Hebrew Acronyms: Identification, Expansion, and Disambiguation

Kayla Jacobs
Computer Science Dept.
Technion
kayla@cs.technion.ac.il

Alon Itai
Computer Science Dept.
Technion
itai@cs.technion.ac.il

Shuly Wintner
Computer Science Dept.
University of Haifa
shuly@cs.technion.ac.il

Acronyms are words formed from the initial letters of a phrase, called its expansion. For example, CIA is a well-known acronym for *Central Intelligence Agency*, though in other contexts it could mean *Culinary Institute of America* or *Cleveland Institute of Art*. Understanding acronyms is important for many natural language processing applications, including search and machine translation.

The collection of acronyms is an open set, with new acronyms constantly being added for company and organization names, technical terms, etc. Thus, while hand-crafted acronym dictionaries exist, they are far from complete and require frequent updates. Most automatic techniques for building acronym dictionaries rely on local acronyms—those that appear in the same documents as their expansions, often adjacent to each other in text and typically in parentheses. For example, CIA is a local acronym, with different expansions, in each of the following sentences:

- "The Central Intelligence Agency (CIA) released its budget."
- "She's applying to the CIA (Culinary Institute of America)."
- "After graduating from the Cleveland Institute of Art, I'm a proud CIA alumnus."

These automatic methods miss non-local acronyms, written without their expansions with the (frequently incorrect) assumption that the reader can easily understand the acronyms' intended meanings.

We developed a new machine learning method to automatically build an acronym dictionary from unstructured text documents. To our knowledge, this is the first such technique that specifically includes non-local expansions that do not necessarily appear even in the same documents as their acronyms, let alone adjacent to them. We applied the technique to a large (80 million tokens) corpus of Modern Hebrew texts, and using easily-calculated linguistically-motivated features we trained a classifier to identify acronym-expansion pairings, achieving an F-score of 82%.

We also enhanced the dictionary with contextual information based on Latent Dirichlet Allocation (LDA) topic modeling to help disambiguate acronyms—determining which of the (possibly multiple) dictionary expansions is most appropriate for the specific context. As a way of extrinsically evaluating our dictionary's quality, we compared its performance on the acronym disambiguation task to that of dictionaries built with the leading two previous methods. Our dictionary performed significantly better than acronym dictionaries constructed both manually and using the leading automatic technique (using local acronyms involving parentheses).

Lastly, while acronyms have a long history in Hebrew, and have previously been investigated from a linguistic perspective, they have never before been studied quantitatively. We discovered new statistically-based linguistic insights about acronym usage in Modern Hebrew texts, of interest to Hebrew language aficionados and developers of Hebrew natural language processing systems.