Analyzing behavioral activities among the multilingual big-data on the complex network with the help of suggested potential monitoring: Altai family languages in China

Doniyorbek K. Ahmadaliev, Chen Xiaohui, Sardor U. Dadabayev

***Abstract*— There are myriad cultures and languages in China which require establishment of a multi-language monitoring platform. That is well be developed here, this is not only to control the Complex network Behaviors of multi-language big data but also to ease their handling. We will mainly depend on combination of text mining and natural language processing for development of our proposed system. This will deal with language namely: Uyghur, Kazakh, Mongolian, Uzbek, and Turkmen. The proposed system may be scaled for public uses. The emergence of the online and mobile state of complex networks motivate us to work on developing a versatile and drastic platform for screening and monitoring the languages texts on the communications tools.**

***Index Terms*—Natural Language Processing, Content Intelligence System, data-mining, multi-language monitoring.**

I. INTRODUCTION

This study as the background based on the present status of the cross language network, common behavior of multi-language and cultural integration in China. In order to find for the social stability, possible impact behavior in cross language communication is the research target. There are a wide range of methods and techniques can be appropriate to approach [7][8]. The analytic method of large complex multi language data is based on the network technology, data mining and Natural Language Processing (NLP). As dealing with complex network requires working next to big data and through analyzing the linkage of the network. Thankfully, on the social objects clustering [3] and link analysis [4], there are being achieved prestigious results by researchers. And the outcome is the construction of multi-language complex network behavior monitoring, analysis and prediction application platform, which is necessary to realize abnormal behavior prediction and intervention experiment. Here in we will work on building a multi-language security and intelligence information monitoring platform.

Manuscript received June, 2015.

***Doniyorbek K. Ahmadaliev**, Informatics, Andijan State University., Andijan, Uzbekistan, +8613244316022*

***Chen Xiaohui**, School of Computer Science and IT, Northeast Normal University, Changchun, China.*

Sardor U. Dadabayev, *Informatics, Andijan State University., Andijan, Uzbekistan*

II. TRUE SIGNIFICANCE OF THE STUDY

The total population around the world about 8 billion with annual increase by 70 million approximately. Chinese occupy around third of the world, China not only contains the majority of population worldwide but also, much great diversity of cultures and ethnics. This discrepancy required a smart system for dominating harmony and peace among those people. Which subsequently rise the national feeling regardless their own cultures. This also can prevent terrorists or mentally confused people form affecting China that set back china progress. Although the great advances in the police and army forces, there are many problems in China ascribed to the culture diversity such as violence, protesting, and widespread discrimination feeling. To prevent these bad events we have to find an intelligent system to control such ethnics. Therefore, this motivates us to work on developing a versatile and drastic platform for screening and monitoring the languages texts on the communications tools. Our proposed system may allow Chinese government to observer any dangerous messages and circumvent any bad consequences. Moreover, we can grasp the people behavior, attitude, and thinking ways via filtering those messages. For instances us, numbers of searching for movies or goods, or articles give us a statistic for people attitude. Hence we are fully enthused to work on this topic for the beforehand reasons.

III. CURRENT STUDY AND RELATED WORKS

Previous attempts to build multi-lingual data processing have led to the creation of a multitude of word-nets such as MultiWordNet [11], Predicting the compositionality of multiword expressions [12], Arabic WordNet [13], the Multilingual Central Repository. However, while providing lexical resources on a very large scale, these do not encode semantic relations between concepts denoted by their lexical entries. The research closest to ours should explain and clarify certain language context behavior processing. The pretreatment of civil service interfaces are to be count such as iFLYTEK deployment Uighur speech recognition service and Dimension/Kazakhstan part of speech tagging interface, the interface of the call testing on various environments. In collaboration with the Xinjiang Party committee propaganda

ISSN: 2278 – 1323

department, in 2014 December, there was a test run on Uighur in "Tianshan net search engine service". The engine server uses Ubuntu Linux clustering, Hadoop MapReduce, and Nutch software framework. Design and implementation of Uighur net data acquisition system is now on the stage of text collection.

IV. BACKGROUND

In recent years, research in Natural Language Processing (NLP) has been steadily moving towards multilingual processing [2]. The availability of ever growing amounts of text in different languages, in fact, has been a major driving force behind research on multilingual approaches. Morphosyntactic [1] and syntactic semantic is the clear phenomena to high-end tasks like textual entailment and sentiment analysis.

A. Comparison of similar languages

Breaking up similarity level of the same family and closer languages, such as Altai family languages including Uyghur, Kazakh, Uzbek, Turkmen, Mongolian and others, causes to make it easier to exam and transform between their entries and texts. Mentioned languages are very similar in their writing system and spoken style [3] (Table 1).

