Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus

Kahaerjiang Abiderexiti, Maihemuti Maimaiti,
Tuergen Yibulayin, Aishan Wumaier

School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, 830046, China

Xinjiang Laboratory of Multi-Language information Technology, Urumqi, Xinjiang, 830046, China

OUTLINE

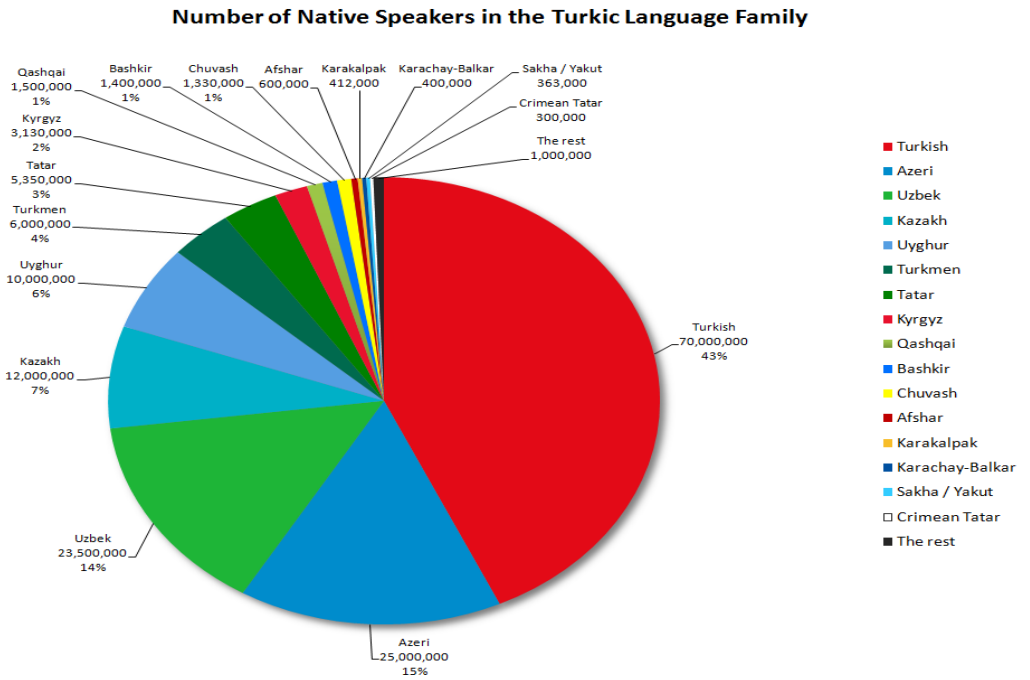
1. Introduction
2. UyNe Annotation Schemes
3. UyNeRel Annotation Schemes
4. Corpus Sampling and Annotation
5. Conclusions

1.Introduction

Uyghur language :

➢is one of the official language as well as Mandarin (Chinese).

➢fifths big language among Turkic languages



1.Introduction

Uyghur language :

