Rami Khater

Georgetown University

rk299@georgetown.edu

Cloud Computing Based Statistical Translation and The Monolithic Narrative

The intersection of technology and translation has always had immense potential, but up until the past five years this fusion had not yielded the returns originally imagined. With the development of cloud computing and the digitization of human knowledge, translation has begun to see major breakthroughs by moving from rule to statistical based translation algorithms. Any individual, application or device is now enabled to translate content in endless languages, facilitating communication and breaking a major barrier – language.

Cloud computing has moved translation into the age of statistical translation-learning algorithms which "learn" a language based on the content it receives over time. If the content fed into the algorithm is the coursework, then the creator of the coursework is the teacher. *The monolithic narrative is the result of the teacher being the hegemon*; the algorithm, like a child, believes what it is told. Its coursework is the narrative of the dominant, the current majority creator, keeper and interpreter of human knowledge.

The art and science of translation is one that has improved dramatically over the last century, but is still largely bound by problems of efficiency and accuracy. Professional translators do excellent work, but that work does not scale well - it is held back by human time - they simply cannot do more work than their collective man-hours allow. Neither can their trade be instantly applied to works they have never encountered; their effort is static and only significant to one document, sentence, or word.

The birth of the computer has brought in visions of instant translation of any language, and computer scientists believed this to be a moderately achievable goal well within technologies they possessed at the time. However, it turned out that linguistic analysis and accurate translation was indeed more difficult than they had imagined and required the most powerful computers of their day. Early translation algorithms consisted of a rule-based structure, a collection of if-then-else conditional statements[1] that would translate a word at a time, rarely within the content of a sentence or phrase.

The past five years have seen a dramatic increase in the efficiency and accuracy of computer-mediated translation because of a change away from rule-based to statistically based algorithms and the virtually unlimited processing power available via cloud computing. Google is at the forefront of the new translation movement but was a latecomer to the field. Their research began in earnest in 2004, and the key decision was that translation should follow the same paradigm as search.

[1] If-Then-Else statements are universal in all programming languages, some call them rules and others simply classify them as conditional statements

Generally speaking, search results are ranked by how many other pages link to them, creating a natural peer-produced ranking and validation of websites. There are no experts involved ranking the websites: rather, the position is evaluated via an algorithm that collects statistics from all over the web. This peer-produced ranking is at the heart of Google's PageRank algorithm[2] and sets it apart from the first-generation search engines such as HotBot and Yahoo. Google's ranking is therefore completely organic and does not require human intervention: it is combing the *collective conscious* of the world's Internet users by gathering statistics about their independent decisions and actions.

By adopting this approach to translation, Google has been able to avoid the pitfall of "understanding" languages out of the box; the algorithm learns how a language works by analyzing content from all corners of the Internet. The more content available in a language the more learning the algorithm will have done and therefore the more accurate its translations and analysis of the language will be. The algorithm is also fed translations between languages so it may learn how others have translated and interpreted text previously.

[2] This basic method is now at the core of every other major search engine as well, the variation is in the variables surrounding this core analysis. While the exact The nature of the PageRank algorithm is a closely guarded trade secret, the general nature of the algorithm has been discussed at numerous events by Google employees

This approach is distinct from traditional rule-based language translation in that it takes the unconscious crowd-sourced use of a language and its translations to build up its understanding. Google and others like them have essentially used the wisdom of the crowds as peer-produced language instructors. Every blog, article, and word on the Internet is an unconscious “vote” on how that language works and more importantly decides how others will view the meaning of the content. The more data these statistical translation systems receive the more likely they are to give better results in the future (Helft 2010). The era of Big Data is truly upon us.

The term Big Data has come to mean many things, but recently it has taken on a more optimistic and positive tone. Researchers have found that with access to huge amounts of information they are able to find trends, patterns, and facts they could never have found in smaller datasets. The issue is not about skill; regardless of the science, the work and time put into it, the *amount* of data is the key factor (Anderson 2008; Bollier 2010). If the key is large amounts of data, perhaps the statistical translation systems have a chance, but in this case the source of the data is the issue, not the amount.

