NLP-driven Data Journalism: Time-Aware Mining and Visualization of International Alliances

Xavier Tannier
LIMSI, CNRS, Univ. Paris-Sud,
Université Paris-Saclay
F-91405 Orsay, FRANCE
xavier.tannier@limsi.fr

Abstract

We take inspiration of computational and data journalism, and propose to combine techniques from information extraction, information aggregation and visualization to build a tool identifying the evolution of alliance and opposition relations between countries, on specific topics. These relations are aggregated into numerical data that are visualized by time-series plots or dynamic graphs.

1 Introduction

Information and communication technologies have provided tools and methods to make the production of information more democratic. As a result, a vast amount of content is available, which arguably creates more noise than knowledge at the end of the day. Without hierarchical organization and contextualization, users may lack perspective to understand and assimilate the multiplicity of events that they come across every day, and to link them to related events in the past.

In this context, journalists and technologists developed the notion of "data journalism", which takes advantage of the growing popularity of Open Data, the development of structured knowledge bases such as DBPedia [Lehmann *et al.*, 2013], YAGO [Hoffart *et al.*, 2013] or OpenCalais and many others, as well as recent work in data visualization, to facilitate information analysis and access a variety of points of view. However, knowledge is still far from being entirely represented in structured databases, and much information remains available in text format. For this reason, natural language processing has a lot to offer to modern journalism.

In this paper, we present a tool that automatically identifies alliance and opposition relations between countries, on a specific subject defined by a user query (*e.g.* situation in Syria, nuclear proliferation, North Pole ownership). The evolution of the relations over time are then illustrated by a time-series plot (see Figure 3) or dynamic graphs and maps (Figure 4).

2 Related Work

The idea of automatically finding topically related material in streams of newswire data goes back to Topic Detection and Tracking evaluation campaigns [Allan, 2002].

Our work contributes to this research effort and explores the quantitative aspects of knowledge that can be extracted from textual documents. With that respect, as well as on the topic of alliance and opposition relations, this can be connected to opinion mining. [Chambers *et al.*, 2015] also consider these relations, based on Twitter data.

We also use well-known techniques of feature-based, supervised relation extraction. The aim is to identify and classify relations between two entities in the same sentence, by learning these relations on a training, manually annotated dataset. Examples of such works are [Miller *et al.*, 2000], [Kambhatla, 2004], [Boschee *et al.*, 2005], [Zhou *et al.*, 2005], among many others. Our approach differs from most works in the fact that each relation is associated to a date, which makes the classification time-aware.

Also, we bias our classifier towards precision. This approach relies on linguistic variation and redundancy in a large amount of documents to ensure a good coverage. It is related to works in question-answering [Dumais *et al.*, 2002], temporal information aggregation [Kessler *et al.*, 2012] or opinion mining [Turney, 2002].

3 System Overview

We apply information extraction techniques to a large amount of newswire textual documents, in order to acquire enough data to make significant statistics on them. These data can then be accessed by a query-based visualization tool.

Relations that we extract are opposition (*NEG*) or alliance (*POS*) relations between two countries, explicitly expressed in a same sentence, such as:

(1) **Indonesia** voiced support for **East Timor**'s bid to join the ASEAN. → *POS(Indonesia, East Timor)*

(2) **London**'s recent condemnations of **Libyan leader Moamer Kadhafi**'s bloody crackdown [...]. → *NEG(U.K., Libya)*

(3) **Chavez** has stoop up for his longtime ally **Kadhafi**, [...]. → *POS(Venezuela, Libya)*

The alliance or opposition can be made explicit mainly by an action verb (*protested*), an event noun (*condemnations*) or a state noun (*ally*). Countries (relation arguments) can be designated by their actual names, the name of their capital or of a person representing this country.

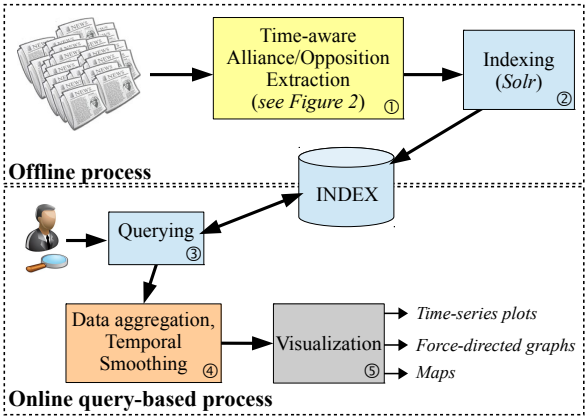

Figure 1: System overview.