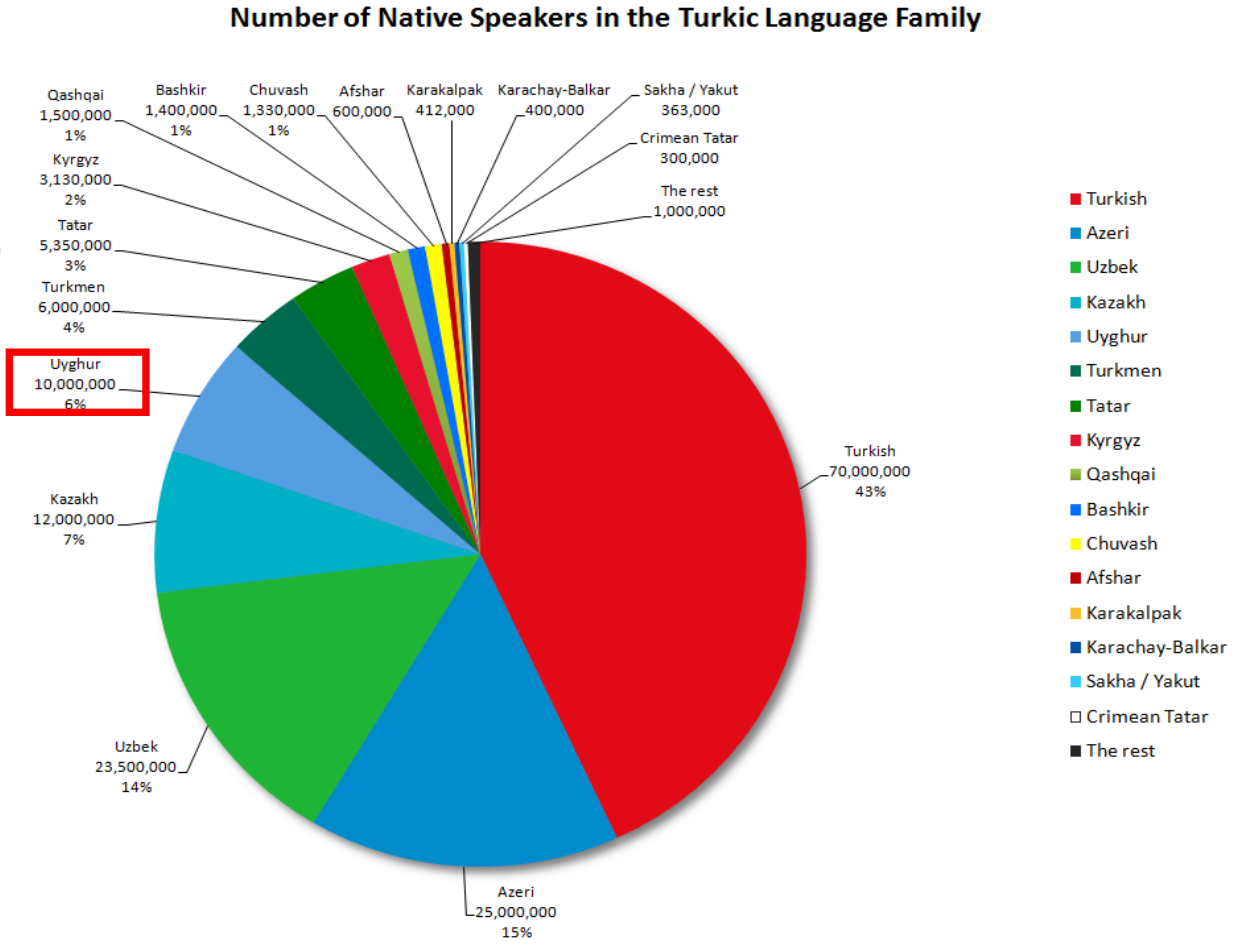
- is one of the official language as well as Mandarin (Chinese).
- fifths big language among Turkic languages

1.Introduction

Uyghur language :

- is one of the official language as well as Mandarin (Chinese).

- fifths big language among Turkic languages



http://en.wikipedia.org/wiki/Turkic_languages

1.Introduction

- **Entity**

  An object or set of objects in the world.

- Entity Type:

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation:

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type:**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation:

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

LOC

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

LOC

GPE

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang

LOC

GPE

1.Introduction

- **Entity**

  An object or set of objects in the world.

- **Entity Type**

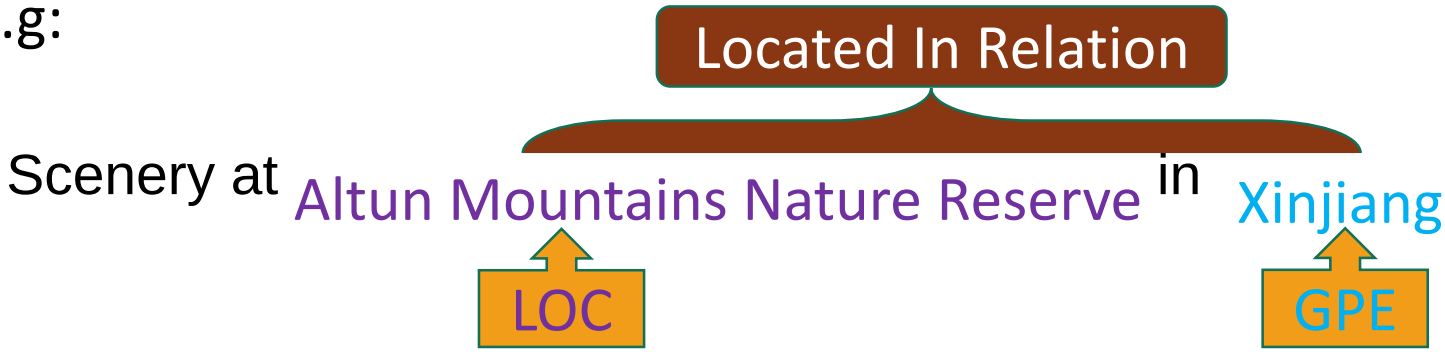
  Person name(PER), Organization name(ORG), geo-graphical/geopolitical entity name(GPE),location(LOC) etc.