With large amounts of data it would appear that the statistical approach is flawless and neutral in nature; a language is defined by all the content gathered about it. This makes perfect sense for the hegemon but what becomes of the other, the subaltern? What of the voice of the global south? In a world overflowing with content produced by the dominant, the narrative and language structure this content suggests will become the monolithic narrative of the globe via a positive feedback affect. If a person constantly reads news, websites and articles that are translated into their native tongue then that translation will in and of itself become part of that native tongue. The user will begin to mimic the language of the translation tool in their writings, which feed back into the system, and further the discourse the translation was interpreted from.

This is the potential of the monolithic narrative; a self-propagating piece of knowledge, which is presented to users repeatedly and therefore becomes part and parcel of their knowledge and story. For the subaltern this is another blow, pushing their knowledge and views out of the mainstream via translation. Even if the original article was the narrative of the other, it may not be so when it is translated, as majority content created by the dominant content creators has taught the algorithm how to think.

A native Arabic-language speaker may read a translation of an English newspaper article in Arabic, but from which perspective? The concept of *contextual objectivity* put forth by Iskandar and El-Nawawy is based on the belief that content is always created from an objective viewpoint within the cultural and historical background of the audience for which it was intended. But, what is objective for one audience may not be so for another and therefore objectivity is a contextual matter and cannot truly exist across borders with distinct audiences who harbor asymmetric beliefs and views (Iskandar and El-Nawawy 2004).

If the article translated was about an incident where a young Palestinian boy was killed by an Israeli soldier, would that boy be a “civilian” from the context of the West or a “martyr” from the context of the Arab world? Translating this as “civilian” to a native Arabic speaker living in the Middle East may cause confusion, whereas “martyr” would be the correct translation for that audience.

The Meedan[3] website is trying to facilitate communication between the West and Arab world by translating articles to and from English and Arabic while providing a forum and comments section for people to interact. It incorporates the power of automated translation with editors who review the results and make changes where necessary. Meedan understands the sensitive nature of the topics that are being discussed and is not interested in creating problems for themselves or others; rather, its goal is to get people talking. The fact that Meedan has active editors to double-check automated translation means it understands that every translation needs to be as non-inflammatory as possible. To further their cause Meedan posts all the steps in their translation and any adjustments in a publicly available Wiki document, so all may follow along while they work (Singel 2010; Phillips 2010).

These new translation systems based on cloud computing are not limited to websites such as Meedan nor are do they require powerful machines. The ethos of the cloud paradigm is supercomputer-like processing from any device, as long as there is connectivity to the

[3] http://www.meedan.com

cloud. An iPhone, watch or picture frame could have access to the same level translation as a university or government. Due to this shift of computing power from the fringes to the center, translation is now embedded in many applications and services once thought to be in the domain of science-fiction films and books.

Google Goggles is an application for smartphones that allows the user to identify objects of interest: the user points the device at the object and the software will identify it by sending a picture back to Google's cloud (Paul 2010). Using Goggles to identify the soda can in front of you is not of much use, but the potential for its use to identify plants, insects and any other number of things is enormous. Goggles recently gained the ability to translate text it sees in an image in real time and show the translated text to the user (Paul 2010). While imperfect, the main issue is getting Goggles to recognize the text, rather than the translation itself. The technology is truly impressive and it is only in its nascent stages.

Microsoft and Google are among the leaders in real-time speech translation and have both demoed their prototypes of the technology (Hachman 2010). This would allow people speaking on a mobile phone to have a conversation in different languages, and have it translated into the native tongue of the other speaker, without any knowledge of the other language or even knowing what language it was originally spoken in.

The previous examples are just the beginning of what will be the application of translation to every type of medium and process. Via Application Program Interfaces (APIs) to translation tools offered by cloud providers, every individual, business, group and government may include instantaneous translation as part of their offering without any knowledge of the languages. *Language is no longer the barrier as long as you agree with the narrative and interpretation of the translation.*

A future composed of a monolithic narrative is not an inevitable fate: over time the situation may arise where the Other has created enough content that statistical translation based algorithms will learn their narrative as part of the dominant core, but can these algorithms be trusted? In another paper (*Digital Protectionism: Preparing for the coming Internet Embargo)*, I have used the term Virtual Infrastructure to describe Internet-based digital services which have become part of the software infrastructure of the Internet. The Other becomes so dependent on the digital services infrastructure of the hegemon that it cannot function without it. Therefore the Other must build its own virtual infrastructure to compete with and balance the power of the hegemon.

