Construction of Uyghur Named Entity Relation Corpus

Kahaerjiang Abiderexiti, Maihemuti Maimaiti, Tuergen Yibulayin, 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, mahumtjan@xju.edu.cn, turgun@xju.edu.cn, hasan1114@xju.edu.cn

Abstract

The Uyghur language is a minority language in China, and it is one of the official languages in the 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, lack of an Uyghur named-entity and named-entity relation corpus constrains Uyghur language extraction applications. First, we propose such a Uyghur named-entity and named-entity relation annotation specifications based on existing guidelines and experience in other languages for Uyghur corpus construction. Then, we have developed annotation software for these specifications. Finally, we have constructed the first Uyghur named-entity relation annotation corpus by manual annotation. This corpus is released under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The corpus can be accessed from `https://github.com/kaharjan/UyNeRel`.

Keywords

Uyghur; named entity; relation; annotation; corpus construction

1. Introduction

As the Internet has developed as an International information sharing platform, ethnic minority languages have emerged, increasing the linguistic diversity found on the world wide web. As the result, there is a new demand for information extraction in ethnic minority languages. However, the lack of corpora in ethnic minority languages constrains information extraction applications for these languages. The Uyghur language, an agglutinative language with a subject–object–verb word order, 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 one of the main communication language. Around the world, the Uyghur language is spoken by over 10 million people[1]. Although there are several Uyghur corpora (Aibaidula and Lua 2003; Abaidulla et al. 2009; Kuerban et al. 2009; Abiderexiti et al. 2015; Wushouer et al. 2016), until now there are no reports on constructing an Uyghur

[1] `http://www.ethnologue.com/18/language/uig/` accessed November 12, 2017

named-entity relation corpus. As the result, Uyghur suffers from a lack of various information extraction systems.

Information extraction is the task of extracting the essential elements from structured text or knowledge from unstructured text, such as indentifying entities and extracting relationships between entities. 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. As shown in Table 1, *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*.

Table 1: Examples of Entities and Relationship Between them

Scenery at Altun Mountains Nature reserve in Xinjiang		
Argument1	Relationship	Argument2
Altun Mountains Nature Reserve	Located In	Xinjiang

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. We also provide handful examples for reasons of these annotation schemes. Then we describe a corpus sampling, annotation process and annotation tool for constructing a Uyghur named-entity relation corpus. Finally, we give the format of the annotated corpus and its statistics.

2. Related Work

Since the Message Understanding Conference (MUC) (Sundheim 1995) began including named-entity recognition tasks in 1995, much research has been conducted in the field of named-entity recognition and relation detection tasks (Aguilar et al. 2014). Many corpora[2,3], annotation tools (Stubbs 2011) (Usbeck et al. 2015) and evaluation benchmarks (Doddington et al. 2004) (Kulick et al. 2014) 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 contributions 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 be regarded as simplification of ACE (Doddington et al. 2004), differing from ACE in terms of separate goals regarding scope and replicability (Aguilar et al. 2014). The ERE program has been developed further to become more complex than previous attempts (Song et al. 2015) (Mott et al. 2016).

[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`

In Uyghur language processing, there are several existing methods for morphological analysis (Wumaier et al. 2009) (Aili et al. 2012) 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 (Aibaidula and Lua 2003), and the knowledge base (including various dictionaries and treebanks (Abaidulla et al. 2009) (Ebeydulla et al. 2011)) are initially constructed by Xinjiang Normal University. Building a Modern Uyghur balanced corpus (Turgun·Ibrahim and Baoshe 2011), Uyghur dependency tree bank (Aili et al. 2016), FrameNet (Kuerban et al. 2009), paraphrase (Abiderexiti et al. 2015), ontology (Yilahun et al. 2015) and grammatical information dictionary (Wushouer et al. 2016) were explored, construction had been initiated by Xinjiang University. The survey on Uyghur person-name recognition (Nizamidin et al. 2016) 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.

3. Annotation Overview

The idea of the ACE (Doddington et al. 2004), ERE (Aguilar et al. 2014) and Uyghur knowledge base (Abaidulla et al. 2009) 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" (Pustejovsky and Stubbs 2012). At the same time, new tags will need to be introduced to address the properties of Uyghur.

