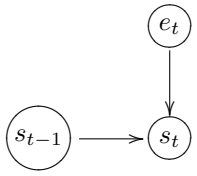
1 The Premises

The starting point is to create a simple model of the generation of syntagms in social media. The general idea is the following.

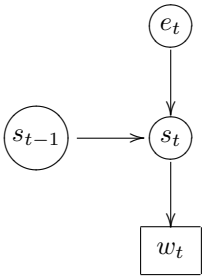
We are "chasing" a specific syntagm $\lceil w_1 \cdots w_n \rfloor$ that we have observed being present at the end of the period of observation using the method in [2]. We observe the comunication during a period $T = [1, \ldots, n]$: we asume that at $t = 1$ the syntagm has not been created, and at $t = n$ has been. We are interested in determining its evolution, in particular, in determining the time t at which we can assume that the syntagm has been created. Our observations are a sequence of documents D_t, $t = 1, \ldots, n$, one for each time step (this will be in general the union of various documents, for example, all the tweets written in a certain group on a given week; for the purpose of our discussion, we merge all the documents into one); the words contained in the documents will form our measurements, as we shall see later on.

Let us place ourselves at a time t at which the syntagm has not yet emerged. The creation of the syntagm is due to the ocurrence of an external *event* (for example, the syntagm "fake news" was related to the beginning of the 2016 presidential campaign). The event is something that happens at a given point in time and does not necessarily repeat itself. On the other hand, the event creates a *situation*, that is potentially self-supporting in time that is, it may be present after the end of the event. Let is consider two stochastic processes, e_t and s_t. The first is the variable that models the ocurrence of the event at time t, the second is the variable that models the existence of a situation conducive to the creation of the syntagm at time t. Both take valoes in $\{0, 1\}$ (1 means that the event ocurred or the situation is present, 0 means the opposite). The existence of the situation at time t depends on the existence at time $t - 1$ or the ocurrence of the event at time t. The diagram representing the dependence of the randon variables at time t is the following:



(1)

The existence of the situation detemines the propability of observing the meaningful syntagm. Let w_t be the event of the syntagm being formed (again, $w_t = 1$ if the syntagm is present, $w_t = 0$ if not). The presence of the syntagm depends directly only on the existence of the *situation* at time t. We have therefore the following scheme:



(2)

(The square indicates the element that we observe). Here, to use the notation of [2], it is

$$w = \lceil w_1, \ldots, w_n \rfloor \tag{3}$$

that is w is the syntagm that we are chasing, composed of n words.

Note that we are fixating on a single syntagm, which we assume to have identified at the end of the temporal sequence, and we are interested in seeing where it was generated estimating e_t in such a way to explain the observations. At time t, we make an observation π_t (we shall see later how this is actually determined). We actually needs two measures, one that determines the information in favor of the hypothesis that at time t the syntagm was not expressed (call it π_t^0) and the other expressing the information in favor of the hypothesis that the syntagm was expressed. The direct dependence of w (we shall omit the inex t when no confusion arises) is on s, so we can write

$$P(w=i) = \sum_j P(w=i|s=j)P(s=j) \quad \text{· No entiendo los sumatorios} \tag{4}$$

or, developing further

$$P(w_t=i) = \sum_j \sum_k \sum_h P\big[w_t=i\big|s_t=j\big]P\big[s_t=j\big|s_{t-1}=k,e_t=h\big]P\big[s_{t-1}=k\big]P\big[e_t=h\big] \tag{5}$$

The observed variable w (the syntagm) depends on the *latent variables*

$$\text{zeta} \quad \zeta_t = (s_t, s_{t-1}, e_t) \tag{6}$$

which are the ones whose probability distribution, at each time t, we have to determine. The problem is solved by minimizing the functional

¿Variables latentes?
¿Minimizar la función?
¿Porqué desaparece el logaritmo?

