WHITEPAPER

Text Analytics – Named Entity Extraction

Executive Summary

Text Analytics is all about knowing who's talking, what they're saying and how they feel about it. In the last few years understanding what's being said and how people feel about it has gotten a lot of press, but if you don't know who's talking then knowing what they're saying probably isn't going to gain you all that much insight.

Named Entity Extraction is one of the oldest and best understood of the Text Analytics technologies, it's been around in one form or another since the late 1990s and is one of the core features of processing unstructured content. There are a bunch of different ways to do entity extraction and this paper examines a few of those including the techniques that are favored in the Lexalytics Salience engine.

It's also important to extract entities so that you can then associate other things with the entity – such as sentiment or important concepts and themes.

Step 1: What is an Entity?

One of the trickiest problems in the text analytics field is getting clear concise definitions of the features that various text engines provide, so to avoid any confusion let's start with a clean and simple definition of what we mean when we say "Entity Extraction" or "Named Entity Extraction".

A named entity for the purposes of this paper is simply a proper noun that fits into a commonly understood type such as a "Person" or "Company" or "Product". Using this definition, examples such as "iPhone" "Tiger Woods" or "Microsoft" would all be types of entities, while words like "solar power" or "baseball" would not be entities because they are generic nouns and not proper nouns. This definition allows us to define the process of extracting entities from unstructured text like press releases or even from short form content like Tweets.

It is possible to consider other things to be entities – URLs, addresses, phone numbers and the like are also perfectly reasonable entities even if they don't fit the traditional definition of a "proper noun".

Step 2: How does it work?

There are a variety of techniques for extracting named entities but most of them begin with the step of understanding the basic parts of speech of the text being processed. This technique is typically called Part of Speech tagging or PoS tagging for short. It is a well understood discipline within text analytics systems and depending on the technique selected can be accurate between 95% and 98%. Lexalytics uses a variant of the most common of these taggers, "the Brill tagger". An example of the PoS tags derived from the PoS tagger follows:

Original Text:

Nautilus Footwear is different. Nautilus paved the way over the last few years in designing ergonomically designed footwear to protect workers in more ways than ever before. They have been an industry icon on ergonomics and have validated their claims with independent test results on shock absorption that clearly demonstrate the ergonomic features of the Nautilus products vs. competitors.

POS Tagged Text:

Nautilus Footwear is different. Nautilus paved the way over the last few years in designing ergonomically designed footwear to protect workers in more ways than ever before. They have been an industry icon on ergonomics and have validated their claims with independent test results on shock absorption that clearly demonstrate the ergonomic features of the Nautilus products vs. competitors.

The key items in this markup are the items in blue which represent the nouns and proper nouns. Remembering our definition from earlier, proper nouns are the types that are candidates to be named entities.

Now that we understand where the proper nouns are, how do we go about identifying the entities and their types? In the example above the type is clearly "Company". It turns out there are a variety of different ways to identify entities in text, including:

- Simple lists → Famous people, company names, etc…
- Regular Expressions → Regex is good for phone numbers, zip codes, etc…
- CRF Models → Trained Models for generic company, people, place, etc…

For the purposes of this paper we will focus our attention on CRF models as they form the core of Lexalytics advanced entity recognition.

What is a CRF model?

Put simply, it's a whole bunch of human expertise rolled into a database. A human with decent grammar skills consumes a lot of content (think news stories, product reviews, tweets, etc…) and hand marks the location of the entities (let's stick with companies in this case) in each document. The model is then trained with this human tagged content so that it can learn the patterns that companies follow in text. For example, the phrase "works for" often precedes the name of a company, so that knowledge is built into the model. When the phrase "works for" appears before a proper noun there is a higher likelihood that this entity is a company. Given enough of these clues the model will make its best guess about what type of entity each proper noun represents.

To build a viable CRF model for a generic type of entity like people or companies you need 3,000 to 5,000 hand tagged sample documents of various types (the more the better). A model built on a good volume of training data will have relatively high accuracy (measured via an F1 score). This paper is not going to dig into the detail of computing an F1 score, but simply stated, F1 is the standard measure for accuracy in things like search results and entity recognition. For the Salience Engine's 3 primary entity types of People, Companies and Places we have F1 scores between 84 and 88 (scale of 1 to 100) which are very good but certainly not perfect.

