Political Footprints: Political Discourse Analysis using Pre-Trained Word Vectors

Christophe Bruchansky
Plural think tank

How political opinions are spread on social media has been the subject of many academic researches recently, and rightly so. Social platforms give researchers a unique opportunity to understand how public discourses are perceived, owned and instrumentalized by the general public. This paper is instead focussing on the political discourses themselves, and how a specific machine learning technique - vector space models (VSMs) -, can be used to make systematic and more objective discourse analysis. Political footprints are vector-based representation of a political discourse in which each vector represents a word, they are produced thanks to the training of the English lexicon on large corpora of text. This paper describes a simple implementation of political footprints, some heuristics on how to use them, and their application to four cases: the U.N. Kyoto Protocol and Paris Agreement, the 2008 and 2016 U.S. presidential elections. The reader will be given some reasons to believe that political footprints produce meaningful results, suggestions on how to improve them and validate the results.

Context and Methodology

Vector space models (VSMs) represent words in a continuous vector space where semantically similar words are mapped to nearby points (*TensorFlow website*, 2017). VSMs are made possible thanks to algorithms that can analyse large corpora of text and determine how likely two words appear in a same passage (words co-occurence): the more often two words appear together, the closer these algorithms will place their vectors. The resulting vector space models are not only statistically significant, but also have a semantic value. This is due to the distributional hypothesis stating that words appearing in the same context share semantic meaning (D. Turney & Pantel, 2010).

Vector space models allow machines to classify documents by meaning, detect opinions, understand natural language, and more interestingly in our case create "semantic word clouds" (Xu, Tao, & Lin, 2016). They “can be used to provide a bird’s-eye view of different text sources, including text summaries and their source material. This enables users to explore a text source like a geographical map” (Heuer, 2015).

Publicly available pre-trained vector space models have made their appearance in recent years. Examples are word2vec models from Google (Mikolov, Chen, Corrado, & Dean, 2013), GloVe from Stanford University (Pennington, Socher, & D. Manning, 2014) and fastText from Facebook (Bojanowski, Grave, Joulin, & Mikolov, 2016). All have accelerated even further the use of the technology.

Cultural and political sciences are a natural fit, and researches have started using VSMs to analyse political opinions. A large proportion focus on social media and allow for instance the categorization of election-related tweets (Vijayaraghavan, Vosoughi, & Roy, 2015), others focus on the political discourse itself, such as argument based (Hirst & Wei Feng, 2015) and semantic word clouds analysis (Chah, 2017). This paper belongs to the latest category.

A political footprint is vector-based representation of a political discourse in which each vector represents a word. Political footprints are computed using machine learning technologies, which allows for systematic and more objective political analysis.

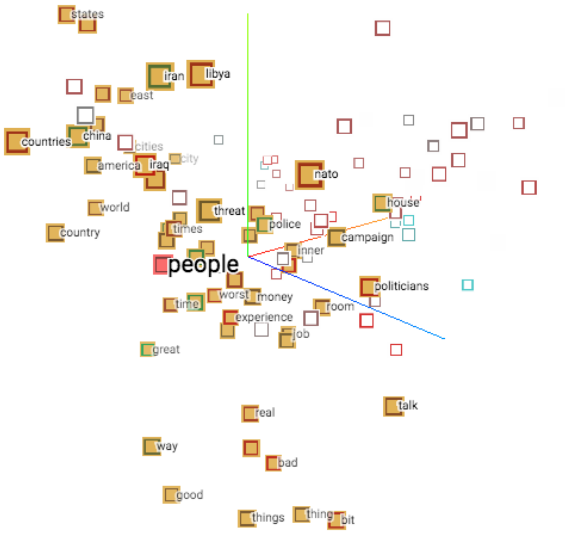
Political footprints are unbiased in the sense that they are not relying on the researcher’s political knowledge or beliefs. They are however very much dependent on the corpus they were trained with (Wikipedia, Google News, etc.), and more generally on the cultural context any political discourse is originating from.

Political footprints focus exclusively on what a statement or speaker says. They are, in this sense, very different from other popular word cloud analysis that focus on news and social media trends: the emphasis is on what a speaker has in his or her control.

