Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 339 – 343

Procedia Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Big data versus small data: the case of 'gripe' (flu) in Spanish

Antonio Moreno-Sandoval[ab]*, Esteban Moro[ac]

[a]*Instituto de Ingeniería del Conocimiento, Cantoblanco, Madrid 28049, Spain*
[b]*Autonomous University of Madrid, Department of Linguistics, Madrid 28049, Spain*
[c]*Universidad Carlos III de Madrid, Department of Mathematics, Leganés 28911, Spain*

Abstract

Big data is a broad term for data sets so large and complex that traditional data processing applications are inadequate. A new field, Predictive Analytics, is trying to extract value from those big (unstructured) data. In Corpus Linguistics, researchers usually deal with small data. In this paper, we compare the amount and the quality of information with respect to a single topic (flu) in Twitter and in MultiMedica (a corpus of medicine texts).

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: big data; small data; prediction; Google Flu Tool; content analysis; Twitter; medicine corpora

1. The starting point

This paper's main objective is to explore with texts in Spanish the following statement from Lazer et al. (2014:1203):

'Big data hubris' is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

* Corresponding author. Tel.: +34-91-497-5250; fax: +34-91-497-4472.
E-mail address: antonio.msandoval@uam.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.452

In that paper, the authors analyze Google Flu Tool's (GFT) failure to predict accurately the flu season. The previous year, in February 2013, the journal *Nature* reported that GFT was predicting more than double of doctor visits for influenza than the USA medical authorities had registered. Effectively, if in the first 2009 version of GFT, "big data were overfitting the small number of cases" and "GFT was part flu detector, part winter detector", the new GFT version "has been persistently overestimating flu prevalence for a much longer time". Lazer et al. (2014) attribute those errors to 'big data' overestimation (versus the 'small data' that we can find in our language corpora) and to the algorithm dynamics, which pollute and manipulate data by expanding rumors and trending topics.

2. What is Big Data?

Big Data is a term for large and complex data sets, from music downloads to medical records and social media messages. Big Data is usually described by the Four V's by IBM (http://www.ibmbigdatahub.com/infographic/four-vs-big-data):

1. Volume: scale of data (2.3 trillion Gb created each day)
2. Velocity: analysis of streaming data
3. Variety: different forms of data
4. Veracity: uncertainty of data

But the relevant question here is what are 'Big LANGUAGE Data'? In terms of volume, in 2015 these could be some figures:

- 30 billion pieces of content are shared on Facebook every month
- 4 billion hours of video on YouTube each month
- 400 million tweets are sent per day by about 200 million monthly active users
- and in dozens of different languages

In this paper we want to address how can we linguistically process big language data. To explore the problem we replicated Google's experiment using the Twitter messages in Spanish that were geolocalized and included the word 'gripe' (flu).

3. The experiment in Twitter

GFT (http://www.google.org/flutrends/) is based on flu-related searches in Google. They describe the algorithm in the web page:

"We have found a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for "flu" is actually sick, but a pattern emerges when all the flu-related search queries are added together. We compared our query counts with traditional flu surveillance systems and found that many search queries tend to be popular exactly when flu season is happening. By counting how often we see these search queries, we can estimate how much flu is circulating in different countries and regions around the world."

Our approach is basically the same, but instead of counting the searches (obviously we do not have access to those data) we counted the times that 'gripe' is mentioned in Twitter. Three are the conditions for retrieving the messages:

1. Only tweets geolocalized in Spain,
2. That include the word 'gripe'

3. In a given time span: from January 2012 to August 2014.

This way we collected the Spanish gripe Corpus in Twitter, which consists of 2759 tweets (including RT), 327072 tokens (i.e. words and other strings), sent from locations in Spain.

Our hypothesis is that the increased number of messages with the word 'gripe' is a predictor of an approaching peak of case. In order to verify or discard the hypothesis, we checked the prediction against the reported cases by the Spanish Health System, which register the real data sent by doctors in health centers and hospitals.

After observing the data, the first finding was that noise should be eliminated. For instance, all the institutional or press messages should be discarded (1):

"100.000 personas aún no han podido vacunarse contra la gripe" (1)
'100,000 people have not yet got a seasonal flu shot'

Those messages were contaminating the prediction, because they are not identifying a person who was actually sick, which are the good data that we were looking for:

"Estoy en la cama con gripe" (2)
'I'm in bed with the flu'

Therefore, we deleted all messages with an URL from the corpus, and generated a graphic with the prediction and the real information from the flu surveillance system. Fig. 1 shows the results. The peaks of flu predicted by the Twitter messages are several weeks before the actual seasonal epidemic (for instance, between week 80 and 100, or after week 140).

4. Discussion

The messages in Spanish on Twitter also magnify the real cases of flu, as the Google Flu Tool, but because of different causes. To our knowledge, the problem is the 'Veracity' in data: an analysis of the Flu Corpus must be done to discover which factors contribute negatively to the prediction. In other words, we must distinguish 'good data', as in (2), from 'bad data'.

A first analysis shows us that the figurative language is a source of error. Message (3) provides a sample of a joke, rather frequent in Twitter:

"No sé si es un constipado virulento o una gripe virurápida" (3)
'I don't know if it is a virulent cold or a viru-fast flu'

"Virurápida" is an invented word in Spanish, where the user plays with the contrast between the ending "-lento" ('slow') and "rápida" ('fast'). It is, of course, difficult to separate real cases from figurative ones, and it requires a good management of pragmatic aspects (intention, irony, metaphor).

