Analytical Method and Research of Uyghur Language Chunks Based on Digital Forensics

Yasen Aizezi[1], Anwar Jamal[1], Dilxat Mamat[1], Ruxianguli Abdurexit[1], and Kurban Ubul[2(✉)]

[1] Xinjiang Police Institute, Urumqi, Xinjiang 830013, China
[2] School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
kurbanu@xju.edu.cn

Abstract. In the digital forensics process based on Uygur language information, stem, affix, synonym mark and other characteristics are added on the features-based English and Chinese chunks and according to relevant characteristics of Uygur language. In terms of performance evaluation indexes in this paper, accuracy rate, recall rate and F value are adopted. The test indicates that the scale of Uyghur chunks database has great effects on the model performance.

Keywords: Conditional random fields (CRF) · Uighur · Chunk analysis

1 Introduction

Chunk is a syntactic structure between word and sentence, which is also called as shallow parsing or partial parsing. It is used to identify some elements with relatively simple structure and comparatively important function and significance in a sentence. However, complete syntactic analysis tree is not served as the goal, so as to simplify complexity of the analysis and improve performance of the analysis. Abney was the first person to propose the thought of chunk analysis in 1991[1].

In recent years, people gradually attach importance to the research on Chinese chunk analysis. In 1996, Zhou Qiang made a research on chunks and basic phrases of Chinese language [5]. In 1999, Zhao Jun and Huang Changning made a research on the definition and automatic identification of basic noun phrases in Chinese language [6]. Li Sujian from the Institute of Computing Technology, Chinese Academy of Sciences proposed 12 types of Chinese chunks. Moreover, he obtained a chunk database through transformation according to the corresponding relation between these chunk types and phrase types of the Chinese tree database of the University of Pennsylvania [7]. Zhou Qiang made a large-scale research on notes of chunks in the Chinese language database [5], established a complete chunk division system, and constructed a chunk balance language database with 2,000,000 Chinese characters [8]. Zhang Yujie and others also researched the analysis on Chinese chunks [9], and [10] proposed an integral analysis model, and [11] suggested a chunk analysis method based on the division strategy.

© Springer International Publishing Switzerland 2015
V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 258–266, 2015.
DOI: 10.1007/978-3-319-23222-5_32

As the research on the natural language processing technology of Uyghur language started late and the lexical analysis technology failed to reach an available level, the research on the syntactic analysis technology of Uyghur language is basically in a primary stage. In this paper, a research is made onto the analysis on chunks of Uyghur language. Moreover, conditional random fields are used to establish a chunk analysis algorithm.

2 Definition of Chunk and Establishment of Language Database

2.1 Definition and Division Principles of Chunk

According to the definition of Abney, chunk of Uyghur language is defined as follows in this paper:

Definition 1: chunk refers to the non-recursive, non-overlapped and non-nested phrase between word and phrase, which is more complicated than word and simpler than sentence, and which has certain syntactic functions.

A detailed interpretation of the above definition: a chunk consists of word sequence, with syntactic functions marked. Moreover, it is non-recursive and non-nested. Generally, in the chunk is a non-nested phrase. Chunk is defined in strict accordance with syntactic form; while semantics or functionality is not reflected. The purpose of chunk analysis is to identify some elements with relatively simple structure but important significance in the sentence and to erect a bridge between lexical analysis and complete syntactic analysis, so as to simplify syntactic analysis and improve the performance of syntactic analysis.

2.2 Formulation of Uyghur Language Chunk Tag Set

Prior to research on and preparation for notes, marks and specifications of the language tree database, an in-depth research is made on the establishment process of English tree database and TCT tree database. Moreover, a comparative research is made with the syntactic structure of Uyghur language. Based on the research and analysis, notes tag-set is prepared according to the following steps:

S1: preliminarily prepare a set of modern Uyghur language phrase tags collection;

S2: select from the language database 100 sentences with relatively great differences in sentence structure;

S3: automatic mark the 100 sentences, and register phenomena that existing tag-set cannot be used to mark correctly;

S4: in case that existing mark collection cannot be used to mark correctly, the mark collection should be analyzed and modified;

S5: in case that there is no any question with the tag-set, we should inspect whether the number of sentences automatic marked is 500. If the answer is no, then turn to S2; if the answer is yes, then turn to S6;

S6: end the mark phase

According to the above steps, the tag-set is repeatedly prepared and modified. Finally, 37 Uyghur language phrase structure tag-sets and 8 functional chunk tag-sets are specified. In this paper, 18 chunk mark types are defined from 37 phrase tag-sets according to relevant characteristics of chunk analysis.

