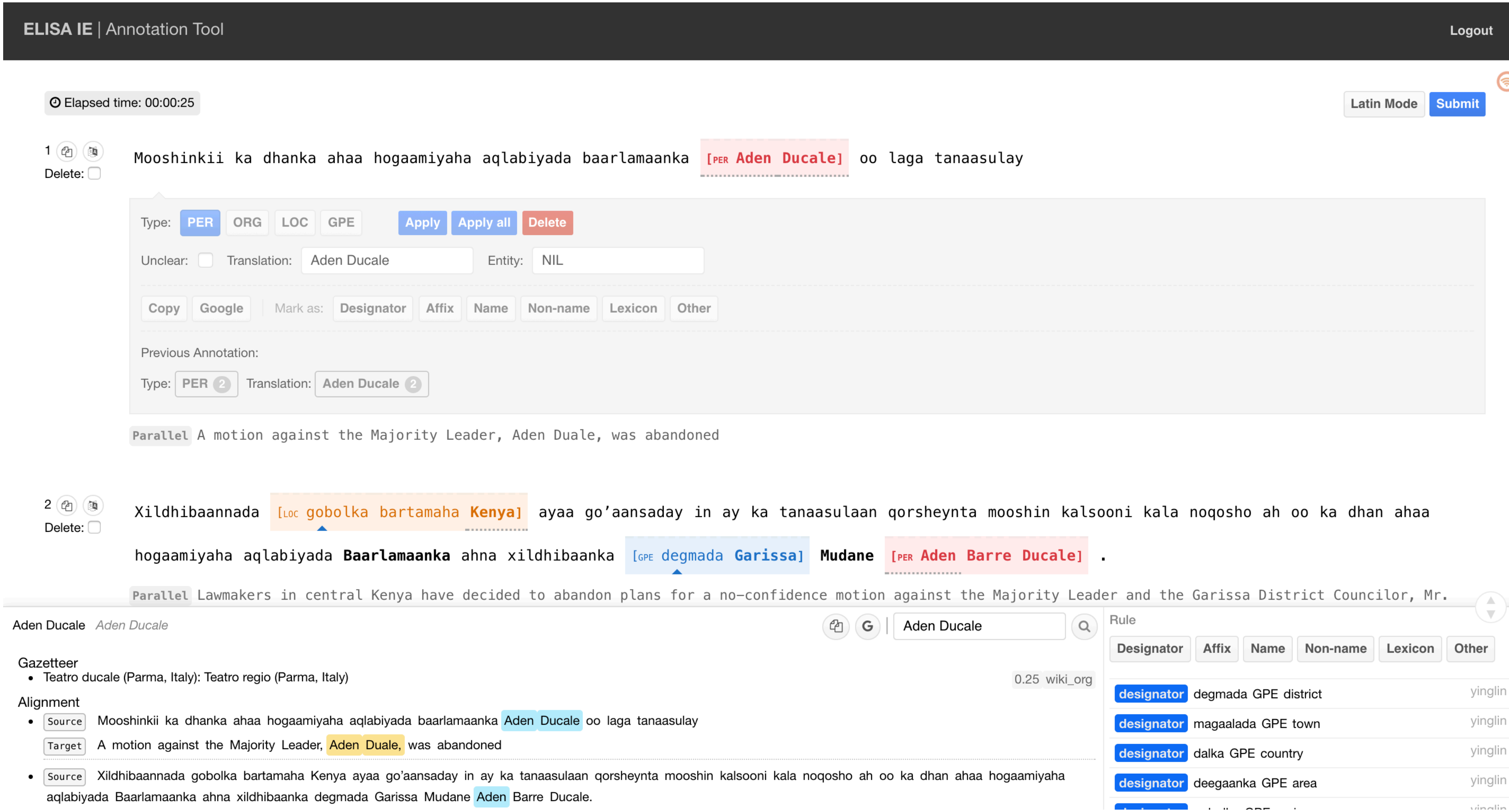
Exploring Multi-level Distributional Semantics for Cross-lingual Entity Discovery and Linking

Boliang Zhang, Xiaoman Pan, Lifu Huang, Ying Lin, Heng Ji

jih@rpi.edu

Rensselaer

Noisy Training Data Acquisition 1: Chinese Room



Elapsed time: 00:00:25

1 Delete: □

Mooshinkii ka dhanka ahaa hogaamiyaha aqlabiyada baarlamaanka [PER Aden Ducale] oo laga tanaasulay

Type: PER ORG LOC GPE

Unclear: □ Translation: Aden Ducale Entity: NIL

Previous Annotation:

Type: PER 2 Translation: Aden Ducale 2

Parallel A motion against the Majority Leader, Aden Duale, was abandoned

2 Delete: □

Xildhibaannada [LOC gobolka bartamaha **Kenya**] ayaa go’aansaday in ay ka tanaasulaan qorsheynta mooshin kalsooni kala noqosho ah oo ka dhan ahaa hogaamiyaha aqlabiyada **Baarlamaanka** ahna xildhibaanka [GPE degmada **Garissa**] **Mudane** [PER Aden **Barre Ducale**] .

Parallel Lawmakers in central Kenya have decided to abandon plans for a no−confidence motion against the Majority Leader and the Garissa District Councilor, Mr.

Aden Ducale *Aden Ducale*

Aden Ducale

Gazetteer

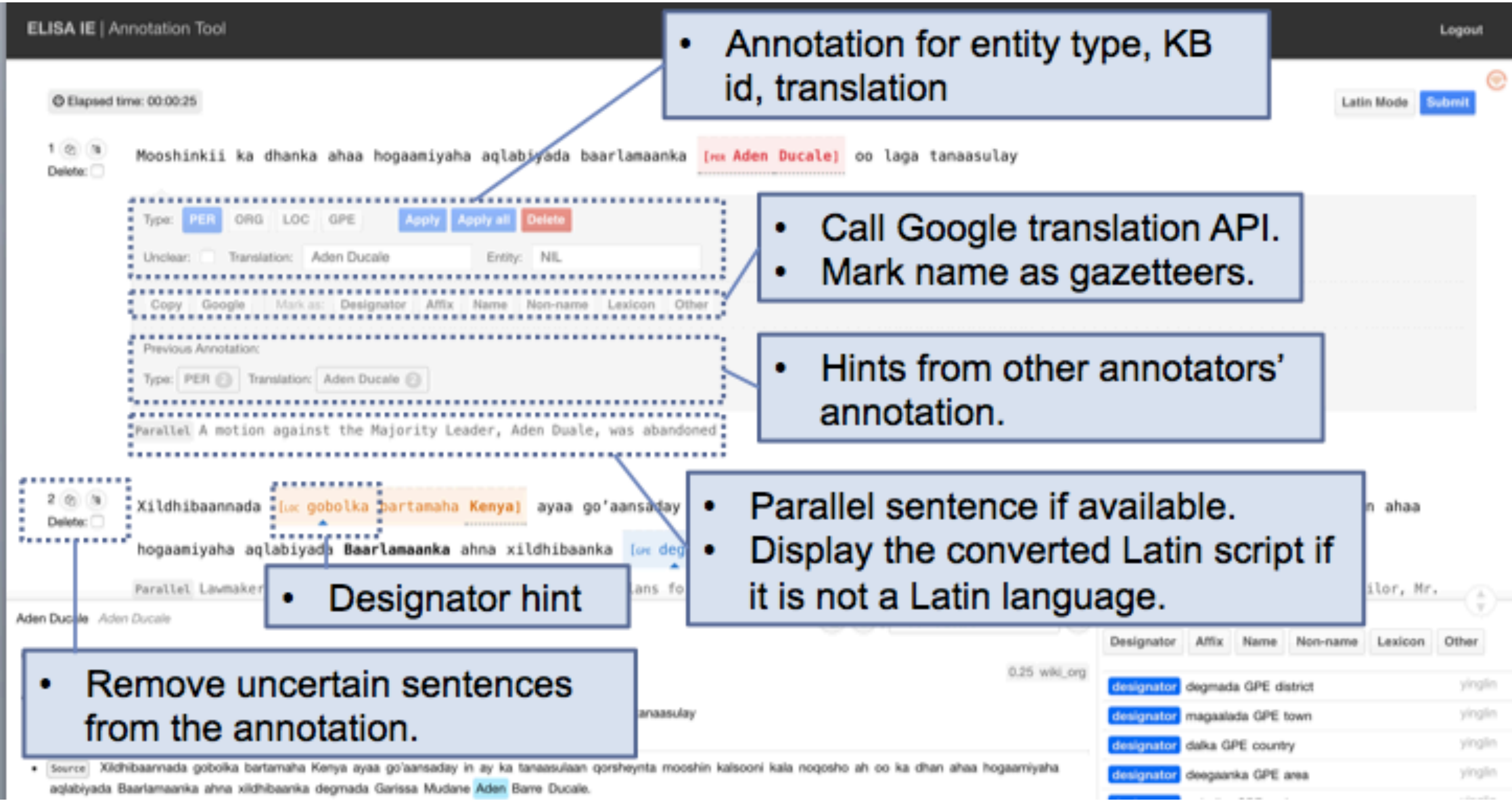
- Teatro ducale (Parma, Italy): Teatro regio (Parma, Italy) 0.25 wiki_org

Alignment

- Source Mooshinkii ka dhanka ahaa hogaamiyaha aqlabiyada baarlamaanka Aden Ducale oo laga tanaasulay
  Target A motion against the Majority Leader, Aden Duale, was abandoned
- Source Xildhibaannada gobolka bartamaha Kenya ayaa go’aansaday in ay ka tanaasulaan qorsheynta mooshin kalsooni kala noqosho ah oo ka dhan ahaa hogaamiyaha aqlabiyada Baarlamaanka ahna xildhibaanka degmada Garissa Mudane Aden Barre Ducale.

designator	degmada GPE district	yinglin
designator	magaalada GPE town	yinglin
designator	dalka GPE country	yinglin
designator	deegaanka GPE area	yinglin

Noisy Training Data Acquisition 1: Chinese Room



ELISA IE | Annotation Tool

Logout

Elapsed time: 00:00:25

Latin Mode Submit

1 Delete: □

Mooshinkii ka dhanka ahaa hogaamiyaha aqlabiyada baarlamaanka [PER Aden Ducale] oo laga tanaasulay

Type: PER ORG LOC GPE Apply Apply all Delete

Unclear: □ Translation: Aden Ducale Entity: NIL

Copy Google Mark as: Designator Affix Name Non-name Lexicon Other

Previous Annotation:

Type: PER Translation: Aden Ducale

Parallel A motion against the Majority Leader, Aden Duale, was abandoned

2 Delete: □

Xildhibaannada [LOC gobolka bartamaha Kenya] ayaa go'aansaday … n ahaa hogaamiyaha aqlabiyada Baarlamaanka ahna xildhibaanka [LOC deg…

Parallel Lawmaker… ans fo… ilor, Mr.