Table 1. Some of the Turkic languages similar writing system and spoken style

	Common meaning	Turkish	Turkmen	Tatar	Kazakh	Kyrgyz	Uzbek	Uyghur
Body parts	Head	Baş	Baş	Baş	Bas	Bash	Bosh	Bash
	Hair	Kıl	Gyl	Qıl	Qıl	Kıl	Qil	Qil
	Eye	Göz	Göz	Küz	Köz	Köz	Ko‘z	Köz
	Nose	Burun	Burun	Borın	Murın	Murun	Burun	Burun
	Finger	Parmak	Barmak	Barmaq	Barmaq	Barmak	Barmoq	Barmaq
	Calf	Baldır	Baldyr	Baltır	Baltır	Baltyr	Boldir	Baldir
	Foot	Ayak	Aýak	Ayaq	Ayaq	Ayak	Oyoq	Ayaq
Animals	Horse	At	At	At	At	At	Ot	At
	Dog	İt	It	Et	Ït	It	It	It
	Fish	Balık	Balyk	Balıq	Balıq	Balık	Baliq	Beliq
	Louse	Bit	Bit	Bet	Bït	Bit	Bit	Pit
Other nouns	Bridge	Köprü	Köpri	Küper	Köpir	Köpürö	Ko‘prik	Kövrük
	Arrow	Ok	Ok	Uq	Oq	Ok	O‘q	Oq
	Ash	Kül	Kül	Köl	Kül	Kül	Kul	Kül
	Water	Su	Suw	Su	Su	Suu	Suv	Su
	Sun/Day	Gün(eş)	Gün	Kön	Kün	Kün	Kun	Kün
	Cloud	Bulut	Bulut	Bolıt	Bult	Bulut	Bulut	Bulut
	Bride	Gelin	Gelin	Kilen	Kelin	Kelin	Kelin	Kelin
	Mother	Ana/Anne	Ene	Ana	Ana	Ene	Ona	Ana
	Person	Kişi/Şahıs	Kişi	Keshe	Kisi	Kishi	Kishi	Kishi
	Heart	Yürek	Ýürek	Yöräk	Jürek	Jürök	Yurak	Yürek
	Blood	Kan	Gan	Qan	Qan	Kan	Qon	Qan
	Father	Ata	Ata	Atta	Ata	Ata	Ota	Ata

We can also see the sentence alignment examples written by different Graphic characters of Mongolian. They are used today in Inner Mongolia, Xinjiang, and Mongolia respectively. With the other set of languages Uyghur, Kazakh, Kyrgyz have consists of a sequence of entries. Yidemucao (2013) explains that when performing machine translation (MT), for example, from Chinese to Uyghur or to Kyrgyz, the method may be a rapid way. This similarity allows applying smart natural language processing algorithms and use them efficiently further desire.

It is essential that multilingual information fusion, identifying groups quickly and determining abnormal behavior intervention. The general principles are based on cross-language complex network behavior evolution dynamics.

B. Applying data mining

Text mining is concerned with the detection of patterns in natural language texts, just as data mining is concerned with the detection of patterns in databases. Information processing applications can benefit from having access to both structured information, as found in databases, along with unstructured information, traditionally found in documents or unstructured text fields within databases. When accessing this textual information, applications can also benefit from a more detailed linguistic analysis of the text, as opposed to a shallower "word based" analysis. There are a wide range of techniques that can be applied to analyzing these multi-language entries, as reflected in the considerable amount of research in the field of natural language processing.

C. Data categorizing and analyzing

We can start the procedure with information

ISSN: 2278 – 1323

categorizing which is one of the most popular applications of text mining. Primary, we consider the analysis of textual information and categorization in the context of an capplication for processing text data. In this context, the textual information is dominated by descriptions entered by multilingual incomings. The texts that are encountered are highly constrained with respect to their semantics. These texts reference entries and relationships contained in similar language entry taxonomies. The texts themselves may be highly fragmented and may make use of numerous abbreviations and acronyms. As a result of the constrained nature of the textual information, we are able to leverage the information contained in similar language entry taxonomies5. The process of automated entity monitoring is outlined in Figure 1.