Our purpose is to make a proof of concept: to use existing technologies, implement a simple version of political footprints and examine if the results are in any way meaningful. It is based on the following technologies:

- IBM Watson (*IBM Watson Natural Language Understanding*, 2017): returns a list of entities and keywords included in a text, with for each a relevance, sentiment, and emotion score.

Figure 1. Donald Trump's political footprint during 2016 U.S. general election debates, with the closest words to "people" highlighted.



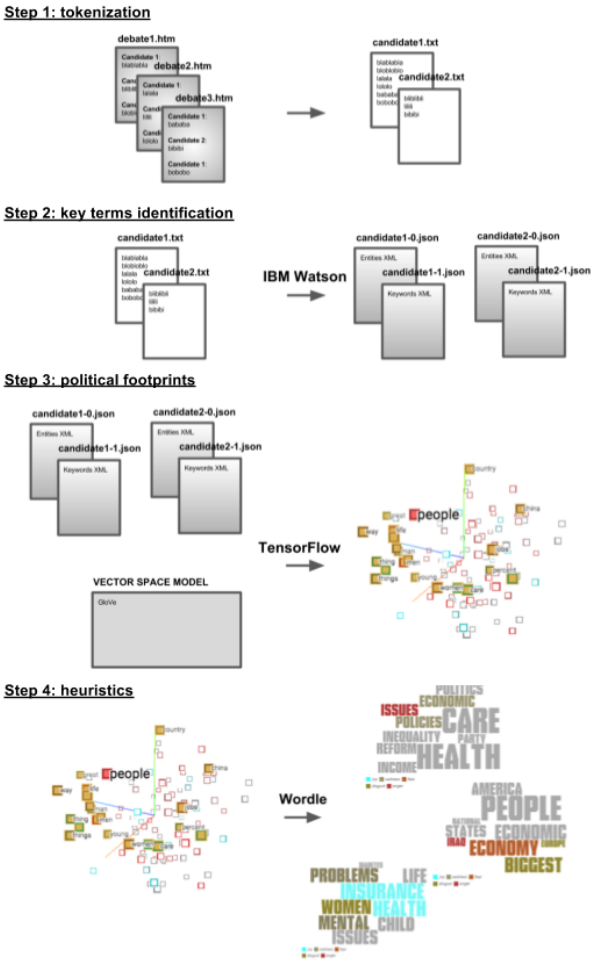
- GloVe from Stanford University (Pennington et al., 2014): package featuring word vectors trained using Wikipedia (2014) and Gigaword 5 (2011).
- FastText from Facebook (Bojanowski et al., 2016): alternative to GloVe that supports many non-English languages.
- TensorFlow and TensorBoard (*TensorFlow*, 2017): open-source software library for machine intelligence, its TensorBoard interface is particularly useful for VSM quick visualisation.
- Wordle (Feinberg, 2017): free online tool to render word clouds.

Here are on figure 2 the steps that we have followed in this paper. The first is to process raw data (tokenization), the second is to identify key terms for each speaker, the third is to compute political footprints, and the last to interpret them using some heuristics. All scripts and data are available on a Github page (Bruchansky, 2017b).

Raw data

Four use cases were chosen to prove our concept: the Kyoto Protocol, the Paris Agreement, the 2008 and 2016 U.S. presidential elections. They Kyoto Protocol and Paris Agreement were taken from the United Nations website (*United Nations Framework Convention on Climate Change*, 2017), and the U.S. elections televised debates from the American

Figure 2. Steps to compute and analyse political footprints.



Presidency Project (Peters & T. Woolley, 2017), a nonprofit and nonpartisan source of presidential documents hosted at the University of California.

They have been chosen for several reasons. Firstly, most readers will be familiar enough with the topics to compare the results with their own intuition. Secondly, they are accessible to anyone wishing to consult the data.

Thirdly and in the case of political elections, televised debates were chosen over rallies and other public declarations because they are assumed to be representative of the U.S. political landscape: processes defining which candidates were invited to the televised debates, how long they could speak, and on what matter are assumed to be objective enough for our purpose. This way, we can easily apply the same method to any election. However, the reader should be aware that no debate can perfectly reflect a political landscape. In the case of the U.S. elections, this lead us to exclude independent and third party candidates, which is arguably regrettable.

