Named Entity Recognition

Entity Lists, Part of Speech Patterns, Regular Expressions, and Machine Learning Models
Introduction

Named Entity Recognition is the process of identifying people, places and things within a piece of text. Natural Language Processing (NLP) systems combine simple entity lists and complex machine learning models to automate this process. A human would take hours to extract entities from 10,000 tweets; a computer can do it in seconds.

Named Entity Recognition is also commonly referred to as Entity Extraction or Entity Recognition, or simply abbreviated to NER.

TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>2 - 4</td>
</tr>
<tr>
<td>Entity Lists</td>
<td>5</td>
</tr>
<tr>
<td>Part of Speech Patterns</td>
<td>6 - 7</td>
</tr>
<tr>
<td>Regular Expressions</td>
<td>8 - 9</td>
</tr>
<tr>
<td>Machine Learning Models</td>
<td>10 - 11</td>
</tr>
<tr>
<td>Summary</td>
<td>12</td>
</tr>
<tr>
<td>About Lexalytics</td>
<td>13</td>
</tr>
</tbody>
</table>
Entity Recognition is useful on two levels. The first is for simply understanding who and what is mentioned within a document. For example, you use high-level entity recognition to quickly see which of your company’s products get the most mentions in a collection of tweets or reviews.

The second level of entity recognition goes deeper. Once we’ve identified the entities within a piece of text, we use other NLP models to evaluate the sentiment, themes and topics associated with them. This approach reveals not just who and what is mentioned within a text document, but also the words people use to describe those subjects and how they feel.

Some entity recognition systems are limited to people, companies and some products. Lexalytics automatically recognizes the following entity types:
In addition, our plug-and-play Industry Packs come with pre-built entity lists. Our Hotels pack, for example, includes lists of hospitality companies, loyalty programs, and booking and review sites. You can also use our configuration tools to define custom entities: particular diseases, specific models within your product line, and more. If it’s an entity to you, it’s an entity to us.

Lexalytics also distinguishes proper nouns from generic nouns, which usually represent larger concepts. So “Bill Gates,” “Niagara Falls,” and “iPhone” all classify as entities, while “leader,” “nature,” and “technology” qualify as themes (see our white paper on Theme Extraction and Context Analysis for more about themes).

The rest of this white paper will review the four tools Lexalytics uses for entity recognition: entity lists, part of speech patterns, regular expressions, and machine learning models.
ENTITY LISTS

Entity lists are just that: simple lists of people, places and things, like car manufacturers, CEOs, or diseases. Once you’ve built your list, the entity recognition system looks for exact matches in your documents.

Lists work best when you have specific entities to look for. A media analyst for a sports franchise, for example, would add each of their team’s players to an entity list to look for mentions of them in the news. Similarly, Microsoft’s Customer Market Research team uses entity lists (among other techniques) to track more than 1,000 brands and products across social media using Lexalytics’ Semantria Storage & Visualization product.

Entity lists are also useful when used in concert with each other. A Medical Affairs Liaison at a pharmaceutical company might add her firm’s products to one entity list and some Key Opinion Leaders (KOLs) to another list. Then she can use these lists together to make sure that KOLs aren’t violating FDA regulations in the language they use to talk about her firm’s drugs and therapies.

Entity lists are easy to understand and simple to set up. Unfortunately, this also limits their applicability. Long entity lists take tedious labor to set up and require constant maintenance. For example, every time your company launches a new product or service, you’ll need to add its name and all possible variants of that name into the appropriate entity list.

What’s more, list-based entity extraction only recognizes exact phrase matches. Any tangential references, including pronouns, are overlooked because the system doesn’t know any better.
PART OF SPEECH PATTERNS

One solution to the problem of tangential references lies in Part of Speech Patterns.

Part of Speech tagging is a core text analytics function that determines whether something represents a noun, a verb, an adjective, or something else entirely. Among other things, Part of Speech tagging creates a useful foundation for entity recognition because any phrase that involves a noun (a “noun phrase”) is statistically quite likely to represent an entity.

Being able to identify the Parts of Speech of a given n-gram tells us a lot about the tokens involved. In an adjective-noun combination, for example, the noun is almost certainly an entity, and the adjective describes it (often in terms that represent a theme and carry sentiment).

Amazing phone | Broken TV

**Adjective-Noun**

Similarly, noun phrases that form subject-verb-object triplets can be assumed to represent an entity acting on another entity.

Apple released the iPhone

**Subject (noun) – Verb – Object (noun)**
Things get more complicated when we bring pronouns into the mix. Imagine a customer review complaining about the delivery of their new television.

The mailman must’ve dropped the box or something. They left me broken headphones!

An entity recognition system based on an entity list that contains “mailman” will identify the entity in the first sentence but completely overlook the second. A Part of Speech-based entity recognition system, on the other hand, will recognize that the first pronoun in the second sentence (“they”) refers to the noun in the previous phrase (“mailman”).

From there, the system will know to associate the other verbs, adjectives and nouns with that base subject. This deeper understanding adds more layers of context and levels of insight to the output. In this example, the negative sentiment of “broken” will result in a theme, which is itself associated with the entity of “headphones,” and which influences the sentiment score of the entity “mailman.”
REGULAR EXPRESSIONS

A Regular Expression is essentially a search term for a pattern. Regular expressions are useful for defining atypical entities that follow particular forms. Entering all possible phone numbers into an entity list, for example, would be impossible. Nor is it realistic to try to identify them by Part of Speech patterns. But a phone number is usually written in just a few ways:

```
(###) ###-####
#######
### ### ####
```

Adding these patterns as regular expressions allows you to quickly and effectively extract phone numbers as named entities.