Aden Ducale *Aden Ducale*

0.25 wiki_org

…anaasulay

Designator Affix Name Non-name Lexicon Other

designator	degmada GPE district	yinglin
designator	magaalada GPE town	yinglin
designator	dalka GPE country	yinglin
designator	deegaanka GPE area	yinglin

- Source Xildhibaannada gobolka bartamaha Kenya ayaa go'aansaday in ay ka tanaasulaan qorsheynta mooshin kalsooni kala noqosho ah oo ka dhan ahaa hogaamiyaha aqlabiyada Baarlamaanka ahna xildhibaanka degmada Garissa Mudane Aden Barre Ducale.

- Annotation for entity type, KB id, translation

- Call Google translation API.
- Mark name as gazetteers.

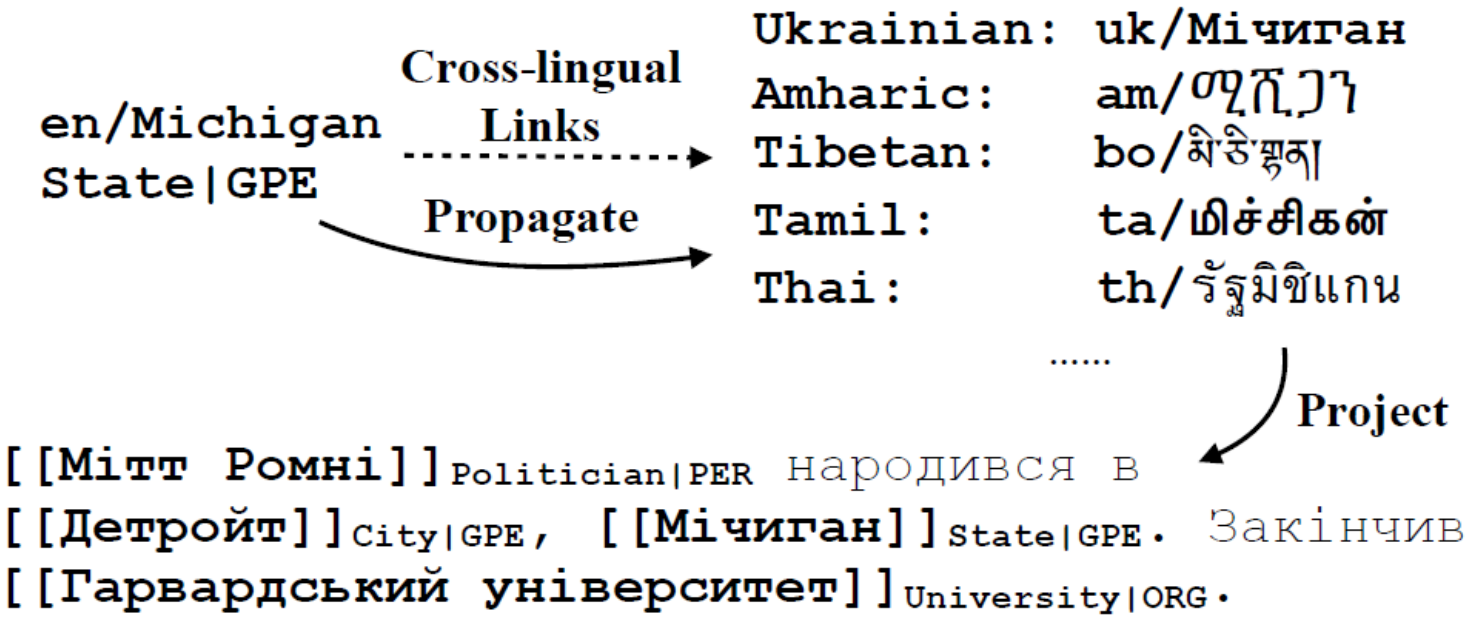
- Hints from other annotators' annotation.

- Parallel sentence if available.
- Display the converted Latin script if it is not a Latin language.

- Designator hint

- Remove uncertain sentences from the annotation.

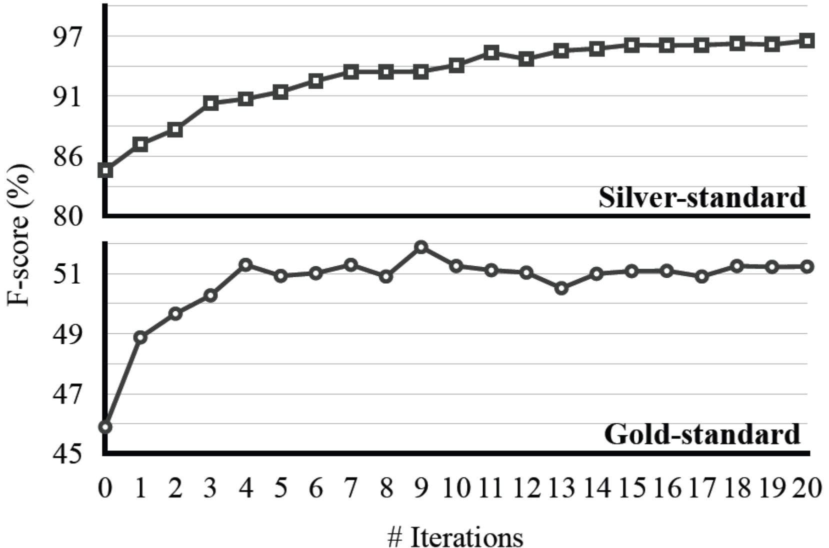
Noisy Training Data Acquisition 2: Wikipedia Mining



[[Мітт Ромні]]Politician|PER народився в [[Детройт]]City|GPE, [[Мічиган]]State|GPE. Закінчив [[Гарвардський університет]]University|ORG.

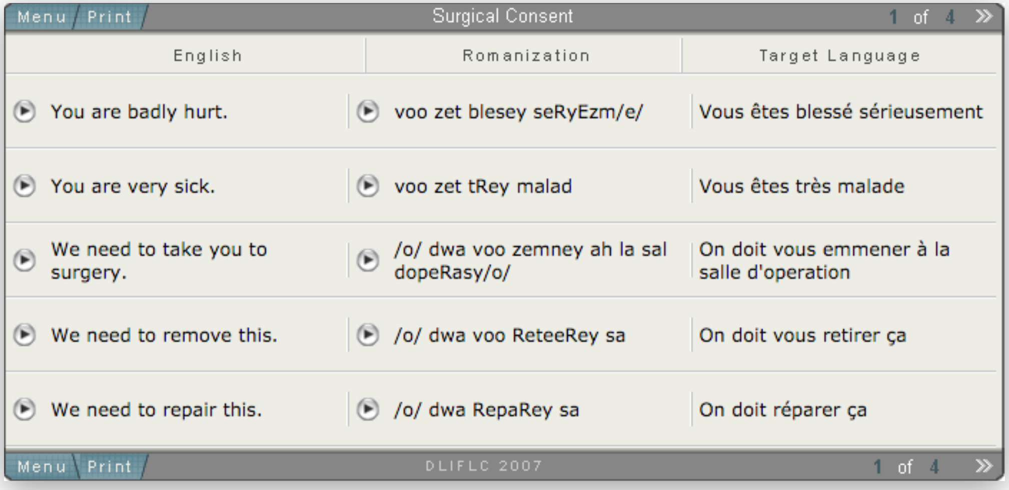
(**Mitt Romney** was born in **Detroit**, **Michigan**. He graduated from **Harvard University**.)

- Generate "silver-standard" training data automatically
- Apply self-training to make training data for complete and consistent



Exploit Non-traditional Universal Linguistic Resources

- Grammar books from Lori Levin's bookshelf and CIA Names from DARPA PM's bookshelf
- Unicode Common Locale Data Repository, Wikitionary, Panlex, Multilingual WordNet, GeoNames, JRC Names, phrase pairs mined from Wikipedia
- Phrase Books from Language Survival Kits and Elicitation Corpus
- Ignored by NLP community