The goal of this study is to focus exclusively on a political

discourse, not on its context or how it has been received. We have thus removed any audience reaction from the transcripts (i.e. "applause", "laugh", etc) and all questions from the moderators. We lose some important information in doing so, making it sometimes difficult to understand what a candidate's answer was about, but it's the price to pay to only take into account candidates' own words.

Key terms identification

Our choice to use IBM Watson is a convenient one: it allows to run our analysis on any personal computer, and with fast results. It comes with a cost though: there is not much control or explanation on how the terms are selected and weighted. For instance, IBM Watson Natural Language Understanding generates several json files per text, including one for its entities and one for its keywords. It's not clear how the two lists are created, and it is assumed that entities are more relevant than keywords since they are more structured objects (entities have types and subtypes). If a term exists in both files, our scripts only keep the entity version.

IBM Watson's emotion detection has not been totally reliable. A better solution could be to use an emotion detection mechanism that can adapt to political language (Rheault, Beelen, Cochrane, & Hirst, 2016). In any case, it is important to keep in mind that an emotion attached to a word is not necessary targeting that word. An angry feeling detected when using the word "people" doesn't mean that the discourse is necessary expressing anger towards people, but that speaking about people generates some anger. Sarcasm is another case that makes it difficult to understand a candidate's emotions.

No other commercial (i.e. Google CloudPlatform) or open source solutions (i.e. NLTK or Stanford CoreNLP) have been tested as part of this paper. Words selected by the default IBM Watson natural language understanding API were for the most part corresponding to our intuition. They were considered good enough for our proof of concept, in the sense that if using a generic tool such as IBM Watson can provide meaningful data, more sophisticated implementations could only improve our results.

Let's look for instance at key terms used in the Kyoto protocol and the Paris agreement. The following word clouds have been computed by IBM Watson and rendered using Wordle, see figure 3 and figure 4 .

Here are few indications of their accuracy:

- The two texts have a neutral sentiment, with perhaps a slightly more positive tone in the Paris agreement, which is what one would expect for international agreements.

- All the terms with a high relevance score have a direct connection with climate change.

Figure 3. Most relevant terms used in Kyoto Protocol.

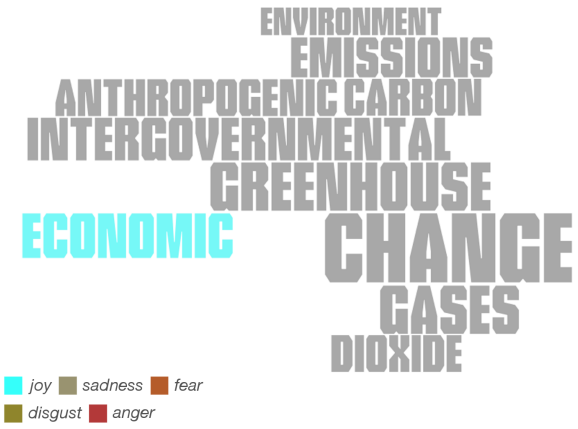


Figure 4. Most relevant terms used in Paris agreement.

- Four out of nine terms appear in both word clouds. The former cloud gives emphasis to change and intergovernmental actions, and the later to economics and sustainability. Both remain, however, consistent even though they are based on different texts.

Political footprints

Our choice of pre-trained vector space model was GloVe6B, which is based on Wikipedia 2014 and Gigaword 5. We've made this choice primarily because of its public availability, its popularity among researchers, and some convenient features, such as being able to quickly run our scripts using a 50 dimensions space, before extending it to 300 dimensions.

We have however encountered some issues. The first is that this space is word based and doesn't include tokens such as "Wall Street" or "New York Times". This wasn't so much an issue in our use case since we were familiar with the data. We could easily guess the associations: we knew that "street" was a relevant word in Bernie Sanders's politi-

cal footprint because of Wall Street. However this requires unnecessary interpretation from the researcher and could be easily avoided with better suited vector space models.

A second issue is the date of the reference corpora (2014 and 2010) compared to the dates of the U.S. elections (2008, 2016). We could only assume that words meaning and usage remained unchanged during that period. An improvement would be to have access to yearly updates of vector space models. But this would create new issues, such as how to compare texts written in different years if we don't base our analysis on a single reference model.

