Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus

Kahaerjiang Abiderexiti*†, Maihemuti Maimaiti*†, Tuergen Yibulayin*†, and Aishan Wumaier*†✉
*School of Information Science and Engineering,
Xinjiang University,Urumqi, China
†Xinjiang Laboratory of Multi-Language information Technology,
Xinjiang University,Urumqi, China
kaharjan@xju.edu.cn, mahmutjan@xju.edu.cn, turgun@xju.edu.cn, hansan1479@xju.edu.cn

***Abstract*—The Uyghur language is a minority language in China, and it is one of the official languages in Xinjiang Uyghur Autonomous Region of China. Approximately 10 million people use Uyghur in their daily lives and regular use is even found on the internet. However, the lack of an Uyghur named-entity relation corpus constrains Uyghur language extraction applications. In this paper, we will propose such an Uyghur named-entity and Uyghur named-entity relation annotation specifications based on existing guidelines and experiences in other languages for Uyghur corpus construction. By sampling raw text from Uyghur language websites, a small experiment has been conducted concerning the practicality of our annotation schemes using an annotation tool. After review, we conclude that this method has practical future applications for other resource-poor minority languages of the world. This schemes will provide a basis for further studies of entity relation corpus construction.**

***Keywords*-Uyghur; named entity; relation; annotation;corpus**

I. Introduction

As the internet has developed as spread, minority languages have emerged, increasing the linguistic diversity found on the world wide web. However, the lack of corpora in minority languages constrains potential information extraction applications for these languages. The Uyghur language, a Uyghuric branch Turkic language of China, is one of these emerging languages of the Internet. Uyghur language communities are primarily situated in the Xinjiang Uyghur Autonomous Region in China, where it is an official language. Around the world, the Uyghur language is spoken by over 10 million people, and it is fifth most-spoken Turkic language [1]. Although there are several Uyghur corpora[1][2][3][4][5], until now there are no reports on constructing an Uyghur named-entity relation corpus. As the result, Uyghur suffers from lacking of various information extraction systems.

In this instance, an "entity" is defined as an object or set of objects in the world. If an entity is referenced by name, it is called a "named-entity". Named-entities include personal names, organizational names, cartographic and geopolitical names and even titles. For example: Obama, China, United Nation, Tarim River, Admiral. Named-entity relations refer to targeted relations between entities. Relations are ordered pairs of entities. For example, in the sentence, *Scenery at Altun Mountains Nature Reserve in Xinjiang*, there is a relation between entity *Altun Mountains Nature Reserve* (argument 1) and *Xinjiang* (argument 2). The type of this relation would be *argument 1* located in *argument 2*.

In this work, we first describe an annotation scheme for named-entity and named-entity relations in Uyghur. These schemes are grounded in existing research and consider the proper morphosyntactic characteristics of Uyghur. Then we describe a corpus sampling & annotation process for constructing an Uyghur named-entity relation corpus. Finally, we examine the practicality of this annotation plan through inter-annotator agreement.

II. Related Work

Since the Message Understanding Conference (MUC)[6] began including named-entity recognition tasks in 1995, much research has been conducted in the field of named-entity recognition and relation detection tasks[7]. Many corpora[2,3], annotation tools[8][9] and evaluation benchmarks[10][11] have been developed for entity and relation tasks in many of the major languages of the world (i.e. English, Chinese, Arabic, etc.). Among them, Automatic Content Extraction (ACE) programs made important contribution to the information-extraction literature. In a sense, it is a continuation of the MUC, and defines entity, relation, and event extraction tasks. The Linguistic Data Consortium (LDC) developed annotation specifications, annotation tools and corpora for English, Chinese, Spanish and Arabic[4] to support the ACE program. Recently, LDC has been developing the ERE (Entities, Relations, Events) program[5] under the DARPA's Deep Exploration and Filtering of Text (DEFT) program. The ERE can

✉Corresponding author.

[1]http://en.wikipedia.org/wiki/Turkic_languages

[2]http://www-nlpir.nist.gov/related_projects/muc/

[3]http://www.itl.nist.gov/iad/mig/tests/ace/

[4]http://www.ldc.upenn.edu/collaborations/past-projects/

[5]http://www.ldc.upenn.edu/collaborations/current-projects

be regarded as simplification of ACE[10], differing from ACE in terms of separate goals regarding scope and replicability[7]. The ERE program has been developed further to become more complex than previous attempts[12].

