Multilingual Stemming and Term extraction for Uyghur, Kazak and Kirghiz

Mijit Ablimit*, Sardar Parhat*, Askar Hamdulla*, Thomas Fang Zheng†
*Xinjiang University, Urumqi, China
Email: mijit@xju.edu.cn
†Tsinghua University, Beijing, China
Email: fzheng@tsinghua.edu.cn

***Abstract*— Stemming and term extraction is an important and difficult step on NLP for low resource languages. Inflectional structure and noisy data aggravate this problem for less popular agglutinative languages. A morphological analyzer with the longer context can utilize resources and provide reliable semantic and syntactic information, and effectively reduce ambiguity on stemming and term detection. There are some previous works on stemming on Uyghur texts based on simple morphology like affix and manually collected rules. But the limited information of smaller context and lower segmentation accuracy cost the reliability. We developed a sentence level multilingual morphological processing tool for Uyghur, Kazak, and Kirghiz languages. This tool can provide sentence level morpheme extraction with 98% accuracy, and further analysis like word embedding and longer context modelling can extract remaining unseen and infrequent stems reliably. Combined with word embedding this tool provides a more reliable way of term extraction from large number of noisy text available from internet.**

I. INTRODUCTION

Less resource, noisy data, and inflectional morphology make natural language processing (NLP) tasks difficult for Uyghur, Kazak, and Kirghiz languages. Sub-word units provide better coverage and improved performance in many NLP tasks, like ASR and MT for these languages [1-2]. The theoretically infinite vocabulary size can be separated into an open stem set and a closed affix set, and dramatically decrease the lexicon size to less than 1/3 of word lexicon. Stemming and term extraction works are heavily dependent on morphological analysis in these languages.

Previous works [3-5] are based on single word morphological analysis based on affix and some rules, and suffer from ambiguity and uncertainty. Only some manually collected rules, which are often elusive and uncertain, are used to locate a possible stem. Sentence level context information is ignored. And the specific volume ratio of notional stems is unclear in previous works. A particular problem is the acoustic harmony and disharmony which cause alteration in morphology and needs sentence level context analysis, and better be learnt automatically.

We develop a compact extendible tool for segmenting word sequences into morpheme sequences for these three languages. It is extendible in terms of both functions and languages. Based on an aligned word-morpheme parallel training corpus, this program will learn the various surface forms and their acoustic rules from the training data. Segmentation program will export all possible segmentation forms for each candidate. An independent statistical model can be incorporated to select the best result and N-best results. This toolkit provide basis for stemming, term extraction, and information retrieval tasks.

Longer context or sentence level analysis will utilize available resources, decrease the ambiguity and improve reliability. Extracting notional units: stems is dependent on other functional auxiliary elements. As functional parts also play vital important rule in language. Thus the unseen notional stems are reliably extracted based on context information.

In this paper, we develop a compact and extendable framework to improve minority language NLP. Our goal is to provide a standard interface to perform various NLP tasks for multiple minority languages. With this framework, the basic functions will be published, and developers can contribute using the same API. The present implementation includes text normalization, stemming, and morphological analysis. Note that some researchers have developed some tools for Uyghur language [6]. By incorporating tools like word embedding, stems and terms are easily extractible.

Our work is part of the Multilingual Minorlingual Automatic Speech Recognition (M2ASR), which is supported by the National Fundamental Science of China (NFSC). The project is a three-party collaboration, including Tsinghua University, the Northwest National University, and Xinjiang University. The aim of this project is to construct speech recognition systems for five minor languages in China (Tibetan, Mongolia, Uyghur, Kazak and Kirgiz). However, our ambition is beyond that scope: we hope to construct a full set of linguistic and speech resources and tools for the 5 languages, and make them open and free for research purposes. We call this the M2ASR Free Data Program. All the data resources, including the tools published in this paper, are released on the website of the project http://m2asr.cslt.org.

II. MULTILINGUAL MORPHOLOGICAL ANALYSIS AND STEMMING

A multilingual morphological processing tool is implemented for the three languages. The stems are independent semantic units and an open set while the suffixes are auxiliary functional units and a closed set. Because of this agglutinative nature, the number of words of these languages can be almost infinite, and most of the words appear very rarely in the text corpus. Modeling based on a smaller unit

like morpheme can provide stronger statistics hence robust models.

2.1 Sentence level multilingual morphological analysis

This tool is designed to reduce repeated programming as much as possible. There are roughly three parts in this framework as in Figure 1. The main decoder block utilizes learnt linguistic resources and independent statistical models to segment raw text into morphemes.

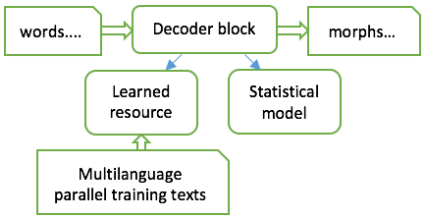

Figure 1 structure of multilingual morpheme segmenter tool.

