Linguistic Rapid Response Teams

Chris Dyer, Lori Levin, Jaime Carbonell

DARPA pilot project on low resource languages

- May through December 2014
- Named entity recognition
  - Sorani Kurdish (four months)
  - Tajik (two days)
  - Only monolingual data
- Machine Learning Staff
  - Kartik Goyal, graduate student
  - Swabha Swayamdipta, graduate student
- Linguistics Staff
  - David Mortensen, visiting faculty
  - Patrick Littell, graduate student, UBC
  - Alexa Little, undergraduate student, Yale

Unsupervised with Linguists in the Loop (scenario without native speaker)

- In a rapid response scenario there may not be enough time for corpus annotation for supervised NER.
- Unsupervised clustering
  - Clusters reduce dimensionality
  - linguists identify clusters that seem to consist of names, locations, organizations
- Unsupervised morphology
  - Morphological analysis reduces sparsity
  - linguists identify correct and incorrect morphemes based on information from
    - reference grammars

Clustering in Sorani Kurdish

- Data (Language Pack)
  - 2.2M words
  - ~145 sentences
  - 208,985 types
- Used the training section of the language pack
- Cluster sizes: 50, 100, 500, 1000
- Gold standard for evaluation was produced by linguists who did not speak Kurdish (more later).
  - Includes NE types: PER, LOC, ORG, TTL, TIME

NER in Sorani

- Experiments
  - Cluster size: 50, 100, 500, 1000
  - Gazeteer: Y/N
  - Morphological analysis: Y/N
  - Foreign: Y/N
    - Map Sorani words to Kurmanji words (more later)
    - Use features of Kurmanji words such as capitalization and whether the word is in Wikipedia

Sorani NER Results

	Gaz	Morph	Foreign	Recall	Precision	F-score	Ans/633	Tags-6295 /correct/ correct IOB	Ratio of class features	Ratio of Foreign features	
N	N	N	Y	0.4123	0.6941	0.5173	376	5734/5741	0	0.01	
500	N	N	N	0.4755	0.7016	0.5669	429	5789/5798	0.35	0	
500	N	N	Y	0.5039	0.7169	0.5918	445	5811/5824	0.348	0.009	
500	Y	N	Y	0.4897	0.7506	0.5927	413	5919/5832	0.346	0.008	
1000	Y	N	Y	0.5039	0.7559	0.6047	422	5829/5844	0.396	0.008	
100	Y	N	Y	0.4976	0.7647	0.6029	412	5825/5833	0.216	0.009	
500	Y	Y	Y	0.5087	0.7506	0.6064	429	5825/5837			
1000	Y	Y	Y	0.4976	0.7975	0.6128	395	5844/5854			

