Analyzing Foreign-Language Social-Media Reaction to Televised Speeches: Lessons Learned

Brian Amanatullah
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

Steven Minton
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

Matthew Michelson
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

Greg Barish
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

Kane See
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

***Abstract*—This paper updates our approach to analyzing social media response to speech events, such as the President's State of the Union address, presenting some new lessons learned in deploying this system. We previously described how we determine which specific lines in the speech resonate with different cohorts of the audience on social media (and how that alignment takes place, etc). In this paper, we describe an update to our approach that allows us to incorporate foreign languages into this analysis. That is, given a speech in English, such as the State of the Union, we describe how we can incorporate Tweets in foreign languages and still perform the clustering. From our development and deployment, we learned two important lessons to share: first, how to bootstrap the data collecting when supporting foreign language analysis, and second, how to order the clustering and translation operations to support a wide variety of languages with minimal cost (in terms of man hours).**

I. Introduction

In our previous work [1] we discussed our approach to clustering social media responses to live speech events, such as the US President's State of the Union address. We designed the approach to allow analysts, social scientists, and policy researchers to measure public reaction to various talking points (and visuals) in a speech. One can measure which points generate the most reaction, including those that may be surprising to the speaker and his/her staff. Further, marketers, journalists, political junkies and the general public can also use the system (and its related, public facing website) to better understand the speech's effect on different groups of people, their opinions, etc.

At its core, our approach is a clustering engine. It takes an input a set of social media posts (Tweets in this case), and a speech, and it creates clusters (e.g., groupings of Tweets that share some common topic in the language). It then classifies each cluster into one of three groups. In turn, our classifications are defined along two axes, which we call "referent" and "temporal." If a cluster refers specifically to a line (or lines) in the speech, we call this cluster "referent." This is in contrast to clusters whose topic is not directly related to the speech itself, such as reaction to what the viewers are seeing on the screen at that moment (e.g., the President's tie). We call these clusters "non-referent." By "temporal," we mean clusters whose Tweets occur in bursts, i.e., there is high volume of Tweets in a short interval. For instance, if many of the Tweets in a cluster happen to fall within a short time-window of one another, we call them "temporal." This usually happens when there is reaction to something specific in the speech's time-line, such a line of the speech that resonates with a large group or when a specific action happens on the screen, such as the camera panning to an audience member frowning. We contrast this behavior with "non-temporal" clusters, which are clusters that refer to the speech generally. For instance, a large number of Tweets commenting on the President's attire or appearance may be grouped together, but occur throughout the speech, since they are not tied to a particular point in time. By definition, we note that non-temporal clusters are also non-referent.

This leaves us with three classifications for a cluster: Temporal/Referent, Temporal/Non-Referent and Non-Temporal (since Non-Temporal cannot be Referent). This classification is important in understanding the reaction to the speech. For instance, if the cluster is Temporal/Referent, then we know it refers to that part of the speech at that time, and therefore that part sparked social reaction. If a cluster is Temporal/Non-Referent, then something in the broadcast outside of the speech itself, such as what is on-screen at that time, prompted reaction. Finally, we can exclude Non-Temporal clusters from the timeline analysis, since they would provide broad color (possibly), but not provide much deeper temporal analysis. We note that in previous work we demonstrated how we perform this classification [1]. Examples of each type of cluster are given in Table I.

Table I makes the clustering clear with a few examples taken from the 2014 State of the Union speech. The first cluster in the table is non-temporal. It reflects a number of Tweets from users watching the State of the Union speech, and Tweeting that they are doing so. The Tweets occur at various times throughout the speech, and do not, as a whole group, refer

