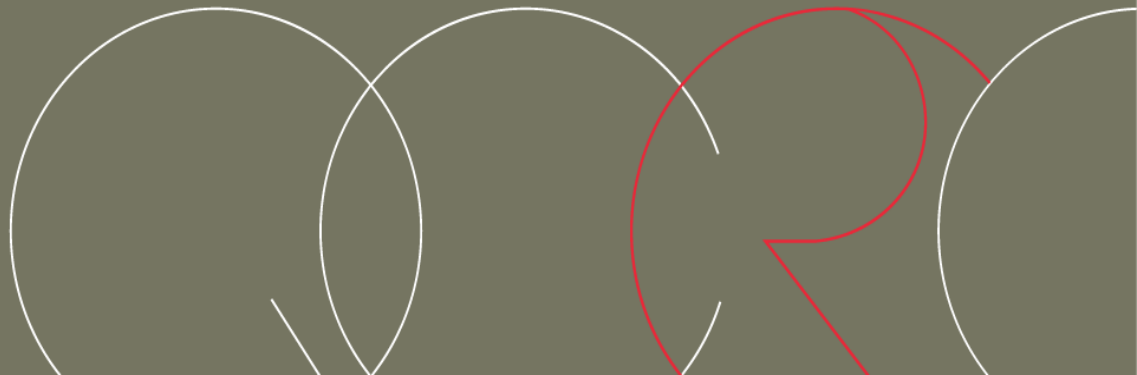
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

NER using Cross-Lingual Resources: Arabic as an Example

Kareem Darwish



Motivation

- Named entity tagging uses features such as:
  - Orthographic features (ex. Capitalization)
  - Contextual features (ex. President *X*)
  - POS tagging (ex. NE's not Verbs)
  - Character level features (ex. Paki*stan*, Bloom*berg*)
  - Gazetteers

Motivation

- Some languages (ex. English) are NER friendly:
  - Indicative features (ex. Capitalization)
  - Good knowledge bases (ex. Large Wikipedia (4.2M+), Freebase, DBPedia)
  - Good language resources (ex. POS taggers)
- Other languages (ex. Arabic) are not so lucky:
  - No strong orthographic features
  - Poor knowledge bases (Ar Wikipedia < 250k)
- Can we use friendly features in one language for another language?

Proposed Solution

- Use cross-lingual features to make use of advantaged language:
  - What is the likelihood that a translation of a word is capitalized?
  - What is the likelihood that a translation of a word is transliterated?
  - What are the knowledge base tags associated with the translation of a word?

Arabic NER Features

- Some of the Arabic properties pertaining to NER:
  - Has no capitalization feature
  - Character level features (leading and trailing letters) are effective – can be substitutes for POS tagging
  - Public Arabic gazetteers are small (few thousand entries – Benajiba et al. 2008) – stemming can improve gazetteer coverage

Our Baseline Setup

- Features in baseline setup
  - Word, previous word, next word
  - Leading & trailing 1, 2, 3, and 4 characters
  - Stemmed version of the word
  - Whether the word appears in a publicly available gazetteer (Benajiba et al. 2008)
- Decoding using CRF sequence labeler (CRF++)

Training and Test Sets

- Training set:
  - 80% of ANERCORP dataset, containing 120k tokens
    - News article from same source and time period
- Test sets:
  1. 20% of ANERCORP dataset, containing 30k tokens
  2. New NEWS test from over a dozen news sources, containing 15k tokens
  3. New Arabic TWEETS test set, containing 26k tokens from randomly selected tweets from Nov. 23-27 2011

Baseline Results

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	93.6	83.3	88.1
ORG	83.8	61.2	70.8
PERS	84.3	64.4	73.0
Overall	88.9	72.5	79.9

Training/test parts cover same time period, same genre, same source

(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	84.1	53.2	65.1
ORG	73.2	23.2	35.2
PERS	74.8	47.1	57.8
Overall	78.0	41.9	54.6

Training on ANERCORP does not generalize well to new news texts

(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	79.9	27.1	40.4
ORG	44.4	9.1	15.1
PERS	45.7	27.8	34.5
Overall	58.0	23.1	33.1

