Morphological Analysis Corpus Construction of Uyghur

Gulinigeer Abudouwaili[1,3], Kahaerjiang Abiderexiti[1,3], Jiamila Wushouer [1,3,*], Yunfei Shen[2,3], Turenisha Maimaitimin[1,3] and **Tuergen Yibulayin[1]**
[1]School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[2]School of Software, Xinjiang University, Urumqi, Xinjiang 830091, China
[3]Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China
{107556518131, shenyunfei}@stu.xju.edu.cn, {jiamila, turgun}@xju.edu.cn,
kaharjan@aliyun.com, turanisa11@163.com

Abstract

Morphological analysis is a fundamental task in natural language processing, and results can be applied to different downstream tasks such as named entity recognition, syntactic analysis, and machine translation. However, there are many problems in morphological analysis, such as low accuracy caused by a lack of resources. In this paper, to alleviate the lack of resources in Uyghur morphological analysis research, we construct a Uyghur morphological analysis corpus based on the analysis of grammatical features and the format of the general morphological analysis corpus. We define morphological tags from 14 dimensions and 53 features, manually annotate and correct the dataset. Finally, the corpus provided some informations such as word, lemma, part of speech, morphological analysis tags, morphological segmentation, and lemmatization. Also, this paper analyzes some basic features of the corpus, and we use the models and datasets provided by SIGMORPHON Shared Task organizers to design comparative experiments to verify the corpus's availability. Results of the experiment are 85.56%, 88.29%, respectively. The corpus provides a reference value for morphological analysis and promotes the research of Uyghur natural language processing.

1 Introduction

Morphological analysis is the process of dividing words into different morphologies or morphemes and analyzing their internal structure to obtain grammatical information. It is an essential step in lexical analysis. Therefore, Straková et al. [2015] showed that the related research of morphological analysis has also attracted the attention of most researchers. In natural language processing based on deep learning, researchers have found that if modeling to words directly, it is easy to ignore the relationship within the words, which will also bring limitations to the model. Thus, the model's input has changed from a single word to a character, subword, and morpheme or morphology. When the word is split into different granularities (Zhuang et al. [2018], Üstün et al. [2019], Zhu et al. [2019]), the performance of the model is improved. Among these words' segmentation methods, morphology or morpheme relies on the morphological analysis tagger or manual annotation. In languages with mature natural language processing technologies such as English, Finnish, and Chinese, there are many manually annotated morphological analysis corpus, and the lexical analysis (including morphological analysis) technology of these languages has reached a high level. To further promote the lexical analysis of minority languages in Xinjiang, many researchers have obtained preliminary results in lexical analysis. Ibrahim and Baoshe [2011] constructed a Uyghur lexical tagging corpus of 1.23 million words; Enwer et al. [2015] built a Uyghur stemming corpus of 10,000 sentence sets; Altenbek et al. [2014] annotated about 30,000 Kazakh sentences and studied a lexical analysis of Kazakh; Osman et al. [2019] constructed a character-level Uyghur morphological collaborative word corpus and annotated morphological analysis for 3500 sentences; for the first time, Abudukelimu et al. [2018] released the Uyghur language based on morpheme sequence for morphological segmentation corpus; the corpus includes about 20,000 Uyghur words, including word-level and sentence-level corpus; based on the statistics of Uyghur noun affixes, Munire et al. [2019]

*:Corresponding author

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

proposed a hybrid strategy for the Uyghur noun morphological Re-Inflection model; Maimaitiming et al. [2020] constructed a small-scale stemming corpus for Uzbek, which includes 7435 words and 568 sentences in total.

Up until now, no reports have been published about the publicly available Uyghur of morphological analysis corpus.In addition, because there is no unified reference standard, the private morphological analysis datasets created by previous studies have fewer or incomplete features.This is not only directly affecting the development of the language's morphological analysis technology but also indirectly affecting the performance of downstream tasks, such as named entity recognition Güngör et al. [2019], syntactic analysis Vania et al. [2018], text classification Parhat et al. [2019], and machine translation Bisazza and Tump [2020]. Therefore, to alleviate the shortage of Uyghur language morphological analysis resources, this paper refers to the format of universal morphological feature schema (UniMorph Schema), analyzes Uyghur words from 14 dimensions and 53 features, constructs a Uyghur language morphological analysis corpus, and provides lemma, part of speech. morphological analysis tags, morphological segmentation, and lemmatization. This can be used for various lexical analysis tasks, like morphological analysis, stemming, and other downstream tasks of natural language processing.

