

Forecasting Conflicts Using N-Gram Models

Camille Besse, Alireza Bakhtiari and Luc Lamontagne

Computer Science & Software Engineering Department
Laval University, Quebec, QC, Canada

Abstract

Analyzing international political behavior based on similar precedent circumstances is one of the basic techniques that policymakers use to monitor and assess current situations. Our goal is to investigate how to analyze geopolitical conflicts as sequences of events and to determine what probabilistic models are suitable to perform these analyses. In this paper, we evaluate the performance of N-grams on the problem of forecasting political conflicts from sequences of events. For the current phase of the project, we focused on event data collected from the Balkans war in the 1990's. Our experimental results indicate that N-gram models have impressive results when applied to this data set, with accuracies above 90% for most configurations.

Introduction

Analyzing international political behavior based on similar precedent circumstances is one of the basic techniques that policymakers use to monitor and assess current situations. These analyses are based on international interactions and events that occur between political actors throughout different periods of history.

Modeling and analysis of geopolitical crises are difficult tasks due to the uncertainty inherent to these situations. Crises are normally highly unstable and a large spectrum of events, ranging from cooperative actions to hostile confrontations, might take place without specific orders. Moreover, geopolitical crises can involve multiple parties leading to the formation and disintegration of coalitions. And the evolution of conflicts over time might be erratic as long periods of relative calm can suddenly lead to active periods of violence without observable transitions.

Some tools are required to capture characteristics that might influence crisis evolution as well as the fluctuations of conflicting events. In this paper, we present experiments we conducted on the usage of probabilistic models for forecasting crisis situations. More specifically, we evaluate the performance of N-gram models on the problem of forecasting political conflicts from sequences of events. For the current phase of the project, we focused on event data collected from the Balkans war in the 1990's. In the next section, we give an overview of the approach we adopted. Then we present the

corpus and the coding schemes used for representing conflicts events. Finally we discuss on the results we obtained by applying N-grams to the Balkans war data set.

An Approach for Crisis Analysis

Our goal is to propose algorithms to analyze geopolitical conflicts as sequences of events and to determine what probabilistic models are suitable to perform these analyses.

Crises are not directly observable from their theater of operations. Hence information must be gathered from various sources such as news agencies to follow their evolution. To exploit information provided by the news reports, some extraction systems are required. Events resulting from the extraction process could include the following features:

- *Source agent*: an actor that has instigated an action. It might be the military forces of a country (ex. Serbia), an organization (ex. United Nations), a faction (ex. Hezbollah) or groups of individuals.
- *Target agent*: The agent upon which an action is taken by the subject agent. As for the source agent, it might be an organization, a nation or a group of individuals.
- *Action descriptor*: it describes the nature of the interaction that occurred between the source and target agents. Various classifications are available in the literature to categorize actions involved in geopolitical conflicts. In our experimentations, we make use of the WEIS action classification (McClelland 1976).
- *Date of the event*: it indicates either when the action was perpetrated or when it was reported to the general public.

Given a sequence of events, forecasting consists of determining the likelihood that some specific patterns might occur in a near future. In this paper, we investigate how we can forecast episodes of violence within a limited prediction window (ex. in 1 month), which is a specific forecasting task for conflict analysis.

Preparation of Conflict Data Sets

To perform our experiments, we made use of a data set depicting the evolution of conflicts in the Balkans during the period of 1970-2003. This data set contains over 70 000 events automatically extracted from news reports using the Kansas Event Data System (KEDS) project (Philip A. Schrod and Weddle 1994). These data sets are series

of events formalized as pair wise interactions involving two participants, one acting against the other (a dyad). Formally, a conflict is described as an event sequence where each event contains:

- a time-stamp t_i (a number representing the date);
- a subject s_i (the source of the action);
- an object o_i (the target of the action);
- an event code c_i (the event/action type).

The events are defined over a spectrum ranging from cooperative to violent actions. The categories of events are given according to the World Event/Interaction Survey (WEIS) which roughly assigns higher codes to more hostile events. Events are distributed in 22 categories, inside of which they may be clustered into other subcategories. We make use, in our experiments, of two sets of event symbols:

- *22 symbols set*: a simplified list of WEIS symbols comprising some actions like blame, threats, opposition, demands, violence, armed conflict, praise, promise, cooperation, and rewards.
- *4 symbols set*: this set corresponds to an aggregation of the 22 symbols in 4 large categories: Strong cooperation, mild cooperation, mild hostility and strong hostility (Schrodt 2006).