2.3 Construction of Uyghur Language Chunking Corpus

At present, there are 3,000 sentences in the completed Uyghur language tree database. In this paper, chunks are selected from this database to construct the Uyghur language chunk database. Besides, in the right-side produced elements are selected for automatic calibration from the Uyghur language mark tree database and the production collection containing only non-terminal symbols and terminal symbols. Then, sentences marked by chunk in the right-side produced elements are selected. At present, there are totally 31,184 chunks in the Uyghur language chunk database established. For instance, the process of extracting chunks from the marked sentences is described as follows:

Table 1. Generative grammar classification results

Production of non-terminal symbol	mixed production	production of terminal symbol
SS->NP NP VP	UP->CP Idi	CP->bësip chüshken
VP->UP	UP->qërindashliqni UP	NP->Aq köngüllük
NP->NP NP		NP->Uning Öyidiki

The chunk shown in Table 2 is a high-frequency chunk in the Uyghur language tree database, accounting for 90.40% of all chunks.

Table 2. Ten main chunk statistics

	Mark	chunk of 2 words	chunk of 1 word	total chunks	average length
Noun chunk	CNP	7776	3924	11700	1.6646
Adjective chunk	CAP	1023	729	1752	1.5839
Verb chunk	CVP	2112	1608	3720	1.5677
Gerund chunk	CGP	2430	1251	3681	1.6601
Participle chunk	CCP	909	999	1908	1.4764
Coverb chunk	CBP	1680	822	2502	1.6715
Pronoun chunk	CPP	39	228	267	1.1461
Quantifier chunk	CQP	825	375	1200	1.6875
Numeral chunk	CMP	1116	249	1365	1.8176
Adverb chunk	CDP	66	30	96	1.6875

3 Chunk Analysis Algorithm Based on Statistical Learning Model

3.1 Description on Relevant issues of Chunk Analysis

Chunk analysis can be regarded as a machine learning process. Its task is to automatically divide blocks of sentences input and to mark the types of blocks divided under the given chunk definition and category. It can be formally described as follows:

Specify sample set W=w1,w2,…,wn and category set C=c1,c2,…,cm, seek for a model f:W×C→Boolean (mapping rules) from sample set W to category set C, and then judge the category of the new input sample by using the relation model obtained from this study; to be specific, set a sentence composed of word sequence W=w1,w2,…,wk, divide the sentence into several chunks, mark each word wi with chunk mark ti; T=t1,t2,…,tn stands for the chunk mark sequence. Relevant results of the chunk analysis are shown as follows:

$$W=...[wi,wi+1,\ldots,wi+m][wi+m+1,\ldots,wi+m+h]\ldots \quad (1)$$

$$T=...[ti,ti+1,\ldots,ti+m][ti+m+1,\ldots,ti+m+h]\ldots \quad (2)$$

Mapping rules of the chunk analysis are classification laws and judgment rules automatically generated by the system according to characteristic information of each sample of machine learning. In the analysis, this mapping is one-to-one single mark classification mapping.

3.2 Research and Analysis on Chunk Analysis Methods

The issue of chunk analysis can be transformed into the issue of sequence mark. However, methods or models available for sequence mark include methods based on error transition, hidden Markov model (HMM), maximum entropy model, support vector machine (SVM), conditional random field(CRF) model, etc. In above methods or models, it is the CRF model that has optimal conditions. Therefore, CRFs are used in this paper to establish the Uyghur language chunk analysis model.

3.3 Establishment of Characteristic Space

The key of discriminant statistical model is to find out various characteristics contribute to eliminate ambiguity, to use such characteristics to combine different characteristic templates, to verify the effectiveness of such characteristic templates by means of experiment and, and to select the optimal characteristic template. On this basis, this paper establishes the characteristic space of Uyghur language chunk analysis with reference to relevant characteristics used for chunk analysis algorithms of English, Chinese and other languages based on CRFs.