2 Related Work

2.1 Agglutinative Language and Morphological Analysis

The world's languages can be roughly classified into four types YE and XU [2006],based on their morphology: isolated language, agglutinative language, fusional language, and polysynthetic language. The main feature of agglutinative language is that there is no inflection inside the word. A word is composed of several morphemes, and a morpheme is the smallest grammatical unit Abudouwaili et al. [2019]. According to the different roles of morphemes in words, they are divided into root and affix. Among them, affix can be subdivided into word-forming affix and inflectional affix. Root and word-forming affix can form new words; the inflectional affix to words can change the grammatical category. The following will take Uyghur as an example:

ئوقۇ (read)	+	غۇچى (morphological affix)	=	ئوقۇغۇچى (student)
ئوقۇغۇچى (student)	+	لار (configuration affix)	=	ئوقۇغۇچىلار (students)

In the example, " ئوقۇ " is the root of the word, combined with the word-forming affix "غۇچى" will form a new word "ئوقۇغۇچى", and then combined with the plural affix "لار" will change the grammatical category and meaning of "ئوقۇغۇچىلار" change from a student to students.

Morphology Vania [2020] refers to the study of the internal structure of words and how they are formed. It refers to recognition of words' lemma Jurafsky and Martin [2000], part of speech (POS), and morphological features of a word. Lemma refers to the condition where the word is not connected to affixes. For example, in English, words "write", "writes", "written" and "writing", all have the same lemma "write"; POS is defined according to syntactic function and morphological function. If words have a similar syntactic function, they can appear in similar contexts. Or they have affixes with similar morphological functions, and then they can be classified into one category. The morphological feature of a word is the grammatical categories (related grammatical information) attached to the lemma, such as number, case, tense, aspect, mood, person, and so on.The example is shown in Table 1.

Word	shadows
Lemma	shadow
POS	V (verb)
Morphological Analysis	shadow;V;SG;3;PRS

Table 1: Lemma, POS, Morphological analysis for "shadows"

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.

2.2 Morphological Analysis Corpus

The corpus is integral part of natural language processing tasks and is a collection of written or spoken material stored ZONG [2013]. Early corpus research dates back before the 1950s, but its rise and development period can trace on the 1980s. In 1964, W. Nelson Francis and Henry Kučera of Brown University released the first machine-readable corpus and the first parallel corpus: the Standard Corpus of Present-Day Edited American English (the Brown Corpus) [0]. However,the Brown corpus was built early and roughly, it has always been the standard for English parallel corpus. Geoffrey Leech and Stig Johansson released the original version of the Lancaster Oslo / Bergen corpus (LOB corpus)[1] and the annotated POS version[2] in 1976 and 1986, respectively. In 1993, the University of Pennsylvania released the Penn Treebank (PTB) Marcus et al. [1993], which mainly annotated POS and syntactic component analysis; NEGRA corpus[3] is a German syntax annotated corpus constructed by Saarland University in Germany, and the second version has been released. There are about 350,000 words(20,603 sentences), mainly annotated POS, MSDs(morphosyntactic descriptions), the grammatical function in the directly dominating phrase, and the category of nonterminal nodes edge labeling. In 2013, Jan Hajič[4] released the Czech Morphological Dictionary, a spelling checker and lemmatization dictionary. The description contains not only traditional morphological categories but also some semantic, stylistic, and derivational information. Almaty Corpus of Kazakh language[5] is an online corpus containing about 40 million words. The corpus texts were marked utilizing the automatic morphological analyzer, 86% of word forms of the corpus were parsed. In terms of other agglutinative languages, Finnish has also released different dependency treebanks, which also annotate the morphological tag of words. For example, Turku Dependency Treebank (TDT) [6] and Finn Treebank (FTB), etc. TDT collects more than 10,000 sentences and about 180,000 words in different fields, respectively annotating lemma, POS, and MSDs; there are three versions of the FTB dependency treebank, the first version FTB1 is a manually annotated corpus, including 19,000 sentences or sentence fragments and about 160,000 words (including punctuations), and can be provided Online services. The corpus is mainly marked with lemma, POS, MSDs, dependency relationship, and sentence component analysis; the second version FTB 2 improves the first version, sentences and words are more than the previous version. For these, uses the same labeling format as the first version is adopted, the sentence components are manually annotated, and the MSDs uses three different analyzers, and finally, the results are manually verified; the third version, FTB 3, contains about 4.36 million sentences and about 76.36 million words. It has the functions of automatic morphology and dependent syntax analysis. The main difference between TDT and FTB is that Silfverberg et al. [2016] FTB includes various grammatical examples, while TDT is more of daily Expressions.