Table 1 shows an excerpt of the dataset after simplifying the event codes and adding the 4 event types, where the first column shows the time-stamp, the second represents the object, the third represents the subject, the fourth and fifth columns show the WEIS code and the simplified WEIS code. The last column is the 4 event type code. Note that we also removed the participants involved in less than one hundred events, and denoted them by '—', hence indicating that the other participant may be any other party.

TIME	OBJECT	SUBJECT	CODE	22C	4C
32551	CRO	UNO	42	4	1
32552	BOSSER	MOS	150	15	3
32554	KSV	—	95	9	1
32556	BOS	USA	41	4	1

Table 1: Event dataset samples

From a computational point of view, a model with 4 symbols is expected to be less expressive and less computationally demanding than a model with 22 symbols. As reported later in this paper, we conducted experiments with both encoding of the data set to compare their forecasting potential.

To exploit this data set, we had to make some additional transformations to adapt the sequences to probabilistic modeling. Multiple events can occur during the same day as different parties are involved this regional conflict and multiple interactions can occur over a short period for a specific dyad, i.e. a specific pair (source agent, target agent). Some authors recommend aggregating multiple events occurring in one time period as a single occurrence. The aggregated value would correspond to the average intensity of the aggregated actions. This conversion can be adopted to preserve constant time intervals between events. However we decided to keep the multiple observations in the data set in order to exploit the maximum of information that was made available to us.

Conversely, it is possible that no event occurs during one day. A non-event observation can be used (null observation) to account for days without activities. This type of padding can be important to account for noise (ex. non-reported event) and to preserve the regularity of the estimates. Sometimes probabilistic models can be sensitive to high densities of null observations. However some preliminary results clearly indicated that the insertion of null events has no significant impact on the results and we decided to conduct our experimentations without any form of padding. In order to define the vocabulary needed by the probabilistic models, we chose to compare two types of event encoding schemes. In the first one, introduced by (Schrodt 2006), we only consider a single actor at a time. This means that each event is split into two separate events; one considering the subject acting upon anybody, the other considering the object "being acted upon" by anybody. For instance, the event

32551 CRO UNO 42 4 1

would be divided into the events

32551 CRO — 42 4 1
32551 — UNO 42 4 1

As a result, such a scheme results in two codes for each actor, one when he is the subject, and one when he is the object. This type of encoding, which we named the low interaction scheme (LO), roughly considers how a specific actor interacts with everybody else. The other type of coding scheme simply considers all possible pairs of interactions, thus requiring much more codes but representing the possible interactions in a better way. We refer to it as the high interaction scheme (HI).

Therefore, a different code was assigned for all interactions of each actor and for each event type. For instance, for the Balkans dataset where the top 11 actors were considered, the low interaction scheme with 4 event types results in 88 codes, while the high interaction scheme with 22 event types gives us 2904 codes. We also varied the number of actors and created datasets for 4, 6, 11, 50 and 99 actors, each of which were chosen according to their level of involvement in the conflict.

Forecasting Conflict Periods

To perform forecasting using probabilistic models, we address this task as a binary classification problem. We make use of two different models: one for detecting low-conflict episodes and one for detecting high-conflict periods. The parameters of the models are learnt from a portion of the data set containing sequences of conflict events.

To make a decision, we evaluate the probability of each sequence of interest with both models. And a decision is made by selecting the model providing the higher probability. Hence forecasting in this setting corresponds to a binary classification for selecting the model that best fits the sequence of events to be analyzed.

Forecasting using N-gram models

N-gram models were used to learn sequences of recurrent patterns in the event database. Formally speaking, an N-gram model is defined as a probability distribution over a sequence of items that reflects how frequent sub-sequences

of length N are within the whole sequence. In our context, the sentences that make up the dataset are sequences of 100 consecutive events based on the coding scheme chosen.

In our experiments, separate event sequences were prepared for each different actor set (i.e. sets containing the top 4, 6, 11, 50, or 99 actors) and for every coding scheme (high/low interactions with 4 or 22 possible event codes). For every actor-coding scheme combination, we built different event datasets for groups of sentences leading to high conflict or low conflict weeks, with 1 month, 3 months and 6 months forecast periods. This resulted in a total of 120 different event datasets. Each of the event datasets were further divided into training and test sets based on a 5-fold cross validation procedure. Training sequences were randomly selected to form each of the cross validation partitions.

Using the event datasets, N-gram models were estimated for all of the possible combinations described above using the SRI Language Modeling toolkit (SRILM) (Stolcke 2002). For each coding scheme and forecast period chosen, separate models were trained over event datasets corresponding to high conflict and low conflict weeks.

