Motivation

Task: Fact extraction from massive corpora (e.g., news, tweets, papers) to facilitate heterogeneous information network construction

Given a sentence “President Blaise Compaoré’s government of Burkina Faso was founded...”, ...

Task 1: (entity, attribute name, attribute value)-tuple extraction
(Burkina Faso, president, Blaise Compaoré) (Burkina Faso, population, 17 million) (Burkina Faso, age, 65)

Task 2: (entity type, attribute name, value type)-tuple extraction
(SLOCATION.COUNTRY, president, $PERSON.POLITICIAN) (SLOCATION.population, $DIGIT.DIGITUNIT) ($PERSON.age, $DIGIT)

Idea: Discovering a group of synonymous “meta patterns” to find facts.

The MetaPAD Framework

Meta patterns:

1. Meta pattern generation via context-aware segmentation
   - president $PERSON.POLITICIAN’s government of sCOUNTRY was founded... SLOCATION.COUNTRY
   - (United States, Barack Obama)

2. Adjust types for appropriate granularity
   - (Burkina Faso, president, Blaise Compaoré) (U.S., president, Barack Obama)

3. Group synonymous meta patterns
   - (CSLOCATION, [president], $POLITICIAN)
   - (Burkina Faso, Burkina Faso, president, Blaise Compaoré)
   - (U.S., president, Barack Obama)

Step 1: Meta pattern quality assessment and segmentation

A rich set of features:
- Frequency
- Concordance: “$PERSON’s wife”
- Completeness: “SLOCATION president” vs “SLOCATION president $POLITICIAN”
- Informativeness: “$PERSON and $PERSON” vs “$PERSON’s wife, $PERSON"

Regression Q(·): random forest with only 300 labels

Q(·)

Step 2: Grouping synonymous meta patterns

Meta patterns

- $PERSON, $DIGIT
- $PERSON, $DIGIT -year-old
- $PERSON, $DIGIT
- $PERSON, $DIGIT
- $PERSON, $DIGIT

Step 3: Adjusting types for appropriate granularity

Meta patterns

- $PERSON, $POLITICIAN
- $PERSON, $POLITICIAN
- $PERSON, $POLITICIAN
- $PERSON, $POLITICIAN
- $PERSON, $POLITICIAN

Entity

- United States
- Barack Obama
- Russia
- Vladimir Putin
- France
- Francois Hollande
- Burkina Faso
- Blaise Compaoré
- Apple
- Tim Cook
- Facebook
- Mark Zuckerberg
- Hewlett-Packard
- Carly Fiorina
- Infor
- Charles Phillips
- Afghan Citadel
- Roya Mahboob
- corynebacterium striatum
- BM467
- methicillin-resistant S aureus
- vancomycin
- multidrug-resistant enterobacteriaceae
- gentamicin
- afrodisiacic acid therapy
- Paget’s disease of bone
- bisphosphonates
- osteoporosis
- calcitonin
- Paget’s disease of bone

Meta patterns

- SLOCATION
- SPERSON
- $POLITICIAN
- $SDIGIT
- $BDIGIT
- $STCT
- $LOC
- $UNIT
- $BDIGIT
- $COUNTRY
- $STCT
- $.politicann

Entity

- $PERSON
- $POLITICIAN
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON

Attribute

- $COUNTRY
- $POLITICIAN
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON

Meta patterns

- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON
- $PERSON

Step 0: Preprocessing

“President Blaise Compaoré’s government of Burkina Faso was founded...”

Phrase mining (SegPhrase by Liu and Han et al. SIGMOD’15)

“president blaise_compaoré’s government of burkina_faso was founded...”

Entity recognition and typing with Distant Supervision (CluType by Ren and Han et al. KDD’15)

“president $PERSON’s government of $LOCATION was founded...”

Fine-grained typing (PLE by Ren and Han et al. KDD’16)

“president $PERSON.POLITICIAN’s government of $LOCATION.COUNTRY was founded...”