Coarse Lexical Semantic Annotation with Supersenses

An Arabic Case Study

Nathan Schneider[P] • Behrang Mohit[Q] • Kemal Oflazer[Q] • Noah A. Smith[P]
Carnegie Mellon University, { Pittsburgh, PA / Doha, Qatar }

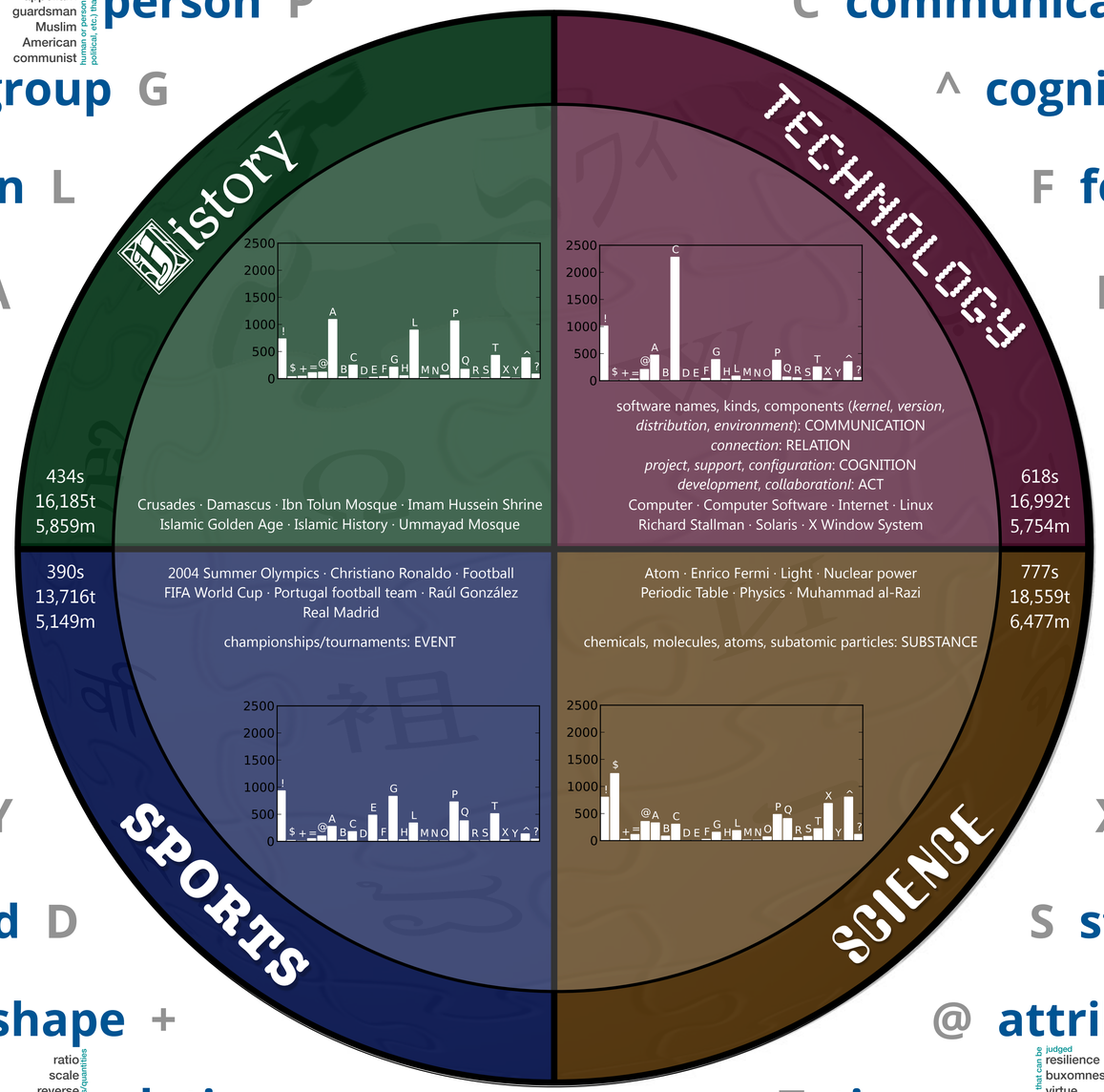
Supersense categories offer a coarse, wide-coverage set of lexical semantic classes for **things and events.** We propose that they be used as a simple first pass of semantic **annotation**.

In this work we annotate a small, multi-domain corpus of **Arabic Wikipedia** articles with nominal supersenses. We share this dataset and the tagging guidelines, which should be readily adaptable to **new languages, domains, and genres**.

	granularity	simple annotation	token coverage
ontological word senses	high	✗	mixed
named entities	low	✓	low
supersenses	**low**	✓	**high**

يعتب كتاب جينيس للأرقام القياسية أن جامعة القيروان في فاس المغرب أقدم جامعة في العالم حيث تم تأسيسها في سنة 859 ميلادي .
TIME ACT LOCATION GROUP LOCATION ARTIFACT COMMUNICATION
'*The Guinness Book of World Records* considers the University of Al-Karaouine in Fez, Morocco, established in the year 859 AD, the oldest university in the world.'