Entity recognition of @mentions is another area where regular expressions come in handy. On Twitter, we can safely assume that an @ symbol followed by a string of characters represents a Twitter account. Therefore, that @-string represents an entity. Adding every single @ handle into an entity list is impossible. But searching for the @ symbol itself will return lots of irrelevant content. Instead, a regular expression can automatically pick up all @mentions that represent entities, without requiring much labor from your end.

Similarly, a trending hashtag takes on a life of its own on social media and becomes an entity itself in the eyes of a social media analyst. But you’d need extrasensory perception to know to add a hashtag to an entity list before it starts trending. A regular expression, on the other hand, will already be on the lookout and can catch and analyze associated content even before you’re aware of its relevance.
Finally, regular expressions are also useful for entity recognition of street addresses. Adding dozens of addresses and all their possible variants to an entity list would be daunting and require constant upkeep. But addresses usually follow a pattern. In the United States, for example, they're generally written as Number-[Adjective/Noun]-[St/Ave/Blvd/Rd/Ln].

There’s always some variation, but address forms are generally standardized by country. Adding your region's address patterns as regular expressions requires far less work than an entity list and is more efficient than trying to train a machine learning model to identify them.

### Address Forms by Country

<table>
<thead>
<tr>
<th>Country</th>
<th>Address Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNITED STATES</td>
<td>24 Beacon St, Boston, MA 02133</td>
</tr>
<tr>
<td>FRANCE</td>
<td>82 boulevard de Clichy 75018 Paris</td>
</tr>
<tr>
<td>ARGENTINA</td>
<td>Piedras 623, Piso 2, depto 4 C1070AAM, Capital Federa</td>
</tr>
</tbody>
</table>
MACHINE LEARNING MODELS

Lists, part of speech patterns and regular expressions are good at identifying named entities that follow predictable forms or patterns. But all of these methods can stumble when faced with irregular entities, ambiguous words and tangential references. Machine learning models are often the perfect solution to this problem.

Lexalytics ships all of our solutions with pre-trained entity recognition models that automatically recognize six entity types:

Person – Place – Company – Product – Job – Title

To train and maintain these models, we use semi-supervised learning: first we feed the model an annotated data set and give it feedback until it matches our precision and accuracy requirements. Then we use that model to recognize patterns within another set of unmarked data.

Our semi-supervised approach goes above and beyond traditional methods. By looking for patterns in the data, we’re able to draw connections and understand complex semantic relationships that dramatically improve the accuracy and precision of our entity recognition models.
For example, our models learned that the phrase “works for” often precedes an entity, such as the name of a company. So, when the models find the phrase “works for” (or a variant of it) in front of a proper noun, it recognizes that the proper noun in question is the name of a company. On a similar note, part of speech-based pattern recognition helps the model identify pronouns and understand which entities they refer to:

Dennis Muilenburg worked for Boeing as CEO until he was fired.


Lexalytics also uses machine learning models to reliably recognize ambiguous company names. The word “Apple,” for example, can represent:

<table>
<thead>
<tr>
<th>A type of food and the tree it comes from</th>
</tr>
</thead>
<tbody>
<tr>
<td>A technology company and its executives</td>
</tr>
<tr>
<td>A record label founded by The Beatles</td>
</tr>
</tbody>
</table>

Entity lists and regular expressions won’t be able to correctly identify when “Apple” actually represents a named entity, let alone which type of entity it is. Lexalytics trains custom machine learning micromodels to solve data-related entity recognition problems like this.

To learn more, read our data sheet, Tuning and Machine Learning Services, or see our white paper, Machine Learning for Natural Language Processing.
SUMMARY

Named Entity Recognition (also known as Named Entity Extraction) is the process of identifying people, places and things within a piece of text. No single entity extraction method can solve every one of your needs. That’s why Lexalytics combines four natural language processing tools for entity recognition: entity lists, part of speech patterns, regular expressions, and machine learning models.

If it’s an entity to you, it’s an entity to us.

Our hybrid approach delivers accurate, reliable entity extraction that is completely customizable to meet your needs. Whether you’re adding a simple list of specific named entities or combining regular expressions with machine learning models to define entirely new types of entities unique to your line of business, Lexalytics supports you.

Once we’ve identified all of the named entities within your text documents, we go the extra mile by performing multi-layered sentiment analysis on those entities and showing all of the topics and themes associated with them. We help you understand not just who and what is being discussed, but also the words people use to describe those subjects, and how they feel.

Contact us today to discuss how Lexalytics can meet your entity recognition needs.
Lexalytics processes BILLIONS of text documents every day, GLOBALLY.

We help our clients unlock the full value of their text data through innovative text analytics solutions delivered through the modular Lexalytics Intelligence Platform® and “semi-custom” enterprise applications.

Our systems combine natural language processing, semi-structured data parsing, and machine learning to reveal context-rich patterns and insights within comments, reviews, contracts, medical files and other complex text documents.

Enterprises and data analysts across a range of industries rely on Lexalytics to improve customer experiences, reduce employee turnover, manage regulatory compliance, improve process automation, and more by gaining deeper insights and greater value from their data.

For more information, visit www.lexalytics.com or call 1-800-377-8036