Dealing with Context in Text Mining

Whitepaper
The ultimate goal of any text analytics tool is to answer three critical questions:

- Who is being discussed?
- What is the context of that conversation?
- What is the tone or sentiment of that conversation?

That second question is vital for attaining actionable insights into the conversations contained in text, and Lexalytics offers the best context determination.

The basis for our context extraction is the noun. Nouns are the best part of speech for determining context — more specifically, our software looks for the nouns that aren’t identified as entities. While entity extraction deals with proper nouns, context determination is based around the more general, non-proper nouns. So “Cessna” and “airplane” get classified as entities, but “transportation” is considered a theme.

Here’s why context is important:

“Governor Smith’s hard-line stance on transportation cost him votes in the election.”

Picking out entity sentiment alone from the sentence above indicates that “Governor Smith” is associated with negative sentiment. That’s all well and good, but why is the Governor being mentioned negatively? Remember, in a real-world scenario you will be processing thousands of tweets similar to the above, not reading each individually. Entity sentiment by itself shows you the general tone of a document, but it can only go so far.

Once you look at the themes in the sentence, however, things become a lot more clear. There are two themes present: “hard-line stance” and “budget cuts”, both of which have negative sentiment. The ability to reliably pick out abstract themes and concepts in addition to entities, and assign accurate sentiment to all, is another reason why Lexalytics text analytics are the best in the business.

For more on entities and sentiment analysis, stop by our website. This whitepaper introduces and explains the methods Lexalytics uses for determining context.
Lexalytics supports four methods of context determination, each with its merits and disadvantages, for optimizing to your needs. These four techniques are:

- N-grams
- Noun phrase extraction
- Themes
- Facets

We’ll start with the first and work our way down.

N-grams: Simple Detail

N-grams are combinations of one or more words that represent entities, phrases, concepts, and themes that appear in text. Generally speaking, the lower the value of “n”, the more general the phrase or entity. N-grams form the basis of many text analytics functions, including our most powerful context-determination method: themes. We’ll discuss themes later, but first it’s important to understand what an n-gram is and what it represents.

There are three common levels of n-gram, referred to as:

1 word = monogram
2 words = bi-gram
3 words = tri-gram

Monograms vs. bi-grams vs. tri-grams

To get an idea of the relative strengths and weaknesses of monograms, bi-grams, and tri-grams, we’re going to run a few items through the system:

Consider these phrases: “crazy good” and “stone cold crazy”
And this sentence: “President Barack Obama did a great job with that awful oil spill.”

<table>
<thead>
<tr>
<th>Monograms</th>
<th>Bi-grams</th>
<th>Tri-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrases</td>
<td>Extracted (crazy good, stone cold crazy)</td>
<td>crazy (2)</td>
</tr>
<tr>
<td>Phrases</td>
<td>Extracted (President Obama)</td>
<td>a awful</td>
</tr>
</tbody>
</table>

The monograms (single words) aren’t specific enough to offer any value. In fact, monograms are rarely used for phrase extraction and context; they provide other value instead, often as entities and themes.

Tri-grams aren’t much use either, for the opposite reason: they’re too specific. Specificity can be valuable, if you happen to be looking for very particular phrases, but tri-grams are usually too narrow a lens to look through. Tri-grams do get used for phrase extraction, but not as frequently as bi-grams.
If n-grams are bowls of porridge, then bi-grams are the “just right” option. As demonstrated above, two words is the perfect number for capturing the key phrases and themes that provide context for entity sentiment.

However, n-grams risk gathering a lot “noise”: useless phrases such as “with that” offer no value in context determination, and do little more than clutter your view.

Stop That Word!

Left alone, the n-gram algorithm will grab any and every n-gram it finds. To avoid unwanted entities and phrases, our n-gram extraction includes a filter system.

Stop words are a list of terms you want to exclude from analysis. Classic stop words are “a”, “an”, “the”, “of”, and “for”. In addition to these very common examples, each domain has a set of words that are statistically too common to be interesting.