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_i \pi_t^i \log P(w_t=i|\theta) \\ &= \sum_i \pi_t^i \sum_{j,k,h} P\big[w_t=i\big|s_t=j\big]P\big[s_t=j\big|s_{t-1}=k,e_t=h\big]P\big[s_{t-1}=k\big]P\big[e_t=h\big]\end{aligned} \tag{7}$$

Here

$$\theta = (P\big[w_t=i\big|s_t=j\big], P\big[s_t=j\big|s_{t-1}=k,e_t=h\big], P\big[s_{t-1}=k\big], P\big[e_t=h\big]) \tag{8}$$

are the parameters over which we must optimize. Since we have to manipulate these parameters, it is convenient to use a simpler notation. Define

$$\begin{aligned}\alpha_j^i(t) &= P\big[w_t=i\big|s_t=j\big] \\ \beta_{kh}^j(t) &= P\big[s_t=j\big|s_{t-1}=k,e_t=h\big] \\ \nu^k(t) &= P\big[s_{t-1}=k\big] \\ \text{nu} \quad \text{gamma} \quad \gamma^k(t) &= P\big[e_t=h\big])\end{aligned} \tag{9}$$

Note that we put in superscript the value of the conditioned variable and in subscript that of the condintioning variable(s).

2 The method

We use the general method of the EM (Expectation Maximization) algorithm [1]. I will skip over it a bit; if necessary I will put more details in a following draft. The function $\mathcal{L}$ is bounded from below by

$$g(\theta,\theta') = \sum_i \pi^i \mathbb{E}_{\zeta|w=i,\theta'}\big[\log P(w=i,\zeta=(j,k,h)|\theta\big] + \sum_i \pi^i H\big[P(\zeta|\theta')\big] \tag{10}$$

Here, during the iteration, θ' is the set of parameters determined at the previous step, and θ the set we solve for. the second term is independent of θ, so we can optimize

$$\begin{aligned} Q &= \sum_i \pi^i \mathbb{E}_{\zeta|w=i,\theta'}\big[\log P(w=i,\zeta=(j,k,h)|\theta\big] \\ &= \sum_i \pi^i \sum_{j,k,h} P(\zeta=(j,k,h)|w=i,\theta') \log P(w=i,\zeta=(j,k,h)|\theta) \\ &= \sum_i \pi^i \sum_{j,k,h} P(\zeta=(j,k,h)|w=i,\theta') \log P\big[w_t=i\big|s_t=j\big]P\big[s_t=j\big|s_{t-1}=k,e_t=h\big]P\big[s_{t-1}=k\big]P\big[e_t=h\big] \\ &= \sum_{i,j,k,h} \pi^i \tau_i^{jkh} \log \alpha_j^i \beta_{kh}^j \nu^k \gamma^h \end{aligned} \tag{11}$$

where we have set

$$\tau_i^{jkh} = P(s_t=j, s_{t-1}, e_t=h|w=i) \tag{12}$$

In the **E** step, we determine Q, that is, in practice, we determine τ_i^{jkh}, which is given by

$$\tau_i^{jkh} = \frac{\alpha_j^i \beta_{kh}^j \nu^k \gamma^h}{\sum_{jkh} \alpha_j^i \beta_{kh}^j \nu^k \gamma^h} \tag{13}$$

For the **M** step we maximuze Q subject to the constraints

$$\sum_i \alpha_j^i = 1 \qquad \sum_j \beta_{kh}^j = 1 \qquad \sum_h \gamma^h = 1 \tag{14}$$

We apply the Lagrange multipliers and minimize

$$\mathcal{F} = \sum_{i,j,k,h} \pi^i \tau_i^{jkh} \log \alpha_j^i \beta_{kh}^j \nu^k \gamma^h - \sum_j \lambda_j \left(\sum_i \alpha_j^i - 1 \right) - \sum_{k,h} \mu_{kh} \left(\sum_j \beta_{kh}^j - 1 \right) - \epsilon \left(\sum_h \gamma^h - 1 \right) \tag{15}$$