In terms of the sequence of part of speech W=w1,w2,…,wk, characteristics of prefix and suffix are added into the Chinese chunk analysis model by selecting windows with the width of 5 in both English and Chinese chunk analyses, and extracting characteristics (e.g. morphology, part of speech, affix, chunk mark, etc.) of the present word wi and two words before and after this word respectively. In this paper, above characteristics are retained. Moreover, stem, affix, first-class mark of part of speech, second-class mark of part of speech, synonym mark and other elements are also added in this paper to establish the characteristic space according to characteristics of words in Uyghur language.

3.4 Establishment of Synonym Mark Database

SY (synonym) in the above characteristic space stands for synonym mark. Relevant contents of this mark are interpreted above in details. At present, there are few Uyghur language tree databases established. It is easy for inaccurate parameter estimation caused by data rarefaction by using statistical model. If words with completely the same meanings can be marked with certain marks or numbers, the issue of decreasing analysis performance resulted from the scale of language database can be remitted to a certain extent. Therefore, a Uyghur language synonym mark dictionary is established with references to existing Uyghur language synonym dictionaries. This dictionary originally has 9,902 entries, in which 1,778 are compound words. In order to guarantee the accuracy rate of synonym mark, 4,623 synonyms with strictly the same meanings are selected from the other 8,104 synonyms. Moreover, a synonym database with part-of-speech notes is established in this paper. All synonyms are classified based on meaning and part of speech. Besides, each classification is allocated with a mark. Finally, the dictionary with 971 synonym marks is established.

The CRF model is a guidance-oriented machine learning model. At first, a certain scaled mark language database should be used to evaluate relevant parameters of the model. Then, the model trained can be used for decoding, i.e. to mark unmarked language materials. The L-BFGS algorithm is used for the model training. Besides, BeamSearch algorithm is used to search for u. The width is 5. The CRF model is subject to CRFComLib training and testing.

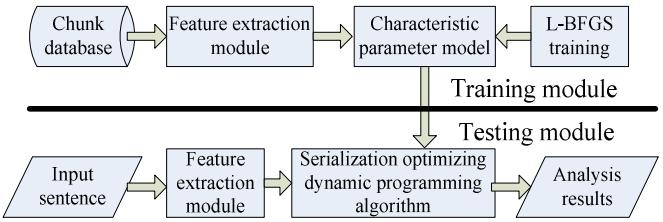

Fig. 1. CRF based Uyghur Chunk parsing System Structure

4 Test and Analysis

In this paper, 3,000 marked sentences are selected as the training and testing language database. Due to the small scale of the language database, the cross validation method is used for testing. International precision (P), recall(R) and F-measure value are adopted as performance evaluation indexes of the chunk analysis algorithm in this paper.

4.1 Feature Selection

Section of characteristic module and feature selection are key steps in judging training and application of the learning model. Features are extracted from training samples, which directly reflect various types of knowledge and cases in chunk texts. Characteristic model and feature description capability selected have direct effects on performance of the analysis system. For different language processing tasks, features selected will be different as well. Generally, there are two methods for feature selection:

- 1) According to experiences summarized from linguistic knowledge of linguists and statistical information of texts, formal characteristic templates are defined based on characters and marked in texts. Moreover, characteristic templates are used to extract characteristics from the texts, or it is called instantiation of characteristic template.
- 2) Additional information and marked are given to texts according to summary of linguistic knowledge by linguistics, such as language rules, grammatical rules, dictionaries, resource bases and other external information.
- 3) As characteristic task correlation, targeted & task-driven characteristic template and characteristic definition are always helpful for the analysis system. On the contrary, ineffective characteristics will reduce performance of the system.
- 4) According to experimental results in [93], word form, stem, affix, part of speech, synonym mark and others are used in this section to structure an atomic characteristic space. On this basis, different characteristic templates are combined for an experiment.

In order to test contribution of word form, part of speech, stem and other characteristic information, a characteristic template shown in Table 6-8 is established based on the summary in [93]. Template A is the word form template. Template B is stem information added, which can be used to observe impacts of part of speech on performance of the model. On the basis of Template B, only affix characteristic is added to Template C. In Template D, first-class part of speech mark is introduced. Both first-class and second-class part of speech marks are used in Template E. Synonym mark is introduced in Template F. In order to observe impacts of first-class and second-class marks on performance of the model, only second-class mark is used in Template G.