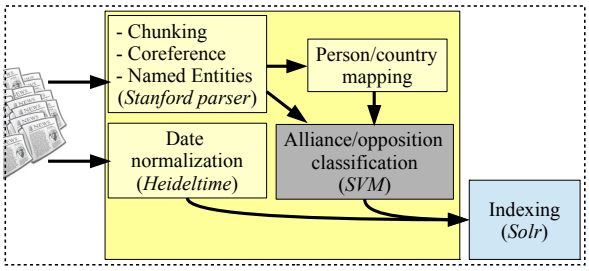

Figure 2: Offline processing details.

Figure 1 shows the general architecture of the system. At indexing time, the corpus is processed (step ① and Figure 2) with chunking, coreference resolution, named entity tagging, temporal information normalization and association between some people names and their countries. A classifier extracts *POS* and *NEG* relations in each sentence.

At query time, all sentences that are relevant to the query are retrieved (③). For each day, *POS* and *NEG* relations are aggregated and a tendency of this day is obtained, as well as a trend over time (④). A graphical visualization of this trend is then proposed to the user (⑤ and Figures 3, 4).

4 Alliance and Opposition Classification

4.1 Resources

We used a corpus of English newswire texts provided by the AFP French news agency. The English AFP corpus is composed of 1.79 million texts that span the 2004-2013 period (635 documents/day in average and 600 millions words).

Entities. Country, nationality and capital name lists were extracted from the linked data repository of the CIA World Factbook, and linguistic variations of these names were collected from DBPedia. All these elements were considered as entities potentially parts of an alliance/opposition relation.

Alliance/Opposition Gazetteers. We constituted manually a restrained lexicon of "relation triggers" containing 110 words used in the corpus to express alliance and opposition (and not a more general "polarity"). These words are verbs (*agree*, *support*, *accuse*, *condemn*, *slam*...), event nouns (*congratulation*, *accusation*, *sanction*...) and other nouns (*ally*...).

Date Normalization. Our aim is to associate alliance and opposition relations to dates, in order to observe the evolution of these relations in time. We used Heideltime [Strötgen and Gertz, 2013] to normalize dates inside the documents. Normalization is the operation of turning a temporal expression (*e.g.*, *July 26th*, *yesterday*, *Last Tuesday*) into a formated, fully specified representation (*2013-07-26*).

Chunking. Parse trees are obtained from the Stanford parser [Klein and Manning, 2003], and we extract minimal chunks (NPs, VPs) from these trees.

4.2 Classification

It is important to note that we do not claim to build a generic classifier of such relations. Our aim is to extract enough relations to be able to produce realistic, substantial and significant data about international relations between countries. At classifier level, a good precision will be favored, because we rely on the redundancy of information in the collection to achieve an appropriate coverage (*i.e.*, recall) of relations. Therefore, we will learn a classifier to identify relations that are expressed explicitely, in the same sentence.

Training Set. We randomly selected sentences containing at least two entities separated by less than 15 chunks, and at least one relation trigger between the two entities.

Each pair of entities satisfying these constraints are an instance of the classifier, which means that one sentence can lead to several instances. *e.g.*, in:

(4) In Jerusalem, Prime Minister **Silvio Berlusconi** pledged **Italy**'s firm *support* for **Israel** and urged effective *sanctions* against **Tehran**.

Entities are in bold and relation triggers in italics. Relations to classify are then {Italy (S. Berlusconi), Israel}, {Italy (S. Berlusconi), Iran (Tehran)}, {Italy, Israel}, {Italy, Iran (Tehran)} and {Israel, Iran (Tehran)}. The resulting relations would be *NEG(Italy, Iran)* and *POS(Italy, Iran)*.

We annotated 2105 such instances with relation NIL (no relation, 1463 instances), NEG (opposition, 349 instances) and POS (alliance, 293 instances).

Classifier. Sentences do not differ in their structure whether they express alliance or opposition. Entities have the same kinds of interactions with each other. Only the polarity of trigger words matters. Therefore, we implement a two-step SVM classification:

A. Filtering out pairs that have no relation at all, *i.e.* classifying between *NIL* and *non-NIL* relations;

B. Among non-NIL relations, classifying between *POS* and *NEG* relations;

These classifiers were evaluated by a 10-fold cross-validation. Results are presented in Table 2. The final model used for next steps is trained with all annotated instances.