(1) Learning block will learn stems, affixes, and tules from a word-morpheme parallel corpora. The parallel corpora are the aligned sentences of word sequence and morpheme sequence, separately prepared for each language. While the word-line provide surface forms, the morpheme-line provide standard forms. Thus the acoustic alterations and morphological changes are automatically learned from these aligned units. As the affix is closed set, which can be optionally prepared beforehand or automatically collected from training data with their various surface forms. Table 1 shows the prepared training corpora for these languages.

TABLE 1
MULTILINGUAL WORD-MORPHEME PARALLEL CORPORA

Language	Uyghur	Kazak	Kirghiz
word-morpheme parallel corpus (sentences)	10 000	5 000	3 000
suffix set (types)	124	124	124

For some applications, pseudo-morphemes are sufficient. For example, in speech recognition, pseudo morphemes can be used as the basic units for language modeling, without the necessity to obtain precise morpheme segmentations. In this case, the semi-supervised morpheme segmenter can provide a good trade-off between precision and efficiency.

(2) A language independent decoding block, which will generate all possible segmentation results with their standard forms for every word in a sentence, and incorporate an independent statistical model to rank them. First each word of a sentence is split into all possible combinations of stems and affixes. Then these candidate units are augmented with all possible ways, and ranked by their probability to from N-best results. To boost the speed, ranking and filtering is processed for every next candidate word until the end of sentence. An independent statistical model block incorporated to calculate and select N-best or the best results on a sentence level.

(3) An independent statistical model block. This block works on selecting best results from a bunch of candidate sequences. Text lines are segmented to all possible ways and sent to this block as unit ID numbers, so this block is language independent. Users can add any available model like N-gram, neural network etc.

Several constraints for the multilingual morpheme segmenter reduce uncertainty and improve reliability. First, the suffix set is a closed set, and all the surface forms are known or learnt. Second, for each language a parallel corpus of aligned word-morpheme sequences is prepared. And an optional stem list can be added. These requirements ensure the quality of the learning and are mostly reasonable in practice.

2.2 Implementations stemming and term extraction

Reliable stemming and term extraction is critically important for low resource language, especially when the raw corpora are often noisy. Table 2 shows some ambiguous examples which can only be disambiguated by longer context measures like a sentence.

TABLE 2
EXAMPLES OF ABIGUOUS PARTICLES

variants (English)	Variants (Arabic/Latin)	Suffix
bAr(give) / bar(go)	bar/bAr → ber+iN	iN
person's name/stand up	tur → turdi, tur+di	di
shoot/horse	at→ at+ti, at+tAk	ti, tAk

Stemming and term extraction have to deal with a lot of unseen and low frequent stems not words for inflectional languages. Mostly, these low frequent notional stems are newly coined and uncertain in terms of spelling and often combined with suffixes.

Recently, deep neural network and representation learning [7] provided better efficient ways of text representation and possibility for solving the problem of data sparsity. Mikolov et al. [8] proposed a text representation method, and used the idea of deep learning and vector operations to simplify the processing of text content into N-dimensional vector space by training, and used the similarity of vector space to represent semantic similarity of text, and seeks a deeper level of feature representation for the text data. This method greatly simplify context measuring especially for stems and term extraction tasks of inflectional languages.

First, we select some low frequent nouns which have similar context measure as unseen stems, and automatically augment them by repeated using of word-embedding method in the training parallel corpora. Then collect most frequent common suffixes as the feature of possible affix for an unseen candidate word. Because only certain affixes are combined with the certain terms. Except from these suffix feature, sentence level context also provides reliable extraction. Figure 2 shows the flow chart.

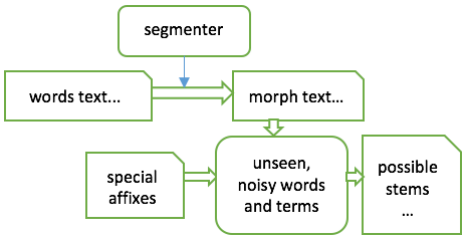

Figure 2 flow char of multilingual stemming.

Word embedding group some similar and related but slightly different terms in spelling. Our multilingual text processing tool provide syllable and acoustic analysis to determine possible noisy expressions of the same stem, and try to correct these expressions.

Morpheme based training text automatically group related terms by using word embedding method much efficiently compared to word units. As the morphological analyzer split words with 98% accuracy, almost all unseen words are naturally filtered out as new stems or out of vocabulary words (OOV).

III. Implementations and Corpus Construction

A multilingual morphological processing tool is implemented for the three languages. The tool has been tested on Uyghur, Kazak, Kirghiz, and the results show that many spelling mistakes are detected by this tool, most detections are new stems. When the morphemes are merged to a word, the phonemes on the boundaries change their surface forms according to the phonetic harmony rules. Morphemes will harmonize each other, and appeal to each other's pronunciation. When the pronunciation is precisely represented, the phonetic harmony can be clearly observed in the text. We train a statistical model using a word-morpheme parallel corpus. Best segmentation result is obtained for Uyghur language with 98% accuracy in terms of morpheme numbers.