Take the phrase “cold stone creamery” as an example. Most stop lists would let each of these words through unless directed otherwise. Remember that if you stopped the phrase “cold stone creamery”, your results would look very different than if you stopped “cold”, “stone”, and “creamery” individually.

To explain:
- If you stop “cold stone creamery”, the phrase “cold as a fish” will make it through and be decomposed into n-grams as appropriate.
- If you stop “cold”, “stone”, and “creamery”, the phrase “cold as a fish” will be chopped down to just “fish” (as most stop lists will include the words “as” and “a” in them).

N-gram stop words generally stop entire phrases in which they appear. For example, the phrase “for example” would be stopped if the word “for” was in the stop list (which it generally would be).

**Advantages of N-grams**
- Offers thematic insight at different levels (mono, bi-, tri-grams)
- Simple, easy to conceptually understand

**Drawbacks to N-grams**
- Indiscriminate: requires a long list of stop words to avoid useless results
- Simple count does not necessarily give an indication of “importance” to text or of its importance to an entity
A form of n-gram that takes center stage in the context determination process is the noun phrase. Noun phrases are just part-of-speech patterns that include a noun. They can include whatever other parts of speech make grammatical sense, and can include multiple nouns. Some common noun phrase patterns are:

- Noun
- Noun-Noun...... -Noun
- Adjective(s)-Noun
- Verb-(Adjectives-)Noun

So “red dog” is an adjective-noun part of speech pattern, and “cavorting green elk” is a verb-adjective-noun pattern.

Other part-of-speech patterns include verb phrases (e.g. “Run down to the store for some milk”) and adjective phrases (such as “brilliant emerald”). Verb and adjective phrases serve well in determining sentiment, but nouns are generally the most useful in understanding the context of a conversation. If you want to know “what” is being discussed, nouns are the go-to. Verbs help with understanding what those nouns are doing to each other, but in most cases it is just as effective to only consider noun phrases.

Noun phrase extraction takes part-of-speech type into account when determining relevance. Many stop words are stopped simply because they are a part of speech that is uninteresting from a standpoint of understanding meaning. Because you’re being very specific about
classes of words that are interesting, most common stop words are instantly eliminated automatically. Stop lists can also be used with noun phrases, but it’s not quite as critical to use them as it is with n-grams.

Consider this article:

Yahoo wants to make its Web e-mail service a place you never want to -- or more importantly -- have to leave to get your social fix. The company on Wednesday is releasing an overhauled version of its Yahoo Mail Beta client that it says is twice as fast as the previous version, while managing to tack on new features like an integrated Twitter client, rich media previews and a more full-featured instant messaging client.

Yahoo says this speed boost should be especially noticeable to users outside the U.S. with latency issues, due mostly to the new version making use of the company’s cloud computing technology. This means that if you’re on a spotty connection, the app can adjust its behavior to keep pages from timing out, or becoming unresponsive.

Besides the speed and performance increase, which Yahoo says were the top users requests, the company has added a very robust Twitter client, which joins the existing social-sharing tools for Facebook and Yahoo. You can post to just Twitter, or any combination of the other two services, as well as see Twitter status updates in the update stream below. Yahoo has long had a way to slurp in Twitter feeds, but now you can do things like reply and retweet without leaving the page.

If asynchronous updates are not your thing, Yahoo has also tuned its integrated IM service to include some desktop software-like features, including window docking and tabbed conversations. This lets you keep a chat with several people running in one window while you go about with other e-mail tasks.


There are scads of noun phrases in this text — which ones are the important ones? You could count them by frequency, but that wouldn’t give any indication of whether they were lexically important or not (meaning, are they representative of the main topics of the content, or are they tangential).

Advantages of Noun Phrase Extraction

• Restricts to phrases matching certain part of speech patterns, fewer stop words needed

Drawbacks to Noun Phrase Extraction

• No way to tell if one noun phrase is more contextually relevant than another noun phrase
• Limited to words that occur in the text
Themes are noun phrases with contextual relevance scores: themes are themselves noun phrases, and are identified and scored through a process called Lexical Chaining.