TABLE I
DIFFERENT CLUSTER CLASSIFICATIONS

Non-temporal cluster
RT @tjholmes: Unless ur watching CSPAN, u might not know President of the United States is delivering State of the Union address n 50 mi...
What did you think of President Obama's State of the Union address?
#NowWatching President @BarackObama's "State Of The Union Address"...& you should be too!
Referent & Temporal cluster: reflects specific part(s) of the speech
Reacting to the line: "So lets get immigration reform done this year."
The tepid response to @BarackObama mentioning immigration reform tells me it probably won't happen this year.
Reps. Gutierrez (D-IL) & Diaz Balart (R-FL), partners in immigration reform, first to jump to feet after POTUS calls passing this year
RT @JuveMeza: Let's make this a year of action. Congress pass immigration reform or Obama should bypass you. #DACA4ALL #SOTU
Immigration reform has been tried since early in Obama's first term. With a partisan Congress, it's unlikely to happen this year, either.
Non-Referent & Temporal: Does not reflect specific part(s) of the speech
Theodore Roosevelt's 1941 #SOTU address? The Repubs are gonna hammer @WhiteHouse for that during mid terms.
Whoever runs the enhanced live stream sucks. They misspelled televised and showed a picture from "Theodor" Roosevelt's 1941 speech. #sotu
livestream needs a new fact checker. Pretty sure Theodore Roosevelt didn't deliver the 1941 state of the union address.
@WBLittlejohn Their slideshow is a riot. Lots of typos. Showed picture of SOTU from "Theodore Roosevelt" in 1941.

to a specific time period in the speech. The second cluster reflects Tweets about a specific topic that occurs at a specific time period in the speech. That is, the cluster is both temporal and referent (it refers to a line about immigration reform). The final example cluster in the table is temporal but non-referent. Instead, the time period reflected by the cluster is a point, just before the speech started, when the on-screen scroll showed an apparently incorrect historical fact.

While the main contribution of our approach is this ability to cluster the reactions, our previous work was limited because we could only analyze the reaction for the English speaking population. However, social and political scientists might also want to understand the reaction for non-English speaking audiences. Translation is a challenging capability to integrate for a number of reasons. For instance, where to fit the translation into the pipeline can have different implications (e.g., is it done early or late in the processing?). Another implication involves the initial data gathering. For instance, how do we bootstrap the collection process taking into account that our goal is to collect social media in foreign languages? Therefore, the crux of this paper is how we tie automated translation into our approach, focusing on our design choices and their implications (and how we addressed these questions).

Our overall approach for analyzing the social media response to a speech is given in Figure 1, which has been updated to emphasize the new translation capability. Briefly, the full system works as follows. During a speech, the system sources, collects and then cleans a set of Tweets. Next, the focus of this paper, we then translate the Tweets. We will discuss this architectural choice and its implications later. The Tweets are then clustered by topic, and also broken down by cohort, where each sub-cohort represents a group of users responding (for instance each cluster is further sub-divided into cluster members provided by men and those provided by women). Finally, the Tweet clusters are aligned temporally. This temporal alignment is the temporal classification we will focus on in this paper. Once the data is processed, we display the results on a webpage[1] where users can explore and analyze the results.