Another source of problems is when the text contains 'gripe' but the speaker doesn't suffers from it, as in (4):

"Acabo de ver un anuncio de Gelocatil gripe y me he acordado de @" (4)
'I just saw an ad of XXX flu and I have remembered @'

Cases like (4) are also hard to discard by automatic means. The reader must remember that both GFT and our 'gripe' detector are programs, because humans, in a reasonable time frame, cannot process big data. We estimate that cases

like (3) and (4) are over 100 in our Flu Corpus, and all contributing to overestimation.

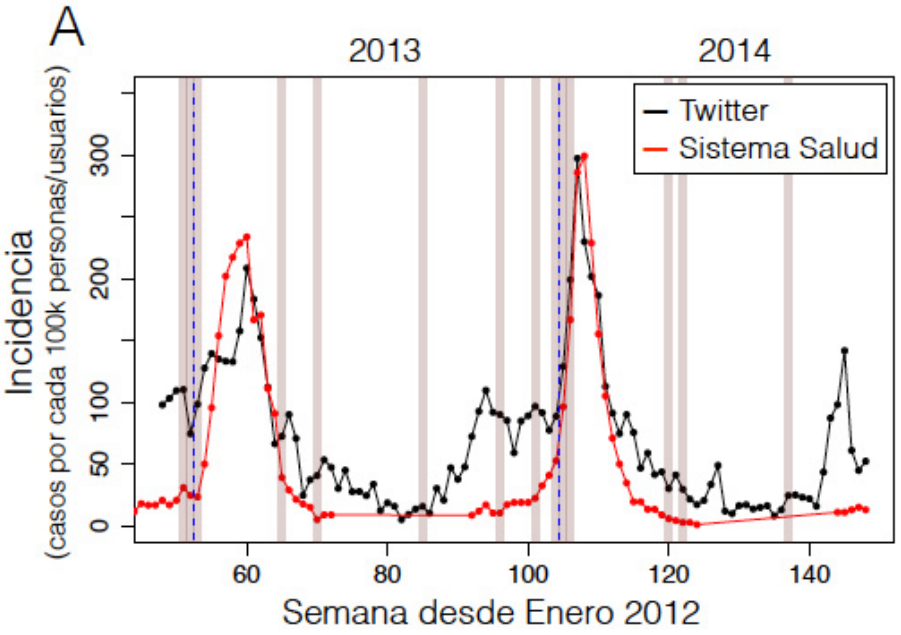

Fig. 1. Twitter prediction against the Health System surveillance

4.1. Introducing some refinements in the predictor

In order to improve the precision of our program, we incorporate to the corpus samples containing synonyms or variants, such as "griposo" ('flu-like'), "gripazo" ('bad flu'). But the most useful improvement has been the detection of sequences of words (n-grams) that indicate that someone has caught the disease (5-8):

"con la gripe en casa" (5)
'with the flu at home'

"menudo gripazo he pillado" (6)
'I caught a bad flu'

"con tos y moqueando" (7)
'cough and runny'

"la gripe me mata" (8)
'this flu is killing me'

However, our most important finding was that there are a few lexical and syntactic patterns such as those in the Twitter corpus. Fig. 2 shows collocations of TENER (to catch) and GRIPE.

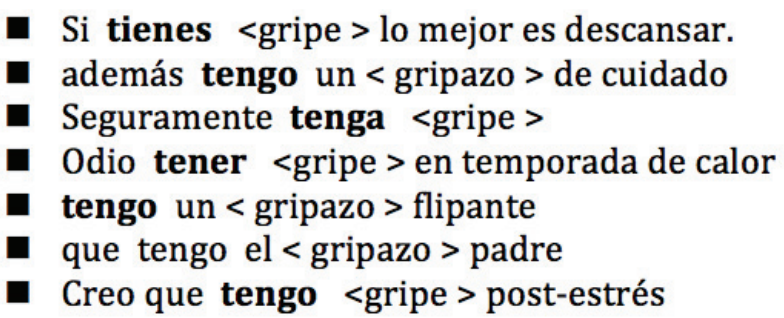

Fig. 2 Collocations of TENER & GRIPE in the Twitter corpus

4.2. Flu patterns in a small data set

The next step in the experiment was to replicate the search in a corpus of medicine texts. For this, we use the Spanish set of the MultiMedica corpus (Moreno & Campillos 2013), which consists of 4,031,174 words in 4,204 documents. We analyzed the corpus with The Sketch Engine (Kilgarriff et al 2014) using the Word Sketch function.

The query returned only 341 occurrences of "gripe", in contrast with the 2759 cases in the Twitter corpus. However, the patterns were richer, since 'gripe' collocates with:

- Nouns: 'virus', 'brote' (outbreak), 'caso', 'estación' (season), 'azote' (scourge), 'epidemia' (epidemic), 'vacuna' (vaccine), 'temporada' (season).
- Verbs: 'padecer' (get), 'sufrir' (suffer), 'tratar' (treat), 'superar' (overcome).

These collocations counterbalance the 'scarcity' of information in the Flu Corpus on Twitter.

5. Conclusions

Our data support the hypothesis of Lazer et al. (2014) that states that "instead of focusing on a 'big data revolution,' perhaps it is time we were focused on an 'all data revolution,' where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world." In terms of corpora, small but well selected collections of linguistic data should be combined with large repositories from internet and social networks, since sometimes 'small data' offer information that is not inferred from 'big data'.

References

Kilgarriff et al. (2014). The sketch engine: ten years on. *Lexicography ASIALEX*, 1, 7 - 36.

Lazer, D. et al. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, 343(6176), 1203 - 1205.

Moreno, A., and Campillos, L. (2013). Design and annotation of MultiMedica: a multilingual text corpus of the biomedical domain. *Procedia. Social and Behavioral Sciences*, 95, *Selected Proceedings of the 5th International Conference in Corpus Linguistics* (pp. 482-489). Amsterdam: Elsevier.