Don’t worry, it’s not as scary as it sounds. Lexical chaining is just a process that relates sentences via related nouns. Consider the following text:

*I like beer. Miller just launched a new pilsner ale. But, because I’m a beer snob, I’m only going to drink pretentious Belgian beers.*

The three sentences above are related through nouns in each (beer → ale → beers). Even if those sentences don’t appear adjacent to each other in a block of text, they are still related to each other lexically and can be associated with each other by Lexalytics’ algorithms.

The strength (score) of a lexical chain is directly related to the length of the chain and the strength of the relationships between the chaining nouns (same words, antonyms, synonyms, etc.). A long chain of closely-related nouns will have a higher score, while tangentially-related nouns are assigned a lower score.

**Theme Extraction and Scoring**

Potential themes are extracted based on part-of-speech patterns. When the lexical chains have been scored, themes that belong to the highest-scoring chain are assigned the highest scores. If fewer than four chains are identified, Lexalytics simply scores them by count.
Let’s return to the article from earlier:

Yahoo wants to make its Web e-mail service a place you never want to -- or more importantly -- have to leave to get your social fix. The company on Wednesday is releasing an overhauled version of its Yahoo Mail Beta client that it says is twice as fast as the previous version, while managing to tack on new features like an integrated Twitter client, rich media previews and a more full-featured instant messaging client.

Yahoo says this speed boost should be especially noticeable to users outside the U.S. with latency issues, due mostly to the new version making use of the company’s cloud computing technology. This means that if you’re on a spotty connection, the app can adjust its behavior to keep pages from timing out, or becoming unresponsive.

Besides the speed and performance increase, which Yahoo says were the top users requests, the company has added a very robust Twitter client, which joins the existing social-sharing tools for Facebook and Yahoo. You can post to just Twitter, or any combination of the other two services, as well as see Twitter status updates in the update stream below. Yahoo has long had a way to slurp in Twitter feeds, but now you can do things like reply and retweet without leaving the page.

If asynchronous updates are not your thing, Yahoo has also tuned its integrated IM service to include some desktop software-like features, including window docking and tabbed conversations. This lets you keep a chat with several people running in one window while you go about with other e-mail tasks.


This article’s top 5 themes are:

<table>
<thead>
<tr>
<th>Theme</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud computing technology</td>
<td>4.11</td>
</tr>
<tr>
<td>Including window docking</td>
<td>2.976</td>
</tr>
<tr>
<td>Mail service</td>
<td>2.672</td>
</tr>
<tr>
<td>Top users requests</td>
<td>2.66</td>
</tr>
<tr>
<td>Rich media previews</td>
<td>2.635</td>
</tr>
</tbody>
</table>

You can see that those themes do a good job of conveying the context of the article. Contextually scoring information is highly useful in determining what’s really important in the text, and is handy in comparing many articles across periods of time to identify emerging trends and patterns.

Lexalytics text analysis solutions include sentiment scoring of themes. This is key in distinguishing sentiment directed at different entities in a single chunk of text. Remember this text:

“President Barack Obama did a great job with that awful oil spill.”

Lexalytics’ scoring of individual themes differentiates between the positive perception of the President and the negative perception of the theme “oil spill”. If you use a lesser text analytics tool, you’ll be left in the sentimental dark.

Advantages of Theme Extraction and Scoring
- Restricts to phrases that match certain part-of-speech patterns
- Scored based on contextual importance
- Includes sentiment scores for themes

Drawbacks to Theme Extraction and Scoring
- Limited to words in the text (true for all algorithms)
LEXALYTICS FACETS: A NEW WAY

Facets were introduced with Salience 5.0 and with our latest updates are even better than before. Facets represent a totally new way of extracting meaning from text.

Before we begin it should be noted that while all of the other techniques discussed in this paper work with Lexalytics’ document-at-a-time processing, Facets only work when processing collections of documents. Please see the Collections page on the Lexalytics website for more detail on collection processing.

Themes present the best combination of intelligence and scoring for noun-phrases. But sometimes your text doesn’t include a good noun phrase to work with, even when there’s valuable meaning and intent to be extracted from the document. Facets are built to handle these tricky cases where even Lexalytics’ powerful theme processing isn’t suited for the job.