We detemine the parameters by setting to zero the derivatives of $\mathcal{F}$

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \alpha_j^i} &= \sum_{hk} \tau_i^{jkh} \pi^i \frac{1}{\alpha_j^i} - \lambda_j = 0 & \alpha_j^i &= \frac{\pi^i}{\lambda_j} \sum_{k,h} \tau_i^{jkh} \\ \frac{\partial \mathcal{F}}{\partial \beta_{hk}^j} &= \sum_i \tau_i^{jkh} \pi^i \frac{1}{\beta_{kh}^j} - \mu_{kh} = 0 & \beta_{kh}^j &= \frac{1}{\mu_{kh}} \sum_i \tau_i^{jkh} \pi^i \\ \frac{\partial \mathcal{F}}{\partial \gamma^h} &= \sum_{ijk} \tau_i^{jkh} \pi^i \frac{1}{\gamma^h} - \epsilon = 0 & \gamma^h &= \frac{1}{\epsilon} \sum_{ijk} \tau_i^{jkh} \pi^i \end{aligned}$$

The Lagrange multipliers are determined by normalization

$$\begin{aligned}\alpha_j^i &= \frac{\pi^i \sum_{kh} \tau_i^{jkh}}{\sum_i \pi^i \sum_{kh} \tau_i^{jkh}} \\ \beta_{kh}^j &= \frac{\sum_i \pi^i \tau_i^{jkh}}{\sum_{i,j} \pi^i \tau_i^{jkh}} \\ \gamma^h &= \frac{\sum_{ijk} \pi^i \tau_i^{jkh}}{\sum_{i,j,k,h} \pi^i \tau_i^{jkh}}\end{aligned} \tag{16}$$

With the new vales of the parameters we calculate τ_i^{jkh} using (13), then new values of the parameters, etc. until convergence. This gives, among other things, the value $\gamma^1 = P(e_t = 1)$, which gives us the probability that the event that caused the syntagm to be created happened at time t.

The calculation depends on the value ν^k, which is equal to $P(s_{t-1} = h)$. In order to run the approximation over a span of time, it is necessary to compute $P(s_t = h)$. From

$$P(s_t = j) = \sum_{k,h} P[s_t = j | s_{t-1} = k, e_t = h] P[s_{t-1} = k] P[e_t = h] \tag{17}$$

we derive

$$\nu^j(t) = \sum_{k,h} \beta_{kh}^j \nu^k(t-1) \gamma^h \tag{18}$$

3 The algorithm

We consider a time span $\Delta = [1, \dots, T]$ over which we want to determine the emergence of the syntagm. At each time t we have a set of document, which we use to determine the measurements $\pi^i(t)$ ($i \in \{0, 1\}$) that determine the information supporting the hypothesis that the syntagm is not present ($\pi^0(t)$) and that is present ($\pi^1(t)$). We shall consider how to determine $\pi^i(t)$ in the next section. For the moment, let us consider that a sequence $[\pi^i(1), \dots, \pi^i(t)]$ is avaiable.

We run the estimation time-wise, from $t = 1$ to $t = T$, using at each time the values $\nu^j(t)$ in order to run the computation at $t + 1$. In order to do this, we must initialize $\nu^h(0)$. We can assume that the situation is not present at time 0 and set $\nu^0(0) = 1$, $\nu^1(0) = 0$, or use a small probability p for the existence of the situation. The algorithm is in Figure 1:

Convergence is determined by the stabilization of the parameters, any number of criteria that measure the change in the parameters between iterations can be used. The list Ψ contains, for each time t, the probability that the event that caused the syntagm to be generated happened at time t.

4 The measure

The final piece of the puzzle that has to be put into place is the measurement $\pi^i(t)$. We have defined it informqlly as the information that supports the hypothesis that the syntagm has been created (resp., has not been created). This is quite a fuzzy definition, and probably can be satisfied in many different ways. This is one possibility.

We are "chasing" a syntagm $\lceil w_1 \cdots w_n \rfloor$ and, as evidence, we have, at each time t, a document D_t. The data that we extract are essentially:

i) the number of observations of the individual words $n(w_1), \ldots, n(w_n)$;

ii) the number of observations of the whole sequence of words that composes the syntagm $n(w_1, \ldots, w_n)$;

iii) the "quality" of the syntagm, $q(\lceil w_1 \cdots w_n \rfloor)$, determined as in [2].

