NLP for political tagging

Attribute extraction for campaign main topics identification

N. J. Hurst[1]

[1]School of Engineering and Sciences
Instituto Tecnológico y de Estudios Superiores de Monterrey

Computational Techniques for Machine Learning
October 2017

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

Problem description

- Many Mexican EdoMex governor election candidates use social networks to react to ongoing events
- Each reaction reflects somehow topics comprised on the candidate's campaign
- Experts classify topic attitudes as follows:
  1. Proactive
  2. Reactive
  3. Aggressive
  4. Pro-vote

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

Hypothesis

1. Previous attribute extraction based on an ensemble of natural language techniques and classifiers will improve overall attitude classification precision over unseen tweets.
2. These techniques are enumerated in 'Proposed techniques' on page 12 and each of those will be tailored to a single aspect of the problem.
3. Due to the open nature of the corpus recall should be ignored.
4. We will test the extracted attributes on a bag-of-words simple classifier to measure the degree of improvement. We expect to improvement reaches at least 10% more *precision*.
5. The contribution is that in Mexico this kind of attribute extraction using spanish and political corpuses has never been done.

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

Objectives

1.

S.M.A.R.T. objective

Specific

Improve classification precision by adding a new set of features to each tweet tuple.

Measurable

Crossover validation will be used to measure precision deltas. So far, no parametrization is needed.

Assignable

Will be done by the student. Periodic weekly reviews on Fridays.

Realistic

An ensemble of 3 classifiers will be used.

Time-related

Weekly reviews will be made until project delivery on November.

Each classifier extract a set of attributes

What are the current trending topics? + The candidate tweets about the topic + The candidate mentions some other candidates

Bag-of-words

What is the attitude of a tweet? (regardless of who is the candidate)

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

Proposed machine learning techniques

- Word2Vec
  - Evaluates context though shallow 2-layer neural networks
  - Requires a very large corpus, over 10 million words
- Template matching
  - Uses syntactical structure proposed by Noam Chomsky
  - Can also be learned
- Bag of words
  - For evaluating other candidates tweet mentions

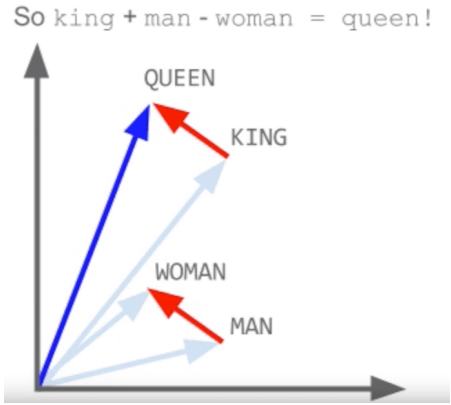

Figure: Word2Vec

Tecnológico de Monterrey

Proposed machine learning techniques

- Word2Vec
  - Evaluates context though shallow 2-layer neural networks
  - Requires a very large corpus, over 10 million words
- Template matching
  - Uses syntactical structure proposed by Noam Chomsky
  - Can also be learned
- Bag of words
  - For evaluating other candidates tweet mentions

Type	Template	Example	Frequency
Verb	NP_1 *Verb* NP_2	X established Y	38%
Noun–Prep	NP_1 *NP Prep* NP_2	X settlement with Y	23%
Verb–Prep	NP_1 *Verb Prep* NP_2	X moved to Y	16%
Infinitive	NP_1 **to** *Verb* NP_2	X plans to acquire Y	9%
Modifier	NP_1 *Verb* NP_2 *Noun*	X is Y winner	5%
Noun-Coordinate	NP_1 (**,** | **and** | **-** | **:**) NP_2 *NP*	X-Y deal	2%
Verb-Coordinate	NP_1 (**,**| **and**) NP_2 *Verb*	X, Y merge	1%
Appositive	NP_1 *NP* (**:**| **,**)? NP_2	X hometown : Y	1%

Figure 22.3 Eight general templates that cover about 95% of the ways that relations are expressed in English.

Figure: Template matching

Tecnológico de Monterrey

Proposed machine learning techniques

- Word2Vec
  - Evaluates context though shallow 2-layer neural networks
  - Requires a very large corpus, over 10 million words
- Template matching
  - Uses syntactical structure proposed by Noam Chomsky
  - Can also be learned
- Bag of words
  - For evaluating other candidates tweet mentions

Bag of Words Model: Binary vectors

- First, normalize (in this case, lowercase)
- Second, compute vocabulary and sort
  - a arrived damaged delivery fire gold in of shipment silver truck

	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
Shipment of gold damaged in a fire	1	0	1	0	1	1	1	1	1	0	0
Shipment of gold arrived in a truck	1	1	0	0	0	1	1	1	1	0	1

6

Figure: Bag of words

Tecnológico de Monterrey

Contents

Tecnológico de Monterrey

References

Books

Peter Norvig. *Artificial Intelligence: A Modern Approach*. 2012. ISBN: 9780123969590. DOI: 10.1016/B978-0-12-396959-0.00001-X.

Articles

Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: (Feb. 2014). arXiv: 1402.3722. URL: http://arxiv.org/abs/1402.3722.

Tecnológico de Monterrey

Collaborators

Post-doctorate professors

1. Miguel Ángel Medina , PhD degree in Artificial Intelligence from the Ciego de Ávila University in 2004

Figure: Migue

Post-doctorate collaborators

1. J. Benito Camiña, PhD. Dissertation. "The Windows-Users and -Intruder simulations Logs dataset (WUIL): An experimental framework for masquerade detection mechanisms"

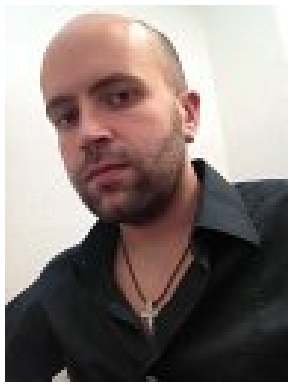

Figure: Benito

Tecnológico de Monterrey

Thanks for your attention! Any questions?

Tecnológico de Monterrey