Results on tweets are horrible

Cross Lingual and English Resources

- True cased Arabic–English phrase table
  - Trained on 3.69 million parallel sentences containing 123.4M English tokens from NIST 2012 MT eval
- Transliteration miner that:
  - Detects the presence of a transliteration between 2 text segments (both could of length 1)
  - Trained on 3,452 parallel word pairs
- Wikipedia cross-language links (254k)
- DBPedia (6M entries): entity + category
  - Ex. NASA: Agent, Organization, & GovernmentAgency

Cross Lingual Capitalization

- Given a phrase table (from MT), what is the likelihood that a translation of a word or a phrase is capitalized?
  - Capitalization is a STRONG feature in English
  - We used an Arabic-English phrase table
  - Feature value =

    Σ(P of CAP'ed translations)/Σ(P of all translations)
  - We favored the longest word sequence with entry in phrase table

Cross Lingual Capitalization Results

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.0/-1.6/-1.7	86.8/3.5/4.2	89.3/1.2/1.4
ORG	82.8/-1.1/-1.3	63.9/2.7/4.4	72.1/1.4/1.9
PERS	86.0/1.7/2.0	75.4/11.0/17.1	80.3/7.3/10.1
Overall	88.4/-0.4/-0.5	78.6/6.1/8.4	83.2/3.4/4.2

(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	82.1/-2.0/-2.4	59.0/5.8/11.0	68.7/3.5/5.4
ORG	68.4/-4.9/-6.6	23.2/0.0/0.0	34.6/-0.6/-1.7
PERS	70.7/-4.0/-5.4	55.6/8.4/17.9	62.2/4.4/7.6
Overall	74.5/-3.5/-4.5	47.0/5.1/12.2	57.7/3.1/5.7

(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	76.9/-3.0/-3.7	27.9/0.9/3.2	41.0/0.5/1.4
ORG	44.4/0.0/0.0	10.4/1.3/14.3	16.8/1.8/11.6
PERS	40.0/-5.7/-12.5	35.0/7.3/26.2	37.3/2.8/8.1
Overall	51.8/-6.2/-10.7	26.3/3.1/13.6	34.9/1.8/5.4

How to read the table:
Value/
Absolute_improvment/
Relative_improvement

- Loss in P, Gain in R
- Overall gain for all collections
- Most R again in PERS

Cross Lingual Transliteration

- Given a phrase table (from MT), what is the likelihood that a translation of a word is also a transliteration?
  - Many NE's (specially PERS & LOC) are transliterated
  - We used an Arabic-English phrase table
  - Given a transliteration model,
    - feature value = $\frac{\Sigma(\text{P of translations that are transliteration})}{\Sigma(\text{P of all translations})}$
  - Ex. حسن ➔ Hasan, Hassan, good
    - Val = (P(hasan)+P(hassan))/(P(hasan)+P(hassan)+P(good))

Cross Lingual Transliteration Results

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.9/-0.7/-0.7	83.5/0.2/0.3	88.0/-0.2/-0.2
ORG	82.9/-0.9/-1.0	61.8/0.6/1.0	70.9/0.1/0.1
PERS	84.5/0.3/0.3	71.9/7.5/11.7	77.7/4.7/6.5
Overall	88.3/-0.5/-0.6	75.5/2.9/4.1	81.4/1.5/1.9

(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	84.9/0.7/0.9	53.6/0.5/0.9	65.7/0.6/0.9
ORG	67.2/-6.1/-8.3	22.9/-0.3/-1.1	34.2/-1.0/-2.9
PERS	72.8/-1.9/-2.6	55.0/7.8/16.7	62.7/4.8/8.4
Overall	75.9/-2.1/-2.6	45.0/3.1/7.4	56.6/2.0/3.7

(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	79.1/-0.8/-1.0	27.1/0.0/0.0	40.3/-0.1/-0.3
ORG	41.8/-2.7/-6.0	9.1/0.0/0.0	14.9/-0.2/-1.1
PERS	40.0/-5.7/-12.5	35.5/7.7/27.8	37.6/3.1/8.8
Overall	51.7/-6.3/-10.9	25.8/2.6/11.3	34.4/1.3/3.9