In 1994, the Association for Computational Linguistics (ACL) established the Special Interest Group on Computational Morphology and Phonology group (SIGMORPHON)[7] and regularly held morphology-related share tasks to promote further the basic research of natural language processing, which mainly models to words, lemma, and MSDs. From the share tasks in the past five years, it is found that most of their data sets are provided by Universal Morphology (UniMorph) and Surrey Morphology Group; among them, UniMorph is used widely. This corpus, published by the Center for Language and Speech Processing (CLSP) of Johns Hopkins University, currently contains a morphological analysis corpus of 118 languages and annotated more than 23 dimensions of meaning with over 212 features, and UniMorph3.0 McCarthy et al. [2020] released in May 2020 is by far the largest high-quality morphological analysis corpus. Also, other share tasks selected the Universal Dependencies (UD) as datasets. UD provides 92 languages of POS, MSDs, and syntactic dependencies. It is an improvement based on Universal Stanford dependencies, Google universal part-of-speech tags, and the Interset interlingua

[0] http://icame.uib.no/brown/bcm.html
[1] http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM
[2] http://korpus.uib.no/icame/manuals/LOBMAN/INDEX.HTM
[3] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
[4] http://ufal.mff.cuni.cz/morfflex
[5] http://web-corpora.net/KazakhCorpus/search/index.php
[6] https://bionlp.utu.fi/fintreebank.html
[7] https://sigmorphon.github.io/

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.

for morphosyntactic tagsets, using CoNLL-X Buchholz and Marsi [2006] format to annotate each word (CoNLL-X format mainly includes word ID, word form or punctuation, lemma, UPOSTAG, XPOSTAG, morphological features FEATS, the central word HEAD of the current word, dependency relationship DEPREL, second-level dependency (head-deprel pair) DEPS and others label MISC). Since the corpus's existence related to morphological analysis, researchers have also released many morphological analyzers, such as Morfessor[8], UDPipe [9], Omorfi, and MorphoDiTa [10]. Morphological analysis corpus currently available for Uyghur, Kazakh, and Uzbek languages. But there are some problems of those languages' corpus, such as the corpus has a one kind of POS, a small amount of data, or incomplete tags, and so on.

3 Construction of Uygur Morphological Analysis Corpus

This chapter will introduce the construction process of the Uyghur morphological analysis corpus. Firstly, a more suitable morphological annotation guidelines is proposed based on Sylak-Glassman et al. [2015] and Tuohuti [2012]. Secondly, the collected dataset is preprocessed. Finally, considering the context, the whole dataset is annotated by human-machine interaction.

3.1 Data Preparation and Preprocessing

We crawled Uyghur news articles from TianshanNet[11] and NurNet[12] as raw corpus and preprocesses them. It is including finance, lottery, technology, tourism, society, fashion, sports, entertainment, and other fields. The preprocessing mainly includes removing web page tags, to punctuate sentences, and word segmentation, and finally selecting 5014 sentences with high quality. These 5014 sentences are used to construct the morphological analysis corpus. These sentences are used to construct the morphological analysis corpus.

3.2 The Annotation Scheme

This paper makes a more fine-grained morphological analysis and annotation of each word based on the previous annotation guidelines. The corpus' annotation scheme will be explained below: the annotation takes sentences as the context environment and words as the basic unit. The annotation mainly includes the current word, lemma, POS, MSDs, morphological segmentation, and lemmatization. The grammatical categories and morphological tags are shown in Table 2.

- Lemma: give the valid word;
- POS: mainly divided into noun, adjective, numeral, classifier, adverb, pronoun, imitation word, verb, postposition, conjunction, particles, interjection, auxiliary verb, in addition to punctuation, additional ingredients, and Latin;
- MSDs: including POS and grammatical features. The grammatical features are expressed through inflectional affixes, mainly including plural affix, possession affixes, case affix, voice affix, aspect affix, negative affix, masdar affix, participle affix, converb affix, tense affix, and mood affix;
- Morphological segmentation: according to different grammatical features, each type of affix is segmented in a fine-grained way;
- Lemmatization: restore the affix obtained by morphological segmentation to obtain a valid affix.

When formulating tags, this article refers to the morphological tags proposed by Sylak-Glassman et al. [2015]. However, these tags are universal tags and do not include special grammatical features. Therefore, new tags are also made for grammatical features that are not in UniMorph.

[8]http://morpho.aalto.fi/projects/morpho/
[9]https://bnosac.github.io/udpipe/en/
[10]http://ufal.mff.cuni.cz/morphodita
[11]http://uy.ts.cn/
[12]https://www.nur.cn/index.shtml

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.

Grammatical Features	Morphological Tags
Parts of speech	N, ADJ, NUM, CLF, ADV, PRO, IMI, V, ADP, CONJ, PART, INTJ, AUX, Y, X, LW;
Case	NOM, GEN, ACC, ALL, LOC, ABL, LQ, LMT, SML, EQUI;
Possession	PSSPNO;
Person	1, 2, 3;
Number	PL, SG
Aspect	PROG, PROSP,ABIL, ITER,DISP,SELF;
Voice	CAUS, PASS, REFL, PECP;
Polarity	NEG;
Participle	PTCP;
Masdar	MSDR;
Converb	CVB;
Tense	PST, NPST;
Mood	IND, INT, COND, IMP, OPT;
Politeness	POL;

