ATLANTIS PRESS
Advances in Engineering Research (AER), volume 131
3rd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2017)

Research on Chinese Semantic Role Labeling Based on Ocean Big Data

Lei-Na JIANG[1, a], Yun-Tao QIAN[2, b], Yan-Jiang SUN[3, c]

[1, 2, 3] College of Information Science and Engineering

Ocean University of China, Qingdao, Shandong, China

[a]jiangleina520@163.com, [b]ytqian@zju.edu.cn, [c]pswqdx616237781@163.com

*Yongquan Yang

Keywords: semantic role labeling (SRL), ocean big data, detailed.

Abstract. Natural Language Processing has reached the stage of semantic analysis. As a tool for semantic analysis, semantic role labeling (SRL) is becoming more and more important. But the SRL is not specifically applied to the marine field in the past. So this paper is a study of SRL based on ocean big data. First of all, this paper populates and updates corpus with data from various categories of ocean. Secondly, (based on the current semantic roles,) this study makes the semantic roles more detailed. Thirdly, we decompose the multi-predicate verbs in a sentence. Finally, SRL takes advantage of the analysis of dependency parsing (DP). In this paper, 1000 sentences of various categories of ocean are selected, and 25,903 words are used to mark the semantic roles, and the accuracy of the annotation is 93.4%, which is revised by human.

Introduction

Recent years, natural language processing has made some progress in the field of word segmentation, part-of-speech tagging, name entity recognition, dependency parsing (DP), but the technology of semantic analysis stage is not yet mature. However, semantic analysis has become very important in order to make the machine understand [1-2] the concepts behind natural language and generate natural language. So this thesis chooses to use the semantic role labeling (SRL) to analyze the sentences semantic.

The study of English semantic analysis has been widely used abroad, comparatively speaking, the Chinese semantic analysis has been developed slowly. The main reason is that the grammar has a big difference between Chinese and English, so the Chinese semantic analysis cannot make use of the English semantic analysis results. But no matter for Chinese or English semantic analysis, the corpus is very important. There is much corpus marked with semantic information, for example: Frame Net (Johnson et al. 2000) [3], BFS-CTC [4], TongYiCi CiLin [5].

As a new research direction, Chinese SRL has a huge improvement since recent years, and has achieved the automatic labeling of Chinese (Liu Ting[6]).Now there are many algorithms that are often used for SRL, for example: SVM algorithm [7],CNN algorithm [8],Maximum Entropy Model [9].Although the research work of SRL develops rapidly, but the research on marine specific fields has not yet been practiced, and the classification of semantic roles is imperfect, therefore, this paper do the SRL based on the above problems in the field of ocean.

SRL refers to: first, given a sentence, you need to find the central predicate verb and clause predicate verb in the sentences; then the sentences are marked respectively with the central predicate verb and the clause predicate verb as the core; finally, we integrate the results of the annotation. The current SRL corpus contains not only the clutter but also the scarcity of the data, because of the narrow and chaotic characteristics that make the domain adaptation of SRL poor. However, the demand for semantic analysis has been demonstrated in the field of ocean big data, so, this paper aims at SRL in the field of ocean. Because the ocean field has itself features: diversity, massive, multi-scale and so on, we divide the data into categories such as marine economy, marine biology, marine chemistry and so on, and then collect the representative vocabulary in each category and put them into the corpus, and in the end generate a corpus that possesses ocean field features and provides the convenience for the next work of SRL.

Chapters are arranged as follows: the second section describes the whole system; the third section gives the experimental results and analysis results; in the end, summarizes the whole paper and gives the next research emphases.

System Description

The overall architecture of the system is shown in Fig 1:

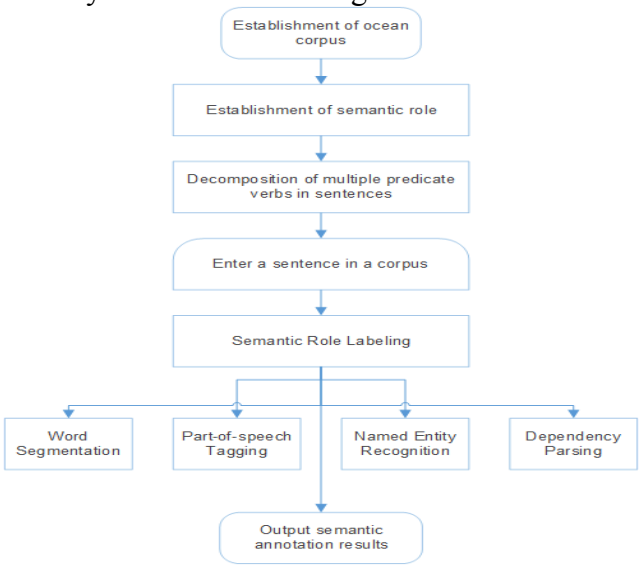

Fig 1. Overall architecture

Establishment of ocean corpus. Ocean big data domain includes categories of marine economy, marine biology, marine chemistry, and marine industry and so on. The marine corpus in this paper needs to include all kinds of representative words in the sea. Since the semantic role labeling (SRL) work is only a part of the entire natural language processing project, so we label semantic role basing on the existing resources in the laboratory. Specific practices are as follows:

- Set up a seed bank and includes 100 authoritative websites about marine domain;
- Use WebCrawler tool to catch 100 websites information of text information, website URL, title and so on, and store data in the hbase database;
- Read the text information from the hbase database, select and pre-process the information, obtains representative vocabularies to fill it into a marine corpus. This corpus is currently used for the research of national marine laboratory, and not available for public, temporarily.
- According to the condition of the crawl, update the ocean corpus constantly.

The ocean corpus describes and marks the part of speech, syntax and semantics.

Establishment of semantic role. SRL aims to mark the target predicate verbs and semantic roles for the given sentences, so that the machine can understand the meaning of sentences.

The main problem about SRL is finding the core semantic role. According to the marine field features, this page makes the semantic roles more detailed on the basis of the original semantic roles, which is much convenient for syntactic and sentence analysis later. In the Table I and Table II, we list all semantic roles that we used:

TABLE I. CORE SEMANTIC ROLES

A0	A1	A2-A5
Agent of action	Impact of action	Different verbs have different meanings

TABLE II. ORIGINAL ADVERBIAL SEMANTIC ROLES

ADV	Adverbial	**EXT**	Extent	**TMP**	Temporal
BNE	Beneficiary	**FRQ**	Frequency	**TPC**	Topic
CND	Condition	**LOC**	Locative	**CRD**	Coordinated arguments
DIR	Direction	**MNR**	Manner	**PSR**	Possessor
DGR	Degree	**PRP**	Purpose	**PSE**	Possess

TABLE III. NOW ADVERBIAL SEMANTIC ROLES THAT ARE ADDED OR MODIFIED

C_PRD	PVC	MOD	NUM	PRN
Core predicate	The predicate in the clause	Modifier	Number	Proper nouns

In the Table III, we can see the third part add some new semantic roles. First of all, owing to the marine field includes many categories, each category has more descriptions of numbers, so we add the semantic roles of numbers; Secondly, some categories, like marine chemistry, marine biological protein and so on, include many proper nouns, for example: potassium sulphate, potassium nitrate. So we add the semantic roles of proper nouns; the last section, it will be explained in the next section, we decompose a long sentence for the situation that a long sentence includes multiple verbs, and these verbs are divided into two parts: core predicate verbs and clause predicate verbs, we will separately mark the sentences semantic roles from the two parts.

Decomposition of multiple predicate verbs in sentences. At present, SRL work is mainly aimed at the situation of single predicate verb in a sentence, but a sentence with multiple verbs also needs to be processed. Facing with this situation, we adopt a proper way to mark this sentence: decompose the sentence. When we get a sentence, the first step is to find the core predicate verbs and clause predicate verbs; the second step is to mark the main sentence and clause sentence with the core of core predicate verbs and clause predicate verbs; the last step is to integrate the results of subject labeling and clause labeling and forms a new labeling results.

There is the difference between single predicate verb and multiple predicate verb, examples are as follows: 1) “Offshore oil is an important energy industry of the country”. In this sentence, there is only one predicate verb “is”, so we call such sentence with a single predicate verb; 2) “This is a kind of molecular structure with regular repetition” .In this sentence, “is” and “have” are the predicate verb, but “is” belongs to the main sentence core predicate, however, “have” belongs to the clause predicate verb, this situation belongs to the multiple predicate verbs.

Semantic Role Labeling (SRL).SRL is based on word segmentation, part-of-speech tagging, named entity recognition and dependency parsing (DP), so we have to ensure the accuracy at each step in the front, once the previous results have faults, it will have a certain effect on the final labeling results.

Word segmentation comes from an open source project. You input a sentence, the system will segment the input sentence based on the ocean corpus, and finally, we output the sentence that are segmented well. The word segmentation aims to simplify sentences in order to understand the sentence semantics in the later stage. So if we want to ensure the accuracy of the word segmentation, our oceanographic corpus should be as comprehensive as possible.

Part-of-speech tagging is based on the word segmentation. This stage is to mark the part-of-speech with the words that are cut, that is to say, a word belongs to the noun, verb or other parts of speech. The purpose is to help better find the predicate verb in the sentence and determine what structural component the word belongs to in the sentence, lay the foundation for the next work.

Named entity recognition stage is to identify people, places and other entities in a sentence, for example: “Premier Zhou visited Shanghai”. In this sentence, “Premier Zhou” belongs to the people, “Shanghai” belongs to the places, so the two words will be marked. The purpose of this stage is to help us find quickly the entity from the sentence, it will provide the convenience for subsequent semantic analysis.