Bridge Language Capitalization Inference

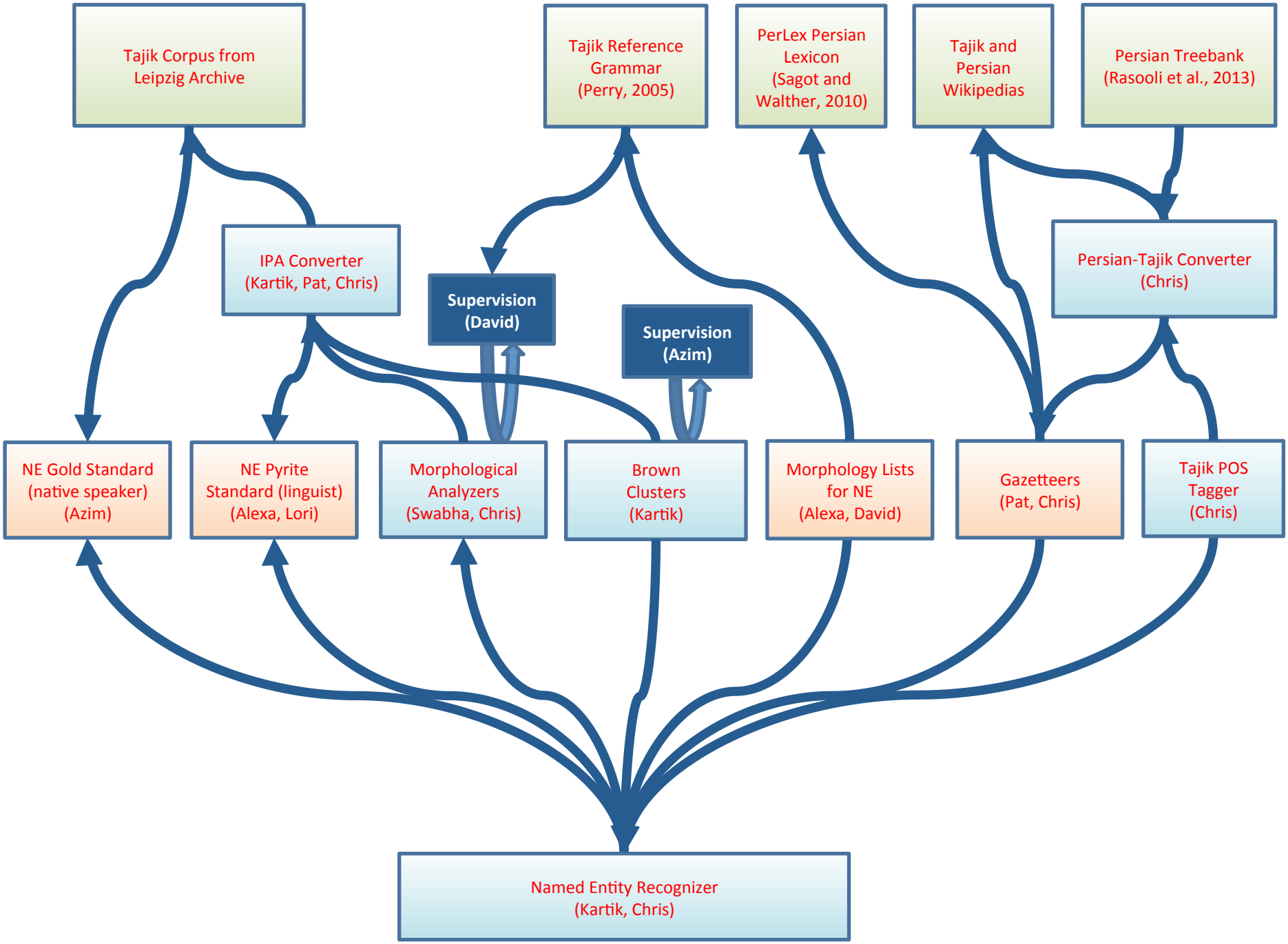
- Sorani Kurdish is written in the Arabic alphabet
  - No short vowels
  - No capital letters
- Kurmanji Kurdish is written in the Latin alphabet
- While the languages share many stems, and often render names similarly, the words are not always the same (especially w.r.t. the vowels), and the languages have notably different morphologies.
  - e.g. Sorani *nɛtɛwɛ jɛkgirtuːɛkan* vs. Kurmanji

IPA-ization and Edit Distance

- Extract Sorani and Kurmanji vocabulary from news stories
  - –Comparable, but not parallel
- Convert Sorani and Kurmanji vocabulary to IPA (International Phonetic Alphabet)
  - –Existing resource: Unitran (Unicode to X-SAMPA mapping)
  - –With improvements
    - Produce and score edits
      - –for a hypothesized pair *<bri:tanja, britanija>*, the edits would be *<b,b>, <r,r>, <i:,i>, <t,t>, <a,a>, <n,n>, <i,∅>, <j,j>, <a,a>.*

Tajik two day exercise

- Seven people at CMU
  - Dyer, Levin
  - Three linguists
  - Two ML people
- Native speaker on Skype
  - six hours each day
- Results on NER
  - P, R, F