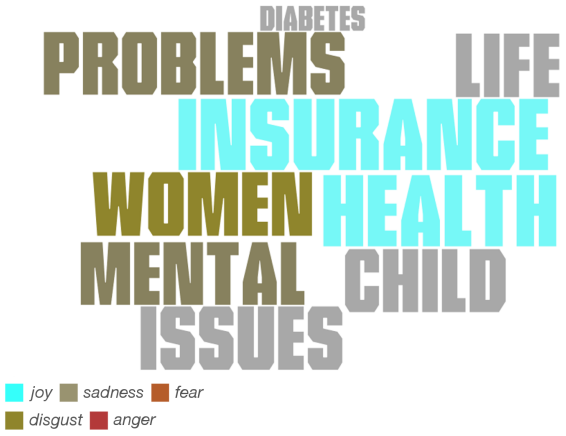
A last problem is that GloVe pre-trained vector space models are currently only available in English. An alternative is to use FastTest from Facebook, which supports many languages and is compatible with our scripts.

Let's now look at how we can practically use political footprints.

Heuristic 1: main themes of a discourse and what they mean

The first heuristic could be described as a clustering technique in which we leverage the relevance score obtained from IBM Watson: we take the most relevant words from a discourse, and select for each the closest words in the vector space model. We obtain a series of clusters each centered around a relevant term. These clusters can then be visualised using Wordle, with word sizes corresponding to their relevance, and word colours to emotions attached.

Figure 5. Hillary Clinton' words that were related to the affordable care act (U.S. election televised debates).



The same heuristic was applied to both 2008 and 2016 U.S. elections and results were surprisingly consistent with our intuition. What makes us confident about the results is the appearance of some terms that have been widely commented during 2016 election: "women" and "health care" by Hillary Clinton (see figure 5), "China" and "Nafta" by Donald Trump (see figure 6). A full analysis is available on the project homepage (Bruchansky, 2017a).

Figure 6. Donald Trump's words that were related to trade deals (U.S. election televised debates).

What's important to understand, and what is at the core of political footprints, is that word similarities have been identified without any human intervention: they have been discovered by machines based on how frequently words appear together, either on Wikipedia or other large corpora of text. This is what allows us to make an unbiased analysis of political discourses. To be clear, there is a strong cultural bias: the one coming from how these words have been used on Wikipedia, news feeds, etc. But it is not coming from the researcher performing the analysis.

According to another of our paper (Bruchansky, 2017a), machine learning techniques such as this one belong to structuralism : word similarities are not inferred from the discourse we analyse but from the large corpus of text that was used to train our words. Comparing distances (cosine similarity) between words doesn't provide any information about the discourse we study, but only about the culture and language it is based on; information about the discourse lies in the choice of these words instead of others, their relevance and emotion attached. According to Claude Lévi-Strauss, often referred as the father of structuralism, cultures are systems analyzed in terms of the structural relations among their elements. Universal patterns in cultural systems are products of the invariant structure of the human mind (*In Encyclopædia Britannica*, 1998).

As a note, this heuristic was performed manually but could be automated using k-means clustering and other unsupervised machine learning techniques. It would then be possible to compare their standard implementation with ones that take advantage of words relevance. Using the latter as centroid candidates would make sense intuitively since they are "at the center" of the discourse.

Heuristic 2: compare how a theme is appropriated by each participant

In this heuristic, we choose a term or theme of our choice. Let's say we are interested in American "values". For this term, we select the 20 closest ones (i.e. "social", "civilization", "inequality", "liberty", etc.) and see how many participants have used them. We choose a couple of terms, ideally those that have been used by many participants and that have many different meanings. "Social" and "interest" were picked in our 2016 US election example, but it could also have been "ethical", "principle" or "freedom": any term that can be used with different sets of words depending on a participant's political views. And we compare their semantic word clouds using our first heuristic.

Here are on figure 7 the main value-related terms used during the 2008 U.S. presidential election. The more a term was shared between presidential candidates (during primaries and general elections), the bigger the word is on our cloud.

Figure 7. Value-related terms used during the 2008 U.S. presidential election (televised debates).