In terms of statistical translation based learning systems, the Other would build its own systems to avoid the possibility of an embargo, as suggested in *Digital Protectionism*, but more importantly so that its narrative and story live on. If both the hegemon and the Other have services that may be used for translation, then somewhat of a balance exists; Internet users and services may decide whose infrastructure and services to leverage.

“Balance” is indeed a loose term, but in this case it means the balance that exists between a dominant and a weaker force by nature. Furthermore, if multiple instantaneous interpretations of a text are accessible, then users can see varied contextually objective viewpoints and may chose from between them

This proposed competition between translation engines is akin to users leveraging multiple search engines for the most accurate results. Websites exist that combine the results of the same queries from multiple engines and present them to the user in a seamless manner[4]. Logically, the same process could potentially work for translation of a text, combining the results from selected translation engines. However, currently there is no process to combine the translation results of two statistical learning systems in a coherent manner with consistent quality (Macherey and Och 2007).

Early reports by users of subpar statistical translation results have been comical in nature, but they exhibit the capability of these systems to learn suspect translations and interpretations based on knowledge gained from content they have received. TechCrunch broke the news that French users had noticed that in translation when the word Vimeo was used in a sentence, it was translated as YouTube (Rao 2010). Vimeo is one of YouTube’s biggest competitors and many rumors spread that this translation was Google’s idea of a joke (Rao 2010) or a real attempt to undercut a competitor.

[4] http://www.dogpile.com

However, French was not the only language affected, nor was this translation of Vimeo to YouTube universal. A comment left by reader Josh Kuhn on the TechCrunch article shows an understanding of statistical learning by content, and said:

> *“This is just a result of Google's translation strategy which is a big ol' (Bayesian) learning algorithm based on a large number of sample texts. This side-effect pops up with things that have a correlation to (a country's) language, even if they aren't actually the same word in the other language. Like translating "America" in an English sentence to "Italy". My guess is that YouTube is more popular to the French than Vimeo.”*

Another commentator noted:

> *“For the past year, Google has translated the word 'English' to 'français' and 'French' to 'anglais'. It also converted the word 'euro' to 'dollar' and vice-versa. It was infuriatingly stupid, and they only fixed it recently.”*

The second commentator believes the automatic translations to be a ridiculous error, but this “error” was achieved through an algorithm that learned enough to understand that when speaking in English most people use the word “dollar” for money, while French speakers use “euro”. The true measure of whether this is a proper translation is the context within which it is used. As noted this is not where automated translation excels. The Vimeo-YouTube issue, while interesting, is more entertaining than it is of concern (unless you work at Vimeo or another YouTube competitor).

However, the findings of native Vietnamese and Russian speakers show the potential for statistical analysis to go truly awry.

The Vietnam War and Cold War are two distinct pieces of world history and depending on which side you sympathized with, your interpretation would be substantially different. Currently, the majority of content on both issues reflects the American interpretation of the conflict. Russia and Vietnam put forth little content, and while the Russian blogosphere and Internet websites are of substantial size they still pale in comparison to the content available that reflects the American point of view.

A Vietnamese user entered “Vietnam won the war” in Vietnamese and had it translated into English as “America won Vietnam”. Further tests showed similar results, where no mention of America in the text to be processed was translated as America beating Vietnam or winning a war[5]. Furthermore, a Russian newspaper noted that when searching for a phrase that ends with “to blame or not to blame” the results differ based on the topic (Topolyanskaya 2010). The Moscow News article states that:

> *“USA is to blame/Russia is to blame/Obama is to blame/Medvedev is to blame" translates as the partially opposite - "US is NOT to blame, Russia is to blame/ Obama is NOT to blame, Medvedev is to blame."*

[5] http://groups.google.com/group/google-translate-general/browse_thread/thread/43a781bfacdd977b/0c21469bfd681f1e#0c21469bfd681f1e

The article continues and notes that the same error occurs with Ukrainian and Belorussian but not with Spanish, German and other European languages.