The supersense representation provides **coarse disambiguation**:
principal (PERSON vs. POSSESSION)
sponge (ARTIFACT vs. ANIMAL)
bank (NATURAL OBJECT vs. GROUP)
order (ACT, STATE, COMMUNICATION, GROUP, or ATTRIBUTE)



person P
mother Arab Roosevelt Persian deity glasscutter kibbutznik firstborn worshiper consumer appellant guardsman Muslim American communist

group G
Mennonite Church trumpet section health profession People's Party U.S. State Department University of California consulting firm Islam (= set of Muslims)
followers of social movements

location L
India Newark interior airspace Côte d'Ivoire downtown stage left New York City
GPE, or other noun functioning as a location or region
groupings of people or objects, including: organizations/institutions;

artifact A
bridge restaurant bedroom stage cabinet toaster antidote aspirin
man-made structures and objects

natural object O
cave nest barrier reef neutron star planet fishpond Mediterranean stepping stone Orion ember sky
natural feature or nonliving object in nature

substance $
krypton mocha atom aluminum hydrochloric acid liquid nitrogen sand cardboard DNA
substances and materials

body B
femur prostate hairstyle connective tissue nervous system fingernail front tooth gland insulin gene
human body parts and associated substances (diseases/symptoms)

animal N
egg cuckoo tapeworm carrier pigeon Mycrosporidia virus tentacle
non-human, non-plant life

plant Y
acorn squash Honduras mahogany genus *Lepidobotrys* Canada violet
plant or fungus

food D
eggnog Swiss cheese rutabaga shrimp cocktail cranberry sauce Guinness (= beverage)
things used as food or drink

shape +
dip hexahedron convex shape lower bound perimeter sine curve groove
2 and 3 dimensional shapes

relation =
ratio scale reverse personal relation exponential function angular position unconnectedness transitivity
relations between entities/quantities

communication C
concert hotel bill alphabet reference broadcast grave accent onomatopoeia Cree language *Book of Common Prayer*
information encoding and transmission

cognition ^
dream hypothesis biomedical science hierarchical structure muscle memory necromancy Platonism
aspects of mind/knowledge/thought/perception; techniques and abilities; fields of academic study/belief systems

feeling F
indifference wonder grudge desperation murderousness astonishment suffering
subjective emotions

motive M
urge incentive compulsion conscience mania reason
abstract compulsive influence on someone

possession H
loan money birthday present tax shelter
term associated with ownership/payment

act !
engineering (= profession) meddling malpractice faith healing dismount carnival football game acquisition
things people do or cause to happen; human professions

event E
ordeal miracle bomb blast crime wave upheaval accident tide
things that happen at a given place and time

process R
oscillation distillation overheating aging accretion/growth extinction evaporation
a sustained phenomenon or one marked by a gradual succession of states

phenomenon X
suction tailwind tornado cold front repercussion effect sound pressure gravity electricity
a physical force or something that happens/occurs

state S
symptom reprieve potency poverty altitude sickness tumor fever bankruptcy infamy opulence
stable states of affairs; diseases and symptoms

attribute @
resilience buxomness virtue gregariousness naughtiness inconclusiveness honesty admissibility valence sophistication temperature
characteristics of objects/people that can be judged

time T
10 seconds Eastern Time 2nd millennium BC velocity frequency runtime middle age basketball season words per minute curfew industrial revolution
a temporal point, period, amount, or measurement

quantity Q
7 cm 1.8 million 12 percent/12% volume (= spatial extent) real number square root
digit 90 degrees handful ounce volt half
quantities and units of measure, including cardinal numbers and fractional amounts

Common Confusions

COMMUNICATION vs. COGNITION
system نظام

ACT vs. COGNITION
measures (~ strategies, policies) الاجراءات

ACT vs. PROCESS
intensification تكثيف

ARTIFACT vs. COMMUNICATION
book كتاب

Boundaries
Fez, Morocco (1 or 2 LOCATIONs?)
World Cup Football Championship (single EVENT, or *Football* as an ACT?)

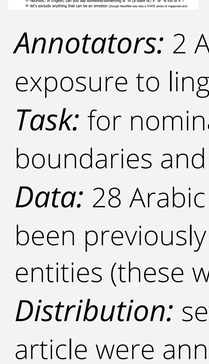
Inter-annotator Agreement

Mentions: 70% F_1
In-mention tokens: 71% rate, 83% F_1, Cohen's κ = .69

Training

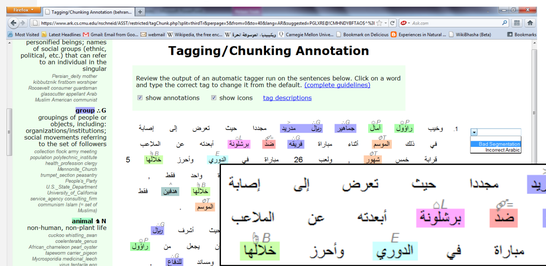
iterative annotation of 4 pilot articles and guideline refinement in consultation with annotators until their agreement reached an F_1 of 75%

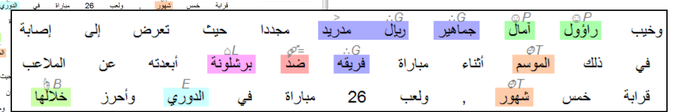
The tagging guidelines developed over the course of annotation are fairly simple, at under 6 pages. They include a 43–rule decision list explaining how to classify different classes of things (e.g., buildings, drugs, geopolitical entities).



Annotation

Annotators: 2 Arabic native speakers, no prior exposure to linguistic annotation
Task: for nominal expressions, mark their boundaries and supersense tags
Data: 28 Arabic Wikipedia articles that had been previously annotated with named entities (these were shown as defaults)
Distribution: several sentences from each article were annotated by both annotators





www.ark.cs.cmu.edu/ArabicSST

Future work

Verb supersenses
Supersense annotation in other languages
Building an Arabic supersense tagger

ACL 2012
Jeju, South Korea

Qatar National Research Fund
NPRP-08-485-1-083

Carnegie Mellon
ARK lti
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar