



3721 Executive Center Drive
Suite 100
Austin, TX 78731-1615
www.cyc.com
(512) 342-4000
info@cyc.com

Cyc Enhancement of Information Extraction

The Challenge

Information extraction (IE) systems are typically fairly limited in the set of entity types they can recognize. In addition, because they have very limited, if any, models of these entities, IE systems usually have no mechanisms for detecting errors in the resulting information extraction. Finally, the merging of references to the same actual entity, either within a single document processed by multiple IE systems or across multiple documents, generally is not adequately addressed.

Cyc's Solution

Cyc's unique and extensive common-sense knowledge base and inference capabilities provide the key to a number of different methods for improving and enhancing the results of existing information extraction systems. In particular, Cyc can support:

- Strengthening of type identification;
- Detection of certain information extraction errors;
- Resolution of references to specific individuals; and
- Unification of references to the same entity.

Stronger Type Identification

A key advantage that Cyc provides in the information extraction task is the rich hierarchy of types (including approximately 30,000 types of entities and events) that can be used as targets for extraction. Using its substantial English lexicon associated with many of these types, Cyc can be used to determine more specific types for many entities extracted by IE systems.

For example, in addition to a class such as "*Person*" available to current information extraction systems, Cyc's extended type system includes many categories of people, including things like `Terrorist`. Thus, if one were searching for documents that refer to both "*terrorist*" and the "*NYSE*", the search would be far more efficient if appropriate entities were tagged as "`Terrorist`" rather than the much larger set that would be tagged as "*Person*".

Below are a few examples from an article about Turkish terrorists. The article was run through two different COTS Information Extraction systems, and the results were merged into one OWL file before they were considered for enhancement by Cyc.

Consider the following tagging resulting from the current IE systems:

Person: "a spice dealer who carried out the bomb attack against HSBC's head office in Levent"

While Cyc will not always be able to determine that the entity referred to in this text is of type `Terrorist`, it will generally be able to strengthen the type of this entity to:

A vendor who sells spices

which Cyc represents as:

```
(SubcollectionOfWithRelationToFn
  Vendor
  sellsProductType
  Spice)
```

Because this tag is compatible with the initial tag *Person*, Cyc can simply add this as another type for that entity. Based on Cyc's understanding of what a `Vendor` is, one could then retrieve a document containing this entity for queries about entities who are, among other things, people, merchants, vendors, and spice vendors.

Another example might be a preliminary tagging such as:

Person: "leader of the terrorist cell"

Rather than the simple *Person* tagging, Cyc could identify this as:

Intelligent Agent that leads a terrorist cell

represented as:

```
(SubcollectionOfWithRelationFromTypeFn
  IntelligentAgent
  hasLeaders
  TerroristCell)
```

As in the previous example, this is not incompatible with the *Person* tag, and so it could be added as another entity-type for this entity. Further, for any given domain, rules of thumb could be added to Cyc that would allow it to determine that this person is also a terrorist. In this case, the rule would be that leaders of Terrorist Cells are themselves terrorists; or more generally, if someone is a leader of a group, and members of the group are of a particular type, the leader is also of that type. (Thus a leader of a monastery is a monk, and a leader of a company is an employee of the company, etc.)

Finally, for a tagging such as:

PoliceAct: "was arrested"
victim: "Fevzi Yitir"

Cyc can determine that the *PoliceAct* is not just a *PoliceAct*, but is in fact an arrest (`ArrestingSomeone`). From this, Cyc can conclude that the role *Victim* corresponds to the more specific role of `arrestee`. This would allow a user to ask questions like "Which people have been arrested?"

which could not have been answered (or even asked) using the simpler types.

Error Detection

Cyc has the ability to detect errors and suggest possible corrections in existing extraction results.

Consider the following erroneous tagging:

Person: "Near East"

Cyc knows that the string *"Near East"* corresponds to its concept `MiddleEast-Region` and knows that this concept does not correspond to a person or group of persons. From this, it could flag the initial extraction as problematic. Furthermore, it might suggest a different meaning, namely the *"Middle East region"*. In general, one would need to be cautious in making such corrections automatically because Cyc's knowledge, though substantial, is nowhere near complete. For example, though Cyc may know about a person named *"Bob Smith"*, if an extraction engine types a reference to *"Bob Smith"* as an organization, it could be that the text refers to *"Bob Smith, Inc."*, an organization not yet known to Cyc. In certain circumstances, it would be possible to define heuristics to determine whether to make such changes automatically, identify potential changes for human approval, or simply identify such taggings as possibly erroneous.

In addition to these sorts of error detection, Cyc can, with its rich representation of predicate and event-roles, represent and take advantage of facts such as "Only people can be arrested" to correct extraction errors. Thus, if the article had typed "Fevzi Yitir" (the arrestee in the section on Stronger Typing) as anything other than a person, Cyc would have been capable of flagging that tag as an error.

Resolution to Individuals

In addition to containing information about a wide variety of types, Cyc also contains a substantial number of individuals in its knowledge base (e.g., approximately 9,000 specific people, as well as thousands of companies and other individual entities). Thus, it can often type particular entities all the way down to the individuals involved. In order to expand the reach of the system, Cyc can be used as a repository for more individuals found by information extraction systems, with such information augmented by hand if desired (e.g., by adding aliases for some of the individuals) using tools that Cycorp has already developed. Additional existing information entry tools can be tailored to meet the particular needs of a given user base and task. Cyc can also interface to existing databases containing specific entities and can use these entries as if they were part of its knowledge base.

For example, where the extraction engine correctly identified:

Municipality: "Mardin"

Cyc knows that this corresponds to the `Mardin` entity in its knowledge base

and it knows, among other things, that it is a province in Turkey. Thus, by connecting this reference to a specific geographic entity, Cyc can utilize any additional information it has about that entity and, combined with its geospatial inference capabilities, correctly respond to queries such as “What incidents occurred in Turkey?”.

Result Unification

The ability to resolve to individuals provides Cyc with the ability to identify the same entity in multiple sources, thereby allowing the system to merge results across multiple documents and across multiple extraction systems.

Consider the following results from the two separate extraction engines:

Country: “israel”, “ISRAEL”

Country: “ISRAEL”

Because Cyc can map each of these references to its concept `Israel`, it can unify the reference both within the first set of results and across the results from the two extraction tools. This is advantageous for several reasons:

- Users of the results now have fewer entities to track;
- Users have access to other information that is present in Cyc about `Israel`. For instance, Cyc knows that `Israel` can also be referred to as “Yisrael”, “Zion”, “ISR”, “IL” (in internet addresses), and “IS”¹; and
- Because there is a canonical representation for Israel, it can be tracked across multiple documents as a single entity.

Conclusions

Cyc has a rich knowledge base of concepts and individual instances and a powerful set of inference mechanisms. These capabilities can be useful to augment the tagging capabilities of current state-of-the-art information extraction tools. Though complete automation of accurate and detailed text tagging is not yet within reach, the quality of automated results, and support for manual validation of tagging results, can increase substantially by augmenting traditional text extraction with the Cyc knowledge base and reasoning capabilities.

¹ There is a substantial amount of additional information in Cyc about many of its entities, such as that Israel borders Syria, Jordan, Lebanon, and Egypt. While we believe this information should be used by systems taking advantage of the richer types offered by Cyc, further work would need to be done to determine how to use this information within the context of existing systems.