Tajik Corpus from Leipzig Archive
Tajik Reference Grammar (Perry, 2005)
PerLex Persian Lexicon (Sagot and Walther, 2010)
Tajik and Persian Wikipedias
Persian Treebank (Rasooli et al., 2013)
IPA Converter (Kartik, Pat, Chris)
Supervision (David)
Supervision (Azim)
Persian-Tajik Converter (Chris)
NE Gold Standard (native speaker) (Azim)
NE Pyrite Standard (linguist) (Alexa, Lori)
Morphological Analyzers (Swabha, Chris)
Brown Clusters (Kartik)
Morphology Lists for NE (Alexa, David)
Gazetteers (Pat, Chris)
Tajik POS Tagger (Chris)
Named Entity Recognizer (Kartik, Chris)

Tajik data

124,000 tokens parallel data extracted from Wiki titles (English-Tajik)
125,000 tokens parallel Farsi-Tajik
1.7M tokens monolingual news web crawl (87,000 types)

Tajik Results: NER

- Cluster size: 500
- Gazeteers
- Morphological analysis
- Foreign word features
- Recall: 0.4421
- Precision: 0.5912
- F-score: 0.5059

Word cluster viewer

Word cluster viewer.

Cluster path (and word type count)	Words (most frequent)
^000000 (252)	Рашт Исфара Тавилдара Дарвоз Айнӣ Варзоб Конибодом Истаравшан Исмоили Восеъ Файзобод Панҷакент Сино Камароб Фархор Ҳисор Панҷ Ғарм Данғара Рудаки Балҷувон Шӯрообод Ашт Спитамен Мурғоб Раштро Свот Мастчоҳ Ишкошим Фарғона Бохтар Нуробод Ҳамадонӣ Хуросон Ёвон Ванҷ Раҳматуллоев Синои Рушд Мӯъминобод Фирдавсӣ Аризона Хазар Сӯх Ховалинг Зафаробод Румӣ Ҷиргатол Фаргона Ванч
^00000100 (120)	Бишкек Кобул Хоруғ Бағдод Тифлис Исломобод Остона Прага Алмаато Брюссел Боку Сингапур Истамбул Женева Вена Берлин Рум Пешовар Сочӣ Порис Истанбул Лиссабон Мумбай Қоҳира Ереван Лоҳур Карочӣ Басра Сараево Афина Торонто Урумчӣ Карбало Мингора Багром Флорида Ҷакарта Грозний Қарабоғ Лисабон Мироншоҳ Тула Махачқалъа Стокҳолм Уфа Копенҳаген Севастопол Зоҳидон Таллин ҳаёташ
^000001010 (48)	Теҳрон Тошканд Вошингтон Вашингтон Пекин Пхенян Сеул Белград маҳофил Токио меҳмонӣ Байтулмуқаддас Химкӣ Деҳлӣ Димишқ фурудгоҳҳо Химки ноумеди сарқонун Пушкин Бейрут ҳамоише ҳуқуқдон Ашқобод қатора Ярославл Санъо Страссбург ғанисозӣ Кишинёв ислоҳотхоҳон Эвин Приштина
^000001011 (62)	Маскав Минск Киев Екатеринбург Варшава Бангкок Ландан Макка Домодедово Шероз Чкалов Париж Ишқобод Мадинаи
^00000110 (204)	Кӯлоб Қӯрғонтеппа Турсунзода Ваҳдат Душанберо Янгон Қурғонтеппа Ӯшро Кӯлобро Қазон Хуҷандро Сарбанд Мазори Курил Хонобод Перм Кант фаластинии Бостон Чоркух ёрон Схинвалӣ Айниро Қӯрғонтеппа Гаага Тирмиз Кӯлоб Табошар академия Мумбайи автомобилӣ Сидней Митровитса Коннектикут Шукнов Гориро зал УБОП Хабаровск Йоханнесбург ИҶШС Ваҳдатро Анталияи Алматӣ Ёнпиёнг шимолро Вирҷинияи Кӯҳдоман Сариоғош таносули
^000001110 (194)	Ӯш Ҷалолобод Қандаҳор Андиҷон Қундуз ҷанубии шимолии Горӣ нооромӣ Уш ғарбии Ҳирот шарқии Талас Бодканд Балуҷистон Смоленски Хайбар Ӯши мухтори Ғазнӣ Мексикаро Смоленск шимолуғарбии Андиҷонро Мексик Бағлон Қазони Қум Вирҷиния Бодканди Исфаҳон Ош соҳилии бандарии Хуҷанди Наҷаф Уши Таласи Маскави Тулаи Қарағанда Челябински Зугдидӣ Екатеринбурги Мухтори Систону Пактико ҷазира Виборги

