Better Content Through NLP (Natural Language Processing) – Whiteboard Friday

Gone are the days of optimizing content solely for search engines. For modern SEO, your content needs to please both robots *and* humans. But how do you know that what you're writing can check the boxes for both man and machine?

In today's Whiteboard Friday, Ruth Burr Reedy focuses on part of her recent MozCon 2019 talk and teaches us all about how Google uses NLP (natural language processing) to truly understand content, plus how you can harness that knowledge to better optimize what you write for people and bots alike.

Click on the whiteboard image above to open a high resolution version in a new tab!

Video Transcription

Howdy, Moz fans. I'm Ruth Burr Reedy, and I am the Vice President of Strategy at UpBuild, a boutique technical marketing agency specializing in technical SEO and advanced web analytics. I recently spoke at MozCon on a basic framework for SEO and approaching changes to our industry that thinks about SEO in the light of we are humans who are marketing to humans, but we are using a machine as the intermediary.

Those videos will be available online at some point. *[Editor's note: that point is now!]* But today I wanted to talk about one point from my talk that I found really interesting and that has kind of changed the way that I approach content creation, and that is the idea that **writing content that is easier for**

Google, a robot, to understand can actually make you a better writer and help you write better content for humans. It is a win-win.

The relationships between entities, words, and how people search

To understand how Google is currently approaching parsing content and understanding what content is about, Google is spending a lot of time and a lot of energy and a lot of money on things like **neural matching** and **natural language processing**, which seek to understand basically when people talk, what are they talking about?

This goes along with the evolution of search to be more conversational. But there are a lot of times when someone is searching, but they don't totally know what they want, and Google still wants them to get what they want because that's how Google makes money. They are spending a lot of time trying to understand the relationships between entities and between words and how people use words to search.

The example that Danny Sullivan gave online, that I think is a really great example, is if someone is experiencing the soap opera effect on their TV. If you've ever seen a soap opera, you've noticed that they look kind of weird. Someone might be experiencing that, and not knowing what that's called they can't Google soap opera effect because they don't know about it.

They might search something like, "Why does my TV look funny?" Neural matching helps Google understand that when somebody is

searching “Why does my TV look funny?” one possible answer might be the soap opera effect. So they can serve up that result, and people are happy.

Understanding salience

As we’re thinking about natural language processing, a core component of natural language processing is understanding **salience**.

Salience, content, and entities

Salience is a one-word way to sum up to what extent is this piece of content about this specific entity? At this point Google is really good at extracting entities from a piece of content. Entities are basically nouns, people, places, things, proper nouns, regular nouns.

Entities are things, people, etc., numbers, things like that. Google is really good at taking those out and saying, “Okay, here are all of the entities that are contained within this piece of content.” Salience attempts to understand how they’re related to each other, because what Google is really trying to understand when they’re crawling a page is: **What is this page about, and is this a good example of a page about this topic?**

Salience really goes into the second piece. To what extent is any given entity be the topic of a piece of content? It’s often amazing the degree to which a piece of content that a person has created is not actually about anything. I think we’ve all experienced that.

You’re searching and you come to a page and you’re like, “This was too vague. This was too broad. This said that it was about one thing, but it was actually about something else. I didn’t

find what I needed. This wasn't good information for me." As marketers, we're often on the other side of that, trying to get our clients to say what their product actually does on their website or say, "I know you think that you created a guide to Instagram for the holidays. But you actually wrote one paragraph about the holidays and then seven paragraphs about your new Instagram tool. This is not actually a blog post about Instagram for the holidays. It's a piece of content about your tool." These are the kinds of battles that we fight as marketers.