Step 3: Hybrid Models

Good but not perfect is currently the state of the art for entity recognizers, but many customers require almost perfect entity extraction for their application. To assist in this problem we bring in lists and rules to augment the CRF model. If, for example, you were processing content about an election cycle and politicians, then you'd want to be very good at picking up the politicians in all processed content. Since the list of elected officials is a finite and not particularly large list you can build a list or "CDL" file containing their name which will guarantee that they are reported as a person regardless of the score they receive in the CRF model. This hybrid approach gives users significant control over the accuracy of discovered entities. Figure 1 below shows how all these pieces work together to return named entities.

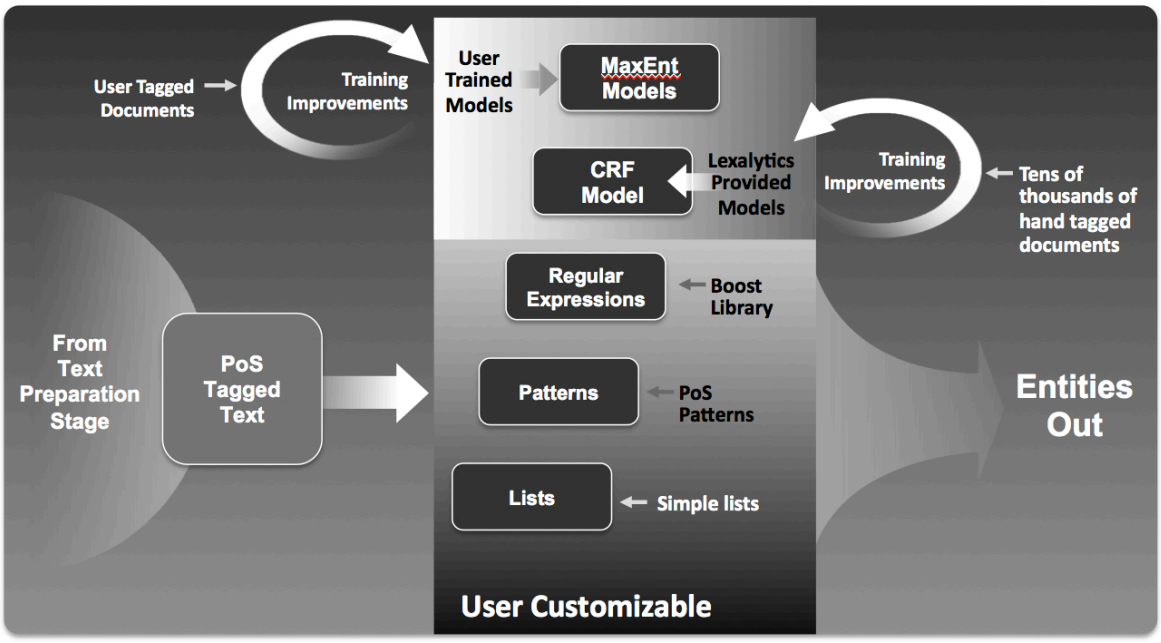

Figure 1 - Various Methods of Entity Extraction

Step 4: User Defined Entity Extraction

Building simple lists and patterns is a good way to extract simple entity types like zip codes, or to enhance our existing recognizers like people or companies, but what if you have a complex type of entity that is not currently supported. If for example you were processing medical content and needed to recognize disease names, you'd be in a tough spot if all you had was user defined lists or regular expressions.

Fortunately Lexalytics provides a toolset to build complex entity recognizers using a different type of statistical technique called a Maximum Entropy model. The tool allows you to import and mark up your own training sets to create new entity types like "Disease" or "Legal Term". Like the CRF model, the Entity Management Toolkit (EMT) is not for the faint of heart, as you'll need to provide hundreds or possibly thousands of examples to build out your custom recognizer.