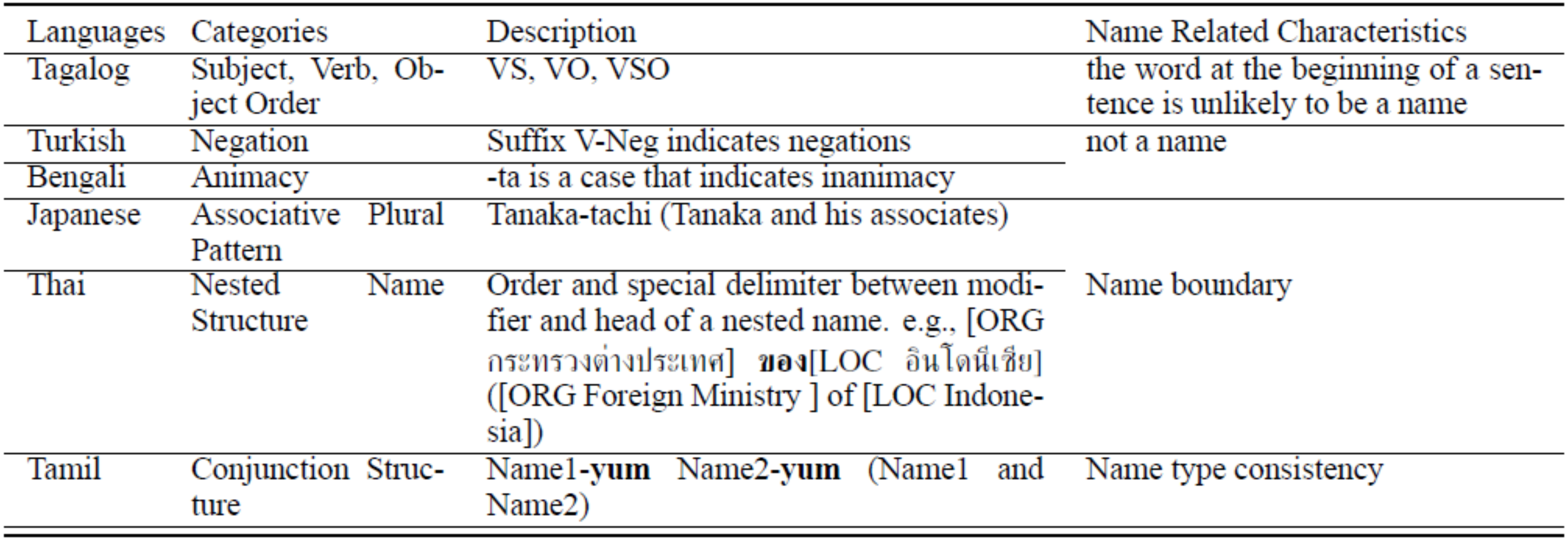


Surgical Consent

English	Romanization	Target Language
You are badly hurt.	voo zet blesey seRyEzm/e/	Vous êtes blessé sérieusement
You are very sick.	voo zet tRey malad	Vous êtes très malade
We need to take you to surgery.	/o/ dwa voo zemney ah la sal dopeRasy/o/	On doit vous emmener à la salle d'operation
We need to remove this.	/o/ dwa voo ReteeRey sa	On doit vous retirer ça
We need to repair this.	/o/ dwa RepaRey sa	On doit réparer ça

DALRYMPLE AND NIKOLAEVA Objects and Information Structure CAMBRIDGE
Comrie Tense Cambridge
Comrie Aspect CAMBRIDGE
TSL 99 Kittilä et al. Case, Animacy and Semantic Roles
HASPELMATH INDEFINITE PRONOUNS OXFORD
Guéron · Lecarme Eds. Time and Modality
De Hoop · De Swart Eds. Differential Subject Marking
Malchukov & Spencer The Oxford Handbook of CASE OXFORD
Lyons Definiteness CAMBRIDGE
OSTL 42 MASSAM Count and Mass Across Languages OXFORD
THE EXPRESSION OF TIME ECC 03
Comorovski von Heusinger Eds. Existence: Semantics and Syntax
THE EXPRESSION OF MODALITY ECC 01
THE EXPRESSION OF NEGATION ECC 04
THE EXPRESSION OF POSSESSION ECC 02

Linguistic Structure from WALS database and Syntactic Structures of the World's Languages

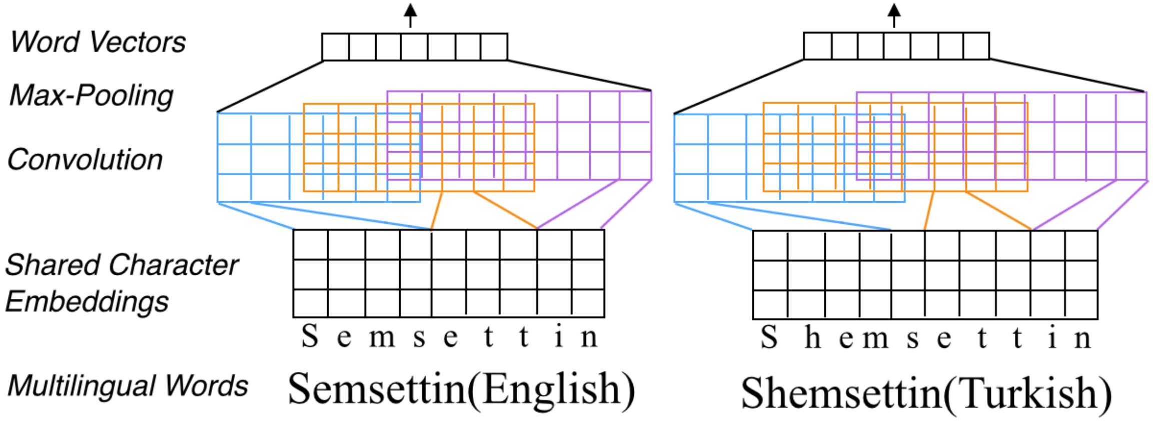


Languages	Categories	Description	Name Related Characteristics
Tagalog	Subject, Verb, Object Order	VS, VO, VSO	the word at the beginning of a sentence is unlikely to be a name
Turkish	Negation	Suffix V-Neg indicates negations	not a name
Bengali	Animacy	-ta is a case that indicates inanimacy	
Japanese	Associative Plural Pattern	Tanaka-tachi (Tanaka and his associates)	
Thai	Nested Name Structure	Order and special delimiter between modifier and head of a nested name. e.g., [ORG กระทรวงต่างประเทศ] **ของ**[LOC อินโดนีเซีย] ([ORG Foreign Ministry] of [LOC Indonesia])	Name boundary
Tamil	Conjunction Structure	Name1**-yum** Name2**-yum** (Name1 and Name2)	Name type consistency

- Universal Morphology Analyzer based on Wikipedia Markups
  - *Kıta Fransası, güneyde [[Akdeniz]]**den** kuzeyde [[Manş Denizi]]**ve** [[Kuzey Denizi]]**ne**, doğuda [[Ren Nehri]]**nden** batıda [[Atlas Okyanusu]]**na** kadar yayılan topraklarda yer alır. (Continental France is located in the south [[Mediterranean Sea]] in the north [[English Sea]] and [[North Sea]] in the east [[Rhine River]] to the west [[Atlantic Ocean]].)*

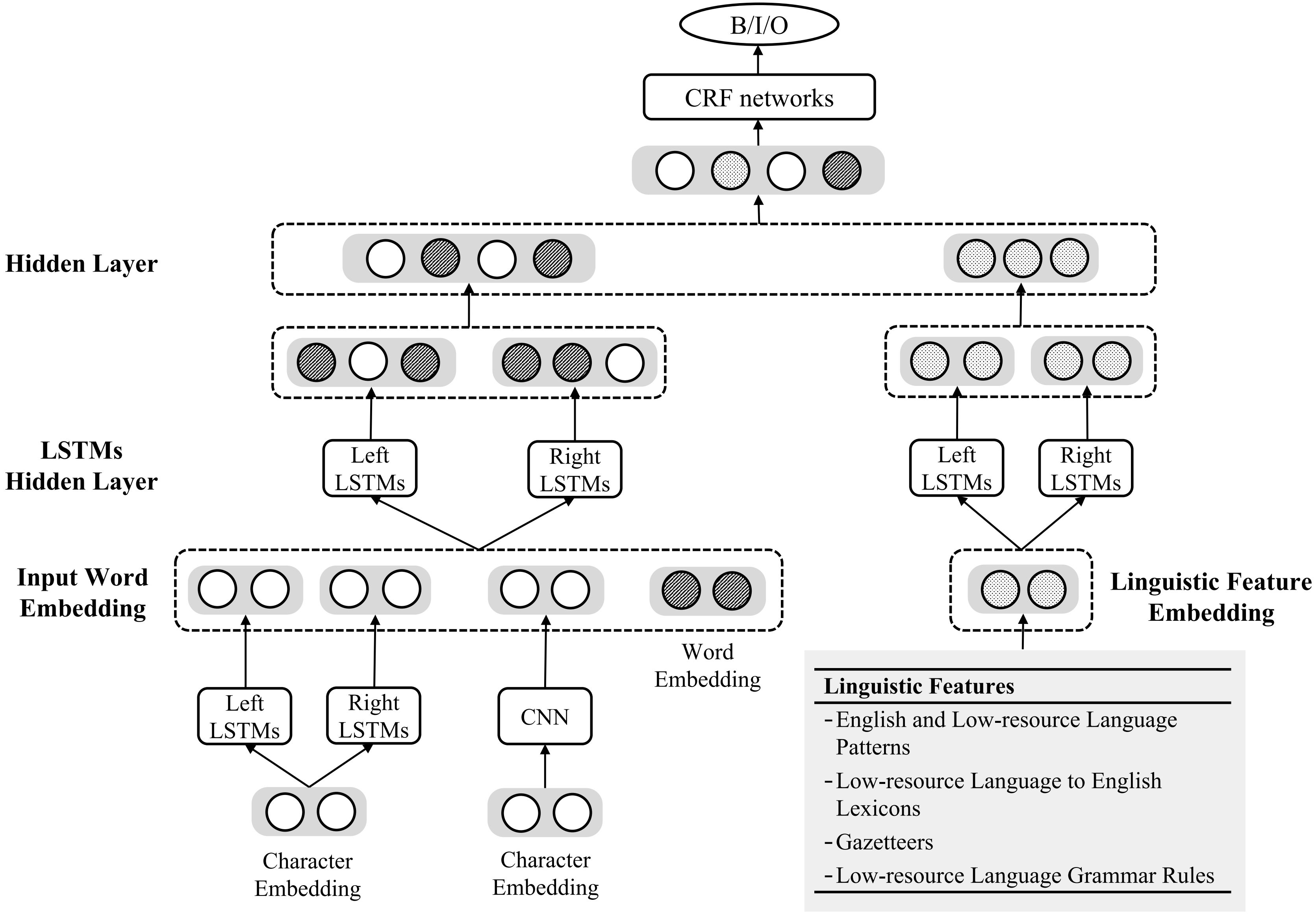
Character-Aware Word Embeddings

- Motivation: mentions of the same concept across languages may share a set of similar characters, e.g., Semsettin Gunaltay (English) = Şemsettin Günaltay (Turkish) = Semsetin Ganoltey (Somali)
- Compose word embeddings from shared character embeddings using Convolutional Neural networks