Training in N-gram models consists of determining the conditional probability that an event can occur given N-1 preceding events have been observed (event history). Estimation of probabilities is done by counting the number of occurrences of a sequence of N events in the whole training set. More formally, an N-gram model is the set of conditional probabilities such that:

$$P(w_n|W_{n-N+1}^{n-1}) = \frac{C(W_{n-N+1}^{n-1}w_n)}{C(W_{n-N+1}^{n-1})}.$$

The Witten-Bell discounting method was used to smooth the probability distributions. Discounting consists of reallocating probabilities in order to account for unseen n-grams during training. A binary classifier was then used to label the event sequences. In other words, every sentence in the test set was evaluated using both the high conflict and low conflict language models, and the perplexity of each model was calculated by counting all input events. Perplexity (PP) is an information theoretic measure proportional to the probability of an event sequence and is estimated as

$$PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i|W_1^{i-1})}}.$$

Intuitively, perplexity represents the average number of events that can be expected at each time step. A model having lower perplexity is considered a better predictor as it reduces the average number of possible events that might be encountered.

Consequently, conflict periods (sequences of 100 consecutive events) were individually classified as leading to a high or low conflict week based on which model gave lower perplexity to that sequence.

Class-based N-gram models

In the next phase of the project, we tried to model events not as interactions among two single actors, but as different coalitions acting against one another. Following the approach in (Brandes and Lerner 2008), actors were attributed

to groups in a way that intra-group negative interactions were rare or nonexistent, while inter-group conflicts were frequent and serious. For instance, two of the main actors in the Balkans war, namely Serbia and Serbs in Bosnia (represented by SER and BOSSER, respectively) appear to have shared a common strategic outlook throughout the conflict, and it seems natural to assign them to a single group. With these criteria in mind, we identified four groups of actors which had quite similar strategic views during the conflict period using the results in (Brandes and Lerner 2008). These consist of:

- the two main groups in the conflict; namely the Serbian forces SER and BOSSER against an opposition group formed of Bosnia, Croatia and Kosovo (BOS, CRO, KSV).
- a mediator group UNO and NATO,
- an international group consisting of all other actors.

Using these groupings, we replaced actors with the corresponding group in which they belonged to, and prepared training and test sets for configurations similar to the ones described for regular N-grams. The IBM class-based N-gram model (Brown et al. 1992) was then applied to build the language models, which (for the simple bigram case) estimates sequences as:

$$p(w_i|w_{i-1}) = p(c_i|c_{i-1})p(w_i|c_i)$$

As with the regular N-gram approach, the models were then used to find the perplexity of individual sentences in the test set, and high conflict or low conflict labels were assigned by comparing the perplexities given by each model.

Experimental Results

We tried different codes for representing actors and event types. Overall, results with 11 actors seem to present a good trade off between involvement and relevance of actors in the Balkans conflict. We then took the one hundred events that happened one, three and six months prior to the start of that week in order to train and test our N-gram models. As a result, we had six datasets of event sequences for every actor-coding scheme combination; one for high conflict weeks and one for low conflict weeks, for each of the three forecast periods. To compare the performance of our models, we computed various correctness measures for the different configurations described above. These measures are essentially based on the total number of sentences predicted as high conflict when we actually have a high conflict week (TP) or a low conflict week (FP), and the total number of sentences predicted as low conflict when we actually have a low conflict week (TN) or a high conflict week (FN). Accordingly, the performance measures calculated were:

- the overall accuracy of each model (the relative number of correctly predicted high and low conflict weeks),
- true-positive and true-negative precisions (how many of the weeks predicted as high/low conflict were actually a high/low conflict week),
- true-positive and true-negative recalls (of all actual high/low conflict weeks, how many were correctly labeled as high/low conflict), and

