Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Time corpora: Epochs, opinions and changes

Octavian Popescu, Carlo Strapparava *

Fondazione Bruno Kessler FBK-irst, I-38123 Trento, Italy

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form 6 March 2014
Accepted 9 April 2014
Available online xxxx

Keywords:
Natural language processing
Large corpora analysis
Sentiment and social analysis
Diachronic analysis
Epoch detection

ABSTRACT

By using large corpora of chronologically ordered language, it is possible to explore diachronic phenomena, identifying previously unknown correlations between language usage and time periods, or epochs. We focused on a statistical approach to epoch delimitation and introduced the task of epoch characterization. We investigated the significant changes in the distribution of terms in the Google N-gram corpus and their relationships with emotion words.

1. Introduction

Traditionally, scholars of history define epochs according to their deep knowledge and understanding of facts over a long stretch of time. Intuitively, in order to define a new epoch, both a big social impact of a series of events and new issues, which arouse the social interest, must be observed. However, it is hard to define what makes a feature "distinctive" or an event a "great change". It is even harder to evaluate and measure the impact of a series of changes in society in an objective way [17]. Since the advent of regular newspapers and the industry of mass media, written information has represented a mirror of the interests of society. A social event is relevant only if people pay attention to it and comment on it. A major change in society is reflected in the frequencies with which a set of topics is mentioned in mass media, some of them becoming mentioned more often than previously, while some others are no more of interest. Furthermore, specific epochs typically develop a particular form of wording or rhetorical style.

In this paper we describe a computational approach to *epoch delimitation* on the basis of word distribution over certain periods of time and analyze the opinion change phenomenon. A big quantity of data, chronologically ordered, allows accurate statistical statements regarding the covariance between the frequencies of two or more terms over a certain period of time. By discovering significant statistical changes in word usage behavior, it is possible to define epoch boundaries. We show that it is possible to distinguish a series of limited periods of time, spanning at most three years, within which non-random changes affect the joint distribution of terms. Between two such short periods (i.e. the boundaries) no statistical significant changes are observed for decades, and thus we can refer to it as an epoch. The distributions of the considered terms before and after boundaries are distinct.

We introduce the task of *epoch characterization*. Certain words carry with them an emotional charge, like *joy*, *fear*, *disgust*, etc. Within a given epoch, we can analyze the distribution of emotion words and their co-occurrences with the set of terms considered indicative for epoch definition. The pattern of these co-occurrences constitutes a blue-print of emotional tendencies with respect to some particular topics in the society within a certain period. Given an arbitrary sample of data from a given, but unknown period of time, the task consists in correlating the emotional pattern of the data with the one of an epoch from which the data comes. The experiments reported here show that this task is feasible and sensible results are obtained. The analysis is extended to individual words in order to understand the phenomenon of opinion change. The distribution of the collocation between a target word and emotion words is analyzed. The opinion change phenomenon is rather rare, but when it occurs it may signal that the society has undergone important changes.

We compiled two set of terms from two different domains: the first set comes from the socio-political domain, while the second comes from sport.[1] The socio-political lexicon contains 761 words, such as: *capitalism*, *civil disobedience*, *demagogue*, *democracy*, *dictator*,

* Corresponding author. Tel.: +39 0461314589.
E-mail addresses: popescu@fbk.eu (O. Popescu), strappa@fbk.eu (C. Strapparava).

[1] www.democracy.org.au/glossary.html, en.wikipedia.org/wiki/Sport.

http://dx.doi.org/10.1016/j.knosys.2014.04.029
0950-7051/

chickenhawk, *education*, *government*, *peace*, *war*, etc. The sport lexicon contains 34 words naming different sports, like *football*, *baseball*, *golf*, *fishing*, etc. The frequency of these terms and their covariance is analyzed over the years and non-random changes are found according to the methodology presented in Section 3. The methodology itself is purely statistical and it does not depend in any way on what the list contains. We could have equally chosen terms from other domains. In general, the boundaries are specific to each domain, but the analysis that we carried out showed that important similarities are observed across domains as well.

Considering all the terms from the cited dictionaries, no weights associated, and then extracting all the Google N-grams containing these terms, guarantees that there is no discriminative bias associated with this analysis. Due to the inherently limited processing power humans have, it is understandable how a human may bias the analysis by considering only some correlations in data and not all of them. The approach developed here overcomes this issue, as the corpus is scanned thoroughly and the whole data is fed to the statistical decision-making mechanism. While it is possible to envision objective filtering methods, for now we consider only the complete data in order to make sure that the analysis is objective and does not bias the conclusions of the statistical tests.