Table 2: Grammatical features and morphological tags

3.3 Corpus Construction Process

After preprocessing the data crawled from news websites, a dataset based on sentences is obtained. To reduce manual annotation workload, POS tagging Maimaiti et al. [2017], and stemming Abudouwaili et al. [2019] are performed on the dataset. Since POS used in this article is not the same as the first-level part-of-speech proposed by Maimaiti et al. [2017], the tagging set is modified. The morphological analysis corpus does not provide stemminng result, but it is valuable for constructing the corpus.

sentence	ساناق مەلۇماتتا سىنا توربنىك پېرىۋوت ساناق مەلۇماتى ئاساس قىلىندى .				
Translation	The data is based on Sina's exchange data .				
word	stem	POS	MSDs	stemming	lemmatization
ساناق	ساناق	ADJ	ADJ;SG;NOM	ساناق	ساناق
مەلۇماتتا	مەلۇمات	N	N;SG;ON	مەلۇمات+تا	مەلۇمات+تا
سىنا	سىنا	N	N;SG;NOM	سىنا	سىنا
توربنىك	تور	N	N;SG;PSS3S;GEN	تور+ى+نىڭ	تور+ى+نىڭ
پېرىۋوت	پېرىۋوت	N	N;SG;NOM	پېرىۋوت	پېرىۋوت
ساناق	ساناق	ADJ	ADJ;SG;NOM	ساناق	ساناق
مەلۇماتى	مەلۇمات	N	N;SG;PSS3S;NOM	مەلۇمات+ى	مەلۇمات+ى
ئاساس	ئاساس	N	N;SG;NOM	ئاساس	ئاساس
قىلىندى	قىل	V	V;PASS;PST;3;SG;IND	قىل+ىن+دى	قىل+ىن+دى
.	_	Y	Y	_	_

Figure 1: An example of annotation

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.

After the labeled data are initially obtained, the data is manually annotated. Three students took part in the manual annotation. First, students should understand annotation scheme. Secondly, they conduct small-scale annotation and check each other. Then refer to the annotation scheme, they discuss the inconsistencies or ambiguous annotations and achieve a consensus. Finally, annotate the data in batches. To ensure the consistency of the annotated data, after the annotation, each student checks the annotation result of the other two students and submits the annotated data after it is true. An example of annotation is shown in Figure 1.

4 Corpus Information Statistics and Evaluation

The statistical distribution of words and sentences in each domain is shown in Figure 2, which including 5014 sentences, 152669 words, 24631 tokens. The average word repetition rate is 1:6, and it can be found that each sentence contains an average of 30 words. Compared with other categories, the number of news in finance is most, sport news has the largest number of words, and entertainment news has more tokens. Figure 3-(a) and Figure 3-(b) respectively represent the distribution of sentence length and word length in the corpus. The abscissa represents the length of the statistics, and the ordinate represents the number of statistics. Figure 3-(a) sentence length statistics figure, most of the sentence length is in the (40,140), the shortest sentence length is 3, the longest is 195. Figure 3-(b) word length statistics figure, most of the word length is (1, 10), the shortest word length is 1, and the longest word length is 33.

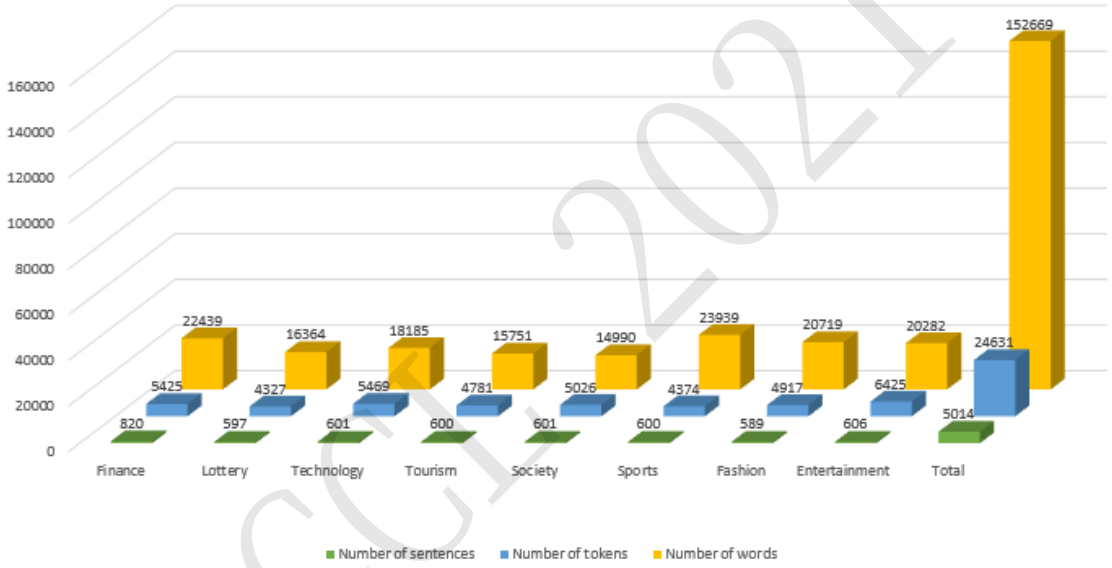
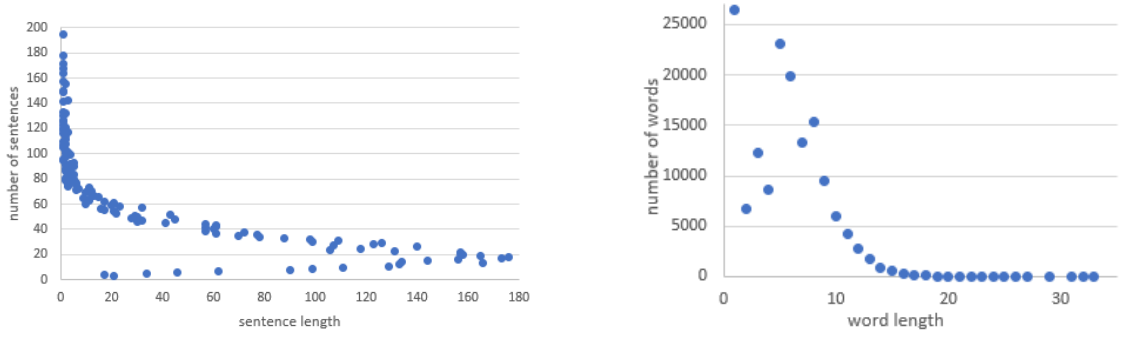