Natural Language Processing (NLP) APIs

Fortunately, there are now a number of different APIs that you can use to understand natural language processing:

Is it as sophisticated as what they're using on their own stuff? Probably not. But you can test it out. Put in a piece of content and see (a) what entities Google is able to extract from it, and (b) how salient Google feels each of these entities is to the piece of content as a whole. Again, **to what degree is this piece of content about this thing?**

So this natural language processing API, which you can try for free and it's actually not that expensive for an API if you want to build a tool with it, will assign each entity that it can extract a salient score between 0 and 1, saying, "Okay, how sure are we that this piece of content is about this thing versus just containing it?"

So the higher or the closer you get to 1, the more confident the tool is that this piece of content is about this thing. 0.9 would be really, really good. 0.01 means it's there, but

they’re not sure how well it’s related.

A delicious example of how salience and entities work

The example I have here, and this is not taken from a real piece of content – these numbers are made up, it’s just an example – is if you had a chocolate chip cookie recipe, you would want chocolate cookies or chocolate chip cookies recipe, chocolate chip cookies, something like that to be the number one entity, the most salient entity, and you would want it to have a pretty high salient score.

You would want the tool to feel pretty confident, yes, this piece of content is about this topic. But what you can also see is the other entities it’s extracting and to what degree they are also salient to the topic. So you can see things like if you have a chocolate chip cookie recipe, you would expect to see things like cookie, butter, sugar, 350, which is the temperature you heat your oven, all of the different things that come together to make a chocolate chip cookie recipe.

But I think that it’s really, really important for us as SEOs to understand that **salience is the future of related keywords**. We’re beyond the time when to optimize for chocolate chip cookie recipe, we would also be looking for things like chocolate recipe, chocolate chips, chocolate cookie recipe, things like that. Stems, variants, TF-IDF, these are all older methodologies for understanding what a piece of content is about.

Instead what we need to understand is what are the entities that Google, using its vast body of knowledge, using things like Freebase, using large portions of the internet, **where is Google seeing these entities co-occur at such a rate that they**

feel reasonably confident that a piece of content on one entity in order to be salient to that entity would include these other entities?

Using an expert is the best way to create content that's salient to a topic

So chocolate chip cookie recipe, we're now also making sure we're adding things like butter, flour, sugar. This is actually really easy to do if you actually have a chocolate chip cookie recipe to put up there. This is I think what we're going to start seeing as a content trend in SEO is **that the best way to create content that is salient to a topic is to have an actual expert in that topic create that content.**

Somebody with deep knowledge of a topic is naturally going to include co-occurring terms, because they know how to create something that's about what it's supposed to be about. I think what we're going to start seeing is that people are going to have to start paying more for content marketing, frankly. Unfortunately, a lot of companies seem to think that content marketing is and should be cheap.

Content marketers, I feel you on that. It sucks, and it's no longer the case. We need to start investing in content and investing in experts to create that content so that they can create that deep, rich, salient content that everybody really needs.

How can you use this API to improve your own SEO?

One of the things that I like to do with this kind of information is look at – and this is something that I've done for years, just not in this context – but a prime optimization target in general is pages that rank for a topic, but they rank on page 2.

What this often means is that Google understands that that keyword is a topic of the page, but it doesn't necessarily understand that it is a good piece of content on that topic, that the page is actually solely about that content, that it's a good resource. In other words, **the signal is there, but it's weak.**

What you can do is take content that ranks but not well, run it through this natural language API or another natural language processing tool, and look at how the entities are extracted and how Google is determining that they're related to each other. Sometimes it might be that you need to do some disambiguation. So in this example, you'll notice that while chocolate cookies is called a work of art, and I agree, cookie here is actually called other.

This is because cookie means more than one thing. There's cookies, the baked good, but then there's also cookies, the packet of data. Both of those are legitimate uses of the word "cookie." Words have multiple meanings. If you notice that Google, that this natural language processing API is having trouble correctly classifying your entities, that's a good time to go in and do some disambiguation.

Make sure that the terms surrounding that term are clearly saying, "No, I mean the baked good, not the software piece of data." That's a really great way to kind of bump up your salience. Look at whether or not you have a strong salient score for your primary entity. You'd be amazed at how many pieces of content you can plug into this tool and the top, most salient entity is still only like a 0.01, a 0.14.