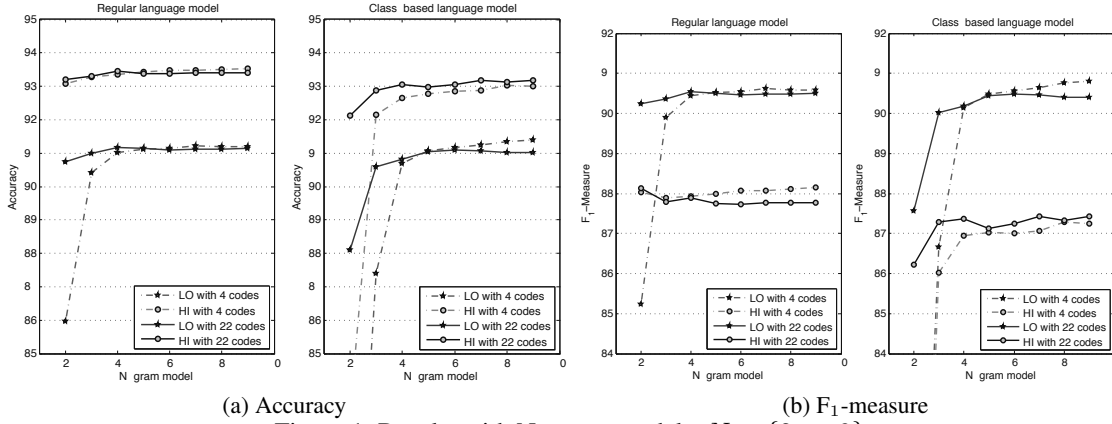


Figure 1: Results with N-grams models. $N = \{2, \dots, 9\}$

- F-measure for values of 0.5, 1, 2 that combines precision and recall to put more emphasis on either one of them.

Due to the large number of performance results obtained for all configurations, only the accuracy and true-positive F1-measure (where precision and recall are evenly weighted) for 11 actors with a 28 day forecast period will be discussed here.

Figure 1 shows how each of our different coding schemes perform when evaluating N-grams for N belonging to 2...9. While the overall performance of N-grams is impressive, there is little improvement in both the accuracy and F1-measures for N greater or equal to 6. Therefore, for this specific actor-coding scheme, it appears that sequences larger than 6 events in length do not carry much more information contributing to our model's performance.

Furthermore, both figures show that class-based N-grams perform as well as regular N-grams (or even do better). Such behavior is also observed in other actor sets containing the top 4, 6, 50 and 99 actors. This interesting result suggests that instead of using a large number of actors, we can achieve the same level of performance with considerably less codes. For example, 80 codes instead of 39600 codes by grouping the top 99 actors into 4 blocks in the HI scheme with 4 event codes.

The complete accuracy and F1-measure results for the class-based 6-gram model are shown in Table 2. We can observe from this table that, in general, low interaction coding schemes (LO) are much better at correctly forecasting low conflict weeks (more TN's), and high interaction schemes (HI) outperform LO schemes when it comes to forecasting high conflict weeks (more TP's). This results in LO schemes having higher precision and recalls with regard to high conflict weeks, which explains the gap between LO and HI schemes in F-measure Figure 1b. On the other hand, since the data is strongly skewed towards low conflict weeks (around 80% of the whole dataset) the number of TN's is larger than the number of TP's, and thus TN's will have more influence over the accuracy of the model. As a result, HI schemes have higher accuracy (F-measure Figure 1a), and higher true-negative F1-measure compared to LO schemes.

Conclusion

In this paper, we discussed the application of N-gram to the problem of forecasting political conflicts. Our results show

	28 days	91 days	184 days
LO interactions	91.16	89.91	89.79
HI interactions	92.84	94.31	94.01

(a) Accuracy

	type	28 days	91 days	184 days
LO	high-conflict	90.56	86.15	86.14
	low-conflict	91.70	92.01	91.92
HI	high-conflict	87.01	77.45	76.21
	low-conflict	95.07	96.74	96.58

(b) F1-measure

Table 2: Results of class-based 6-gram model

that the N-gram models have impressive results when applied to the Balkans war, with accuracies above 90% for most configurations. Analysis of the top frequent N-grams shows some interesting recurrent sequences of events, however, extracting meaningful patterns from the large number of data remains to be done in a future work. These models must also be used with other datasets to analyze their performance in forecasting more complicated conflicts with more involved actors, like the war in central Asia (between Afghanistan, Armenia-Azerbaijan and former Soviet republics). Furthermore, instead of manually assigning actors to groups, clustering algorithms could be used to determine the different groups of actors in conflict.

References

- Brandes, U., and Lerner, J. 2008. Visualization of conflict networks. *NATO security through science series human and societal dynamics* 36:169.
- Brown, P.; Desouza, P.; Mercer, R.; Pietra, V.; and Lai, J. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18(4):467-479.
- McClelland, C. A. 1976. *World Event/Interaction Survey Codebook*. (icpsr 5211).
- Philip A. Schrodt, S. G. D., and Weddle, J. L. 1994. Political Science: KEDS, A Program for the Machine Coding of Event Data. *Social Science Computer Review* 3(12):561-588.
- Schrodt, P. 2006. Forecasting conflict in the balkans using hidden markov models. In Trapp, R., and al., eds., *Programming for Peace*, volume 2 of *Advances in Group Decision and Negotiation*. Springer Netherlands. 161-184.
- Stolcke, A. 2002. Srlm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, 901-904. Citeseer.