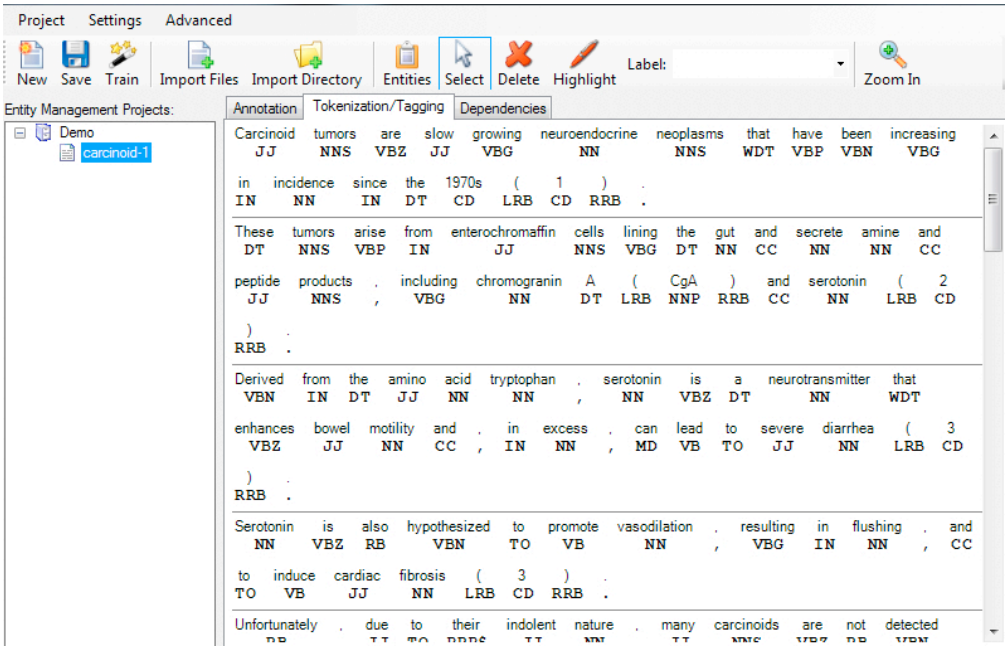

Figure 2 - Entity Management Toolkit

All of these techniques combined give customers the flexibility to extract any type of entity using a range of tools – from lists of companies to highly sophisticated statistical models based on Part of Speech patterns. From there, the rest of the Salience Engine can associate important information about the entity, like sentiment, summaries, themes – so that you can tell what's happening with that entity across millions of items of text.

About Angoss Software

As a global leader in predictive analytics, Angoss helps businesses increase sales and profitability, and reduce risk. Angoss helps businesses discover valuable insight and intelligence from their data while providing clear and detailed recommendations on the best and most profitable opportunities to pursue to improve sales, marketing and risk performance.

Our suite of desktop, client-server and big data analytics software products and Cloud solutions make predictive analytics accessible and easy to use for technical and business users. Many of the world's leading organizations use Angoss software products and solutions to grow revenue, increase sales productivity and improve marketing effectiveness while reducing risk an cost.

About Lexalytics, Inc.

Lexalytics, Inc. is a software and services company specializing in text and sentiment analysis for social media monitoring, reputation management and entity-level text and sentiment analysis. By enabling organizations to make sense of the vast content repositories on sources like Twitter, blogs, forums, web sites and in-house documents, Lexalytics provides the context necessary for informed critical business decisions. Serving a range of Fortune 500 companies across a wide spectrum, Lexalytics partners with industry leaders such as Endeca, ThomsonReuters, Radian 6 and TripAdvisor to deliver the most effective sentiment and text analysis solutions in the industry.

Corporate Headquarters

111 George Street, Suite 200
Toronto, Ontario M5A 2N4
Canada
Tel: 416-593-1122
Fax: 416-593-5077

European Headquarters

Surrey Technology Centre
40 Occam Road
The Surrey Research Park
Guildford, Surrey GU2 7YG
Tel: +44 (0) 1483-685-770

www.angoss.com