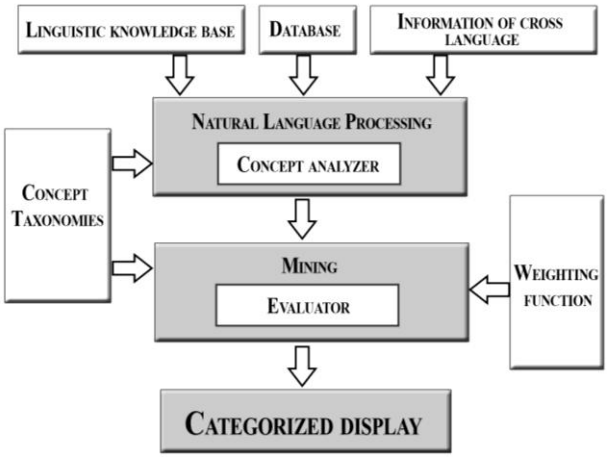

Figure 1. Automated entry monitoring

It illustrates how the output of a natural language processing system, which performs detailed linguistic analysis using domain specific information in the form of Concept Taxonomies, is then used by a mining system to produce.

V. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) deals with the automatic processing and analysis of unstructured textual information. One direction of NLP research relies on statistical techniques, typically involving the processing of words found in texts. Another approach makes use of rule based techniques, leveraging knowledge resources such as ontology, taxonomy, and linguistic rule bases. Statistical human language processing systems require collections of training material which exemplify the desirable (and/or undesirable) relationships and dependencies. Subsequent modification of the system then requires some degree of retraining of the system. Instead of requiring training material, rule based techniques require knowledge in the form of on-line dictionaries, established linguistic theories, and they are able to leverage existing classification systems or taxonomic frameworks. NLP applications may make use of either or both of these techniques, and the decision of which technique to use is often dependent on the availability of training materials, external resources, and the actual text analysis tasks required in the resulting application.

A. Content Intelligence System

The Axonwave Content Intelligence System (CIS) contains core natural language processing systems that perform both rule-based and statistic-based NLP. The CIS is able to leverage existing knowledge sources, in addition provide the capability for unusual.

B. Concept Specification Language

The core technology concerns the matching of "Concepts" which are represented in a Concept Specification Language (CSL). CSL is used to specify rich linguistic patterns that incorporate as fundamental the notion of recursion (embedding) of patterns and various linguistic predicates. CSL and concept matching are embodied in the CIS, which analyzes the structure of words, phrases and sentences (making use of general purpose linguistic rules and dictionaries). Specific information can then be extracted according to rules and concepts formulated with CSL which is organized within various taxonomies. CSL allows the definition of key concepts or terms; and the specification of the interrelationship among concepts in the form of multiple operators.

C. Creating concepts

While it is possible to create very complex and accurate specifications using CSL, this can be a very time consuming task, furthermore, it may require both linguistic expertise, and domain expertise. To facilitate this task, there are pinning control algorithms [5], [10] and we can leverage the linguistic and domain expertise contained within the linguistic rules and knowledge base of a natural language processing system to assist in the creation of new CSL. So, we can bootstrap from an existing system to create a new system that has a richer knowledge base using what we will call text-based concept creation. The text-based concept creation algorithm consists of the following five steps. An example that illustrates each of these steps is then provided in Table 3.

ISSN: 2278 – 1323

Table 3. CSL from text

Algorithm step number	Step function	Examples	
1	Input of text fragments	Uyghur: ك رمه ب ه ب اها كه رو سد يه امريد كا Kazakh/Mongol: Американы Ороска тыйым салуда?	Does America sanction on Russia?
2	Fragments split into words	ك رمه , ب ه ب اها , كه يه ه , روس كا , امري Американы, Ороска, тийим, салуда?	Does, America, on, Russia, sanction,?
3	Selection of relevant words	ك رمه , ب ه ب اها , كه يه ه Американы, Ороска, тийим	America, Russia, sanction
4	Narrowing category on relevant words	Тийим/ كه يه ه	sanction
5	Decision making	The topic of the sentence was related political	

- Input of text fragments. Incoming uncategorized text fragments. These fragments are input to the next step.
- Fragments split into words. The fragments are split into individual words using the Concept Analyzer from Figure 2
- Selection of relevant words. Separation key word matched with database. (Default selection is available).
- Narrowing category on relevant words. In this step information is narrowed and goes matched hyponyms available in Wordnet (or can automatically include them).
- Decision making (accept/analyze/reject/block).

VI. Research Plan and Methodology

Although the specific methods of this research may change during the course of my studies it is important at this point to provide an indication of where my methodological interests lie and how these could be utilized.

