Design of the Uyghur FrameNet Desktop

Mirejiguli Rezhake and Alifu Kuerban

Abstract—This paper presents a case study aimed at adapting FrameNet Desktop to Uyghur FrameNet. On the example sentence of the lexical unit "Donate" (donate), gives a detailed overview of the annotation tasks and process within FrameNet Desktop, which illustrates many functions of the tool. In what follows, we presents lexical unit report and lexical entry report, which are automatically generated based on annotation completed using the annotation Desktop.

Index Terms—Uyghur, Uyghur FrameNet, annotation Desktop.

I. INTRODUCTION

The Berkeley FrameNet project has developed a lexical knowledge base provided background and the power for the existence and using of words and a number of semantically and syntactically annotated sentences [1]. Successfully, it has also been applied to various natural language processing such as question answering systems, information retrieval and machine translation etc.

Comparing with languages such as English, Chinese, Arabic-based Uyghur semantic analysis research and related methods is rare and lagging, is still in the primary state of exploration. Aim at the current state of Uyghur semantic analysis research is not mature and semantic role labeling is bottlelenecks, we chose the Fillmore semantics as theoretical basis, with ideas of the Berkeley FrameNet and the Chinese CFN design system for reference, combined Uyghur source language frame semantics description system, we constructed a word level Uyghur FrameNet (henceforth, UFN).

As same with the FrameNet in other languages such as English, Chinese, UFN has three major components: frame database, lexical database and annotated example sentences [2]. Frame database as a frame to unit, clearly presents a) frame definition and b) semantic roles and corresponding example sentences, and c) describes concept relationships between the frame and other frames, annotated example sentences contains lots of sentences with frame semantic and syntactic tagging information which would be automatically generated based on annotation completed using UFN Desktop annotation tools. The lexical database contains a) formulas which capture the syntactic ways in which elements of the semantic frame can be realized within the phrases or sentences built up around the world; b) links to semantically annotated example sentences which illustrate each of the potential realization patterns identified in the formula.

Construction of UFN is a huge project, Regardless of a) entering and editing frame descriptions and lists of lexical units, b) extracting sentences from the corpus, c) annotating selected sentences with frame-relevant labels, or d) organizing and displaying the results are need to highly human-computer interactive tools. This paper discussed the design and implementation of the Uyghur FrameNet desktop tool.

II. UYGHUR LANGUAGE

Uyghur language is main communication tool of Uyghur people. It is one of the statutory working languages of Xingjiang Uyghur Autonomous Region, China and common language of other ethnic minorities in Xingjiang for communication as well.

Uyghur belongs to the Turkic language family of the Altaic language system. At present, Uyghur is written in Arabic scripts with some modifications. As same with any kind of language, An Uyghur sentence constituted by words, in which words separated by space or punctuation marks. As an agglutinative language, the structure of Uyghur word is "prefix+stem+suffix1+suffix2+...". Because Uyghur has its own characters and nature, It needs approaching some different methods in processing.

The research on the Uyghur information processing in the 20th century can be dated back to the beginning of 1980’s. Up to now, Uyghur language processing work obtained some promising results. However, comparing with Europe language and Chinese, it still needs a lot of work to be done. Especially on semantic studies of information processing still does not have its own theoretical system. In this paper, we have designed “Uyghur FrameNet Desktop” for the first time in order to meet the urgent requirement of the current semantic research. The result of this paper, have high value on semantic research and application development of Uyghur even the other languages of Altaic.

III. UFN DESKTOP ANNOTATION TASK

Initially, the FrameNet Desktop was designed for work on English. Scholars around the world have recognized that it can be across language used and have general value. Since then, it has been adapted by various sites setting up FrameNets for different languages, such as German, Japanese, Spanish and Chinese.

However, there is a difference between languages the way of lexicalization, therefore, the frame and the frame elements need to be a modified according to the characteristic of different languages. English and Uyghur are quite differ from

Manuscript received February 15, 2014; revised April 18, 2014. This work was supported by China Natural Sciences, Foundation project subsidization (authorized number 61163029) and Xinjiang Uyghur autonomous region university scientific research project (authorized number XJEDU2011I107).