Figure 2: The number of words and sentences in each field

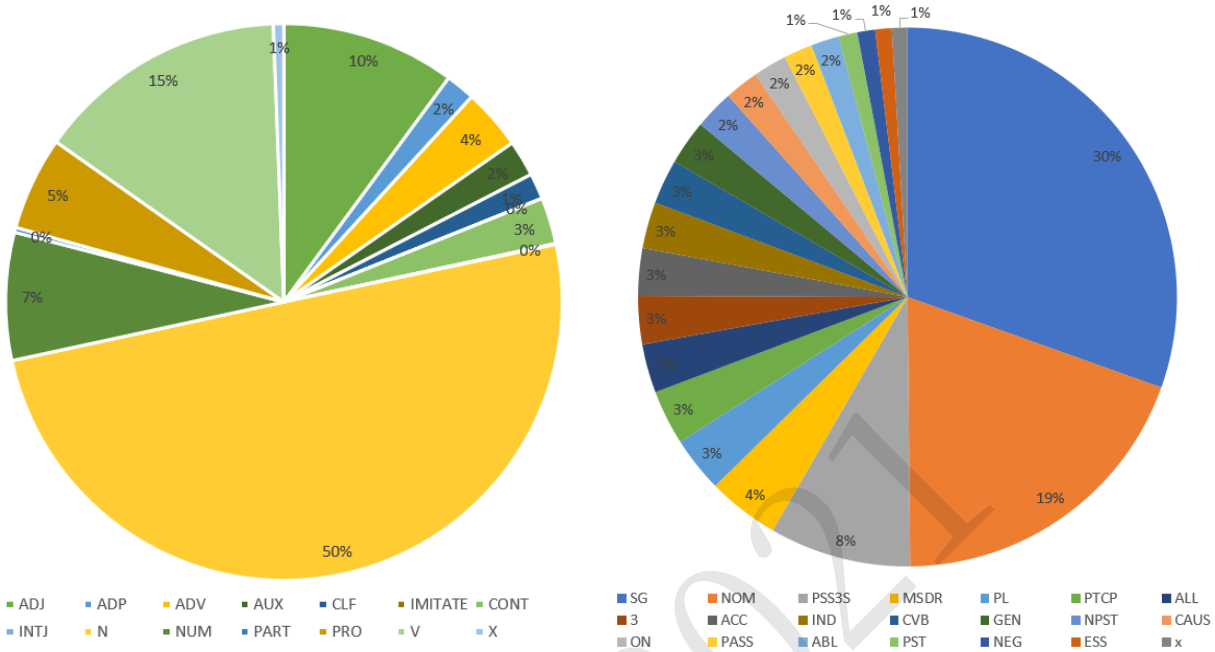


(a) sentence length statistics

(b) word length statistics

Figure 3: Sentence and word length statistics

In general, a sentence contains at least one verb and several nouns, and there are a small number of conjunctions or interjections, etc. Each type of word has a different grammatical category, and as the number of words increases, its grammatical features will increase. According to the morphological labels, the POS distribution, morphological feature distribution, and morphological label distribution of some words are respectively counted, as shown in Figure 4-(a), Figure 4-(b), and Figure 5.