[1]http://www.socialreactiongroup.com

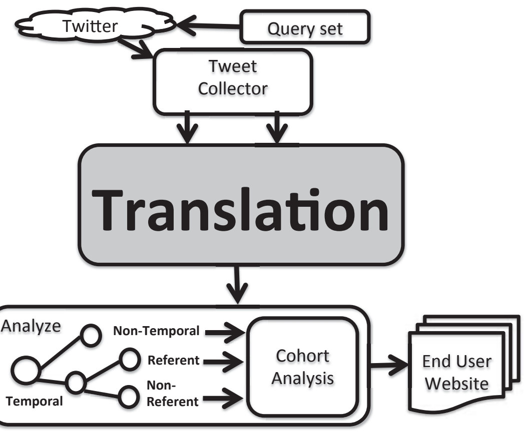

Fig. 1. Our architecture

As mentioned above, we deployed our approach in a public-facing website, where users can select various speeches and analyze the output themselves. A screen shot of this website is given in Figure 2. There are a number of design choices chosen used to make understanding the data more intuitive. The first involves the time aspect of the speech. At the top of the figure, there is a timeline with vertical bars. Each vertical bar represents a cluster linked to some line of the speech. The size the bar reflects the relative volume of Tweets on that line. So, taller bars mean more Tweets on that line (e.g., in the cluster). The bars are also split into colors, showing the proportion that one cohort represents of that cluster, versus the other. Therefore, at a glance, users can hone in on parts of the speech that have high volume, and determine those that looked skewed to a demographic. Then users can click on the bar to get details about that line of the speech and its associated cluster.

In the figure, we have clicked on one of the taller bars in the timeline (it is highlighted in grey) and the site has automatically scrolled to highlight that line of the speech (also highlighted in grey) and updated on the box on the bottom right with information about the cluster. The first item to note is text of the speech transcript on the bottom left. There are two visual aspects to the transcript. First, the size of the font reflects the amount of reaction to that line (e.g., the number of Tweets which is the size of the cluster). Therefore, lines with large text correspond to large clusters (much reaction) and those with small font have smaller clusters (less reaction). We picked a tall bar in the timeline, and as expected the associated line in the speed has huge font (it is one of the biggest clusters). Second, the text is color-coded if it appears that the cluster is skewed to one demographic. In this case, because this cluster skews significantly toward the female cohort, the cluster is colored pink (we color blue for the male cohort).

The second important aspect is the box on the bottom right, titled "Reaction Stats." At the top of this box are two infographics. The speedometer on the right reflects the relative number of Tweets on the speech line. "Going fast" (speedometer to the right) means many Tweets on the topic, and "going slow" (speedometer on the left) reflects few Tweets on the line. This speedometer updates in real-time as one scrolls throughout the speech, giving another view on volume. The pie-chart on the left reflects proportion of each cohort in the cluster for this line of the speech. In the example, we see that most of the cluster was identified as belonging to the female cohort.[2] The pie-chart not only shows the proportion, but also shows the z-score of the associated majority cohort (e.g., female in this case) which is the input to determine whether the cohort is significant or not. Finally, below the pie-chart, scrolling up and refreshing every few seconds, are actual example of random Tweets selected from the cluster.

One of the most interesting and relevant aspects to note for this paper is that the Tweets scrolling on the bottom right are in Spanish, even though the speech itself is in English. That is, we have aligned Spanish language social media reaction to the English speech. In fact, this visualization focuses on Spanish language reaction to the 2014 State of the Union address. As mentioned, Tweets will scroll in the bottom right, but those Tweets will be in Spanish (with their translation included). Therefore, we have allowed social and political scientists to perform deep analysis on language specific reaction to a public speech.

The rest of this paper is organized as follows. Section II describes the lessons that we learned incorporating translation into our analysis pipeline, and Section III contains our conclusions and future directions for this research.

II. Translation and Social Media Reaction

This paper focuses on the lessons we learned integrating translation into our social media analysis pipeline.[3] There were two important lessons, (i) when in the process to perform the translation, and (ii) improved mechanisms for gathering foreign language reaction. Here we discuss those lessons in more detail.

A. *Lesson I: When to translate*

As shown in Figure 1, our process kicks off with set of terms, called the "Query Set" (shown on the top right of the figure above). Ideally, we would collect every feasible Tweet during the speech and then filter out those that are irrelevant. However, this is impractical (both from a data and API service stand point). Therefore, instead we collect queries bootstrapped from terms in the query set. For instance, for the Statue of the Union, the query set might include terms such as "SOTU" (an acronym for State of the Union) or Obama. We are then given large samples of the Tweets that contain these terms. We continue this collection process for the duration of the speech we are analyzing (that is, we being collecting slightly before the speech begins, during its duration, and for a short time period afterward). In this way we ensure that our analysis captures the concurrent aspect of the reaction (though we note that we can align the Tweets to the speech using both topical and temporal analysis, as described above). Once Tweets are collected, they are then clustered and aligned to the speech for analysis.