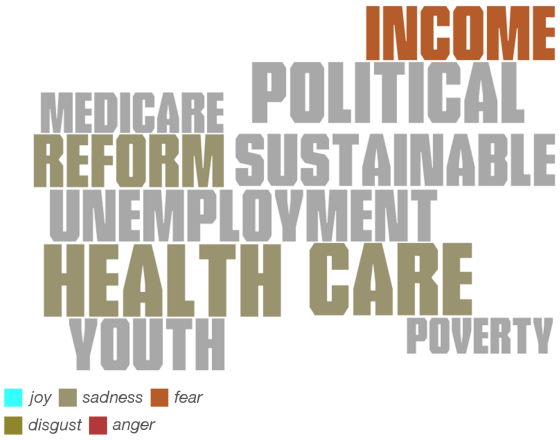
Most of the same terms appear during the 2016 U.S. presidential election, see figure 8. This is an indication of both the robustness of political footprints and the consistency of the terms used from one election to the other.

Figure 8. Value-related terms used during the 2016 U.S. presidential election (televised debates).

We have, with the two word clouds above, the confirmation that social values played an important role in the two elections. Let's now see how two candidates have articulated the social topic (key terms situated nearby on the GloVe vector space model).

Figure 9. John McCain's topics that were related to social matters (2008 primaries debates).

Figure 10. Bernie Sanders' topics that were related to social matters (2016 primaries debates).



Health care was undoubtedly a big theme in the two elections. Focus was on national security and radical Islam for John McCain, see figure 9. Poverty, unemployment, and youth were important topics for Bernie Sanders, see figure 10. These are indications that the semantic word clouds that we have obtained thanks to political footprints are meaningful representations of the 2008 and 2016 U.S. elections. See a full analysis along with more evidences on the project homepage (Bruchansky, 2017a).

Using a unique reference model is both the biggest weakness and strength of our political footprints. On one hand, it doesn't accurately represents what were words meaning in each campaign (years of elections were not the same than the years of our VSM), but on the other hand it provides an

easy way to compare the two: the semantic relation between words have been defined at the GloVe level, we can thus compare texts that haven't necessarily used the same words to address a same issue, and we can use instead the closest ones in each model.

Heuristic 3: sort discourses by style and affinity

The goal of this last section is to test if some general properties of our political footprints could be used for political analysis. For this, we visualized political footprints coloured by relevance, sentiment, and emotions. We calculated their centroids (centers of semantic gravity) and compared their distance with one another.

None of these approaches were, at least in our current implementation, conclusive. A few interesting properties were revealed in our US election example, but not enough to exclude that they were mere coincidences. Bernie Sanders' political footprint was more focussed than the others, see figure 11 : Wall Street was detected as being by far the most relevant of his topics (along with Hillary Clinton). But this might also have been due to a strategy of repeating the same words instead of describing his views in different ways. We can't conclude that Bernie Sanders was fundamentally more focussed. Hillary Clinton's emotions were less visible. But as explained above, this might have been due to her sarcastic tone, and more subtle ways to express her emotions. Finally, except maybe the fact that Bernie Sanders' centroid was the furthest from Donald Trump's, centroids were not corresponding to any of our intentions.

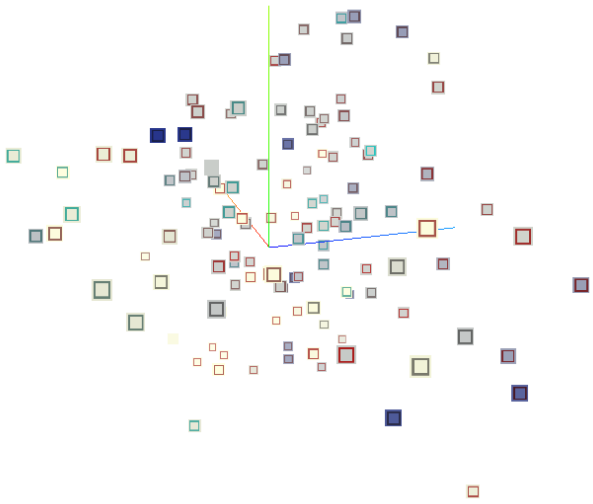
Using centroids is arguably a simplistic way to look at political footprints: they are not taking into account relevance, sentiments and emotions. In this model, seeing Islam as a positive and secondary topic counts exactly the same as seeing Islam as a negative and central topic.