At the early stage of research it is required to represent multilingual efficient heterogeneous data fusion method and query technology. At the following stage, gather the most recent publications about the method of complex network of multi language communication behavior and analyze them to be familiar with current research trend. The current phase of study as scheme can be explained by dividing three: the research object, method, and goal. As an object we select required languages to build multi language model of complex network and community behavior. As the methodology on this stage community division method is based on inter node interaction strength and we will take through temporal and spatial characteristics of topic fusion method based on community.

Meanwhile, the complex network structure characteristics of the abnormal communication behavior can be included in this stage. Extracting of the findings from multi language characteristics of multi language and recognition of cross language abnormal behavior will be taken as a goal. Next is effective prediction and intervention method of cross language communication behavior. On this stage, we will take cross language, communication behavior and intervention strategies as research objects. Then, we may apply more methods, as the followings: temporal and spatial variation characteristics of complex network structure model, combined with the trend model of hot topic, finding the technical calculation algorithm of online complex network structure, automatic designing algorithm of balance information and the node influence of estimation.

VII. Conclusion

Due to the great impact of building multi-language monitoring platform, it is expected to identify abnormal behavior activity and propaganda on humanity progress. As we required building a multi-language monitoring platform, I have tried to explain and suggest some approachable methods and algorithms in this paper. The system that combines both text mining and NLP, and I have explained combining two methods, between NLP and text mining. We strongly believe that the proposed way will play an essential role on dealing with the problem. And we have seen is that it is possible to gain high value by using NLP techniques to map different sequences of natural language text to a relatively small number of high level indicators.

In the future, when more complex network of behavioral language data becomes available, it will be possible to apply additional data-mining techniques to detect previously unknown abnormal scale of the multilingual data of homogenization.

REFERENCES

[1] Dipanjan Das and Slav Petrov. 2011. Unsupervised partof-speech tagging with bilingual graph-based projections. In proceedings of the ACL of ACL-11, pages 600–609.
[2] Navigli, R., & Ponzetto, S. P. 2012, July. Multilingual WSD with just a few lines of code: the BabelNet API. In Proceedings of the ACL 2012 System Demonstrations (pp. 67-72).
[3] Colomer-de-Simon, P., & Boguná, M. 2012. Clustering of random scale-free networks. arXiv preprint arXiv:1205.2877.
[4] Li, M., Deng, S., Wang, L., Shengzhong Feng, & Fan, J. 2014. Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowledge-Based Systems, 65, 60-71.
[5] Popowich, F. 2005. Using text mining and natural language processing for health care claims processing. ACM SIGKDD Explorations Newsletter, 7(1), 59-66.

ISSN: 2278 – 1323

[6] Sun, W., Lü, J., Chen, S., & Yu, X. 2014. Pinning impulsive control algorithms for complex network. Chaos: An Interdisciplinary Journal of Nonlinear Science, 24(1), 013141.
[7] Shuang Wang, Xiaojun Chen, Joshua Zhexue Huang, Shengzhong Feng. 2012. Scalable Subspace Logistic Regression Models for High Dimensional Data. APWeb: 685-694
[8] Yidemucao, D., Tohuti, A., Yu, Q., & Roukeyanmu, M. 2013. A Transformation Approach between the Similar Language Text. Advanced Materials Research, 791, 1716-1720.
[9] Yong Zhang, Francis Y.L. Chin, Hing-Fung Ting, Xin Han, Chung Keung Poon,Yung H. TsinDeshi Ye. Online Algorithms for 1-Space Bounded 2-Dimensional Bin Packing and Square Packing, Theoretical Computer Science, 2014.
[10] Zhengzhong Yuan, at el. 2013. Exact controllability of complex networks. Nature Communications. 4(2447).
[11] Shoaib, U., Ahmad, N., Prinetto, P., & Tiotto, G. (2014). Integrating multiwordnet with Italian sign language lexical resources. Expert Systems with Applications, 41(5), 2300-2308.
[12] Salehi, B., & Cook, P. (2013, June). Predicting the compositionality of multiword expressions using translations in multiple languages. In SEM, Proc. of the Main Conference and the Shared Task: Semantic Textual Similarity (Vol. 1, pp. 266-275).
[13] Rodríguez, H., Farwell, D., Ferreres, J., Bertran, M., Alkhalifa, M., & Martí, M. A. (2008, May). Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. In LREC.

Doniyorbek K. Ahmadaliev has completed his M.Sc. Namangan Engineering-Pedagogical Institute, Presently he is an assistant teacher at Andijan State University, Andijan city, Uzbekistan.

Chen Xiaohui, She is currently professor of Northeast Normal University, PhD, master of education and technology, member of the China educational technology society.

Sardor U. Dadabayev, Informatics, Andijan State University

ISSN: 2278 – 1323