- Further optimized by language model based on Recurrent Neural Networks
  - maximize the prediction of the current word based on previous words

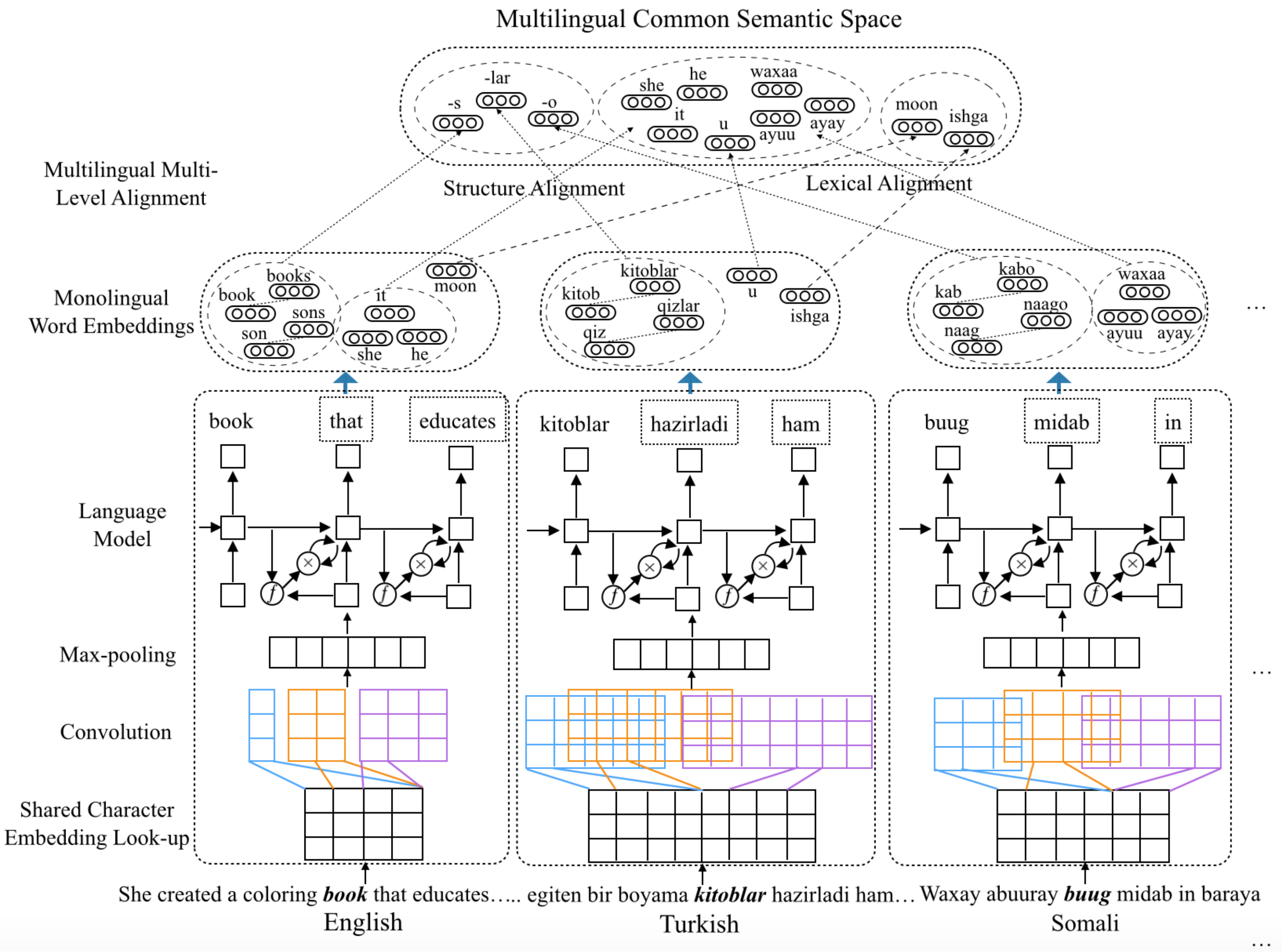
Feed Non-traditional Linguistic Resources into DNN



Linguistic Features

- English and Low-resource Language Patterns
- Low-resource Language to English Lexicons
- Gazetteers
- Low-resource Language Grammar Rules

Common Semantic Space Construction


Multilingual Common Semantic Space
-lar
-s
-o
she
he
waxaa
it
u
ayay
ayuu
moon
ishga
Multilingual Multi-Level Alignment
Structure Alignment
Lexical Alignment
Monolingual Word Embeddings
books
book
sons
son
it
she
he
moon
kitoblar
kitob
qizlar
qiz
u
ishga
kabo
kab
naago
naag
waxaa
ayuu
ayay
…
Language Model
book
that
educates
kitoblar
hazirladi
ham
buug
midab
in
Max-pooling
Convolution
Shared Character Embedding Look-up
…
She created a coloring ***book*** that educates….. egiten bir boyama ***kitoblar*** hazirladi ham… Waxay abuuray ***buug*** midab in baraya
English
Turkish
Somali
…

Construct a Common Semantic Space for Thousands of Languages

- Motivations
  - There are 3000+ languages with electronic record
  - NLP training data only available for several dominant languages
- Goals
  - Build a common semantic space across thousands of languages for resource sharing and richer semantic continuous representation for words, concepts and entities
- Limitations of Previous Attempts (e.g., Upadhyay et al., 2016, Cho et al., 2017)
  - Mostly English-anchored, cannot capture all linguistic phenomena
  - Heavily relied on bilingual dictionaries and parallel data which are not always available
  - Only limited to dozens of languages

Multi-Level Multi-lingual Alignment

- When bilingual word dictionaries are not available, back-off to shared linguistic structures
  - e.g., apposition, conjunction, plural suffix (English (-s / −es), Turkish (-lar / -ler), Somali (-o))
  - Generalized from language universal resources such as *WALS database* and *Syntactic Structures of the World's Languages*
  - Classify languages according to a large number of topological properties (phonological, lexical, grammatical)
  - 2,676 languages, 58,000+ (language, feature, feature value) tuples, e.g., (English, canonical word order, SVO)
- Project monolingual word embeddings into a common semantic space, and align both representations of words and linguistic-structures in the common space

Model Training

- Model training
  - Language model prediction loss

$$Loss_1 = -\sum_{l\in L}\sum_{t=1}^{|C_l|}\log Pr(w_t = W|w_{1:t})$$

$l \in L$: language l in the language set L
$C_l = [w_1, w_2, ..., w_T]$: the sequence of words in the corpus for language l
W : the real t^{th} word in C_t

  - Multilingual alignment loss:

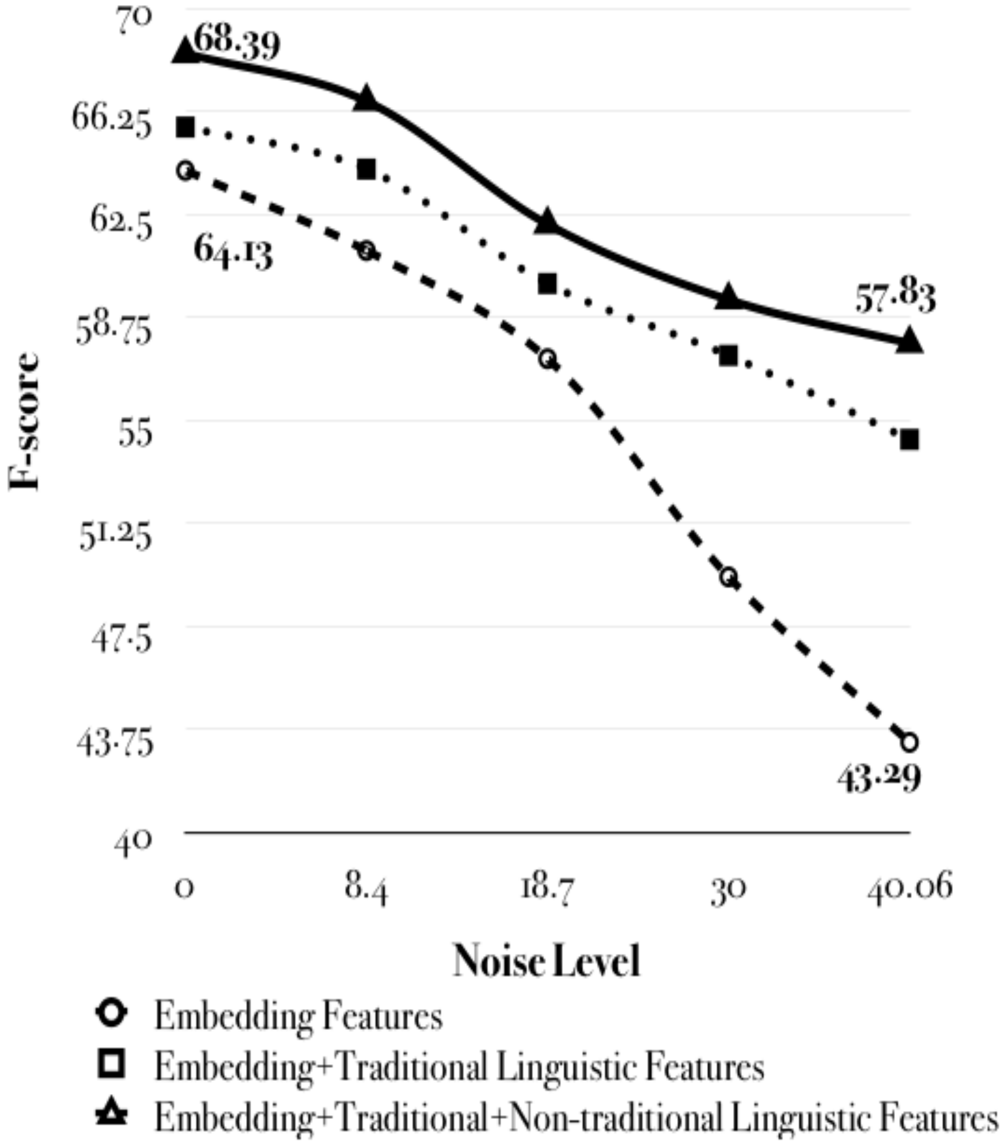
$$Loss_2 = \frac{1}{2}\sum_{i,j\in L}\sum_{(v_m^i,v_m^j)\in Align^{ij}} ||M_i\cdot v_m^i - M_j\cdot v_m^j||$$