A lot of times the API is like "I think this is what it's about," but it's not sure. This is a great time to go in and bump up that content, make it more robust, and look at ways

that you can make those entities easier to both extract and to relate to each other. This brings me to my second point, which is my new favorite thing in the world.

Writing for humans and writing for machines, you can now do both at the same time. You no longer have to, and you really haven’t had to do this in a long time, but the idea that you might keyword stuff or otherwise create content for Google that your users might not see or care about is way, way, way over.

Now you can create content for Google that also is better for users, because the tenets of machine readability and human readability are moving closer and closer together.

Tips for writing for human and machine readability:

What I’ve done here is I did some research not on natural language processing, but on writing for human readability, that is advice from writers, from writing experts on how to write better, clearer, easier to read, easier to understand content.Then I pulled out the pieces of advice that also work as pieces of advice for writing for natural language processing. So natural language processing, again, is the process by which Google or really anything that might be processing language tries to understand how entities are related to each other within a given body of content.

Short, simple sentences

Short, simple sentences. Write simply. Don’t use a lot of flowery language. Short sentences and try to keep it to one idea per sentence.

One idea per sentence

If you're running on, if you've got a lot of different clauses, if you're using a lot of pronouns and it's becoming confusing what you're talking about, that's not great for readers.

It also makes it harder for machines to parse your content.

Connect questions to answers

Then closely connecting questions to answers. So don't say, "What is the best temperature to bake cookies? Well, let me tell you a story about my grandmother and my childhood," and 500 words later here's the answer. Connect questions to answers.

What all three of those readability tips have in common is they boil down to **reducing the semantic distance between entities.**

If you want natural language processing to understand that two entities in your content are closely related, move them closer together in the sentence. Move the words closer together. Reduce the clutter, reduce the fluff, reduce the number of semantic hops that a robot might have to take between one entity and another to understand the relationship, and you've now created content that is more readable because it's shorter and easier to skim, but also easier for a robot to parse and understand.

Be specific first, then explain nuance

Going back to the example of "What is the best temperature to bake chocolate chip cookies at?" Now the real answer to what is the best temperature to bake chocolate cookies is it

depends. Hello. Hi, I'm an SEO, and I just answered a question with it depends. It does depend.

That is true, and that is real, but it is not a good answer. It is also not the kind of thing that a robot could extract and reproduce in, for example, voice search or a featured snippet. If somebody says, "Okay, Google, what is a good temperature to bake cookies at?" and Google says, "It depends," that helps nobody even though it's true. So in order to write for both machine and human readability, be specific first and then you can explain nuance.

Then you can go into the details. So a better, just as correct answer to "What is the temperature to bake chocolate chip cookies?" is the best temperature to bake chocolate chip cookies is usually between 325 and 425 degrees, depending on your altitude and how crisp you like your cookie. That is just as true as it depends and, in fact, means the same thing as it depends, but it's a lot more specific.

It's a lot more precise. It uses real numbers. It provides a real answer. I've shortened the distance between the question and the answer. I didn't say it depends first. I said it depends at the end. That's the kind of thing that you can do to improve readability and understanding for both humans and machines.

Get to the point (don't bury the lede)

Get to the point. Don't bury the lead. All of you journalists who try to become content marketers, and then everybody in content marketing said, "Oh, you need to wait till the end to get to your point or they won't read the whole thing,"and you were like, "Don't bury the lead," you are correct. For those of you who aren't familiar with journalism speak, not burying the lead basically means get to the point upfront, at the top.

Include all the information that somebody would really need to

get from that piece of content. If they don't read anything else, they read that one paragraph and they've gotten the gist. Then people who want to go deep can go deep. That's how people actually like to consume content, and surprisingly it doesn't mean they won't read the content. It just means they don't have to read it if they don't have time, if they need a quick answer.