Conclusions

In this paper, we have presented a very simple implementation of political footprints: one that can be computed on any personal computer and still leverage some of the semantic embedded in word space models. Identifying "real" meaning of a political discourse is in itself an impossible endeavour. There is no such thing as real meaning. Political footprints fit however surprisingly well with the discourse of each US presidential candidate and how they have been covered in the press. They have been obtained with nearly no human intervention, which makes them a potentially very useful tool for political discourse analysis.

A better way to test their validity would be to compare resulting word clouds with those that the public, commentators, and authors of a discourse themselves could draw. We could ask volunteers to recognise a discourse based on its political footprint, and measure what's the success rate. Or we could compare political footprints of a same politician in various context (debates, rallies, public declarations, twitter) and test how consistent they are.

Figure 11. Bernie Sanders political footprint strongly oriented towards Wall Street and Hillary Clinton.



We have suggested various improvements, such as using a domain specific and open source solution instead of IBM Watson, using a pre-trained vector space model supporting multiple words as tokens (i.e. "Wall Street"), and improving our heuristics with existing clustering techniques.

Political footprints are meant to be used by anyone interested in studying political discourse, in highlighting some of the different meanings a word can have in politics, and the underlying semantic tensions underlying any political debate. It is hoped that such tools will help researchers and commentators focus a bit less on public opinion and news trends that nobody fully own, and regain interest in the political discourse itself. Political footprints gives emphasis to what politicians have in their control, and in their responsibility: their own words.

"When a man commits himself to anything, fully realising that he is not only choosing what he will be, but is thereby at the same time a legislator deciding for the whole of mankind – in such a moment a man cannot escape from the sense of complete and profound responsibility" (Sartre, 1946).

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching word vectors with subword information.* `https://github.com/facebookresearch/fastText`.

Bruchansky, C. (2017a). *Political footprints.* `https://plural.world/research/political-footprints/`. Plural think tank.

Bruchansky, C. (2017b). *Political footprints: Analysing political discourses using pre-trained word vectors.* https://github.com/Plural-thinktank/pfootprint. Plural think tank.

Chah, N. (2017). *word2vec4everything.* https://github.com/nchah/word2vec4everything.

D. Turney, P., & Pantel, P. (2010). *From frequency to meaning: Vector space models of semantics* (Vol. 37). doi: http://www.jair.org/media/2934/live-2934-4846-jair.pdf

Feinberg, J. (2017). *Wordle.* http://www.wordle.net/.

Heuer, H. (2015). *Text comparison using word vector representations and dimensionality reduction.* https://arxiv.org/abs/1607.00534.

Hirst, G., & Wei Feng, V. (2015). *Automatic exploration of argument and ideology in political texts.* http://ftp.cs.toronto.edu/pub/gh/Hirst+Feng-ECA-2016.pdf.

Ibm watson natural language understanding. (2017). https://www.ibm.com/watson/developercloud/natural-language-understanding.html.

In encyclopædia britannica. (1998). http://www.britannica.com/science/structuralism-anthropology.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* https://code.google.com/archive/p/word2vec/.

Pennington, J., Socher, R., & D. Manning, C. (2014). *Glove: Global vectors for word representation.* https://nlp.stanford.edu/pubs/glove.pdf.

Peters, G., & T. Woolley, J. (2017). *The american presidency project.* http://www.presidency.ucsb.edu/.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). *Measuring emotion in parliamentary debates with automated textual analysis.* doi: https://doi.org/10.1371/journal.pone.0168843

Sartre, J.-P. (1946). *Existentialism is a humanism.* https://www.marxists.org/reference/archive/sartre/works/exist/sartre.htm. (translated by Mairet, Philip)

Tensorflow. (2017). https://www.tensorflow.org/.

Tensorflow website. (2017). https://www.tensorflow.org/tutorials/word2vec.

United nations framework convention on climate change. (2017). http://unfccc.int.

Vijayaraghavan, P., Vosoughi, S., & Roy, D. (2015). *Automatic detection and categorization of election-related tweets.* http://soroush.mit.edu/publications/pv_sv_dr_icwsm2016.pdf.

Xu, J., Tao, Y., & Lin, H. (2016). *Semantic word cloud generation based on word embeddings.* doi: https://doi.org/10.1109/PACIFICVIS.2016.7465278