(a) POS distribution

(b) morphological feature distribution

Figure 4: POS and morphological feature distributions

From the part of speech distribution in Figure 4-(a), we found that the proportions of noun (N), adjective (ADJ), pronoun (PRO), verb (V), and auxiliary verb (AUX) are more than the proportions of classifier (CLF), interjection (INTJ) and imitate (I). From the distribution of lexical features in Figure 4-(b) and the morphological distribution label in Figure 5, we observed that the proportion of grammatical labels that modify nominal words is the largest, such as SG, NOM, and PSS3S, etc., followed by verbs or auxiliary verbs, such as MSDR, PTCT, and so on. This statement can be further verified from Figure 5. Also, it proves that the morphology of verbal words is more complex and richer than nominal words.

To verify the effectiveness of the morphological analysis dataset constructed in this paper, we select part of the agglutinative or morphologically complex languages and use baseline models provided by the SIGMORPHON Shared Task. And we design a comparative experiment.

Evaluation task definition Given the word, lemma, and morphological label to train model, so that the model predicts the word that by the lemma and morphological labels, such as given the lemma “shadow” and morphological label “V;SG;3;PRS” to predict words “shadows”.

The experiment uses two modes provided by the task organizers, a non-neural baseline Cotterell et al. [2017] and a neural baseline Vaswani et al. [2017]. The neural baseline is a multilingual transformer. The version of this model adopted for character-level tasks currently holds the state-of-the-art on the 2017 SIGMORPHON shared task data. The transformer takes the lemma and morphological tags as input and outputs the target inflection. The non-neural baseline heuristically extracts lemma-to-form transformations; it assumes that these transformations are suffix-or prefix-based. A simple majority classifier is used to apply the most frequent suitable transformation to an input lemma, given MSDs, yielding the output form.

Table 3 shows the model performance of the Czech, Polish, Russian, Sakra, Eibela, Hebrew, and Uyghur datasets on the SIGMORPHON Share Task, with the first six languages provided by the organiz-

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076–1086, Hohhot, China, August 13 – 15, 2021.