  E.g. Trump, United Nation ,China, Tian Shan mountain.

- **Entity Relation**

  relations of the targeted types between entities.

  E.g:



1.Introduction

To the best of our knowledge, currently

No Uyghur Named-Entity Relation Corpora

2.Uyghur Named Entity Annotation (UyNe) Scheme

1) Types of Entities

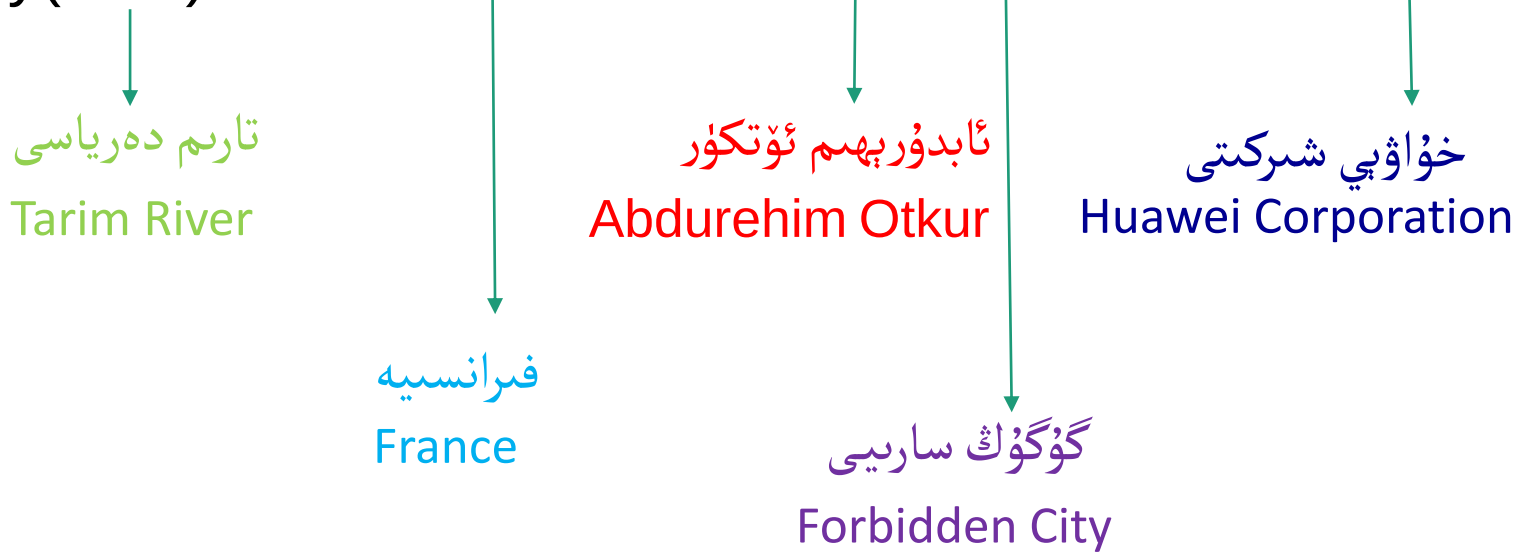
We annotate entities of Person(PER), Organization(ORG), GeoPolitical Entities(GPE), Location(LOC) and Facility(FAC).

2.Uyghur Named Entity Annotation (UyNe) Scheme

1) Types of Entities

We annotate entities of Person(PER), Organization(ORG), GeoPolitical Entities(GPE), Location(LOC) and Facility(FAC).