3.1. UyNe Annotation

3.1.1. 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 proper name not pronominal or nominal. For example: named entity mention $_{\text{PER}}$[تەن گوجۇن] (Tan Guojun) 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 (Abaidulla et al. 2009) 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 morpholog-

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

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

ical analysis (Wumaier et al. 2009) (Aili et al. 2012) which is used in real applications, as mentioned above. Achievements in Uyghur morphological analysis make it possible to do automatic morphological analysis after manual entity and relation annotation. So we chose surface form as the basic entity annotation unit.

3.1.2. UyNe Annotation Rules

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

غەلبە قىلدى.	جېڭىدە	نۆۋەتتا توپ	ORG[برازىلىيە]
won	race	penalty	ORG[Brazil]

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

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

ئەسلىمىسى	PER[كاترىنا كەيفنىڭ]
memoirs	PER[Katrina Kaif's]

مەدەنىيەت	ئۆزگىچە	GPE[قەشقەردىكى]
culture	special	GPE[at Kashgar]

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:

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

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

3.2. UyNeRel Annotation

3.2.1. 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*(Person-Social) *Org-Aff* (Organization-Affiliation). Relation subtypes differ from Rich ERE which have 20 subtypes but in our annotation scheme there are 15 subtypes. The difference is shown in Table 2.

In Rich ERE annotation specification, relation type *Physical* is 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* is 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 in Table 3.

Table 2: Differences between Rich ERE andUyNeRel

Rich ERE 5 types, 20 subtypes		**UyNeRel 5 types, 15 subtypes**	
Types	**Subtypes**	**Types**	**Subtypes**
Physical	OrgHeadQuarter 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

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 in Table 4.

The *Per-Social* (Person-Social) is divied into 4 subtypes. A *Per-Social* relation that is not the subtype of *Business*, *Family* or *Role*, belongs to *Other*. The example is shown in Table 5.

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 eas-

Table 3: Examples of *Physical* Relations

سەمەت قاناستا قاپسلىپ قالدى. (Semet was trapped in Kanas.)		
Type.Subtype	Argument1	Argument2
Physical. Located	سەمەت (Semet)	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 4: 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)

ier, as this is a simple distinction between subtypes in *Org-Aff* . The example is shown in Table 6.

3.3. UyNeRel Annotation Rules

We only annotate relations between two entities within one sentence. This can be seen from table 3 to table 6. 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 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.Since it is not informative about the true relationships

Table 5: Examples of *Per-Social* Relations

ئەلى سادىقنىڭ ئادۋوكاتى ئەنۋەر (Eli Sadiq's lawyer is Enwer)		
Type.Subtype	Argument1	Argument2
Per-Social.Business	ئەلى سادىقنىڭ (Eli Sadiq's)	ئەنۋەر (Enwer)
كېرەم ئابلىزنىڭ ئايالى يۇلتۇز (Kërim Abliz's wife Yultuz)		
Type.Subtype	Argument1	Argument2
Per-Social.Family	كېرەم ئابلىزنىڭ (Kërim Abliz's)	يۇلتۇز (Yultuz)
لى شىڭ پروفىسسور (Professor)		
Type.Subtype	Argument1	Argument2
Per-Social.Role	لى شىڭ (Li Xing)	پروفىسسور (Professor)
ئەمەتنىڭ يۇرتدىشى سەمىجان (Emet's fellow-townsman is Semijan)		
Type.Subtype	Argument1	Argument2
Per-Social.Other	ئەمەتنىڭ (Emet's)	سەمىجان (Semijan)

between entities. For example:

رەنا هازىر بېيجىڭدا ئەمەس.
Rena at now Beijing .not

(Rena is not in Beijing at the moment.)

4. The Annotation Process

4.1. Raw Corpus

Once a 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,

[8] http://uy.ts.cn
[9] http://uyghur.people.com.cn
[10] http://uyghur.news.cn

Table 6: Examples of *Org-Aff* Relations

هىندىستان زۇكتۇڭى مۇكارجى (Indian President Mukherje)		
Type.Subtype	Argument1	Argument2
Org-Aff. Employment	مۇكارجى (Mukherje)	هىندىستان (Indian)
شەندۇڭ ‹‹رۇيى›› گۇرۇھى قەشقەر ۋىلايىتىگە20 ملىيارد يۈەن مەبلەغ 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)

source, and other related info. This corpus construction is still in progress.