$Align^{ij}$: the word/structure alignment for language i and j
(v_m^i, v_m^j) : the vector representation of the m^{th} alignment pair in $Align^{ij}$
M_i, M_j : the projection matrix for language i and j from original semantic space to the common semantic space

  - Overall loss:

$$Loss = Loss_1 + Loss_2$$

Linguistic Features Matter: More Robust to Noise



Uzbek (Zhang et al., 2017)

Impact of Character-Aware Word Embeddings

- Name Tagging F-Score (%)

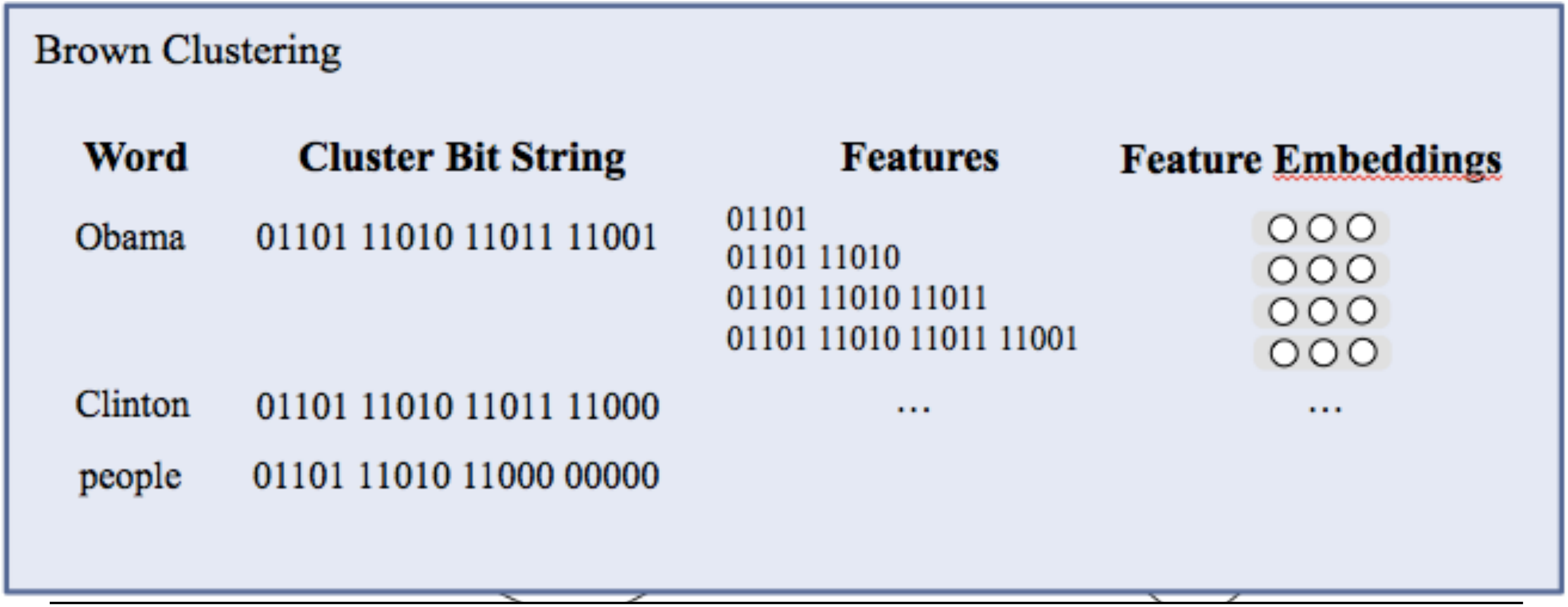
Models	Chinese	English	Spanish
Before	64.1	67.4	64.6
After	68.0	70.9	68.9

Impact of Common Semantic Space

- Chechen Name Tagging

Models	P (%)	R (%)	F (%)
Randomly initialized	46.3	45.31	45.8
Pre-trained	54.8	41.3	47.1
+ Common semantic space word embedding	62.1	50.1	55.4

Something Old: Hierarchical Brown Clustering



Brown Clustering

Word	Cluster Bit String	Features	Feature Embeddings
Obama	01101 11010 11011 11001	01101 01101 11010 01101 11010 11011 01101 11010 11011 11001	○○○ ○○○ ○○○ ○○○
Clinton	01101 11010 11011 11000	…	…
people	01101 11010 11000 00000		

Languages	w/o BC (%)	with BC (%)	Languages	w/o BC (%)	with BC (%)
Albanian	72.4	**74.6**	Northern Sotho	90.2	**90.8**
Chechen	53.1	**55.4**	Polish	49.6	**53.2**
Chinese	66.3	**68.0**	Somali	76.9	**78.5**
English	69.5	**70.9**	Spanish	67.1	**68.9**
Kannada	51.9	**56.0**	Swahili	64.3	**67.8**
Kikuyu	84.2	**88.7**	Yoruba	46.1	**49.5**
Nepali	41.6	**43.9**			

Joint Learning of Word and Entity Embeddings from Wikipedia

- Consider all Wikipedia anchor links as entity annotations, a training corpus can be created by replacing anchor links with unique entity IDs.

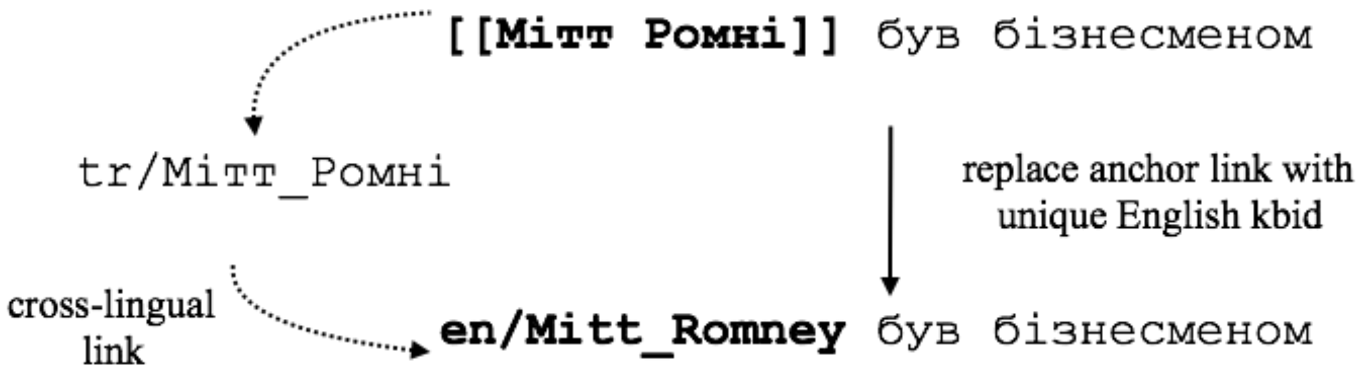
e.g., `[[en/Apple|apple]] is a fruit`
`[[en/Apple_Inc.|apple]] is a company`