Mirejiguli Rezhake and Alifu Kuerban are with the Information Science and Engineering Institute of Xinjiang University, Urumqi, China (e-mail: mirejiguli123@sina.com, Ghalipk@xju.edu.cn).
structure and knowledge representation, so we make more
efforts on modifying the FrameNet Desktop [3] and adjust
them in accordance with Uyghur characteristics.
Computer-aided annotation of example sentences, the goal is
to document the range of semantic and syntactic combinatorial
possibilities valences of each word in each of its senses.
The Berkeley FrameNet project annotation sentences on
layers, and mapping from semantic meaning to syntactic
expressions can be realized easily [3]. Similar approach has
been taken for Uyghur sentence annotation in UFN Desktop.
For a given sentence, target word and it is semantic frame,
annotate to direct dependents of target words, and do not
annotate constituents that are understood only through
context as referring to the filler of a particular frame element
role [3]. These dependents are annotated, on their separate
layers, for frame element name, grammatical function (GF),
and phrase type (PT) and these three types; for example:

```
benf-np-attr > < dt هم پيالی time-tp-adv > < pp جىق tgt-np-sub  
فۇزون كېرىش جەمئىيىتگە خېلي جىق پۇل ئىئبنە قىلىنذى

(CNI)
```

(Every year he donates a lot of money to the Red Cross Society)

In above sentence, target word belongs to (giving) frame, using “tgt” lables annotate to target; agt (agent), time,
benf (benefactor) and pat (patient) are semantic role lables;
np(noun phrase), tp (time phrase) are phrase type lables
(subject), adv (adverbial), attr(attribute) obj (object) are
syntactic function labels. Pp (personal pronoun), dt (time
adverbs), ao (positive adjective), n (nouns) , dm (state adverbs)
and v (verbs) are part-of-speech lables.

It is known from the above analysis, UFN sentence labeled
objects as the target word, frame elements, phrase type,
syntactic function and part-of-speech. Below, we discuss
different terms.

A. Target Word

UFN is lexical resource for Uyghur, based on frame
semantics and supported by corpus evidence. Frame
semantics believe that, frame is the background and
motivation of word understanding. In UFN, Uyghur verbs,
adjectives, and in the event noun can act as a frame. A frame
can contain more than one target word, that is, a frame that
can be activated multiple target word. When we annotate
an example sentence, we always do so from the point of view
one particular target word in the sentence.

B. Frame Element

Frame element is a frame participant and the support which
is a concept role presented in the scheme scene. The frame
element describes the semantics and the syntactic features of
the frame element with corresponding demonstration. It
carries on the concise description, enables the sentence to
be understood, and supplies the regular information for text
labeling. In the different frame the numbers and the concept
roles of frame are different due to the cognitive background
of frame element [2]. The frame elements may be divided
into the core and non-core frame elements according to
participation and the support function of frame element. A
core frame element is one that instantiates a conceptually
necessary component of a frame, and must be showing the
properties of the frame. Non-core frame element can’t show
any individual character of a frame, and it can occur in most
frames.

In the UFN corpus, some core frame elements are
conceptually necessary but do not occur as lexical or phrasal
material; these we call null instantiations. There are three
types of null instantiations-constructional, definite and
indefinite. We will now briefly discuss the three types of
omission. Constructional Null Instantiation (CNI) is licensed
by a grammatical construction, as with the missing subject of
an imperative sentence, which does not depend on the
identity of any particular lexical item [3]. For example, in the
Giving frame, all lexical units happen to allow the core frame
element donor to be omitted under constructional null
instantiation. This is shown for lexical unit Donate in
following sentence 1): With Definite Null Instantiation (DNI)
the oMissing core frame element must be understood in the
linguistic or discourse context [3]. In sentence 2) below, in
the Giving frame, the recipient in not expressed overtly in the
sentence, but it is understood from the context. Indefinite null
instantiation (INI) covers missing objects of certain transitive
verbs that are characterized as used intransitively; the
missing of object of such a verb often has a special
interpretation [3]. In sentence 3) below, in the Giving frame,
the missing object of give is most likely to be audience.

```
1(1) قۇبۇل قىلغۇچي جەمئىيىتگە خېلي جىق پۇل ئىئبنە قىلىنذى
(Lots of money donated to the Red Cross Society. )

2(2) جېرىشچي كېرىش جەمئىيىتگە خېلي جىق پۇل ئىئبنە قىلىنذى
(He donated $thousand.)

3(3) تەختخۇچي كېرىش جەمئىيىتگە خېلي جىق پۇل ئىئبنە قىلىنذى
(It gives an overview of Xinjiang’s history and
daily life and tradition.)
```

C. Grammatical Function (GF)
The grammatical functions (GF) that are assigned describe
the ways in which the constituents satisfy abstract grammatical
requirements of the target word. In annotation, only constituents tagged
with frame elements are assigned grammatical functions.