Ex:



2.Uyghur Named Entity Annotation (UyNe) Scheme

2) Basic Annotation Unit

Surface form as the basic entity annotation unit.

2.Uyghur Named Entity Annotation (UyNe) Scheme

2) Basic Annotation Unit

Surface form as the basic entity annotation unit.

Ex:

2.Uyghur Named Entity Annotation (UyNe) Scheme

3) UyNe Annotation Rules

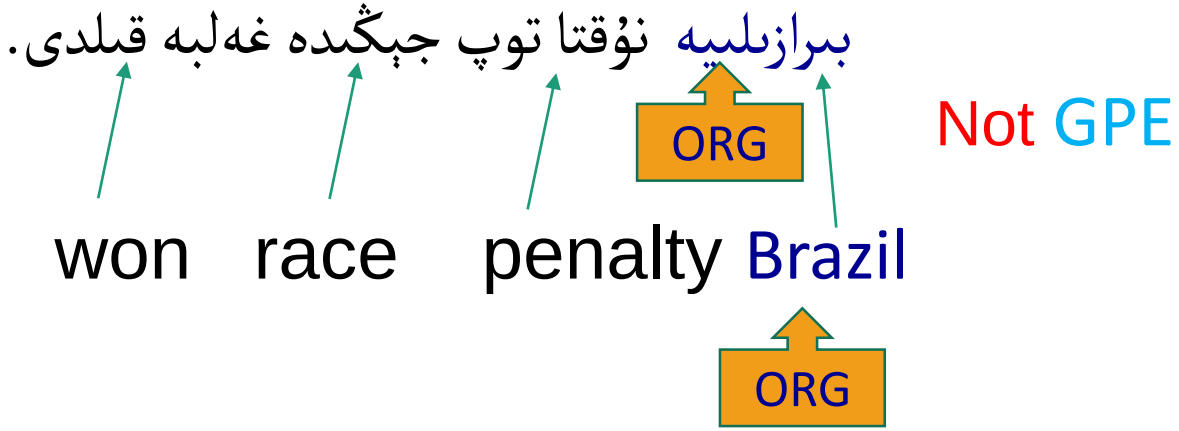
Such as entities are annotated according to usage.

2.Uyghur Named Entity Annotation (UyNe) Scheme

3) UyNe Annotation Rules

Such as entities are annotated according to usage.

Ex:



3.Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

1) Types of UyNeRel

In UyNeRel there are 5 types of relations.These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation). There is 15 subtypes.

3.Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

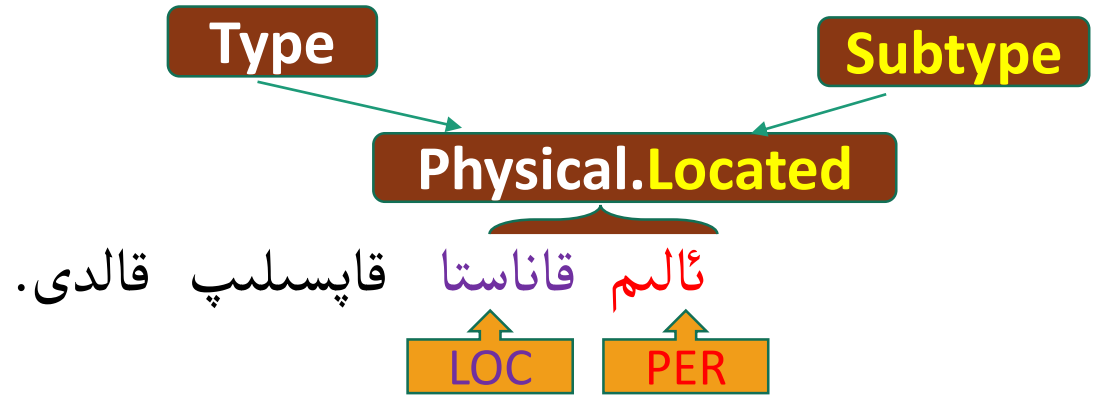
1) Types of UyNeRel

In UyNeRel there are **5 types** of relations.These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation).
There are **15 subtypes** .

3.Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

1) Types of UyNeRel

In UyNeRel there are **5 types** of relations.These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation).
There are **15 subtypes** . Ex:



(Alim was trapped in Kanas.)