Table 3. Close test result

Feature Template	Precision	Recall value	F-measure
A	56.23%	57.01%	56.62%
B	58.87%	59.63%	59.25%
C	61.34%	61.97%	61.65%
D	76.02%	76.54%	76.28%
E	79.11%	78.34%	78.72%
F	81.65%	81.84%	81.74%
G	81.93%	82.30%	82.11%

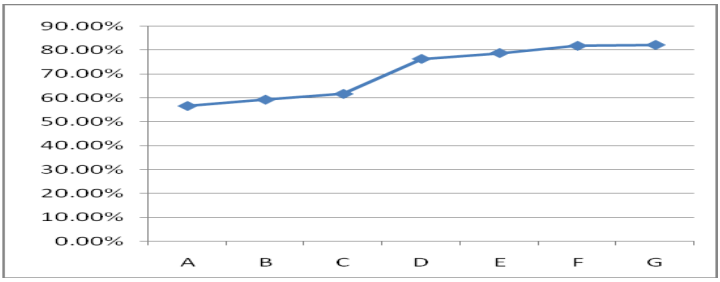

Fig. 2. Test Result Statistical Plotting

4.2 Cross Validation Test

Due to the small scale of the language database, the cross validation method is used in the test. In order to observe performances of language databases in different scales during training, the test is performed for three times:

Test A: the language database is divided into 10 subsets without cross data. Each subset has 300 sentences. The test is performed for 10 times. Finally, the average value of 10 tests is calculated.

Test B: the language database is divided into 5 subsets without cross data. Each subset has 600 sentences. The test is performed for 5 times. Finally, the average value of 5 tests is calculated.

Test C: the language database is divided into 3 subsets without cross data. Each subset has 1,000 sentences. The test is performed for 3 times. Finally, the average value of 3 tests is calculated.

Table 4. Open test result

Test	Proportion	Precision	Recall	F-measure value
A	9:1	80.23%	80.45%	80.34%
B	8:2	76.61%	77.14%	76.87%
C	2:1	66.52%	67.01%	66.76%

According to the experimental results, the scale of language database has great effects on the model. The principal reason is that the scale of language database trained with the model is not strong enough to make the model reach a saturated mode. That is, the expansion of the language database scale cannot improve the status of model performance.

5 Conclusion

In this work, 3,000 marked sentences are used as the language database for training and testing. The cross validation method is adopted in the lab. The precisions of the training and testing language databases are 9:1, 8:2, 2:1; and the recall rates are 80.34%, 76.87%, and 66.76%.

Acknowledgment. This paper is supported by the National Social Science Fund of China (No. 13CFX055), Science Research Program of the Higher Education Institute of Xinjiang (No. XJEDU2013I11, XJEDU2013I34), and Natural Science Fund of Xinjiang (No. 2015211A016).

References

1. Abney, S.P.: Parsing by Chunks. Computation and Psycholinguistics, 257–278 (1991)
2. Sang, T.K., Buchholz, S.: Introduction to the Conll-2000 shared task: chunking. In: Proceeding of CoNLL-2000, Lisbon, Portugal, pp. 127–132 (2000)
3. Kinyon, A.: A Language-independent shallow-parser compiler. In: Proceedings of 39th ACL Conference, Tourouse, France, pp. 322–329 (2001)
4. Hammerton, J., Osborne, M., Armstrong, S.: Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing. Journal of Machine Learning Research (2), 551–558 (2002)
5. Qiang, Z.: Research on Automatic Division and Marking of Phrases in Chinese Language Database, pp. 1–9. Peking University (1996)
6. Jun, Z., Changning, H.: Model for Chinese BaseNP Structure Analysis. Chinese Journal of Computers **22**(2), 141–146 (1999)
7. Sujian, L., Qun, L., Zhifeng, Y.: Chunk Analysis based on Maximum Entropy Model. Chinese Journal of Computers **25**(12), 1722–1727 (2003)

8. Yuqi, Z., Qiang, Z.: Automatic Identification of Chinese Base Phrases. Journal of Chinese Information Processing **16**(6), 1–8 (2002)
9. Chen, W., Zhang, Y., Isahara, H.: An empirical study of chinese chunking. In: Proceedings of the 44th Annual Meeting of ACL, Sydney, Australia, pp. 97–104 (2006)
10. Guanglu, S.: Technical Research on Chinese Chunk Analysis based on Statistical Learning. Harbin Institute of Technology (2008)
11. Qiaoli, Z., Xin, L., Wenjing, L., Dongfeng, C.: Chunk Analysis based on Division Strategy. Journal of Chinese Information Processing **26**(5), 120–128 (2012)