The general idea is that if we see many ocurrences of the whole syntagm ($n(w_1, \ldots, w_n)$ high) and the syntagm is significant ($q(\lceil w_1 \cdots w_n \rfloor)$ high), then there is evidence of its presence ($\pi^1(t)$ high, $\pi^0(t)$ low). Vice versa, if the words are observed individually but not as a whole ($n(w_1), \ldots, n(w_n)$ high, $n(w_1, \ldots, w_n)$ low) or the whole is not a viable syntagm ($q(\lceil w_1 \cdots w_n \rfloor)$ low), then the evidence supports the absence of the syntagm ($\pi^0(t)$ high, $\pi^1(t)$ low).

So, OK, without further ado, one possibility is

$$\begin{aligned} \pi^1 &= \begin{cases} q(\lceil w_1 \cdots w_n \rfloor) \dfrac{n(w_1, \ldots, w_n)}{\max(n(w_1), \ldots, n(w_n))} & \text{if } \max(n(w_1), \ldots, n(w_n)) > 0 \\ 0 & \text{otherwise} \end{cases} \\ \pi^0 &= 1 - \pi^1 \end{aligned} \tag{19}$$

Note that if $\max(n(w_1), \ldots, n(w_n)) = 0$ (no words of the syntagm are observed), we consider that there is sure evidence of the absence of the syntagm. This might not be the best choice: one can consider this case neutral ($\pi^1 = 1/2$) or even discard the information at that time. There things should be tried out.

Quite honestly, this is the part I feel less sure about. The general idea of the information $\pi^i(t)$ is quite clear, as are the general characteristics that these coefficients should have vis-á-vis the observations, but the specific form that these function should have is still something to be worked out.

References

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1--38, 1977.

[2] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang ren, Clare R. Voss, and Jawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825--37, 2018.

psi

1. $\Psi \leftarrow []$
2. for t ← 1 to T do
3. initialize α_j^i, β_{kh}^j, γ^h, τ_i^{jkh}
4. while not convergence do
5. $$\tau_i^{jkh} \leftarrow \frac{\alpha_j^i \beta_{kh}^j \nu^k(t-1)\gamma^h}{\sum_{jkh} \alpha_j^i \beta_{kh}^j \nu^k(t-1)\gamma^h} \quad (i,j,k,h \in \{0,1\})$$
6. $$\alpha_j^i \leftarrow \frac{\pi^i(t) \sum_{kh} \tau_i^{jkh}}{\sum_i \pi^i(t) \sum_{kh} \tau_i^{jkh}} \quad (i,j \in \{0,1\})$$
7. $$\beta_{kh}^j \leftarrow \frac{\sum_i \pi^i(t) \tau_i^{jkh}}{\sum_{i,j} \pi^i(t) \tau_i^{jkh}} \quad (j,k,h \in \{0,1\})$$
8. $$\gamma^h \leftarrow \frac{\sum_{ijk} \pi^i(t) \tau_i^{jkh}}{\sum_{i,j,k,h} \pi^i(t) \tau_i^{jkh}} \quad (h \in \{0,1\})$$
9. od
10. $\nu^j(t) = \sum_{k,h} \beta_{kh}^j \nu^k(t-1)\gamma^h \quad (j \in \{0,1\})$
11. $\Psi \leftarrow \Psi + [\gamma^1]$
12. od

Alfa, beta y gamma se inicializan de forma random entre 0 y 1 (foto)
Se normalizan

La convergencia se da cuando la diferencia de cuadrados del betta/gamma anterior con el actualizado se hace muy pequeña

¿El (t-1) que hay en la fórmula de tau no hace que sea todo 0 al principio del bucle, es decir, cuando t = 1?

¿En alpha, los sumatorios de tau no deberían cancelarse?

¿En el sumatorio de tau, qué hago con la j?

Figure 1: The algorithm for the determination of the probability of having the syntagm-creating event.