The emergence of DP has been of great help to SRL. It is based on context to analyze grammatical relations in sentences, judges the relation between words that belong to “subject predicate relation” or “verb object relation”. By means of DP, we can further understand the structure of sentences and provide the basis for SRL.

SRL bases on the above steps. As shown in Figure 2: input a sentence, the result of SRL is as follows:

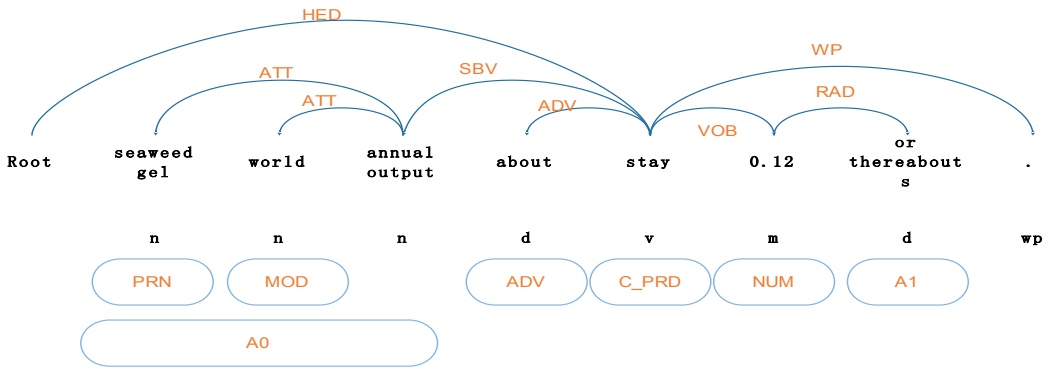

Fig 2. Semantic Role Labeling

In the graphical interface showing above, we can see the input of this sentence as: "Seaweed gel world annual output about stay 0.12 or thereabouts". At the top of this sentence, the result of DP is presented in the form of connected arcs. What is marked on the arc is the grammatical component and see Table IV; the first part of this sentence is the result of the part-of-speech tagging and also see Table IV; the second part is the result of tagging with new semantic roles. In SRL, we mark the semantic roles of the sentences as much detailed as possible, it is much easier for machine to understand natural language.

TABLE IV. DP AND PART-OF-SPEECH TAGGING

Tag	Description	Tag	Description
HED	head	n	General noun
ATT	attribute	d	adverb
SBV	Subject-verb	v	verb
ADV	adverbial	m	number
VOB	Verb-object	wp	punctuation
RAD	Right adjunct		

Experimental and discussion

From the marine economy, marine industry, marine environment, marine platform, marine news, marine biology and so on many kinds, we select 1000 sentences, altogether 25903 words for the test. We use the semantic role labeling (SRL) procedure from the open source program LTP at HIT, which includes all the natural language processing key technology. The bottom layer of the program is written in C++ language, and we use the Java platform to call the C++ development package, in the end, enables the entire program to run in Java's Eclipse.

The improvement to the LTP. We made some changes on the basis of the original LTP:

- Modify the lexicon model, the lexicon in the original lexicon model is rather cluttered, and marine field features are not strong, so we build the corpus of ocean resources in the section as our new thesaurus.
- Modify the semantic model, the semantic roles in the original semantic model are more adaptable when adapts to other domains, but in the marine field, we add some new semantic roles, these new semantic roles can help label semantic role in ocean field.
- The original program is integrated, the original program puts each result separately, but we want results in a form of that each result depends on other results, that is to say, SRL depends on a series of results, such as word segmentation, part of speech tagging and so on.
- Make the program apply to both the Java platform and the Linux platform. Because our program uses the hbase database, so it needs to run on the Linux platform to access data. There are two reasons for using hbase database instead of using MySQL database: first, our data is discrete and also in a large amount; second, hbase database has great flexibility. However, it is really complicated to use because the original program calls different platforms which need different compilation methods, so we compile all the Linux and Java platform. Then we put the compiled files into the program to prevent problems.

Experimental results. The results of experiments is shown as Table V: input a sentence: "Recent year's research discovery, chitosan have resist tumor effect". The results of word segmentation, part-of-speech tagging, named entity recognition, DP and SRL are as followed:

TABLE V. OUTPUT RESULTS

Word segmentation	Recent years| research| discovery|,| chitosan| have| resist| tumor| effect|.|	
Part-of-speech tagging	Recent years_nt| research_v| discovery_v|,_wp| chitosan_n| have_v| resist_v| tumor| effect_n|._wp	
Name entity recognition	o o o o o o o o o	
DP	2:ADV 0:HED 2:COO 2:WP 6:SBV 2:COO 9:ATT 7:VOB 6:VOB 2:WP	
SRL	***Main mark***	1. Type=TMP begin=0 end=0 2. Type=C_PRD begin=2 end=2 3. Type=A1 begin=4 end=8
	Clause mark	1. Type=A0 begin=4 end=4 2. Type=PVC begin=5 end=5 3. Type=A1 begin=6 end=8

Mark all 1000 sentences in the manner shown above, get a total mark result, and then manually check the dimension scale and accuracy, in the end, the accuracy of labeling is up to 93.4%, it is of great significance for further semantic analysis and text extraction.