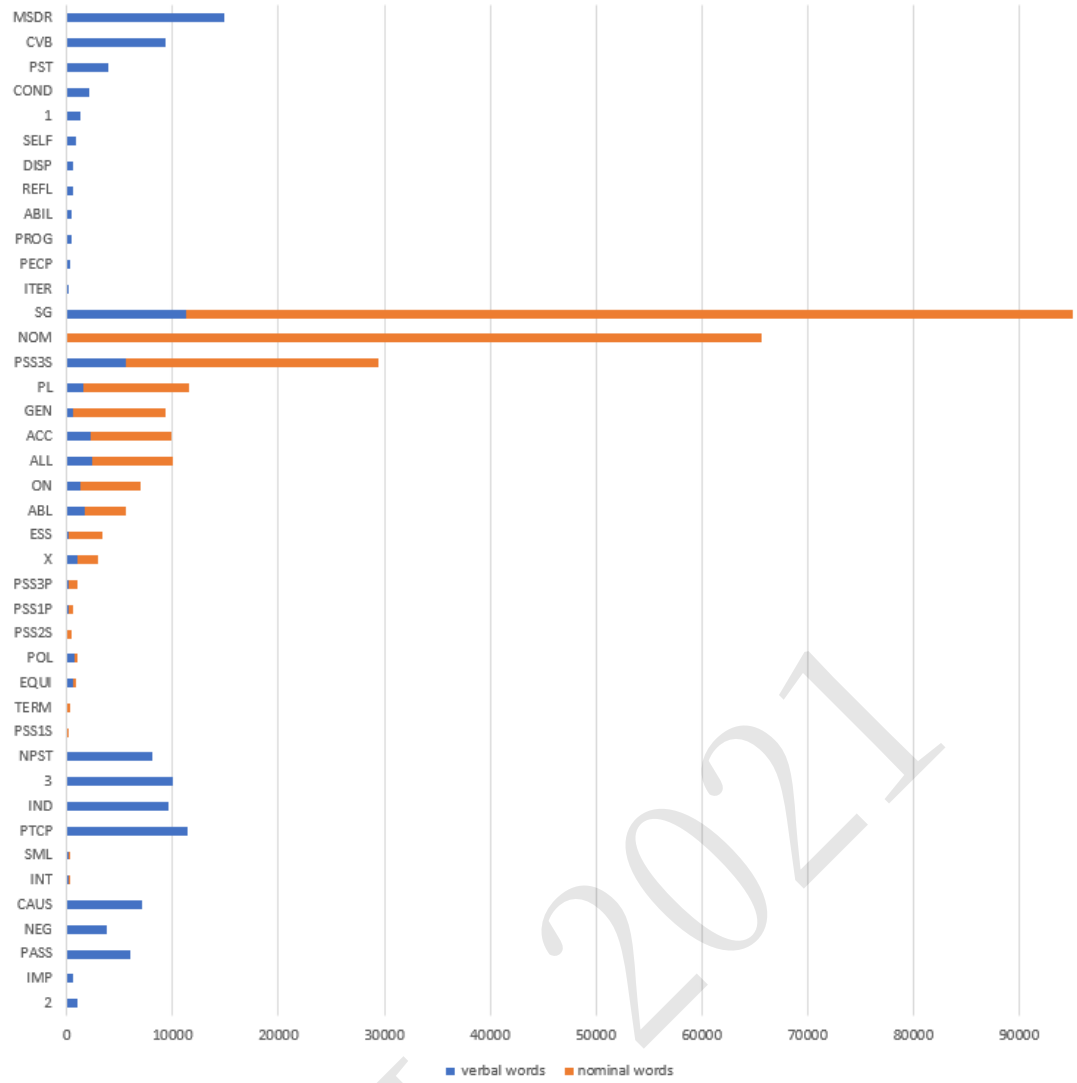


Figure 5: MSDs distribution

ers, the last two datasets constructed in this paper, and nominal words in the dataset filtered out to build the Uyghur nominal words dataset. The number of training sets, test sets, and development sets is Czech datasets (94169:6659:6659), Polish datasets(100039:8023:6023), Russian datasets (100002:7104:7104), and Sakra datasets (100046:6971:7098), Eibela dataset (918:64:66), Hebrew dataset (23204:1640:1642), Uyghur ALL Word dataset (19704:2463: 2463) and Uyghur nominal words dataset (13652:1706:1706).

Languages	Non-Neural Baseline	Neural Baseline	
	Test	Test	Development
Czech	92.81%	96.95%	96.92%
Polish	94.13%	99.54%	99.33%
Russian	88.72%	96.33%	96.34%
Sakha	90.54%	95.51%	95.49%
Eibela	4.68%	4.68%	13.63%
Hebrew	36.16%	99.02%	99.09%
Uyghur(ALL)	85.56%	88.29%	94.00%
Uyghur(nominal words)	89.25%	92.77%	96.00%

Table 3: Result of experiments

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.

Experimental results can be found that 1) the performance of the neural model is superior to the performance of the non-neural model, such as Czech, Hebrew, and Uyghur, which are respectively 4.14%, 62.86%, and 2.76% higher than the non-neural model. 2) In the languages with large datasets, such as Czech, Polish, Russian, and Sakha, the accuracy rates of the non-neural model are 92.81%, 94.13%, 88.72%, and 90.54%, and the accuracy rate of the neural model respectively 96.95%, 99.54%, 96.33%, 95.51%. However, for languages with small datasets, such as Eibela, the effect of the two models was not noticeable. Because the training dataset is more extensive, and coverages is broader, the model has a strong learning ability; when the scale of the training set is large enough, the learning ability of the neural network model will be significantly higher than that of the statistical learning model, which is also more in line with the actual situation and our expectations. 3) From the overall experimental results, regardless of the size of the dataset or the language feature, the Uyghur language dataset in the two models are similar to other languages. It also reflects the dataset effectiveness constructed in this paper. 4) From two Uyghur datasets, the results of nominal dataset higher than all words dataset by 3.69% and 4.48%. Because the nominal dataset is relatively single part of speech, and the morphological changes are not too complicated. Compared with the nominal dataset, all words dataset is richer in morphological feature and have more POS.

5 Summary

In this paper, firstly, we mainly introduce the related work of constructing the Uyghur language morphological analysis corpus, including data preparation and preprocessing, making morphological analysis tags (POS and morphological tags), and corpus construction process. To reduce manual work, we used a POS tagger and stemming tool. Secondly, we statistically analyzed the basic information of the dataset, such as the distribution of the number and length of words and sentences, the distribution of POS, morphological feature, and MSDs. Finally, we designed comparative experiments using the models and datasets provided by the SIGMORPHON Shared task, to analyze and verify the validity of the dataset. The result of experiment shows that the dataset constructed on this paper is similar to other datasets. The morphological analysis corpus provides information about the lemma, POS, MSDs, morphological segmentation, and lemmatization of words. In the following research, we will continue to expand the data set, increase the coverage of words, enrich the language, and study morphological analyzers suitable for agglutinative language, to provide high-quality data sets for downstream tasks.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (grant numbers 61762084); and the Scientific Research Program of the State Language Commission of China (grant number ZDI135-54); the Opening Foundation of the Key Laboratory of Xinjiang Uyghur Autonomous Region of China (grant number 2018D04019).

References

Gulinigeer Abudouwaili, TUERGEN Yibulayin, KAHAERJIANG Abiderexiti, and WANG Lulu. Research on uyghur stemming based on bi-lstm-crf model. *JOURNAL OF CHINESE INFORMATION PROCESSING*, v.33(08):65–71, 2019.