In Uyghur language processing, there are several existing methods for morphological analysis[13][14] which are adequate for real applications like machine translation[6], and orthographic transcription[7]. Some works about Uyghur corpus construction exist, specifically, the Uyghur POS tagged corpus[1], and the knowledge base (including various dictionaries and treebanks[2][15]) are initially constructed by Xinjiang Normal University. Building a Modern Uyghur balanced corpus[16], Uyghur dependency tree bank[17], FrameNet [3], paraphrase [4], ontology[18] and grammatical information dictionary[5] were explored, construction having been initiated by Xinjiang University. The survey on Uyghur person-name recognition[19] informs us that there are only 4 existing studies concerning Uyghur person name identification methods in 2011-2015. However, to the best of our knowledge, there are no reports on annotation schemes for constructing Uyghur named-entity relation corpora in the literature.

III. Outline of Annotation Schemes

The idea of the ACE[10], ERE[7] and Uyghur knowledge base[2] are followed to define Uyghur Named-entity (UyNe) annotation scheme and Uyghur Named-entity Relation (UyNeRel) annotation scheme. This complies with the rule that future studies "be grounded in existing research as much as possible"[20]. At the same time, new tags will need to be introduced to address the properties of Uyghur.

A. UyNe Scheme

1) Types of Entities and Basic Annotation Unit: In the process of defining UyNe annotation specification, simplicity in the ERE is adopted by only annotating entities of Person(PER), Organization(ORG), Geo-Political Entities(GPE), Location(LOC) and Facility(FAC). Like ERE annotation scheme, entities are not divided into subtypes but used Title(TTL), Age and URL as argument fillers in relation. Considering the resource-scarcity and the morphological complexity of Uyghur, the plan is simplified further. In our new plan, the definition of "entity" does not include the pronominal and nominal entities that ACE and ERE included. In other words, we define named-entity mention as the reference of the entity in a text, only indicated by a common noun. For example: named entity mention PER[تاشپولات تيىپ] (Tashpolat Tiyip) is annotated, but the pronominals ئۇ ، ئۇنىڭ ،ئۇنى (He/She, His/Hers, Him/Her) and nominals مۇدىر-نامزاتى (director candidate) which refer to this named entity are not annotated (denoted by strikeout in this example).

Because Uyghur is an agglutinating language (a language which morphologically attaches affixes to phrasal units), the same entity may have a significant number of forms with the connection of various suffixes. Stem-based annotation[2] helps to overcome this issue. However, in named-entity annotation, entity stem-forms give annotators additional work, since one must identify or modify for the correct stem-form. Annotators not only have to be familiar with entity and relation annotation, but also have to be familiar with Uyghur morphological analysis. To cope with this, a trade-off is required between time and efficiency spent on annotating. Furthermore, much work is done on Uyghur morphological analysis[13][14] which is used in real applications, as mentioned above. Achievements in Uyghur morphological analysis make it possible to do automatic morphological analysis after entity and relation annotation. So we chose surface form as the basic entity annotation unit.

2) UyNe Annotation Rules : In defining UyNe annotation rules, ERE entity annotation scheme is adopted. Entities are annotated according to usage. For example:

[برازىلىيە]ORG	نۆقتا توپ	جېگىدە	غەلبە قىلدى.
ORG[Brazil]	penalty	race	won

In the above sentence, Brazil is annotated as ORG, because it referenced as Brazil football team.

As mentioned above, regardless of suffixes, we annotate surface form of entities. For example:

[سەيپىدىن ئەزىزىنىڭ]PER		ئەسلىمىسى
PER[Saypidin Azizi's]		memoirs
[قەشقەردىكى]GPE	ئۆزگىچە	مەدەنىيەت
GPE[at Kashgar]	special	culture

There is a lack of variety in this Uyghur web site. In order to mine every possible relation in the small annotated corpus, we annotate entities in an overlapping manner. For example:

[[[شىنجاڭ]GPE	ئۇنىۋېرسىتتى]ORG	مالىيە	باشقارمىسى]ORG	
[[Xinjiang]GPE	University]ORG	financial	department]ORG	

In addition, in above example, although *Xinjiang* in the *Xinjiang University* express the name of the University, it also express that this university located at *Xinjiang* in this particular example. We describe details of UyNeRel in next section.