apple is a fruit
apple is a company

→

en/Apple is a fruit
en/Apple_Inc. is a company

- Multi-lingual



[[Мітт Ромні]] був бізнесменом

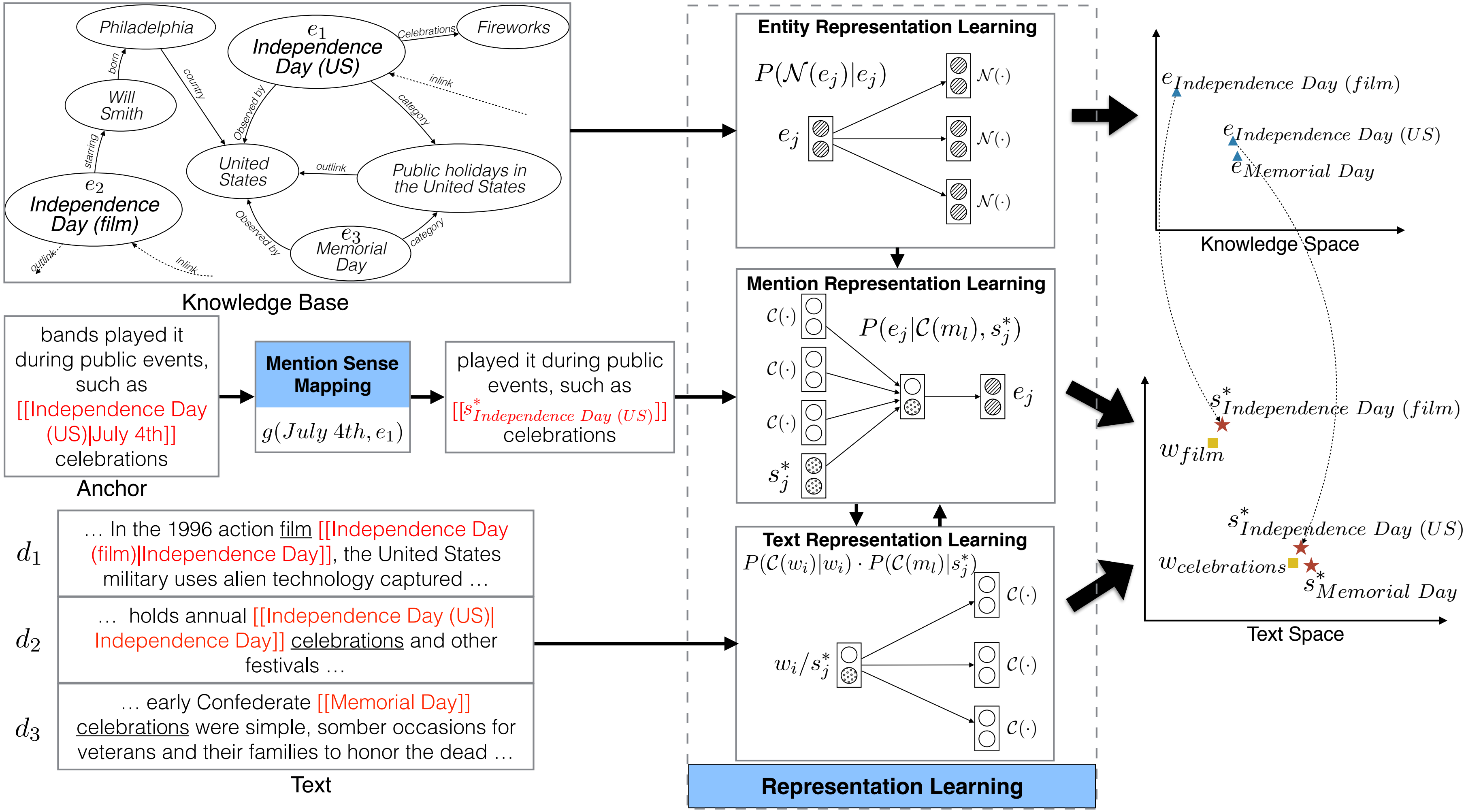
tr/Мітт_Ромні

replace anchor link with unique English kbid

cross-lingual link

en/Mitt_Romney був бізнесменом

Joint Learning of Word and Entity Embeddings from Wikipedia



bands played it during public events, such as [[Independence Day (US)|July 4th]] celebrations

played it during public events, such as [[$s^*_{Independence\ Day\ (US)}$]] celebrations

d_1 … In the 1996 action film [[Independence Day (film)|Independence Day]], the United States military uses alien technology captured …

d_2 … holds annual [[Independence Day (US)| Independence Day]] celebrations and other festivals …

d_3 … early Confederate [[Memorial Day]] celebrations were simple, somber occasions for veterans and their families to honor the dead …

Learning Entity Embeddings from DBpedia

- Construct a weighted undirected graph $G = (E, D)$ from DBpedia, where E is a set of all entities in DBpedia and $d_{ij} \in D$ indicates that two entities e_i and e_j share some DBpedia properties. The weight of d_{ij} , w_{ij} is computed as:

$$w_{ij} = \frac{|p_i \cap p_j|}{\max(|p_i|, |p_j|)}$$

where p_i, p_j are the sets of DBpedia properties of e_i and e_j respectively.

- Apply the graph embedding framework proposed by (Tang et al., 2015) to generate knowledge representations for all entities

Impact of Joint Embeddings on Entity Linking

- Unsupervised entity linking based on salience, similarity and coherence
- Tested on EDL16 perfect English NAM mentions

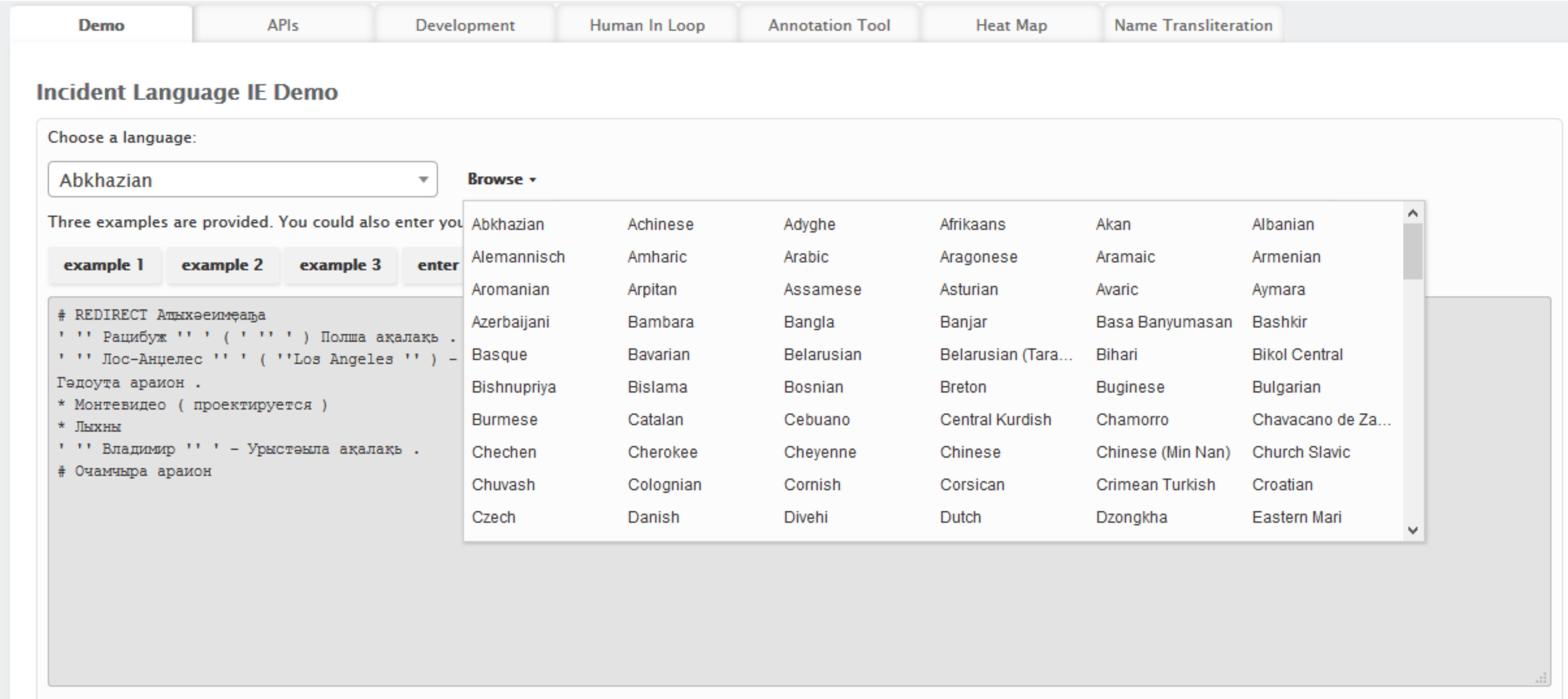
	CEAFm P	CEAFm R	CEAFm F1
Baseline	0.762	0.843	0.801
+ Joint word and entity embeddings from Wikipedia	0.791	0.875	0.831
+ Entity embedding from DBpedia	0.812	0.897	0.852

Resources and Demos

Systems, Data and Resources Publicly Available

- Re-trainable Systems:
  - http://blender02.cs.rpi.edu:3300/elisa_ie/api
  - Source code base available for government users upon requests
  - Tri-lingual EDL is being integrated into CoreNLP and hope to release in 2017
- Data and Resources:
  - http://nlp.cs.rpi.edu/wikiann/
- Demos:
- http://blender02.cs.rpi.edu:3300/elisa_ie
- http://blender02.cs.rpi.edu:3300/elisa_ie/heatmap

Demo 1: Cross-lingual Entity Discovery and Linking for 282 Languages


Demo
APIs
Development
Human In Loop
Annotation Tool
Heat Map
Name Transliteration
Browse ▾
example 1
example 2
example 3
enter
Abkhazian ▾
Abkhazian Achinese Adyghe Afrikaans Akan Albanian
Alemannisch Amharic Arabic Aragonese Aramaic Armenian
Aromanian Arpitan Assamese Asturian Avaric Aymara
Azerbaijani Bambara Bangla Banjar Basa Banyumasan Bashkir
Basque Bavarian Belarusian Belarusian (Tara... Bihari Bikol Central
Bishnupriya Bislama Bosnian Breton Buginese Bulgarian
Burmese Catalan Cebuano Central Kurdish Chamorro Chavacano de Za...
Chechen Cherokee Cheyenne Chinese Chinese (Min Nan) Church Slavic
Chuvash Colognian Cornish Corsican Crimean Turkish Croatian
Czech Danish Divehi Dutch Dzongkha Eastern Mari

Incident Language IE Demo

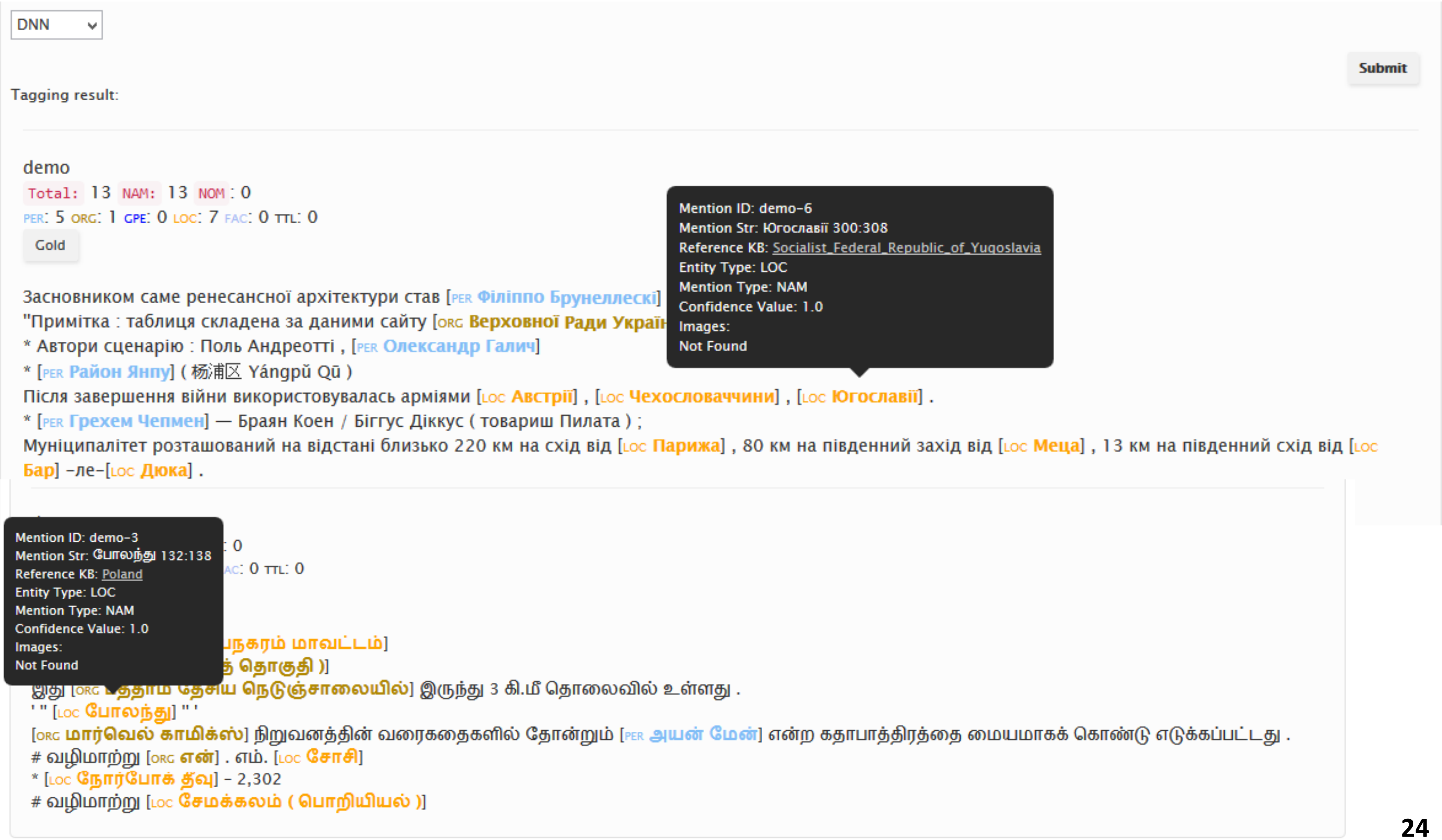
Choose a language:

Three examples are provided. You could also enter you

```
# REDIRECT Аҭыхәеимҿаҕа
' '' Рацибуж '' ' ( ' '' ' ) Полша ақалақь .
' '' Лос-Анцелес '' ' ( ''Los Angeles '' ) -
Гәдоуҭа араион .
* Монтевидео ( проектируется )
* Лыхны
' '' Владимир '' ' - Урыстәыла ақалақь .
# Очамчыра араион
```

- http://blender02.cs.rpi.edu:3300/elisa_ie
- http://blender02.cs.rpi.edu:3300/elisa_ie/heatmap

Cross-lingual Entity Discovery and Linking for 282 Languages (Cont')



Tagging result:

demo

Total: 13 NAM: 13 NOM : 0

PER: 5 ORG: 1 GPE: 0 LOC: 7 FAC: 0 TTL: 0

Засновником саме ренесансної архітектури став [PER Філіппо Брунеллескі]

"Примітка : таблиця складена за даними сайту [ORG Верховної Ради Украї

* Автори сценарію : Поль Андреотті , [PER Олександр Галич]

* [PER Район Янпу] (杨浦区 Yángpǔ Qū)

Після завершення війни використовувалась арміями [LOC Австрії] , [LOC Чехословаччини] , [LOC Югославії] .

* [PER Грехем Чепмен] — Браян Коен / Біггус Діккус (товариш Пилата) ;

Муніципалітет розташований на відстані близько 220 км на схід від [LOC Парижа] , 80 км на південний захід від [LOC Меца] , 13 км на південний схід від [LOC Бар] –ле–[LOC Дюка] .

Mention ID: demo-6
Mention Str: Югославії 300:308
Reference KB: Socialist_Federal_Republic_of_Yugoslavia
Entity Type: LOC
Mention Type: NAM
Confidence Value: 1.0
Images:
Not Found

: 0

AC: 0 TTL: 0

நகரம் மாவட்டம்]

த் தொகுதி)]

இது [ORG தாம தேசிய நெடுஞ்சாலையில்] இருந்து 3 கி.மீ தொலைவில் உள்ளது .

' " [LOC போலந்து] " '

[ORG மார்வெல் காமிக்ஸ்] நிறுவனத்தின் வரைகதைகளில் தோன்றும் [PER அயன் மேன்] என்ற கதாபாத்திரத்தை மையமாகக் கொண்டு எடுக்கப்பட்டது .

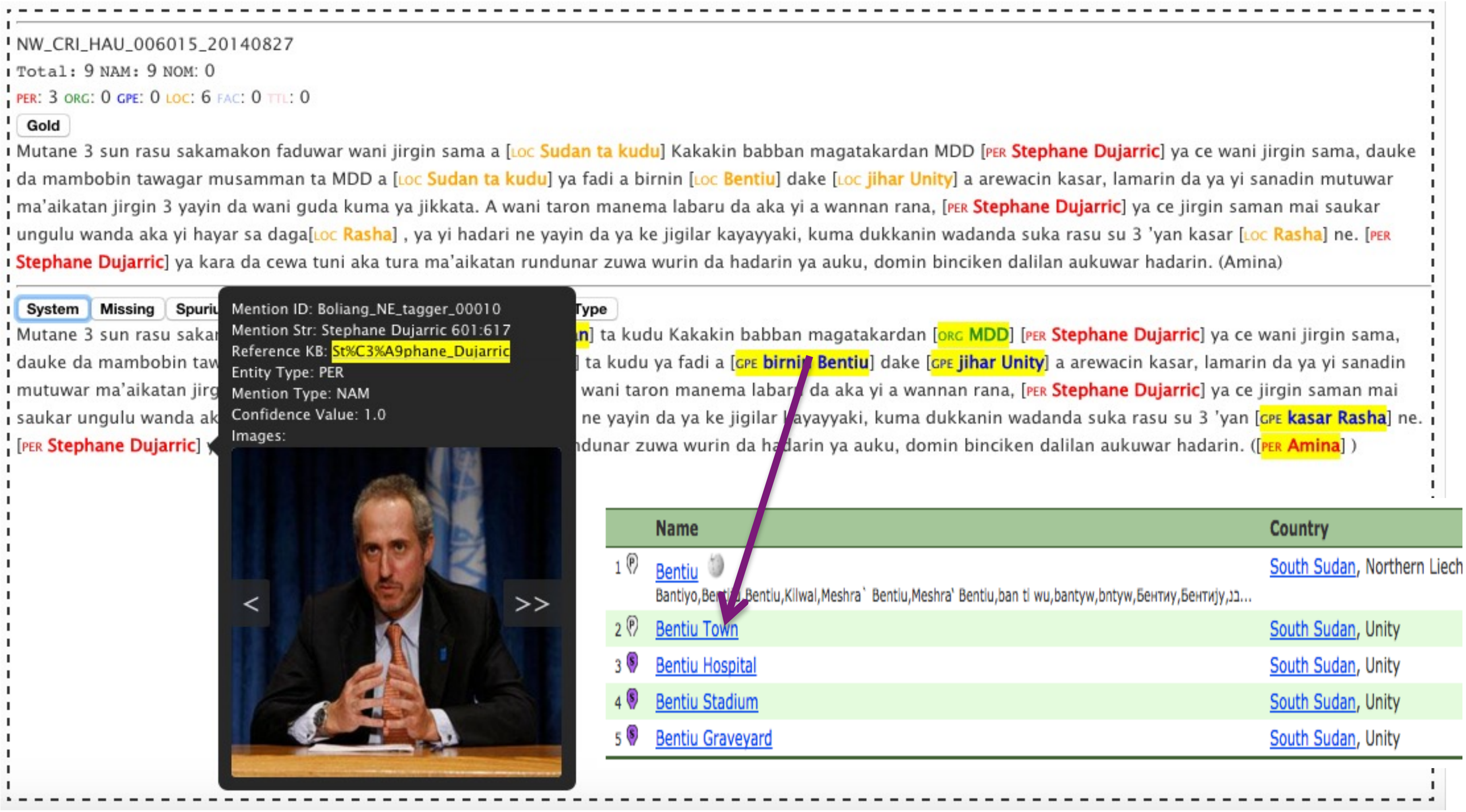
வழிமாற்று [ORG என்] . எம். [LOC சோசி]

* [LOC நோர்போக் த்வு] - 2,302

வழிமாற்று [LOC சேமக்கலம் (பொறியியல்)]

Mention ID: demo-3
Mention Str: போலந்து 132:138
Reference KB: Poland
Entity Type: LOC
Mention Type: NAM
Confidence Value: 1.0
Images:
Not Found

Cross-lingual Entity Discovery and Linking for 282 Languages (Cont')



NW_CRI_HAU_006015_20140827
Total: 9 NAM: 9 NOM: 0
PER: 3 ORG: 0 GPE: 0 LOC: 6 FAC: 0 TTL: 0

Gold

Mutane 3 sun rasu sakamakon faduwar wani jirgin sama a [LOC Sudan ta kudu] Kakakin babban magatakardan MDD [PER Stephane Dujarric] ya ce wani jirgin sama, dauke da mambobin tawagar musamman ta MDD a [LOC Sudan ta kudu] ya fadi a birnin [LOC Bentiu] dake [LOC jihar Unity] a arewacin kasar, lamarin da ya yi sanadin mutuwar ma'aikatan jirgin 3 yayin da wani guda kuma ya jikkata. A wani taron manema labaru da aka yi a wannan rana, [PER Stephane Dujarric] ya ce jirgin saman mai saukar ungulu wanda aka yi hayar sa daga[LOC Rasha] , ya yi hadari ne yayin da ya ke jigilar kayayyaki, kuma dukkanin wadanda suka rasu su 3 'yan kasar [LOC Rasha] ne. [PER Stephane Dujarric] ya kara da cewa tuni aka tura ma'aikatan rundunar zuwa wurin da hadarin ya auku, domin binciken dalilan aukuwar hadarin. (Amina)