Word embedding produce real value vectors, and we can easily find word similarity by calculating the distance between any of the two given embedding word or stem vectors. We can train word vectors quickly and efficiently by using word2vec toolkit.

3.1 Implementations on stemming and text classification

We segmented 630k general topic text into morphemes and extracted all the stems by excluding 130 types of affixes and some functional "stop words" of more than thousand. The stop words are collected by part manually and part automatically using word embedding method. Statistics are shows in Table 3. We can see that the stem vocabulary size decreased to 34% of the word vocabulary size.

Table 3
VOLUME REDUCTION IN STEMS COMPARED TO WORDS

Sentence number	word vocab.	morph. vocab.	stem vocab.	stem/word ratio
630K	466K	160k	158.5K	34%

We compiled a small text classification test by collecting text corpora from internet. Web crawler downloaded texts from official Uyghur language wed sites such as www.tianshan.com and www.Uyghur.people.com. This corpus includes 3 categories: law, finance, and sports; each category contains 500 texts, total 1500 texts and 13K sentences. We used 75%, them as training corpus, and used the rest part as the test corpus. The extracted stems used for CHI-2 feature extraction is 27K. When applying SVM classifier, the accuracy is 95.4%. Should be noted that the stemming approach proposed in this paper enable us to obtain an excellent result in feature dimension reduction.

3.2 Implementations on term extraction

Based on the 630K sentences, word embedding method is used for term extraction for numbers and people names etc. Some examples are shown in Table 4. Some terms as "number" and "title" are surprisingly same terms. While for term like "person names" and "city names" some relevant terms also extracted with some spelling mistakes. And these terms are extracted in simple unsupervised way from a single general text corpus without specific domain knowledge.

Table 4
EXAMPLES OF TERM EXTRACTION

number	person names	city names	title
bir	fAruq	xinjaN	valim
vikki	pirvAwini	juNgo	pAylasop
vUc	mutAwAkkil	tANritaG	vAdib
tOt	qaligon	mAmlikAt	mutApAkkur
nAccA	vimpriyisiniN	HowtikiGa	vOlima

More sophisticated term extraction methods can be designed by comparing and filtering extracted terms. Sentence embedding methods also help to extract more patterns and terms with more training data. But for small training data set simple sentences embedding method did not help much.

IV. Conclusion

Reliable stemming for low resource language with noisy data is vital important for various aspects of NLP. In the paper, we discuss a multilingual morphological analyzer framework for three agglutinative languages. Comparative statistics between word and morpheme units has shown a dramatic decrease in stem lexicon size. For the text classification task, the notional stems are reduced to 1/3 of word lexicon. The sentence level context analysis not only reduce ambiguity, but also separate the noisy and unseen word for further analysis.

We have discussed the system structure of our recent public tools used for multilingual phonetic and morphological processing. Originally, these tools were constructed to facilitate the corpus construction for less popular languages, but they can be used for minority NLP in general. This tool was designed by separating the program from the data. Users can easily use their own language-specific data. We hope these tool will provide a uniformed multilingual information processing platform and assist multiple speech processing

tasks for minority languages, for example multilingual speech recognition.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC; grant 61462085, 61662078, and 61633013).

REFERENCES

[1] M. Ablimit, T. Kawahara, A. Hamdulla, "Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language," Speech Communication, 2014.5.

[2] M.Y. Tachbelie, S. T. Abeta, L. Besacier, "Using different acoustic, lexical, and language modeling units for ASR of an under-resourced language – Amharic," Speech Communication, 2013.1.

[3] Seyyare Imam, Rayilam Parhat, Askar Hamdulla, Zhijunli, performance analysis of different keyword extraction algorithms for emotion recognition from Uyghur text, IEEE 2014

[4] Palidan Tuerxun,Fang Dingyi, Askar Hamdulla. The KNN based Uyghur Text Classification and its Performance Analysis, International Journal of Hybrid Information Technology, Vol.8, No.3(2015).pp.63-72

[5] Zhou, Xi. Text classification model of Uyghur based on improved Bayes[J]. Journal of Computational Information Systems, v 9, n 11, p 4319-4327, June 1, 2013

[6] Mijit Ablimit, Sardar Parhat, Askar Hamdulla, Thomas Fang Zheng. "Multilingual Language Processing Tool for Uyghur, Kazak and Kirghiz". (APSIPA ASC), 2017

[7] Bengio Y, Schwent H, Senécal J S, et al. Neural Probabilistic Language Models [M]// Innovations in Machine Learning. Springer Berlin Heidelberg, 2006

[8] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.