Think about this sentence:

"The bed was hard."

There is no qualifying theme there, but the sentence contains important sentiment for a hospitality provider to know. Now imagine a whole collection of review documents, full of important information like in that sentence, but not obvious through themes alone. Here’s where Lexalytics Facets come into their own.

Noun phrase extraction relies on part-of-speech phrases in general, but Facets are based around “Subject Verb Object” (SVO) parsing. In the above case, “bed” is the subject, “was” is the verb, and “hard” is the object. When processed, this returns “bed” as the facet and “hard” as the attribute.

The nature of SVO parsing requires a collection of content to function properly. Any single document will contain many SVO sentences, but collections are scanned for facets or attributes that occur at least twice.

Facets work especially well with review processing. Here’s an example of two specific facets, pulled from an analysis of a collection of 165 cruise liner reviews:

<table>
<thead>
<tr>
<th>Facet</th>
<th>Positive Documents</th>
<th>Neutral Documents</th>
<th>Negative Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ship</td>
<td>45</td>
<td>127</td>
<td>14</td>
</tr>
<tr>
<td>Food</td>
<td>36</td>
<td>44</td>
<td>0</td>
</tr>
</tbody>
</table>

Top 5 Ship Attributes For “Ship”:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beautiful</td>
<td>22</td>
</tr>
<tr>
<td>Clean</td>
<td>8</td>
</tr>
<tr>
<td>New</td>
<td>8</td>
</tr>
<tr>
<td>Huge</td>
<td>6</td>
</tr>
<tr>
<td>Nice</td>
<td>6</td>
</tr>
</tbody>
</table>

Top 5 Attributes for “Food”:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>14</td>
</tr>
<tr>
<td>Good</td>
<td>12</td>
</tr>
<tr>
<td>Great</td>
<td>10</td>
</tr>
<tr>
<td>Best</td>
<td>4</td>
</tr>
<tr>
<td>Fabulous</td>
<td>4</td>
</tr>
</tbody>
</table>
Judging by these reviews, this is a new ship with great food — the kitchen seems to be doing an excellent job.

Our Facet processing also includes the ability to combine Facets based on semantic similarity via our Wikipedia™-based Concept Matrix. We combine attributes based on word stem, and Facets based on semantic distance.

Here’s an example:

<table>
<thead>
<tr>
<th>Enterprise</th>
<th>Company</th>
<th>Employee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profitable</td>
<td>Fun</td>
<td>Motivated</td>
</tr>
<tr>
<td>Growing</td>
<td>Enjoyable</td>
<td>Smart</td>
</tr>
<tr>
<td>Acquired</td>
<td>Acquired</td>
<td>Intelligent</td>
</tr>
<tr>
<td>Bought</td>
<td>Bought</td>
<td>Smart</td>
</tr>
<tr>
<td>Makes</td>
<td>Makes</td>
<td>Turnover</td>
</tr>
<tr>
<td>Produces</td>
<td>Produces</td>
<td>Attrition</td>
</tr>
<tr>
<td>Scaling</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

You can see how Enterprise and Company are combined into a single facet, providing you with even richer information by combining the attributes from both.
Lexalytics theme extraction and scoring provides a valuable combination of context-scored noun phrases for better insight into what your customers are saying about you. Whether you’re processing individual comments or massive libraries of documents, Lexalytics has the software you need to make informed business decisions. Lexalytics® is the industry leader in translating text into profitable decisions. Lexalytics deploys state-of-the-art on-premise and in-the-cloud text and sentiment analysis technologies that process billions of unstructured documents every day globally, transforming customers’ thoughts and conversations into actionable insights. The on-premise Salience® and SaaS Semantria® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.

Lexalytics is based in Boston, MA, and has offices in the U.S. and Canada. For more information, please visit www.lexalytics.com, email sales@lexalytics.com or call 1-617-249-1049. Follow Lexalytics on Twitter, Facebook, and LinkedIn for updates and insights into the world of text mining.