The same is true with machines. Get to the point upfront. Make it clear right away what the primary entity, the primary topic, the primary focus of your content is and then get into the details. You'll have a much better structured piece of content that's easier to parse on all sides.

Avoid jargon and "marketing speak"

Avoid jargon. Avoid marketing speak. Not only is it terrible and very hard to understand. You see this a lot. I'm going back again to the example of getting your clients to say what their products do. You work with a lot of B2B companies, you will you will often run into this. Yes, but what does it do? It provides solutions to streamline the workflow and blah, blah. Okay, what does it do? This is the kind of thing that can be really, really hard for companies to get out of their own heads about, but it's so important for users, for machines.

Avoid jargon. Avoid marketing speak. Not to get too tautological, but the more esoteric a word is, the less commonly it's used. That's actually what esoteric means. What that means is the less commonly a word is used, the less likely it is that Google is going to understand its semantic relationships to other entities.

Keep it simple. Be specific. Say what you mean. Wipe out all of the jargon. By wiping out jargon and kind of marketing speak and kind of the fluff that can happen in your content, you're also, once again, reducing the semantic distances

between entities, making them easier to parse.

Organize your information to match the user journey

Organize it and map it out to the user journey. Think about the information somebody might need and the order in which they might need it.

Break out subtopics with headings

Then break it out with subheadings. This is like very, very basic writing advice, and yet you all aren't doing it. So if you're not going to do it for your users, do it for machines.

Format lists with bullets or numbers

You can also really impact skimmability for users by breaking out lists with bullets or numbers.

The great thing about that is that breaking out a list with bullets or numbers also makes information easier for a robot to parse and extract. If a lot of these tips seem like they're the same tips that you would use to get featured snippets, they are, because featured snippets are actually a pretty good indicator that you're creating content that a robot can find, parse, understand, and extract, and that's what you want.

So if you're targeting featured snippets, you're probably already doing a lot of these things, good job.

Grammar and spelling count!

The last thing, which I shouldn't have to say, but I'm going to say is that grammar and spelling and punctuation and things like that absolutely do count. They count to users. They don't count to all users, but they count to users. They also count to search engines.

Things like grammar, spelling, and punctuation are very, very easy signals for a machine to find and parse. Google has been specific in things, like the “Quality Rater Guidelines,”that a well-written, well-structured, well-spelled, grammatically correct document, that these are signs of authoritativeness. I’m not saying that having a greatly spelled document is going to mean that you immediately rocket to the top of the results.

I am saying that if you’re not on that stuff, it’s probably going to hurt you. So take the time to make sure everything is nice and tidy. You can use vernacular English. You don’t have to be perfect “AP Style Guide” all the time. But make sure that you are formatting things properly from a grammatical standpoint as well as a technical standpoint. What I love about all of this, this is just good writing.

This is good writing. It’s easy to understand. It’s easy to parse. It’s still so hard, especially in the marketing world, to get out of that world of jargon, to get to the point, to stop writing 2,000 words because we think we need 2,000 words, to really think about are we creating content that’s about what we think it’s about.

Use these tools to understand how readable, parsable, and understandable your content is

So my hope for the SEO world and for you is that you can use these tools not just to think about how to dial in the perfect keyword density or whatever to get an almost perfect score on the salience in the natural language processing API. What I’m hoping is that you will use these tools to help yourself understand how readable, how parsable, and how understandable your content is, how much your content is about what you say

it’s about and what you think it’s about so you can create better stuff for users.

It makes the internet a better place, and it will probably make you some money as well. So these are my thoughts. I’d love to hear in the comments if you’re using the natural language processing API now, if you’ve built a tool with it, if you want to build a tool with it, what do you think about this, how do you use this, how has it gone. Tell me all about it. Holla atcha girl.

Have a great Friday.

Video transcription by Speechpad.com

http://tracking.feedpress.it/link/9375/12994239?utm_medium=feed&utm_source=feedpress.me&utm_campaign=Feed%3A+mozblog