IV. ANNOTATION PROCESS

The UFN Desktop software has deployed that combines
and runs on c# (2.0) application system using SQL2008 for
data storage. We now turn to UFN Desktop annotation
process. Specific process is shown in Fig. 1. In annotation,
our main focus is on the goal of recording the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses. To that end, we extract sentences from the different texts of a corpus. Then we annotate a selection of the extracted sentences in respect to that particular LU. Once annotation is completed, we can examine the semantic and syntactic patterns.

FrameNet annotations are constellations of triples that make up the frame element realization for each annotated sentence, each consisting of a frame element, a grammatical function and a phrase type.

On the basis of the above tasks, the semantic and syntactic Valency pattern of a lexical unit generated automatically in the annotated sentences database. The above example sentence takes the following patterns:

Semantic pattern (read right to left):

\[
\text{بەرگۉچي} \text{(agt)} + \text{ۋاقىت} \text{(time)} + \text{ئېرىشكۉچي} \text{(benf)} + \text{ئوبيېكت} \text{(pat)}
\]

Syntactic pattern (read right to left):

noun phrase+Noun phrase+data phrase+noun phrase

\[\text{[object]} \phantom{x} \text{[attribute]} \phantom{x} \text{[adverbial]} \phantom{x} \text{[subject]}\]

V. THE IMPLEMENTATION OF THE UFN DESKTOP

The system design for UFN annotator, aim at reduce the format and spelling mistakes problems in the process of annotation, realized the lexicon automatically generated, and presentation lexical unit and lexical entry report.

UFN Desktop annotation window as shown in Fig. 2, divided into four workspace: workspace 1) is frame list, contains an expanding list that gives direct access to the main kinds of objects, such as frame, frame elements, and lexical units; workspace 2) is selecting sentences; workspace 3) is annotation layers; workspace 4) is lists of labels available for each of the annotation layers.

Once annotator chose a lexical unit for annotation, a list of sentences appear in the workspace 1). annotator read through the sentences and select a with variety information of semantic and syntactic pattern for annotation. Selected sentence appear in the workspace 2), and which contains 7 seven annotation layers are target, frame element, phrase type, grammatical function, other, sentence and part of speech layers. through mouse dragging annotator select target and its direct dependents, and then use corresponding labels as [4]-[6] to filling each of layers in the workspace 3).

After annotation is complete, through review the lexical unit report and lexical entry report we can examine annotated result and valence patterns they manifested. these reports are automatically generated based on annotation completed using UFN Desktop annotation tools. The two parts of this report are illustrated in Fig. 3 and Fig. 4, respectively, for lexical unit

As displays in Fig. 3, the top of the report is a list of frame element; this is called Frame Element table. The frame has core frame elements such as (donor), (recipient), (theme), and has non-core frame elements such as (circumstance), (means), (place) (purpose) (reason), (time), (description), (means), (purpose)
There are still some shortcomings in the design, and we are adding some Shortcuts and more linguistic information for convenience of annotator. Our next step is using this UFN Desktop tool to build systematic and comprehensive corpora, and actually realize the Uyghur automatic semantic role labeling and question answering.

VI. CONCLUSION

The Desktop annotation software has been designed for the UFN sentence annotation. The aim is to obtain various semantic and syntactic patterns which example sentences manifestation.

The UFN, by taking the large-scale Uyghur text language materials as the basis, and translating the Berkeley FrameNet project frames, and its corresponding information into equivalent frames in the Uyghur FrameNet. So far, approximately 402 frames have been constructed, with more than 4252 lexical unit, including verb, noun, adjective, covers broad semantics domain, annotated more than 18500 sentences.

REFERENCES

Mirejiguli Rezhake was in Turpan, Xinjiang, China, in 1986. She earned her bachelor degree in computer science and technology from Information Science and Engineering Institute of Xinjiang University, Urumqi, Xinjiang, China in June 2011. She is a postgraduate student in Information Science and Engineering Institute of Xinjiang University, Urumqi, China. Her research area are natural language processing.

Alifu Kuerban was born in Korla, Xinjiang, China in 1967. He earned his master’s degree in software engineering, Institute of Software Engineering of University of Electronic Science and Technology of China, Chengdu, China, in 2011. From 1993 to 2012, he worked as a lecturer in Xinjiang University, a professor from 2012-present. His research areas: ethnic minority language processing, database application.

So far he has published more than 20 articles. In 2008 he won the second prize of the “Science and Technology Progress”, China; in 2009 won “Excellent Teacher” in Xinjiang Uyghur autonomous region, China.