4.2. Tool for the Annoation Process

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 (Stubbs 2011). 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 for annotating NAM (NAmed-entity mention, we still set other two values NOM (NOMinal-entity mention) and PRO (PROnominial-entity mention.

[11] https://github.com/keighrim/mae-annotation

We 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 the Arabic script written from right to left, the tool is slightly modified to display Uyghur letter. However, after the preliminary annotation, it is found that this software does not handle well the Arabic script and suffer from slow speed. So we have developed our own annotation tool using C#. The interface of this process shown in Figure2. In order to alleviate work amount of annotators we also add some functions to our annotation software. These functions include automatic suggestion about entities and relations that previously annotated. We also use different colors for different types of entities.

Figure 2: Interface of the UyNeRel Annotation Software

4.3. Results

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 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 0.7, and UyNeRel 0.6; We assume that this should remain stable as the size of the corpus increases. So the annotation is

conducted by these two annotators using our annotation software. We have already finnished 571 documents. The raw statistics about corpus is as shown Table7 .

Table 7: Statistics about UyNeRel Corpus

Documents	Sentences	Words	Tokens
571	6173	27846	2384397

The corpus is released under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.It can be accessed from `https://github.com/kaharjan/UyNeRel`. For the annotation format, we have applied character-based annotations (Pustejovsky and Stubbs 2012). This format uses the character offsets information to place tags in docments shown Figure3 by *spans*. Although it is a little hard to clearly see the actual annotation text by tags in xml file.The important thing is that in this annoation format, annotation tags and original text are separated, it would be possible to have nested annotations within the same context without interfering with one another. It also will not change the original text of documents.

Figure 3: Sample of annotated xml file

5. Conclusions and Future Work

In this work, we intend to construct Uyghur named entity and named entity relation corpus.For this we first investigate theoretical foundations. Then we mine raw text from Internet sources. Finally trained annotators tag the resulting collected data according to our annotation guidelines. In the process of defining Uyghur named entity and Uyghur named entity relation annotation process, we have utilized existing research on other languages, along with existing corpus annotation experience for Uyghur. 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 is needed, and so it has developed. We have released our annotated corpus, which includes nearly 600 documents. The corpus is can be accessed from `https://github.com/kaharjan/UyNeRel`. We think releasing data not only promotes Uyghur information extraction research in the NLP community but also improves the

quality of our data by feedback. In the future, we will enhance functionality of our annotation tool to accelerate efficiency of the annotation process, and will expand the corpus size and improve its quality.

6. Acknowledgments
This work has been supported as part of the NSFC (61462083, 61331011 and 61262060), 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

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

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

Aguilar, J., Beller, C., McNamee, P. and Van Durme, B., 2014, 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*, pp. 45–53.

Aibaidula, Y. and Lua, K. T., 2003, The development of tagged Uyghur corpus, in Ji, D. H. L. K. T., editor, *17th Pacific Asia Conference on Language, Information and Computation*, pp. 228–234.

Aili, M., Jiang, W.-B., Wang, Z.-Y., Yibulayin, T. and Liu, Q., 2012, Directed graph model of Uyghur morphological analysis, *Journal of Software*, vol. 23, no. 12, pp. 3115–3129.

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

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. and Weischedel, R. M., 2004, The automatic content extraction (ace) program-tasks, data, and evaluation, in *LREC*.

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

Kuerban, A., Kuerban, W. and Abdurusul, N., 2009, Research on Uyghur framenet description system, in *2009 Oriental COCOSDA International Conference on Speech Database and Assessments,*, pp. 160–163.

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

Mott, J., Song, Z., Bies, A. and Strassel, S., 2016, Parallel Chinese - English Entities, Relations and Events Corpora, in *LREC 2016*, pp. 3717–3722.

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

Pustejovsky, J. and Stubbs, A., 2012, *Natural Language Annotation for Machine Learning*,

O'Reilly Media, Inc.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N. and Ma, X., 2015, From light to rich ere: Annotation of entities, relations, and events, in *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 89–98.

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

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

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

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

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

Wushouer, J., Abulizi, W., Abiderexiti, K., Yibulayin, T., Aili, M. and Maimaitimin, S., 2016, Building Contemporary Uyghur Grammatical Information Dictionary, in Murakami, Y. and Lin, D., editors, *Worldwide Language Service Infrastructure*, pp. 137–144, Springer International Publishing, Cham, doi:10.1007/978-3-319-31468-6_10.

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