Contrast experiment. The tagging accuracy of SRL in marine field is more accurate than that in LTP, in the test cases of 1000 sentences and 25903 pairs of words, the contrast results are shown at the following Table VI:

TABLE VI. CONTRAST RESULTS

	Sentences	**Words**	**Labeling rate (%)**	**Precision (%)**
This paper	1000	25903	98.2	93.4
LTP	1000	25903	90.8	82.4

As you can see from the table, the final tagging results are relatively high because the ocean corpus contains more comprehensive words.

Accuracy curve. Because the training corpus is large enough and able to obtain a learning curve. We perform test basing on the 15 ocean categories. With the increase of training set, the accuracy of the system is improved continuously. Therefore, in the current state, the increase of training data is still helpful to improve the accuracy of the system.

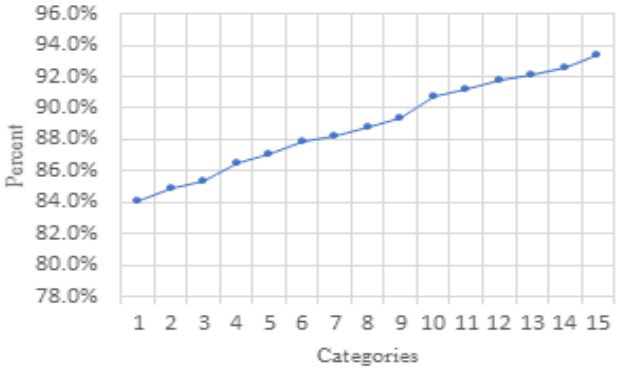

Fig 3. Accuracy Curve Effecting Of Training Set

From the Fig 3, because the original corpus is not big and it will lead to the accuracy not higher. When this paper adds the 15 categories in turn, this corpus will become very big and detailed, so the learning curve will have a rising trend.

Conclusion

This paper sets up a semantic role labeling (SRL) system that bases on the ocean big data. Comparing with the existing SRL system, it pays more attention to the domain adaptability.

Specifically, first, this paper builds an ocean resource corpus for ocean field, and performs SRL for sentences in the corpus; at the same time, this research adds some new semantic roles and makes the semantic labeling more decomposed; finally, the study deals with the situation that a sentence includes multiple predicate verbs so that the machine can better understand the results of SRL. Results show that it is feasible to do SRL in the ocean domain, and it will make the annotation result more accurate. Although we make much progress, we still have a long way to go in the future. Constantly improving SRL and having a deep semantic analysis are the focuses of our research in the future.

Acknowledgement

This work is suppoted by The Aoshan Innovation Project in Science and Technology of Qingdao National Laboratory for Marine Science and Technology(No.2016ASKJ07)

References

[1] Wang Ya-bin. Research on Semantic Annotation Based on Ontology［D］. LanZhou: LanZhou University of Technology, 2010.

[2] BECHHOFES,et al. The semantics of semantic annotation ［M］. Irvine,California,2002.

[3] Jonson, C.R.,Fillmaor,C.J. The framenet tag set for frame-semantic and syntactic coding of predicate-argument structure[C]. Seattle, WA .Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics,2000:56–62.

[4] LIU Yingying, LUO Senlin, FENG Yang. BFS-CTC:A Chinese Corpus of Sentential Semantic Structure [J].JOURNAL OF CHINESE INFORMATION PROCESSING,2013,1.

[5] MEI Jiaju, ZHU Yiming, GAO Yunqi. TongYiCi CiLin(the second Edition) [M]. ShangHai : Shanghai Lexicographical Publishing House,1996.

[6] LIU Ting, CHE Wan-Xiang, LI Sheng. Semantic Role Labeling with Maximum Entropy Classifier [J].journal of software, 2007, 3: 565-573.

[7] Jing Zhiqiang. A Method of Chinese Text Classification Based on the Expansion of VSM[D].Harbin Engineering University 2010.

[8] Qin Jun, Xu Fei. Gradual Incremental Learning Algorithm of Support Vector Machine Based on Hull Vector [J].Journal of South-Central University for Nationalities. 2011(03).

[9] XU Yanyong, ZHOU Xian-zhong, JING Xiang-he. Chinese Sentence Parsing Based on Maximum Entropy Model[J].acta electronica sinica. 2003(11).