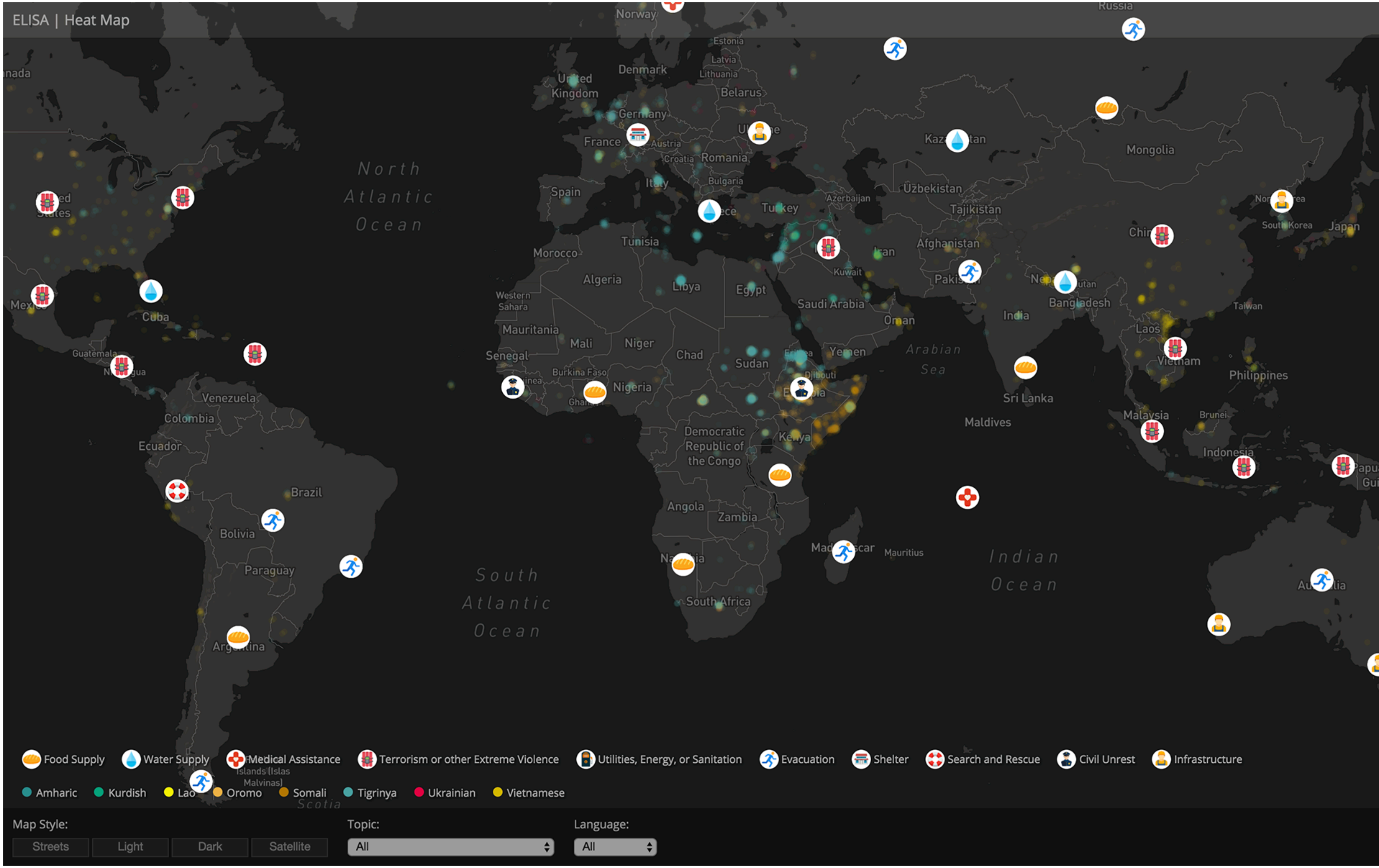
System | Missing | Spuri... | ...Type

Mutane 3 sun rasu sakar ... n] ta kudu Kakakin babban magatakardan [ORG MDD] [PER Stephane Dujarric] ya ce wani jirgin sama,
dauke da mambobin taw ...] ta kudu ya fadi a [GPE birnin Bentiu] dake [GPE jihar Unity] a arewacin kasar, lamarin da ya yi sanadin
mutuwar ma'aikatan jirg ... wani taron manema labaru da aka yi a wannan rana, [PER Stephane Dujarric] ya ce jirgin saman mai
saukar ungulu wanda ak ... ne yayin da ya ke jigilar kayayyaki, kuma dukkanin wadanda suka rasu su 3 'yan [GPE kasar Rasha] ne.
[PER Stephane Dujarric] ... ndunar zuwa wurin da hadarin ya auku, domin binciken dalilan aukuwar hadarin. ([PER Amina])

Mention ID: Boliang_NE_tagger_00010
Mention Str: Stephane Dujarric 601:617
Reference KB: St%C3%A9phane_Dujarric
Entity Type: PER
Mention Type: NAM
Confidence Value: 1.0
Images:

	Name	Country
1	Bentiu Bantiyo,Bentiu,Bentiu,Kilwal,Meshra` Bentiu,Meshra' Bentiu,ban ti wu,bantyw,bntyw,Бентиу,Бентиjу,בנ...	South Sudan, Northern Liech
2	Bentiu Town	South Sudan, Unity
3	Bentiu Hospital	South Sudan, Unity
4	Bentiu Stadium	South Sudan, Unity
5	Bentiu Graveyard	South Sudan, Unity

IE Application Example: Disaster Relief



ELISA | Heat Map

Food Supply · Water Supply · Medical Assistance · Terrorism or other Extreme Violence · Utilities, Energy, or Sanitation · Evacuation · Shelter · Search and Rescue · Civil Unrest · Infrastructure

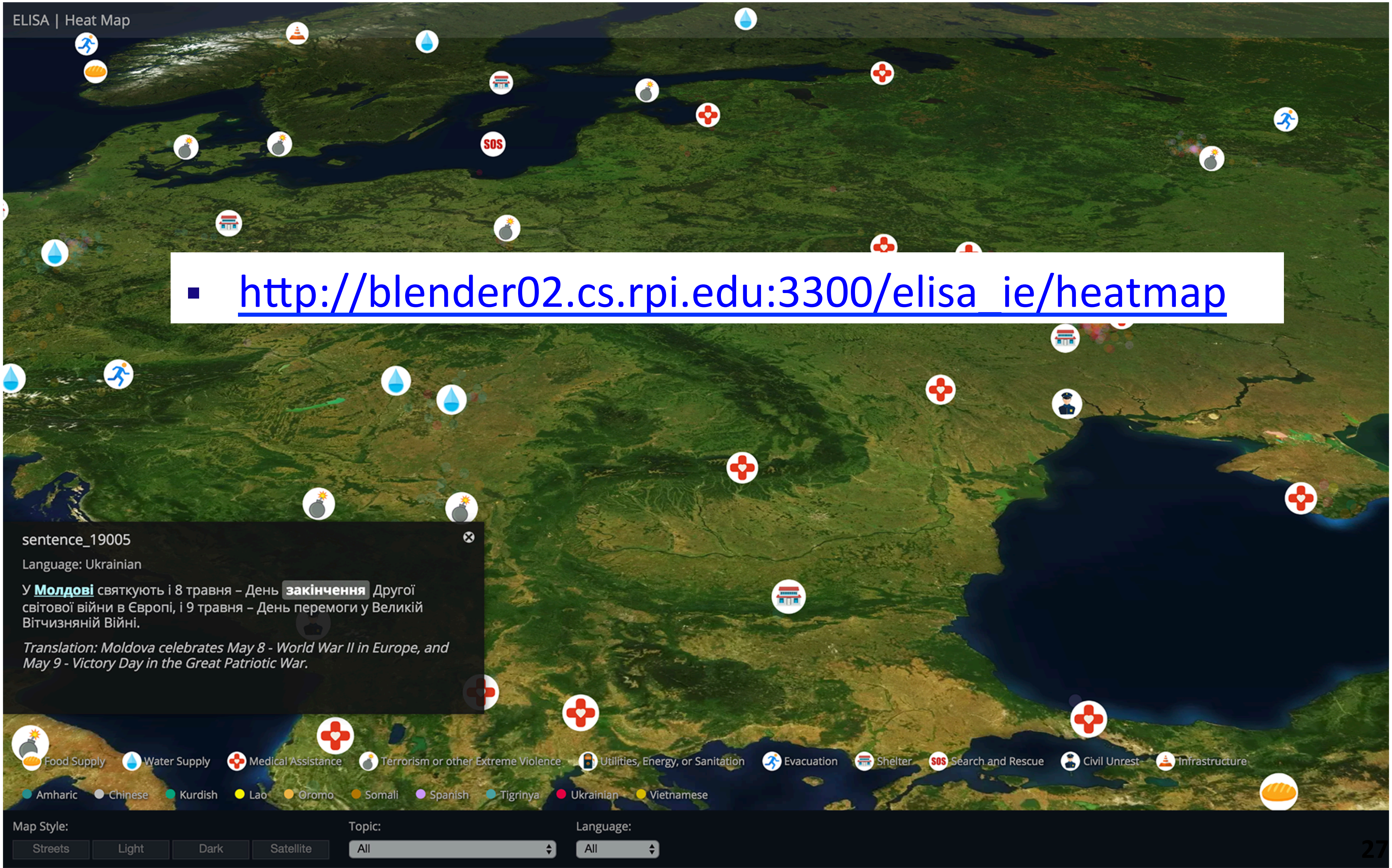
● Amharic ● Kurdish ● Lao ● Oromo ● Somali ● Tigrinya ● Ukrainian ● Vietnamese

Map Style: Streets | Light | Dark | Satellite

Topic: All

Language: All

Cross-lingual Entity Discovery and Linking for 282 Languages (Cont')



- http://blender02.cs.rpi.edu:3300/elisa_ie/heatmap

sentence_19005

Language: Ukrainian

У **Молдові** святкують і 8 травня – День **закінчення** Другої світової війни в Європі, і 9 травня – День перемоги у Великій Вітчизняній Війні.

Translation: Moldova celebrates May 8 - World War II in Europe, and May 9 - Victory Day in the Great Patriotic War.