Halidanmu Abudukelimu, SUN Maosong, LIU Yang, and Abudukelimu Abulizi. Thuuymorph: An uyghur morpheme segmentation corpus. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 32(2):81, 2018.

Gulila Altenbek, Wang Xiaolong, and Gulizhada Haisha. Identification of basic phrases for kazakh language using maximum entropy model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1007–1014, 2014.

Arianna Bisazza and Clara Tump. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pages 2871–2876, 2020. doi: 10.18653/v1/d18-1313.

Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. *Proc. Tenth Conf. Comput. Nat. Lang. Learn. CoNLL-X*, 2006. doi: 10.3115/1596276.1596305.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. Conll-Sigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. *CoNLL 2017 - Proc. CoNLL SIGMORPHON 2017 Shar. Task Univers. Morphol. Reinflection*, pages 1–30, 2017. doi: 10.18653/v1/k17-2001.

Sediyegvl Enwer, Xiang Lu, Zong Chengqing, Akbar Pattar, and Askar Hamdulla. A multi-strategy approach to uyghur stemming. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 29(5):204, 2015.

Onur Güngör, Tunga Güngör, and Suzan Üsküdarli. The effect of morphology in named entity recognition with sequence tagging. *Nat. Lang. Eng.*, 25(1):147–169, 2019. ISSN 14698110. doi: 10.1017/S1351324918000281.

Turgun Ibrahim and YUAN Baoshe. A survey on minority language information processing research and application in xinjiang. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 25(6):149, 2011.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2000.

Maihemuti Maimaiti, Aishan Wumaier, Kahaerjiang Abiderexiti, and Tuergen Yibulayin. Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. *Inf.*, 8(4), 2017. ISSN 20782489. doi: 10.3390/info8040157.

Wumaierjiang Maimaitiming, Abuduwaili Gulinigeer, Maihemuti Maimaiti, Kahaerjiang Abiderexiti, and Tuergen Yibulayin. A comparative study of uzbek stemming algorithms. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 34(1):45, 2020.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskij, Natalya Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. UniMorph 3.0: Universal morphology. *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, 0 (May):3922–3931, 2020.

Muhetaer Munire, Xiao Li, and Yating Yang. Construction of the Uyghur noun morphological re-inflection model based on hybrid strategy. *Appl. Sci.*, 9(4), 2019. ISSN 20763417. doi: 10.3390/app9040722.

Turghun Osman, YANG Yating, and Eziz Tursun. Collaborative analysis of uyghur morpholo, gy based on character level. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 55(1):47, 2019.

Sardar Parhat, Mijit Ablimit, and Askar Hamdulla. A Robust Morpheme Sequence and Convolutional Neural Network-Based Uyghur and Kazakh Short Text Classification. *Inf.*, 10(12), 2019. ISSN 20782489. doi: 10.3390/info10120387.

M. Silfverberg, T. Ruokolainen, Krister Linden, and M. Kurimo. Finnpos: an open-source morphological tagging and lemmatization toolkit for finnish. *Language Resources and Evaluation*, 50(4):863–878, 2016.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. of52nd Annu. Meet. ofthe Assoc. Comput. Linguist. Syst. Demonstr.*, pages 13–18, 2015. doi: 10.3115/v1/p14-5003.

J. Sylak-Glassman, C. Kirov, M. Post, R. Que, and D. Yarowsky. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology*, 2015.

Litifu Tuohuti. *Modern Uyghur Reference Grammar*. CHINA SOCIAL SCIENCES PRESS, 2012.

Ahmet Üstün, Murathan Kurfalı, and Burcu Can. Characters or Morphemes: How to Represent Words? In *Proc. ofthe 3rd Work. Represent. Learn. NLP*, pages 144–153, 2019. doi: 10.18653/v1/w18-3019.

Clara Vania. *On Understanding Character-level Models for Representing Morphology*. 2020.

Clara Vania, Andreas Grivas, and Adam Lopez. What do character-level models learn about morphology? The case of dependency parsing, 2018. ISSN 23318422.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 2017-Decem (Nips):5999–6009, 2017. ISSN 10495258.

FEISHENG YE and TONGQIANG XU. *Essentials of linguistics*. PEKING UNIVERSITY Press, 2006.

Y. Zhu, I Vulić, and A. Korhonen. A systematic study of leveraging subword information for learning word representations. *Conference of the North*, 2019.

Hang Zhuang, Chao Wang, Changlong Li, Yijing Li, Qingfeng Wang, and Xuehai Zhou. Chinese Language Processing Based on Stroke Representation and Multidimensional Representation. *IEEE Access*, 6:41928–41941, 2018. ISSN 21693536. doi: 10.1109/ACCESS.2018.2860058.

Chengqing ZONG. *Statistical Natural Language Processing*. TSINGHUA UNIVERSITY PRESS, 2013.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1076-1086, Hohhot, China, Augest 13 - 15, 2021.