B. UyNeRel Scheme

1) Types of UyNeRel: In UyNeRel there are 5 types of relations different from the ACE and Light ERE, similar to the Rich ERE. These are *Physical*, *Part-Whole*, *Gen-Aff* (General-Affiliation), *Per-social*, *Org-*

[6]http://www.tilmach.cn/Home/Translation
[7]http://www.tilmach.cn/Home/Convert

Table I Differences between Rich ERE and UyNeRel

Rich ERE 5 types, 20 subtypes		**UyNeRel 5 types, 15 subtypes**	
Types	**Subtypes**	**Types**	**Subtypes**
Physical	OrgHead-Quarter LocatedNear Resident OrgLocOrigin	Physical	Located Near
Gen-Aff	MORE OPRA PersonAge OrgWebSite	Gen-Aff	PersonAge OrgWebSite
Part-Whole	Subsidiary	Part-Whole	Subsidiary Geographical
Per-Social	Business Family Unspecified Role	Per-Social	Business Family Other Role
Org-Aff	Employment-Membership Leadership Invest-Shareholder Student-Alum Ownership Founder	Org-Aff	Employment Invest-Shareholder Student-Alum Ownership Founder

Aff (Organization-Affiliation). Relation subtypes differ from Rich ERE which have 20 subtypes but in our annotation scheme there is 15 subtypes. The difference is shown by Table I.

In Rich ERE annotation specification, relation type *Physical* divided into four subtypes. However, in Uyghur corpus Organization-Headquarter and OrgLocOrigin (OrganizationLocationOrigin) subtype relation frequency is very low, and the annotated corpus won't be as large as English or other major world languages. We merge this with subtype *Near*. As a result, UyNeRel relation type *Physical* divided into two subtypes. In the *Located* subtype, first argument of the entity should be PER, means Person *Located* in FAC, LOC or GPE. The example is shown by Table II.

In the *Gen-Aff* (General-Affiliation) relation type, considering the simplicity, UyNeRel adopted two subtypes, which is *Person-Age* and *Organization Web Site*. On the contrary, in the *Part-Whole* relation, instead of defining *Membership* subtype, we define *Geographical* subtype, which captures the location of an entity, such as FAC, LOC or GPE in or at or as a part of another FAC, LOC or GPE. The example is shown by Table III.

Table II Examples of *Physical* Relations

ئالىم قاناستا قاپسلىپ قالدى. (Alim was trapped in Kanas.)		
Type.Subtype	Argument1	Argument2
Physical. Located	ئالىم (Alim)	قاناستا (in Kanas)
شىنجاڭ گەنسۇ ئۆلكىسىنىڭ غەرب تەرىپىگە جايلاشقان. (Xinjiang is located in the west of Gansu Province.)		
Type.Subtype	Argument1	Argument2
Physical. Near	شىنجاڭ (Xinjiang)	گەنسۇ ئۆلكىسىنىڭ (Gansu Province)

Table III Examples of *Gen-Aff* Relation

جۇڭگو شىنجاڭ ئۇيغۇر ئاپتونوم رايونى (Xinjiang Uyghur Autonomous Region of China)		
Type.Subtype	Argument1	Argument2
Part-Whole.Geo	شىنجاڭ ئۇيغۇر ئاپتونوم رايونى (Xinjiang Uyghur Autonomous Region)	جۇڭگو (China)

Table IV Examples of *Per-Social* Relations

ئەلى قۇربانىنىڭ ئادۋوكاتى ئەنۋەر (Eli Qurban's lawyer is Enwer)		
Type.Subtype	Argument1	Argument2
Per-Social.Business	ئەلى قۇربانىنىڭ (Eli Qurban's)	ئەنۋەر (Enwer)
ئابدۇكېرىم ئابلىزنىڭ ئايالى رۇقىيە (Abdukërim Abliz's wife Ruqiye)		
Type.Subtype	Argument1	Argument2
Per-Social.Family	ئابدۇكېرىم ئابلىزنىڭ (Abdukërim Abliz's)	رۇقىيە (Ruqiye)
پروفىسسور تۈرگۈن ئىبراھىم (Professor Türgün Ibrahim)		
Type.Subtype	Argument1	Argument2
Per-Social.Role	تۈرگۈن ئىبراھىم (Türgün Ibrahim)	پروفىسسور (Professor)
ئالىمجاننىڭ يۇرتداشى سەمجان (Alim's fellow-townsman is Semijan)		
Type.Subtype	Argument1	Argument2
Per-Social.Other	ئالىمجاننىڭ (Alim's)	سەمجان (Semijan)

The *Per-Social* (Person-Social) is a relation that is not in the category of *Business*, *Family* or *Role* type, but rather belongs to *Other* type. The example is shown by table IV.