Instructions - Please Read

We need your help classifying affixes guessed by a morphological analysis tool on **Tajik** word froms. For each listed hypothesized affix, you will see a list of words where it has been used (possibly incorrectly). Your job is to tell us whether or not each affix is a possible affix in the language or not (or whether you are not sure). Hints:

- Forms should be labeled as **Good** as long as they are correct in some (possibly phonologically or morphologically conditioned) context. For example in English, both **+es** and **+s** should be annotated as correct although we usually think of the later as the "default" form. A form like **+ka** would not be a good English form since this is not a derivational or inflectional affix in English, therefore the option marked **Bad** should be checked if it were to appear in a morpheme list for English.
- Higher quality affixes will probably be higher in the list, but please go through the entire list.
- Examples of analyses containing each affix are listed for informational purposes. It is important to remember that you are **not** judging these anaylses, i.e., whether the affix in question is actually present in the examples, but only whether the posited affix is part of the language.

Forms are in the International Phonetic Alphabet. If you see boxes or other indications of missing symbols, you may need to install an IPA font on your computer. Here is one from SIL.

Task (Don't forget to submit using the button below!)

Bad	Unsure	Good	Affix	Example Analyses with Affix
○	●	○	+i (30860)	/intizɔːm/+**i**+jɑʃ+ɔːn /ʃukuh/+**i** /jɑtim/+**i** /tærk/+**i**+dæ /sæ… /t͡ʃæm/+k+ɔːni+**i** /bɔːnæzɔːkæt/+**i**
○	●	○	+rɔː (10447)	/ʃærɔːit/+æʃ+**rɔː** /bæhs/+hɔː+**rɔː** /dɔːχil/+i+**rɔː** /muhɔːd͡ʒirɔːn… /sɔːhibkɔːr/+**rɔː**+n /ʃær/+n+if+**rɔː** /ræhbær/+i+**rɔː**
○	●	○	+æ (9762)	/ʃunævænd/+**æ**+æʃ /gʃæv/+**æ**+d /durust/+**æ**+nd /æjd͡ʒ/+**æ**… /bɔːzgæʃt/+**æ** /næmɔːnæ/+n+**æ**+d /rɔːb/+b+ik+**æ**
○	●	○	+ɔːn (6826)	/intizɔːm/+i+jɑʃ+**ɔːn** /ʃæχv/+æt+æt+**ɔːn** /ræfik/+**ɔːn** /d͡ʒɔːn/… /deːχæ/+æm+**ɔːn**+rɔː /vɔːr/+**ɔːn**+eː+ʒ /χæjr/+**ɔːn**+æm
○	●	○	+eː (6058)	/kætræ/+**eː**+st /sɔːziʃ/+**eː**+rɔː /vɔːr/+ɔːn+**eː**+ʒ /vɔːqeːæ/+**eː**… /rəːhɔːnijɔn/+**eː** /kəːʃiʃ/+**eː** /ɔːtæʃsəːz/+i+**eː**
○	●	○	+hɔː (5538)	/bæhs/+**hɔː**+rɔː /dæstægul/+**hɔː** /χudsæri/+**hɔː**+i /sæhnæ/+i… +vu /blɔːqaeːr/+**hɔː**+i /ideːɔːlɔːqiiq/+**hɔː** /kɔːnveːntsi/+iq+**hɔː**

500	Y	Y	Y	0.4421	0.5912	0.5059	181/242	4926/493		

MURI and LRRT

- The linguistic core for LRRT was the use of IPA.
- LRRT uses linguists in active learning
- MURI uses more linguistic resources