA), 2014 17th Oriental Chapter of the International Committee for The*. Utilizing social media data through similarity-based text normalization for lvcsr language modeling, (2014), pp. 1–6. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Utilizing+social+media+data+through+similaritybased+text+normalization+for+lvcsr+language+modeling&btnG=
12. S Kasuriya, V Sornlertlamvanich, P Cotsomrong, S Kanokphara, N Thatphithakkul, in *Oriental COCOSDA*. Thai speech corpus for Thai speech recognition, (2003), pp. 54–61. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Thai+speech+corpus+for+Thai+speech+recognition&btnG=
13. A Chotimongkol, N Thatphithakkul, S Purodakananda, C Wutiwiwatchai, P Chootrakool, C Hansakunbuntheung, A Suchato, P Boonpramuk, in *Oriental COCOSDA Held Jointly with 2010 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2010 International Conference*. The development of a large thai telephone speech corpus: Lotus-cell 2.0, (2010). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=The+development+of+a+large+thai+telephone+speech+corpus%3A+Lotus-cell+2.0&btnG=
14. A Chotimongkol, V Chunwijitra, S Thatphithakkul, N Kurpukdee, C Wutiwiwatchai, in *Oriental COCOSDA Held Jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference*. Elicit spoken-style data from social media through a style classifier, (2015), pp. 7–12. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Elicit+spokenstyle+data+from+social+media+through+a+style+classifier.&btnG=
15. P Chootrakool, V Chunwijitra, P Sertsi, S Kasuriya, C Wutiwiwatchai, in *Oriental COCOSDA Held Jointly with 2016 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2016 International Conference*. Lotus-soc: A social media speech corpus for Thai lvcsr in noisy environments, (2016). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Lotus-soc%3A+A+social+media+speech+corpus+for+Thai+lvcsr+in+noisy+environments.&btnG=
16. V Sornlertlamvanich, N Takahashi, H Isahara, in *Proc. Oriental COCOSDA 1998*. Thai part-of-speech tagged corpus: ORCHID, (1998), pp. 131–138. http://www.academia.edu/1215347/ORCHID_Thai_part-of-speech_tagged_corpus
17. C Haruechaiyasak, S Kongyoung, in *in Proc. of SNLP*. Tlex: Thai lexeme analyser based on the conditional random fields, (2009). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Thai+lexeme+analyser+based+on+the+conditional+random+fields&btnG=
18. T Joachims, *Advances in kernel methods*, (1999), pp. 169–184. Chap. Making Large-scale Support Vector Machine Learning Practical. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Advances+in+kernel+methods+Joachims&btnG=
19. JD Lafferty, A McCallum, FCN Pereira, in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, (2001), pp. 282–289. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Conditional+random+fields%3A+Probabilistic+models+for+segmenting+and+labeling+sequence+data.&btnG=
20. S Hochreiter, J Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
21. K Thangthai, A Chotimongkol, C Wutiwiwatchai, in *INTERSPEECH*. A hybrid language model for open-vocabulary Thai LVCSR, (2013), pp. 2207–2211. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+hybrid+language+model+for+open-vocabulary+Thai+LVCSR&btnG=
22. M Yang, H Jiang, T Zhao, S Li, in *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006. Proceedings*. Construct trilingual parallel corpus on demand, (2006), pp. 760–767. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Construct+trilingual+parallel+corpus+on+demand&btnG=
23. G Kikui, E Sumita, T Takezawa, S Yamamoto, in *INTERSPEECH*. Creating corpora for speech-to-speech translation, (2003). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Creating+corpora+for+speech-to-speech+translation&btnG=
24. K Kosawat, M Boriboon, P Chootrakool, A Chotimongkol, S Klaithin, S Kongyoung, K Kriengket, S Phaholphinyo, S Purodakananda, T Thanakulwarapas, C Wutiwiwatchai, in *SNLP*. BEST 2009: Thai word segmentation software contest, (2009), pp. 83–88. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=BEST+2009%3A+Thai+word+segmentation+software+contest&btnG=
25. A Chotimongkol, K Saykhum, P Chootrakool, N Thatphithakkul, C Wutiwiwatchai, in *Oriental COCOSDA*. LOTUS-BN: A Thai broadcast news corpus and its research applications, (2009), pp. 44–50. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=LOTUS-BN%3A+A+Thai+broadcast+news+corpus+and+its+research+applications&btnG=
26. P Boonkwan, Part-of-speech tagging guidelines for Thai. National Electronics and Computer Technology, 1–34 (2012)
27. DP Kingma, J Ba, Adam: A method for stochastic optimization. CoRR. **abs/1412.6980** (2014). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+method+for+stochastic+optimization&btnG=
28. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, J Silovsky, G Stemmer, K Vesely, in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. The Kaldi speech recognition toolkit, (2011). https://infoscience.epfl.ch/record/192584
29. L Bahl, P Brown, P de Souza, R Mercer, in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86*. Maximum mutual information estimation of hidden markov model parameters for speech recognition, vol. 11, (1986), pp. 49–52. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Maximum+mutual+information+estimation+of+hidden+markov+model+parameters+for+speech+recognition&btnG=
30. A Stolcke, in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*. SRILM - an extensible language modeling toolkit, (2002), pp. 901–904. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=SRILM+-+an+extensible+language+modeling+toolkit&btnG=

31. R Schwartz, L Nguyen, F Kubala, G Chou, G Zavaliagkos, J Makhoul, in *Proceedings of the Workshop on Human Language Technology*. On using written language training data for spoken language modeling, (1994), pp. 94–98. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=On+using+written+language+training+data+for+spoken+language+modeling&btnG=
32. Y Akita, T Kawahara, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 4*. Topic-independent speaking-style transformation of language model for spontaneous speech recognition, (2007), pp. 33–36. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Topicindependent+speakingstyle+transformation+of+language+model+for+spontaneous+speech+recognition.&btnG=
33. R Masumura, S Hahm, A Ito, in *Interspeech*. Training a language model using web data for large vocabulary japanese spontaneous speech recognition, (2011), pp. 1465–1468. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Training+a+language+model+using+web+data+for+large+vocabulary+japanese+spontaneous+speech+recognition.&btnG=
34. S Burusphat, Speech analysis: Nakhonpathom discourse analysis. Research Institute for Languages and Culture for rural development Mahidol University (1994). http://e-book.ram.edu/ebook/t/TH103/chapter11.pdf
35. S Chodchoey, in *Proc. of the Second International Symposium on Language and Linguistics*. Spoken and written discourse in thai: The difference, (1998). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Spoken+and+written+discourse+in+thai%3A+The+difference&btnG=