Therefore, our design choice hinges upon where to perform the translation. We can collect the Tweets, translate them, and then perform our clustering. Or, we can collect the Tweets, cluster them, and then translate them to both align the cluster to the speech and to understand the driver of that cluster (e.g., the anchoring terms in the clustering).

One of the first important lessons that we learned was that it was preferable to translate the Tweets prior to clustering them, rather than vice-versa. This is largely an artifact of *data preparation and understanding.* Specifically, there are three important aspects to the language that dramatically improve clustering. First, it's often beneficial to perform stemming on terms before clustering (we use the Porter stemmer [2]). Stemming is an operation that turns words into their base form. For instance, a plural word might be stemmed into it's singular form, adjectives and adverbs to their base ("terrible" and "terribly" become "terribl"), etc. This allows clustering algorithms to consider variants of the same word to be the same when clustering, so that a high content word, in multiple forms, will yield the same basis for clustering. As an example, both "illegal immigrant" and "illegal immigration" stem to "illeg immigr."

A second important aspect to preparing the data for clustering involves stop words. For instance, clustering in general looks for words in common across text items (Tweets in our case), yet it's not meaningful if two Tweets share the words "and," "the," or "a." These common words that can be ignored are known as "stop words" and they reflect commonly occurring words that are generally not distinctive enough to be useful for tasks such as clustering. Our approach, like most, relies on a common stop word list, and knows to ignore

[2] Again, we note we always treat one pair of mutually exclusive cohorts at a time, e.g., men/women, red-state/blue-state, and users select between them on another page

[3] We thank and acknowledge SDL Language Weaver for helping us leverage their machine translation API in this effort.

Fig. 2. The SocialReactionGroup website

any words on that list. In general, stop words are built by examining word frequencies across a very large language corpus.

Finally, the third aspect is to understand synonymy since this might too form a basis for clustering. For instance, a Tweet that references "clapping" might also be clustered with one about an "ovation" (for instance, when talking about a Temporal but Non-Referent cluster, such as people describing audience reaction). Our synonym knowledge base included both WordNet [3] and some simple synonyms (and word substitutions) we derive from the corpus (such as POTUS is President).

The most important challenge is that while all of these aspects greatly improve clustering, they all depend heavily on the language. This is what drove our choice to translate the Tweets before clustering. Although translation might introduce noise (via imperfect translations), it allows us to leverage our knowledge of synonyms, stop words and stemming, using one common model, regardless of the input language. Otherwise, if we were to cluster prior to translation, while the input language might be cleaner, since it's not translated, it would required us to develop synonyms and a stemming algorithm, for each possible input language, which is challenging (we presume that developing stop word lists is easier to do). This is a costly proposition, requiring both development and testing of the models for each new language.

As clear examples of why this is problematic, consider the following cases from the Spanish language reaction to the State of the Union address. First, Table II shows some Tweets in a cluster centered around the common tokens "una," "por" and "que." We note that these Tweets have little common with one another, but because we did not have a collection of Spanish stop words for clustering, we ended up with this cluster.

Our next example, in Table III demonstrates two Tweets that should have been clustered, but were not. They both focus on investments in the United States versus China. If the system knew that EEUU was a synonym for Estados Unidos ("United States" in Spanish), then the clustering would have been clear. But, because we lack this domain knowledge in Spanish, they are separated.

Both of these serve to illustrate that the lack of language specific stop words, stemming algorithms and synonyms can lead to problematic clustering. Yet, if we translate before clustering, we can continuously re-use our knowledge about English to perform accurate clustering.