In the *Org-Aff* (Organization-Affiliation) relation type, we combine *Employment-Membership* and *Leadership* in Rich ERE to *Employment* in UyNeRel. This makes the annotation task easier, as this is a simple distinction between subtypes in *Org-Aff*. The example is shown in Table V.

C. UyNeRel Annotation Rules

We only annotate relations between two entities within one sentence. This can be seen from table II to table V. Like ACE and ERE annotation rules, we annotate relations according to its usage. This means that if the relation existed in the real world, but not in a single sentence, it will not be annotated. For

Table V Examples of *Org-Aff* Relations

هندىستان زۇكتۇكى مۇكارجى (Indian President Mukherje)		
Type.Subtype	Argument1	Argument2
Org-Aff. Employment	مۇكارجى (Mukherje)	هندىستان (Indian)
شەندۇڭ ››رۇيى‹‹ گۇرۇھى قەشقەر ۋىلايىتىگە20 مىليارد يۈەن مەبلە غ سالدى. (Shandong Ruyi Group has invested 20 billion yuan in Kashgar Prefecture.)		
Type.Subtype	Argument1	Argument2
Org-Aff. Shareholder	شەندۇڭ ››رۇيى‹‹ گۇرۇھى (Shandong Ruyi Group)	قەشقەر ۋىلايىتىگە (Kashgar Prefecture)
ئادىل شىنجاڭ ئۇنىۋېرسىتېتىنى پۈتتۈرگەن. (Adil graduated from Xinjiang University)		
Type.Subtype	Argument1	Argument2
Org-Aff. Student-Alum	ئادىل (Adil)	شىنجاڭ ئۇنىۋېرسىتېتىنى (Xinjiang University)
شەھەرىستان مۇزىكىلىق تاماللار رېستورانىنىڭ خوجايىنى مەخمۇت (The Owner of Sheheristan Music Restaurant is Mahmut)		
Type.Subtype	Argument1	Argument2
Org-Aff. Owner	مەخمۇت (Mahmut)	شەھەرىستان مۇزىكىلىق تاماللار رېستورانىنىڭ (Sheheristan Music Restaurant)
ئالىبابانىڭ قۇرغۇچىسى ما يۈن (The founder of Alibaba is Jack Ma)		
Type.Subtype	Argument1	Argument2
Org-Aff. Founder	ما يۈن (Jack Ma)	ئالىبابانىڭ (Alibaba)

example, in the sentence below, we can annotate قاسىم and قەھرىمان as a Per-Social.Family.

قاسىم ۋە ئۇنىڭ ئىنىسى قەھرىمان بىللە تىجارەت قىلىدۇ.
. do business together Qehriman brother his and Qasim
(Qasim and his brother Qehriman do business together.)

However, in the sentence below we can't annotate قاسىم and قەھرىمان as a Per-Social.Family even if it can be seen from context. Because it is not expressed within one sentence.

قاسىم ۋە قەھرىمان بىللە تىجارەت قىلىدۇ.
.do business together Qehriman and Qasim

(Qasim and Qehriman do business together.)

We will not annotate negative relation. For example:

رەنا ھازىر بېيجىڭدا ئەمەس.
.not Beijing at now Rena

(Rena is not in Beijing at the moment.)

IV. Corpus Sampling and Annotation Process

Once the preliminary annotation schemes for the UyNe and UyNeRel are set up, articles in Uyghur websites are sampled as a source for text. These sites include Uyghur version of Tianshan Net [8] , People's Daily [9] and Xinhua News [10] from which texts are collected by using a combination of automated and manual efforts. The reasons of selection of these sites are that these web sites are government based news agencies and content of news are representative, and authoritative.

