Abstract

- Automatically build acronym dictionary
  - Rank by context match
  - Include acronyms unaccompanied by their expansions
- Improve acronym disambiguation
- Statistically-based linguistic insights

Why We Care

- Acronyms affect NLP applications (search, machine translation, ...)
- Hand-crafted dictionaries incomplete and require constant updating.

Previous Work

- Prior acronym dictionary-building techniques rely on local acronyms (acronyms adjacent to their expansions, often in parentheses).
  - “The CIA (Central Intelligence Agency) released its budget.”
  - “She works at the Culinary Institute of America (CIA).”
  - “Alumni of the Cleveland Institute of Art support the CIA.”
- Only computational work on Hebrew acronyms: HaCohen-Kerner [30,11,12]
  - Disambiguation of Hebrew/Aramaic acronyms in Jewish law domain.
  - Assumes a pre-existing, hand-crafted acronym dictionary.

Hebrew Acronyms

- In Hebrew corpus, acronyms are 1% of word tokens and 3% of types.
- More common in news and encyclopedia genres than in literature.

A never-ending story for unique acronyms:
new acronyms continue to be found as more text is read

Acronym Disambiguation (Extrinsic Eval)

- Given 200 acronyms and their contexts, how many of the correct expansions are in the top r dictionary results for the acronyms?

Future Work

- Try specialized Hebrew genres/domains: military, Jewish legal texts.
- Apply to other languages:
  - Hebrew advantages: Easy acronym identification, widespread acronym use.
  - Hebrew disadvantages: Complex morphology/orthography, poor NLP resources.
- Additional applications: search, machine translation.
- Exploit for multi-word expressions (MWEs) and named entity recognition (NER).