How to read the table:
Value/
Absolute_improvment/
Relative_improvement

- Loss in P, Gain in R
- Overall gain for all collections
- Helps most for PERS

Using DBPedia

- What is the tag of the translation in DBPedia?
  - DBPedia provides meta-information about entities
    - Some categories are too broad (ex. Work, Agent). Both ignored.
    - Some entities have multiple categories. We picked most common category.
  - Translation done using Wikipedia cross-language links and phrase table
    - Favored longest word sequence and used most likely translation
    - Ex. حزب الله ➔ Hezbollah ➔ Organization
      - Feature: حزب :B-Organization; الله :I-Organization

Using DBPedia Results

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.7/-0.9/-0.9	87.1/3.9/4.6	89.9/1.7/1.9
ORG	84.6/0.8/0.9	66.6/5.3/8.7	74.5/3.7/5.3
PERS	87.8/3.6/4.2	69.9/5.5/8.6	77.8/4.8/6.6
Overall	89.8/0.9/1.0	77.2/4.7/6.5	83.0/3.2/4.0

(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	87.8/3.6/4.3	61.8/8.6/16.2	72.5/7.4/11.3
ORG	76.1/2.9/3.9	30.2/7.0/30.1	43.2/8.0/22.7
PERS	83.2/8.5/11.3	54.2/7.1/15.0	65.7/7.8/13.6
Overall	83.5/5.5/7.1	49.5/7.5/18.0	62.2/7.6/13.9

(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	77.4/-2.5/-3.1	30.5/3.5/12.9	43.8/3.4/8.4
ORG	57.0/12.5/28.2	15.9/6.8/75.1	24.8/9.8/64.9
PERS	40.8/-4.9/-10.6	31.7/4.0/14.3	35.7/1.2/3.4
Overall	55.3/-2.6/-4.5	27.5/4.4/19.1	36.8/3.7/11.2

How to read the table:
Value/
Absolute_improvment/
Relative_improvement

- Gain in P & R
- Most gain for ORG
- Tweets set is weird

Putting it all together

Using ALL Features

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.3/-1.3/-1.4	87.8/4.6/5.5	90.0/1.9/2.1
ORG	81.4/-2.4/-2.9	66.0/4.7/7.7	72.9/2.1/3.0
PERS	87.0/2.8/3.3	77.7/13.3/20.7	82.1/9.1/12.5
Overall	88.7/-0.2/-0.2	80.3/7.8/10.7	84.3/4.4/5.5

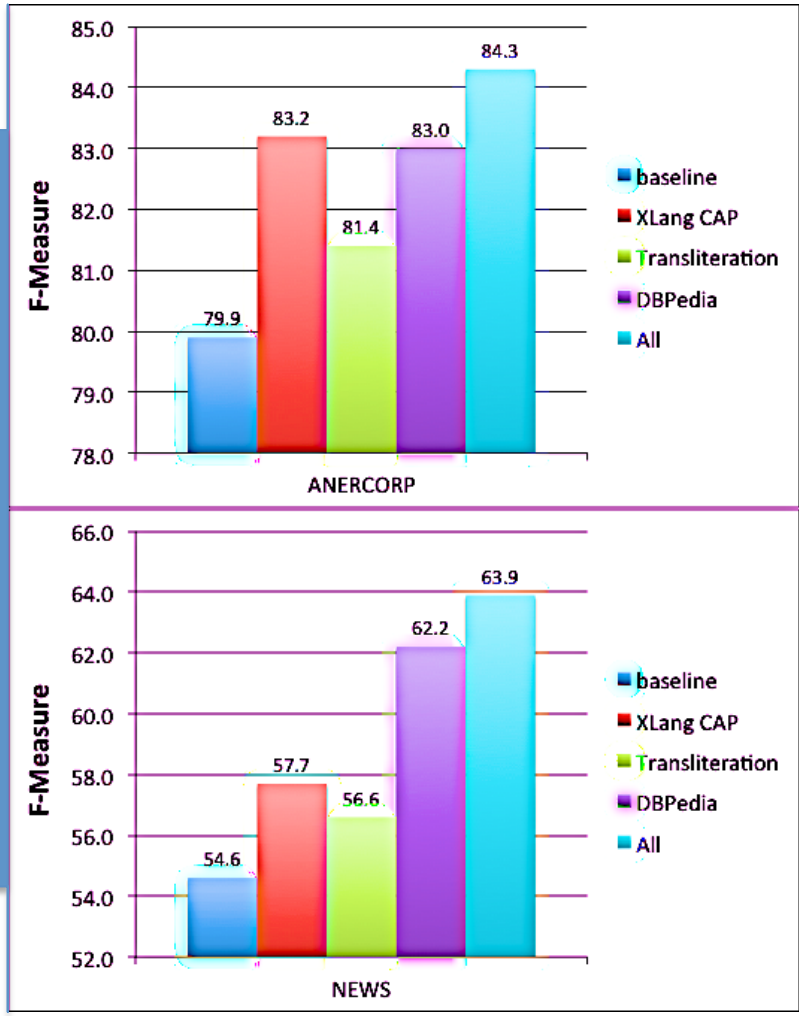
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	85.1/1.0/1.2	64.1/11.0/20.6	73.1/8.0/12.3
ORG	73.8/0.5/0.7	29.4/6.2/26.9	42.1/6.8/19.4
PERS	76.8/2.0/2.7	63.4/16.3/34.5	69.5/11.7/20.2
Overall	79.2/1.2/1.6	53.6/11.6/27.7	63.9/9.4/17.1

(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	81.4/1.5/1.8	33.5/6.5/23.9	47.5/7.1/17.4
ORG	52.1/7.6/17.2	16.2/7.1/78.6	24.7/9.6/64.1
PERS	40.5/-5.2/-11.4	39.2/11.5/41.3	39.8/5.3/15.4
Overall	54.4/-3.6/-6.2	31.4/8.3/35.9	39.9/6.8/20.5

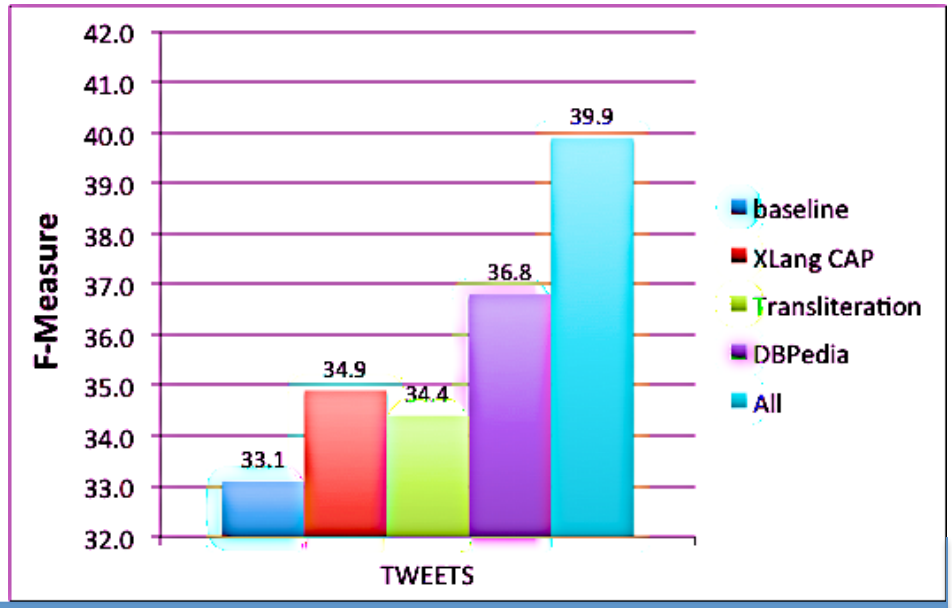
- Small loss in P, BIG gain in R
- Results for ANERCORP are better than the best reported result in lit
- Much bigger gain for NEWS & TWEETS compared to ANERCORP
- TWEETS set still weird (more later)

All Results

- DBPedia gives the most gain
- Cross language capitalization is better than detecting transliterations
- The greater the difference between training set and test set, the greater is the relative gain



All Results



- Tweets have:
  - Abbreviated NE's ("Real" instead of "Real Madrid")
  - NE usually appear at beginning or end: (FED: interest …)
  - Has non-standard text: abbreviations, emoticons, URL's, hashtags, dialectal text
- Tweets need in-domain data and other tweet specific methods

Conclusion

- Cross lingual features significantly improve NER:
  - Make use of useful features in other languages
  - Makes use of large knowledge bases in other languages
  - Detecting transliterations can potentially work between any two language pairs
- Outstanding issues:
  - When using DBPedia, including multiple tags
  - What to do about tweets