To construct the corpus, the general procedure is as follows: first, a set of web pages containing the articles are downloaded from the aforementioned websites; second, HTML tags, image captions, and other advertisement contents are excluded by using the webpage analyzer that is able to identify unique structures of these web sites; third, informations and main contents of the articles are saved to database in order to handle information such as added time, source, and other related info. This corpus construction is still in progress. As Of September 2016, the corpus contains approximately 2000 articles.

Since cleaned articles are obtained, human annotators begin to annotate articles by using MAE 2V annotation tool[11] which is improved version of MAE 0.7V[8]. In the annotation tool, XML Document Type Definition (DTD) is used to define UyNe and UyNeRel annotation scheme. The sample of DTD file is shown in Figure 1. The string !EEMENT indicates tags (FAC, PartWhole,Physical). The #PCDATA indicates that the information about entities and relations will be parsable character data. The !ATTLIST line declares attributes like ID, mention level, comment and possible three value of mention level. In designing this mention level attribute, we consider the long term research. So although in UyNeRel specification, the annotator is only required to annotate named-entity mention (NAM), we still set other two values and make NAM default value by #IMPLIED="NAM".

```
<!ELEMENT FAC ( #PCDATA ) >
<!ATTLIST FAC id ID prefix="FAC" #REQUIRED>
<!ATTLIST FAC mentionLevel ( NAM | NOM | PRO ) #IMPLIED "NAM">
<!ATTLIST FAC comment #CDATA>
<!ELEMENT PartWhole EMPTY>
<!ATTLIST PartWhole id ID prefix="Part.Wh" #REQUIRED>
<!ATTLIST PartWhole subType ( Geo | Subsidiary )>
<!ATTLIST PartWhole comment #CDATA>
<!ELEMENT Physical EMPTY >
<!ATTLIST Physical id ID prefix="Phys" #REQUIRED>
<!ATTLIST Physical subType ( Located | Near )>
<!ATTLIST Physical comment #CDATA>
```

Figure 1 Sample of DTD file.

Although this tool could load UTF-8 text format, because Uyghur script is based on Arabic script written from right to left, the tool is slightly modified to display Uyghur letter. However, after preliminary annotation, it is found that this software does not handle well Arabic script and suffer from slow speed. So we have developed our own annotation tool using C#.

In order to measure our annotation plans initially, a small experiment was conducted. Two college students whose mother language is Uyghur, taught Uyghur language from elementary school to high school, are

[8]http://uy.ts.cn
[9]http://uyghur.people.com.cn
[10]http://uyghur.news.cn
[11]https://github.com/keighrim/mae-annotation

asked to learn the annotation guidelines. A random sample of 10 documents from our raw corpus database will then be annotated by them independently. The result of Cohen's K inter-annotator agreement (IAA) for UyNe is currently 0.7, and UyNeRel 0.6; this should remain stable as the size of the corpus increases. This will mean that the annotation plans are valid for Uyghur. From discussions with annotators about the errors and inconsistency, it can be concluded that more special cases are required for the guideline that prevent annotators from confusion. Annotation is currently under progress, 100 documents currently completed.

V. Conclusions and Future Work

The main aim of this project at the first stage is to investigate theoretical foundations, mine raw text from internet sources, then tag the resulting collected data according to our annotation guidelines. Existing research on other languages has been utilized, along with existing corpus annotation experience for Uyghur, the UyNe and UyNeRel. In this plan, the sparseness of Uyghur corpus is accounted for by eliminating some tags, and adding others. Because Uyghur uses the Arabic script, a new annotation tool was required, and so developed. Preliminary annotation results are promising, indicating that the annotation plan is adequate for the data under consideration. Further corpus sampling and annotation is in progress.

Future plans include: improving annotation tool to alleviate the workload of annotators; and expanding the annotated corpus size. The annotated corpus will be publicly available in the near future[12]. We think this may promote Uyghur information extraction research in the NLP community.

VI. Acknowledgments

This work has been supported as part of the NSFC (61462083, 61262060, 60963018, and 61463048), 973 Program (2014cb340506), National Social Sciences Foundation of China (10AYY006), Autonomous Region Project for Studying Abroad in 2015, Science and Technology Talents Training Project for Young Dr (qn2015bs004).