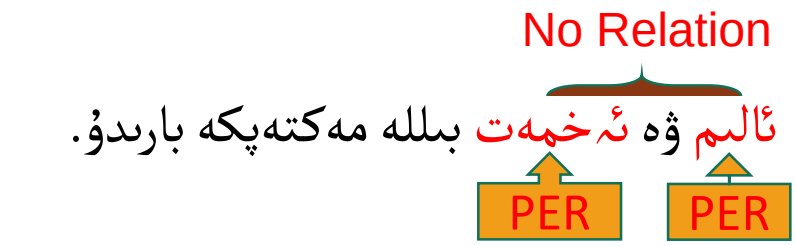
3.Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

2) UyNeRel Annotation Rules

Annotate relations according to its usage.

if the relation existed in the real world, but not in a single sentence, it will not be annotated.

(Alim and his brother Ahmet go to school gether.)



(Alim and Ahmet go to school gether.)

4.Corpus Sampling and Annotation

Crawl raw corpus from websites

- http://uy.ts.cn
- http://uyghur.people.com.cn
- http://uyghur.news.cn

4.Corpus Sampling and Annotation

Annotation Tool

4.Corpus Sampling and Annotation

Annotated XML text Format:

```
ا مەسلىسى بويىچە ئۇچ تەرەپ بېلورۇسىينىڭ پايتەختى مىنسكىدا مۇنداق دېدى: شۇ ئۇكرائىنانىڭ سابىق زوڭتۇڭى كوچىما 31- يانۋار
سكى سۆھبىتىنىڭ مەغلۇب مۇزاكىرە قىلىپ، ئۇكرائىنا مەسلىسى بويىچە ئۇچ تەرەپ هوللاندى، مېركىل ۋە پروشىنكو ئۇكرائىنا ۋەزىيىتىنى
]]></TEXT>
 <TAGS>
  <GPE id="GPE0" spans="723~730" text="مىنسكى" mentionLevel="NAM" comment="" />
  <GPE id="GPE1" spans="722~731" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
  <GPE id="GPE2" spans="665~674" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
  <GPE id="GPE3" spans="628~637" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
  <PER id="PER0" spans="619~627" text="پروشىنكو" mentionLevel="NAM" comment="" />
  <PER id="PER1" spans="608~615" text="مېركىل" mentionLevel="NAM" comment="" />
  <PER id="PER2" spans="598~606" text="هوللاندى" mentionLevel="NAM" comment="" />
  <GPE id="GPE4" spans="559~568" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
  <ORG id="ORG0" spans="551~578" text="ئەل ئىچى قۇراللىق تەشكىلاتى" mentionLevel="NAM" commer
  <LOC id="LOC0" spans="528~550" text="ئۇكرائىنانىڭ شەرقىدىكى" mentionLevel="NAM" comment="
  <GPE id="GPE5" spans="528~540" text="ئۇكرائىنانىڭ" mentionLevel="NAM" comment="" />
  <GPE id="GPE6" spans="465~474" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
  <GPE id="GPE7" spans="422~431" text="مىنسكىدا" mentionLevel="NAM" comment="" />
  <GPE id="GPE8" spans="399~412" text="بېلورۇسىينىڭ" mentionLevel="NAM" comment="" />
  <PER id="PER3" spans="381~387" text="كوچىما" mentionLevel="NAM" comment="" />
  <TTL id="TTL0" spans="373~380" text="زوڭتۇڭى" comment="" />
  <TTL id="TTL1" spans="367~380" text="سابىق زوڭتۇڭى" comment="" />
  <GPE id="GPE9" spans="354~366" text="ئۇكرائىنانىڭ" mentionLevel="NAM" comment="" />
```

4.Corpus Sampling and Annotation

Cohen's K inter-annotator agreement (IAA):
Preliminary result:

κ_{UyNe}=0.7

$\kappa_{UyNeRel}$=0.6

Now it is higher that 0.8.
500 documents are annotated.

The data available :
https://github.com/kaharjan/UyNeRel

6. Conclusion

- Investigate theoretical foundations, sample raw text from internet sources for Uyghur language
- Developed Uyghur UyNeRel annotation schemes and tool
- The annotated corpus is publicly available

Thank You !