The emotion words used in epoch characterization come primarily from the NRC Word-Emotion Association Lexicon [10] to which the list of emotion words extracted from WordNet-Affect [14], distributed in the Semeval 2007 Affective Text task [13], has been added. The lexicon is made up of English words to which eight possible tags are attached: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. All in all there are 14,000 words for which at least one affective tag is given.

The paper is organized as follows. In Section 2 we review the relevant literature. Section 3 presents the statistical apparatus employed in epoch determination and epoch characterization. In Section 4 we present the experiments and the results we have obtained. In the last section we highlight the contribution of this paper and make an overview of further immediate work.

2. Related work

In [8], besides a complete introduction to the Google Books corpus, a limited diachronic study of words meaning and form is also carried out. The authors introduce the term 'culturomic' and show that quantitative analyses may lead to interesting results. They show that it is possible to determine censorship and suppression by comparing the frequencies of proper names in bilingual Google books corpora. However, the authors did not proceed to a systematic studies of epochs.

Regarding semantic change, the task of sense disambiguation over the years is introduced in [9]. In their paper, the authors refer to definite periods of time as epochs but the authors considered them priorly defined.

In [16] an analysis of topics over time is performed. The paper focuses on rather fixed topics, which are expressed by frozen compounds, such as "Mexican war", "CVS operation", and determines how these topics evolve during the years. However, because the scope of their paper is not global, the corpus used comes from 19 months of personal emails. It is hard to see how this method could generalize.

In [17] a detailed study of the impact of a piece of news on others is carried out in large corpora. The authors use EM to develop a topic-driven impact analysis. In this way, a statistical data driven approach to social importance of a topic is built. Thus, by using big data corpora and statistical methods, it is possible to obtain objective analyses of the impact of some the social facts. Our paper consolidates this new line of research.

A similar approach is described in [15]. The authors use Latent Dirichlet analysis to facilitate the search into large corpora by automatically organizing them.

In [18], the statistics tests and the google N-gram corpus are used for the semi-automatic creation and validation of a sense pool. The frequencies extracted from the Google N-gram corpus are statistically filtered and verified by humans.

The richness and complexity of cultural information contained in the Google N-gram corpus is analyzed in [5]. By considering the degree of interdependence as a measure for complexity, the author used the 2-gram corpus to analyze the complexity of American culture. However, there is no epoch distinction and statistical support.

Sentiment analysis, text categorization according to affective relevance, opinion exploration for market analysis, etc., are just some examples of applications of this NLP area [11]. While positive/negative valence checking is an active field of sentiment analysis, a fine-grained emotion checking is nowadays an emerging research topic. For example, the SemEval task on Affective Text [13] focused on the recognition of six emotions in a corpus of news headlines.

3. Methodology

The Google N-gram corpus, where N goes form 1 to 5, specifies the number of occurrences of each gram over the years. In Table 1 a snippet from a Google 5-gram file is presented. The first column represents a sequence of 5 consecutive words. The second column represents a year, the third column reports the number of occurrences during that year, the fourth column reports the number of pages and the fifth column reports the number of books in which the respective word sequence appears during that year, respectively. In order to find all the occurrences of a particular word sequence in a particular year, the entire corpus must be read and the figures must be summed up.

The published data has known an exponential growth since the beginning of the twentieth century, notably after 1950. A cursory look at Fig. 1 shows that all the plots seem to share the same basic pattern of an exponential growth. For this reason it is difficult to directly compare the absolute numbers. We use a simple normalization procedure which consists in counting the occurrences of all content nouns,[2] including proper names, and we compute for each term of interest the percentage of occurrences (i.e. frequency) of that term with respect to the sum of all content nouns. The computation is carried out for each year.

While the absolute values may also be informative, for example in Fig. 1 we see an "irregularity" in the exponential growth rate somewhere between 1950 and 2000, the percentage is actually more informative on what the public opinion is concerned about in certain periods. Substantial differences may be observed within a period of time spanning several decades. For example, *democracy* was 25 times less a probable topic at the beginning of the twenty-first century than 50 years before, see Fig. 2.

Further, we present a methodology which determines for a given set of terms the periods in which the percentages differ in a statistical significant way. In the next subsection we introduce an algorithm for epoch determination based on a several statistical tests. The methodology developed here observes the condition of non-generic judgement on sentiment, [2], because the sources and the respective quantitative assessment can be traced down. We also show that the epochs can be also characterized from an emotional point of view, both as a blue-print formed by the distribution of the emotion words, but also that the dominant opinion

[2] Base form count, i.e. *foxes* are count with *fox*, and all words are in lower case.

Please cite this article in press as: O. Popescu, C. Strapparava, Time corpora: Epochs, opinions and changes, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.04.029