LUMINOSO

Case Study

Use Case:
Customer Analytics

Segment:
Voice Of The Customer

Products:
Rosette

Functions:
Text Analytics

Availability:
API or SDK

Shining A Light On Consumer Feedback

Spun out of the MIT Media Lab in 2010, Luminoso quickly drew the attention of major consumer brands. Its flagship product, Luminoso Analytics, digests large volumes of text-based customer feedback, such as online reviews, surveys, and customer service interactions. Instead of just finding key words, Luminoso identifies key concepts, ideas, thoughts, and sentiments that drive consumer choices. Luminoso, with multilingual capabilities, empowers its clients to understand, measure, and act on consumer feedback across any number of channels.

Surprisingly, the color of the item was mentioned in comments as a significant factor in the product experience. This key feature would not have been discovered through guided queries alone.

Luminoso In Action

A multinational consumer goods company was designing a new variant of its most popular men's personal care product. The million dollar question was, what design features actually matter most to their consumers?

Thousands of testers filled out a survey about the new product variant, which asked them to rate attributes on a scale of 1 to 5, and to leave free-form comments. These comments were more text than could be accurately processed manually

The company chose Luminoso to analyze the free-form comments, uncovering distinct product feature discussions, and then measured each one for positive and negative feedback. Luminoso also looked at reviewer rating tiers separately to see what company and product attributes drove both high and low review scores.

Surprisingly, the color of the item was mentioned in comments as a significant factor in the product experience. This key feature would not have been discovered through guided queries alone.

The Challenge

Every time Luminoso adds a new human language to their portfolio, finding a reliable linguistic analyzer in that language is step one.

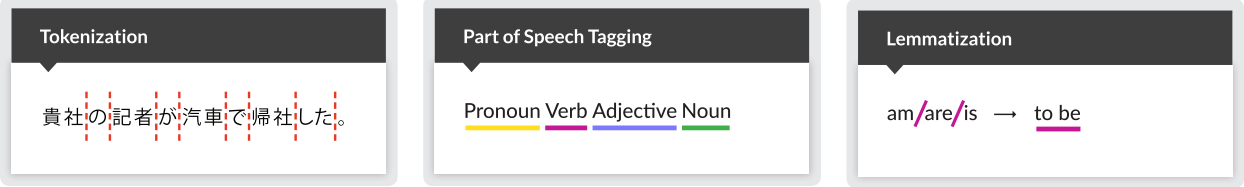
For Luminoso's algorithms to tease out ideas, thoughts, and sentiment from unstructured text, the raw text needs to be enriched and tagged with exacting accuracy. This enrichment enables "tuning out" the "noise" to reveal finely tuned messages within.

"Noise" includes:

Single words with multiple surface forms: These "single words" are the lemmas, the dictionary form of a word.	**child/children** (English singular/plural) **beau/beaux/belle/belles** (French adjective forms: masculine singular/ masculine plural/feminine singular/feminine plural)
Dialectal variations:	**colour/color** (British English vs. American English) **nonante/quatre-vingt dix** ("90" in Belgian French vs. French-French)
Recognizing pronouns:	attributes linked to pronouns out of context are unhelpful

The process of standardizing words begins with careful analysis of the words. Natural Language Processing (NLP) means:

- Finding word boundaries (needed for Chinese, Japanese, and Korean, which don't reliably use spaces between words)
- Tagging parts of speech
- Finding the dictionary form (lemma) of each word



"Whenever someone asks us, does your system work in language x? The answer is 'Yes, it does.'—to the extent that you can put your text into our system and we'll see how it does," Lance Nathan, Senior Linguistics Developer of Luminoso said. "The question really isn't 'does it work in language x', but how well it works in language x." When the linguistic analysis isn't up to snuff, meaningless distinctions like—color v. colour—creep in, muddying the results.

Nathan's job of getting the quality results expected by Luminoso users is easier when the linguistic enrichment is of higher quality.

> Things that Rosette excelled at, according to Nathan, include finding lexicalized phrases. That is, knowing that "nouveaux riches" is a single unit that should be singularized to "nouveau riche," for instance...

The Old Solution

Luminoso started by using open source solutions for linguistic analysis, but in some cases, the results were simply not accurate enough for Luminoso.

In early experiments, Luminoso used stemming, which is one of the most basic tools to remove suffixes from words. However errors like turning "Mount Everest" into "Mount Ever" and stemming "installed" or "installing" (but not "install") to "instal" were unacceptable.

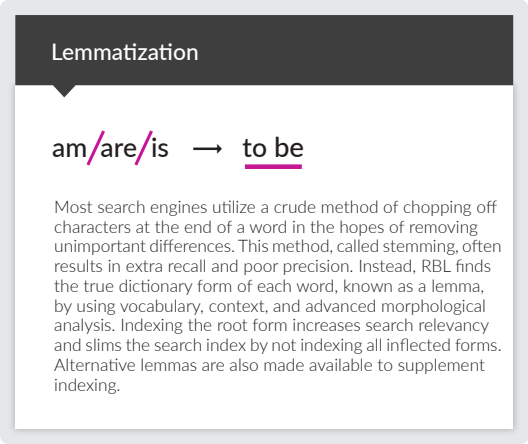
More sophisticated tools made more sophisticated mistakes, like deciding that "crew" is the past tense of "crow" or that "cola" should be singularized to "colon."