In more modern examples, would American forces in Iraq be occupiers or liberators? Is the new Iraqi government a puppet or legitimate? Did America win the war? As these translation systems become more entrenched as virtual infrastructure, issues such as these will give rise to great concern between individuals and governments. Would the Turkish government withdraw its ambassador from the United States based on instantaneous translation of topics that have to do with Armenians in Turkey? The possibilities are endless and as varied as the viewpoints of all human beings. We are just at the beginning of the new narrative, a story of which this document will become a part as well.

These examples show the power of statistical translation and bring to light an important issue. ***The problem*** *with statistical translation engines is that they believe what they are told and take whatever happens to be said the most as the truth*. Perhaps the saying, "A lie told a thousand times is still a lie" is no longer accurate. A lie told more than the truth becomes the narrative of choice.

In conclusion, the purpose of the paper is not to belittle the magnificent accomplishments of cloud-based statistical translation and the promise it has for the future. Rather, it is to inform of the potential for further exegesis of all human knowledge - past, present, and future – more distinctly in the direction of hegemonic norms and beliefs. The subaltern's narrative and voice will potentially be removed from the interpretation of all human history, as our collective knowledge will pass through the filters of these trained algorithms. This reinterpretation will no doubt shape the future, as history will now have only one story to tell, the monolithic narrative.

BIBLIOGRAPHY

Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*. June 23. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+wired/index+(Wired:+Index+3+(Top+Stories+2.

Bollier, David. 2010. The Promise and Peril of Big Data. The Aspen Institute. http://www.aspeninstitute.org/publications/promise-peril-big-data.

Chen, Rita. 2009. Translate documents: sharing across languages and generations. *Official Google Blog*. August 27. http://googleblog.blogspot.com/2009/08/translate-documents-sharing-across.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+blogspot/MKuf+(Official+Google+Blog.

Hachman, Mark. 2010. Microsoft Shows Off Future Product Tech at R&D Day. *PC Magazine*. May 7. http://www.pcmag.com/article2/0,2817,2363500,00.asp.

Helft, Miguel. 2010. Google's Computing Power Refines Translation Tool. *The New York Times*, March 8, sec. Technology. http://www.nytimes.com/2010/03/09/technology/09translate.html?hpw&pagewanted=all.

Iskandar, A., and M. El-Nawawy. 2004. Al Jazeera and War Coverage in Iraq: The Quest for Contextual Objectivity. *Reporting War: Journalism in Wartime, London and New York, Routledge*: 315-332.

Khater, Rami. Digital Protectionism: Preparing for the Coming Internet Embargo presented at the Center for Contemporary Arab Studies 2010 Annual Symposium Information Evolution in the Arab World, Georgetown University. http://bit.ly/902ocN.

Macherey, W., and F. J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 986–995.

Paul, Ryan. 2010. Google Goggles learns to translate, does reasonably well. *Ars Technica*. May 6. http://arstechnica.com/gadgets/news/2010/05/google-goggles-learns-to-translate-does-reasonably-well.ars?utm_source=rss&utm_medium=rss&utm_campaign=rss.

Phillips, Mark. 2010. Bridging the Online Language Barrier: Translating the Internet. *NPR*. April 30. http://www.npr.org/blogs/alltechconsidered/2010/04/30/126420060/bridging-the-online-language-barrier-translating-the-internet.

Rao, Leena. 2009. ¡Nuevo! Google Adds Message Translation To Gmail. *TechCrunch*. May 19. http://techcrunch.com/2009/05/19/google-adds-message-translation-to-gmail/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+Techcrunch+(TechCrunch.

———. 2010. Google Needs French Lessons; Translates "Vimeo" to "YouTube". February 18. http://techcrunch.com/2009/02/18/google-needs-french-lessons-translates-vimeo-to-youtube/?

utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+Techcrunch+(TechCrunch).

Singel, Singel. 2010. Site Hopes Automatic Arabic-English Translation Translates into Peace. *Wired*. February 21. http://www.wired.com/epicenter/2010/02/arabic-english-diplomacy/.

Topolyanskaya, Alyona. 2010. Google Lost in Translation. *Moscow News*. January 28. http://www.mn.ru/news/20100128/55406807.html.

Van Buskirk, Eliot. 2010. Google's Real-Time Voice Translator Could Make Any Language Lingua Franca. *Wired*. February 8. http://www.wired.com/epicenter/2010/02/googles-real-time-voice-translator-could-make-any-language-lingua-franca/.