TABLE II
A CLUSTER AROUND SPANISH STOP WORDS: UNA, QUE, POR

RT @suvi_94: **Una** pregunta a los **que** votaran x el fmln. De aqu a 20aos,votaran **por** el viejo lin para presidente?Si no lo haran,porq van
RT @UniPolitica: Presidente Obama seala **que** las mujeres merecen igual paga **por** igual trabajo Que esto ocurra en 2014 es **una** vergenza #E
Espero **que** sea **por** **una** apuesta su estado por que soy capas de ir a buscarlo y pegarle.
@CFKArgentina sra presidente mientras ud esta de gira el pais es **una** cochinada aumento todo un veinte **por** ciento **que** dicen **por** el dolar
#EEUU Recin recib **una** nota de un amigo; dice **que** lleva +d 8 horas en trfico, en el mismo lugar, **por** la nieve. #Surprise!

TABLE III
TWO UNCLUSTERED TWEETS

Estados Unidos es el pais mas atractivo para invertir, por encima de China: Barack Obama
"@luisjorojas: Obama" China ya no es el mejor pais para invertir, **EEUU** Lo es #SOTU

B. Lesson II: Data Gathering

The second lesson involves collecting the data. As we discussed, we define a query set which acts as a filter, pulling Tweets out of the "Twittersphere" that we believe might be relevant. We emphasize that we are not making an assumption that any returned Tweet will be relevant. Rather, these are potential candidate Tweets that will be deemed relevant if they cluster appropriately.

TABLE IV
QUERY SETS IN ENGLISH AND SPANISH

English Query Set	Spanish Query Set
barack	democrata
barackobama	eeuu
barak	elpresidente
democrat	estado
obama	estado de la union
obama2014	estadodelaunion
potus	estados unidos
president	presidente
republican	republicano
sotu	
sotu2014	
stateoftheunion	
stateofunion	

When trying to perform a foreign language analysis, we again have two options. We can use the general query set, as described above, detect the language of the returned Tweets,[4] and then analyze buckets of Tweets in the target foreign language. While doable, this can be problematic because we are at the mercy of the sampling procedure of the given access methods provided by the data provider (such as the Twitter API). For example, using the query set shown in Table IV as "English query set," we captured 235,105 Tweets, but only 2.3% (5,523) were in our target language of Spanish.

An alternative is to modify the query set to include terms more aligned with the target language. For instance, for the State of the Union, rather than President, we may include the term "Presidente" in the query set (Spanish for President). This helps ensure a majority of the Tweets are in a specific target language. We contrast this approach by using the query set shown as "Spanish query set" in Table IV, which resulted in 25,136 Tweets, more than 99% of which were in our target language of Spanish.[5] We note that these query terms yield interesting results, which can be viewed on the website for this project.

While this is a small change overall, it is an important lesson to impart because it has a strong impact on the end results. If this step is not taken, then there may not be enough Tweets during data collection to provide a suitable analysis.

III. CONCLUSION

This paper describes our approach to incorporating translation into our social media analysis pipeline. In turn, this allows us to examine the social media reaction to speeches, even when the reaction is in a different language than the original speech. As before, we built a user facing website that allows policy analysts and social scientists to view the results and draw conclusions about the reaction.

Much of this paper focused on two important lessons we learned (and therefore shared) from this process. First, we argue that performing translation before clustering is beneficial, because it allows us to leverage our knowledge of language specific improvements to clustering (such as stop word lists, synonyms and language-specific stemming), even at the cost of potential translation errors. Second, we describe a small modification to our gathering of data that allows us to get more reaction in a target language, so we can ensure that we have a larger sample upon which to run our analysis.

REFERENCES

[1] B. Amanatullah, G. Barish, M. Michelson, and S. Minton, "Temporally aligning clusters of social media reaction to speech events," in *Proc. of the International Conference on Artificial Intelligence*, 2013.
[2] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
[3] G. A. Miller *et al.*, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[4] We used SDL's language detection capability.

[5] We note that we did not include some obvious terms from the English query set, such as "SOTU" and "SOTU2014" in the Spanish set since they would have introduced English language Tweets. This helps explain some of the lower volume, since our query set could have surely been improved.