The Rosette Solution

Luminoso is a successful alumnus of Basis Technology's Startup Program, which puts its professional-grade multilingual text analytics into the hands of early-stage, high-impact firms to help them quickly realize their vision.

Things that Rosette excelled at, according to Nathan, include finding lexicalized phrases. That is, knowing that “nouveaux riches” is a single unit that should be singularized to “nouveau riche,” for instance, or that “according to” is just a preposition and not an instance of the verb “accord.”

A case in point being the Portuguese analyzer they were using which sometimes had difficulty splitting contractions such as “pelas” and “pela” which are the preposition “por” plus an article in the context: “reforçada pelas mesas fartas e pela moda de viola.” The analyzer they were using would lemmatize “pela” to “pelar,” meaning “to scale a fish.



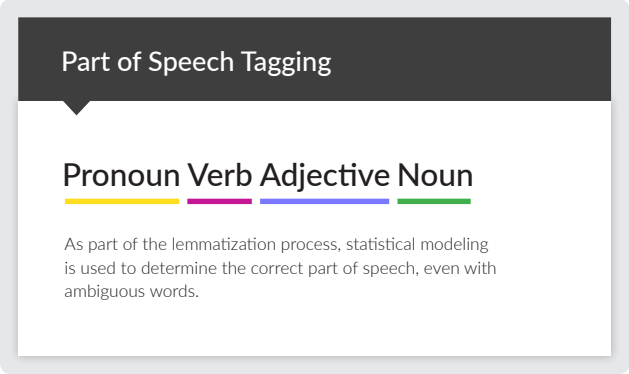
Example: English
Linguistic analysis is useful for every language; lemmatization for English improves recall and precision.

CHALLENGE	QUERY	STEM	LEMMA
Two unrelated words may share a stem.	animals animated	anim	animal animate
Stemming may deliver unintended results.	several	sever	several
Irregular verbs and nouns stump the stemmer.	spoke	spoke	speak (v.) spoke (n.)

Another reason Rosette replaced the open source solution for Portuguese was Rosette’s better recognition of Brazilian Portuguese spellings that allowed it to more frequently return the correct dictionary form.

Part Of Speech Tagging

As part of the lemmatization process, statistical modeling is used to determine the correct part of speech, even with ambiguous words. Each token is then tagged for enhanced comprehension and search relevancy.



In French, Rosette's part of speech tagger usually distinguishes "pas"-the-adverb, which marks something as negative from "pas"-the-noun, which is just a word meaning "step."

"If we had to work solely off tokenized surface forms [in French], we'd never be able to draw that distinction," Nathan said.

Rosette's ability to identify ambiguous parts of speech with greater accuracy, meant that swapping in Rosette increased Luminoso's robustness in the face of spelling errors such as missing accents in French and Spanish.

For the languages that Luminoso has tested so far for its linguistic processing, Rosette has replaced their pre-existing solution.

Still for every new language, Nathan still performs a battery of tests to determine which linguistics solution to go with. For a language Luminoso recently started working on, Nathan began by finding a public corpus in the language that was already tagged with parts-of-speech. Then he ran the untagged corpus through Rosette to see how it did. If Rosette wasn't accurate enough, he was going to keep looking. Fortunately, Rosette's output turned out to be a 97-98% match, so he looked no further than Rosette.

> "Using Rosette gets [the results] to a point that I wouldn't be embarrassed to show to a native speaker of the language," Nathan said.

The Result

Although a few alternatives might be better for single languages, Rosette's accuracy across over 40 languages, and the ease of a unified API for all languages is very valuable to Luminoso, Nathan says.

"I need to be convinced we are doing at least an adequate job in any new language," Nathan said. "Our English is the gold standard. The new language has to be at least 'pretty good' to 'very good' or at a level approaching English."

"Using Rosette gets [the results] to a point that I wouldn't be embarrassed to show to a native speaker of the language," Nathan said.

"Our relationship with Basis Technology means we have a ready source of high quality text analytics in a broad range of languages," Luminoso founder and CEO Dr. Catherine Havasi said. "Using Rosette positions Luminoso well for faster entry into new markets and languages."

The feedback from Luminoso customers is loud and clear. In less than four years, Luminoso has gone from startup to a portfolio of 60 to 70 customers, including Fortune 1000 companies such as MARS, Sony, Intel, Scotts, and ConAgra. With a recent sales relationship with Basis Technology, Luminoso is poised to reach new users worldwide.

"Our relationship with Basis Technology means we have a ready source of high quality text analytics in a broad range of languages," Founder and CEO Dr. Catherine Havasi said. "Using Rosette positions Luminoso well for faster entry into new markets and languages."

Rosette® provides businesses and government agencies text analytics in 55 languages. www.rosette.com