References

[1] Y. Aibaidula and K. T. Lua, "The Development of Tagged Uyghur Corpus," in *17th Pacific Asia Conference on Language, Information and Computation*, D. H. L. K. T. Ji, Ed., 2003, pp. 228–234.

[2] Y. Abaidulla, I. Osman, and M. Tursun, "Progress on Construction Technology of Uyghur Knowledge Base," in *2009 International Symposium on Intelligent Ubiquitous Computing and Education,*, 2009, pp. 554–557.

[3] A. Kuerban, W. Kuerban, and N. Abdurusul, "Research on Uyghur FrameNet Description System," in *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, 2009, pp. 160–163.

[4] K. Abiderexiti, M. Maimaiti, A. Wumaier, and T. Yibulayin, "Construction of Uyghur Initial Paraphrase Corpus," in *Proceeings of the International conference "Turkic Languages Processing" TurkLang 2015*, Russia,Tatarstan,Kazan, 2015, pp. 87–90.

[5] J. Wushouer, W. Abulizi, K. Abiderexiti, T. Yibulayin, M. Aili, and S. Maimaitimin, "Building Contemporary Uyghur Grammatical Information Dictionary," in *Worldwide Language Service Infrastructure*, Y. Murakami and D. Lin, Eds. Cham: Springer International Publishing, 2016, pp. 137–144.

[6] B. M. Sundheim, "Overview of results of the muc-6 evaluation," in *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics, 1995, Conference Proceedings, pp. 13–31.

[7] J. Aguilar, C. Beller, P. McNamee, and B. Van Durme, "A Comparison of the Events and Relations across ACE, ERE, TAC-KBP,and FrameNet Annotation Standards," in *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2014, Conference Proceedings, pp. 45–53.

[8] A. Stubbs, "Mae and mai: Lightweight annotation and adjudication tools," in *Proceedings of the Fifth Law Workshop (LAW V)*. Association for Computational Linguistics, 2011, Conference Proceedings, pp. 129–133.

[9] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, and B. Eickmann, "Gerbil: General entity annotator benchmarking framework," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, Conference Proceedings, pp. 1133–1143.

[10] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel, "The Automatic Content Extraction (ace) Program-Tasks, Data, and Evaluation," in *LREC*, 2004, Conference Proceedings.

[11] S. Kulick, A. Bies, and J. Mott, "Inter-annotator agreement for ere annotation," in *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2014, Conference Proceedings, pp. 21–25.

[12] Z. Song, A. Bies, S. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma, "From Light to Rich ERE: Annotation of Entities, Relations, and Events," in *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, 2015, Conference Proceedings, pp. 89–98.

[13] A. Wumaier, P. Tursun, Z. Kadeer, T. Yibulayin, A. Wumaier, P. Tursun, Z. Kadeer, and T. Yibulayin, "Uyghur Noun Suffix Finite State Machine for Stemming," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*. Beijing: 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 161–164.

[14] M. Aili, W.-B. Jiang, Z.-Y. Wang, T. Yibulayin, and Q. Liu, "Directed Graph Model of Uyghur Morphological Analysis," *Journal of Software*, vol. 23, no. 12, pp. 3115–3129, 2012.

[15] Y. Ebeydulla, H. Abliz, and A. Yusup, "Research on the Uyghur Information Database for Information Processing," in *2011 International Conference on Asian Language Processing (IALP 2011)*. 2011 International Conference on Asian Language Processing, 2011, pp. 26–29.

[16] Turgun · Ibrahim and Y. Baoshe, "A Survey on Minority Language Information Processing Research and Application In Xinjiang," *Journal of Chinese Information Processing*, vol. 25, no. 06, pp. 149–156, 2011.

[17] M. Aili, A. Xialifu, and S. Maimaitimin, "Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools," in *Worldwide Language Service Infrastructure*. Springer, 2016, pp. 124–136.

[18] H. Yilahun, S. Imam, and A. Hamdulla, "A Survey on Uyghur Ontology," *International Journal of Database Theory and Application*, vol. 8, no. 4, pp. 157–168, 2015.

[19] T. Nizamidin, P. Tuerxun, A. Hamdulla, and M. Arkin, "A Survey of Uyghur Person Name Recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 3, pp. 273–280, 2016.

[20] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., 2012.

[12] https://github.com/kaharjan/UyNeRel