5 Time-Aware Aggregation and Visualization

The alliance classifier described in previous section extracts a total of 330,222 instances of relations from the corpus. If a date has been identified and normalized in a sentence, then this date stamps the sentence. If Heideltime was unable to

STEP A
Entities
- Distance between entities, sentence length, positions of entities - Type of E_1 and E_2 (capital, person name, country name) - Number of entities around and inside entities
Lexical features
- Presence, number and type (verb, noun) of triggers around or between entities. - Negation between E_1 and E_2 - Among 20 most frequent prepositions in the corpus: those occurring just before and just after E_1 or E_2
Syntactic features
- Whether E_1 (resp. E_2) is the head of its chunk, size of chunks - Number of verbs, nouns around entities

STEP B
Distinction between NEG and POS triggers, negation
- Number of positive (resp. negative) triggers in the sentence - Number of positive (resp. negative) triggers around entities - Negation marks

Table 1: Features used for the classifiers A and B.

Relation		Precision	Recall	F1
Step A	**non-NIL**	**0.82**	0.76	0.79
	NIL	0.90	0.93	0.91
	average	0.87	0.88	0.87
Step B	POS	0.95	0.99	0.97
	NEG	0.99	0.95	0.97
	average	**0.97**	**0.97**	0.97
A + B	**POS & NEG**	**0.80**	0.73	0.76

Table 2: 10-fold cross validation results for the alliance/opposition classifier.

normalize the date, then the sentence is skipped. Otherwise (no date found), the document creation time stamps the sentence. A typical query is then composed of a few keywords representing the topic, a temporal interval (minimum of maximum dates) and zero or more country names on which the user wants to restrict relation extraction.

For all pairs of considered countries, inside the same day d, the weight for the pair and the day d is:

$$w(d) = log(\frac{1 + P(d)}{1 + N(d)})$$

where $P(d)$ and $N(d)$ are the number of *POS* and *NEG* relations between the two countries. $w(d)$ is a number between $-\infty$ and $+\infty$, where $w = 0$ is neutral, $w < 0$ is an opposition and $w > 0$ is an alliance. The noise is then reduced by a weighted mean smoothing over a temporal window of 5 days.

5.1 Visualization

For bilateral relations (field *countries* containing two items), we provide the user with a time-series plot representing $sw(d)$, and show them on demand the sentences which led to this value. Figure 3 shows an example of results concerning the relations between United States and Russia ("*countries:United States AND Russia*") concerning the situation in Syria ("*keywords:syria*").

When zero, one or more than two countries are specified by the user, we generate a graph of countries, where the distance between vertices reflects the opposition between the countries in a given time span (Figure 4). We use the Barnes-Hut force-directed layout algorithm, where a value between two vertices is considered as a repulsive force. The color of the nodes reflects the their proximity with each other (using a first-neighbor shortest path algorithm).

For this algorithm, we need to transform our weights $sw(d)$ into positive numbers (repulsive forces). We also need to damp the noise and the variations of the weights that are introduced by the volume of data. For example, two positive values of 1 and 3.5 (diff. = 2.5) should be considered as close to each other, while a positive value of 1 should be far from a negative value of -1 (diff. = 2). This kind of effects can be corrected by a "S-shaped", logistic function, that models a level of saturation after an approximately exponential growth (or, in our case, decrease):

$$sw'(d) = 1 - \frac{1}{1 + e^{-sw(d)}} \quad (5)$$

This function levels off high weights (both negative and positive), increases differences between positive and negative values and thus helps reducing noise without having to discretize values arbitrarily. Resulting numbers are all positive, between 0 and 1, where 1 is a strong opposition (then, repulsion in the graph) and 0 is a strong alliance (attraction in the graph), while 0.5 is neutral.

5.2 Evaluation

Evaluating the relevance of the produced trends is very subjective and would require a high level of expertise in every tested domain. Even if this work is carried out in collaboration with journalists, we cannot afford such an effort. That is why we opted for a protocol that is at the same time more objective and easier to conduct:

1. We chose 14 queries with the following information: names of two countries (or unions of countries) involved in the relation, and an optional keyword-based thematic restriction. We selected queries having potentially a high density of extracted relations (*e.g.* North Korea *vs.* South Korea, or Russia *vs.* United Nations on "Syria"), as well as sparser topics (France *vs.* Germany on "austerity", China *vs.* Japan on maritime affairs).
2. On the resulting plot, we selected up to 5 strong peaks — $abs(sw(d)) > 1$ — and 5 weak peaks — $abs(sw(d)) \leq 1$ (total of 97 relations).
3. For each of these peaks, we estimated whether the polarity of the peak was relevant or not. For that purpose, we seeked news articles from a time-stamped web collection that was not part of the tested collection, in order to validate whether the two countries rather agreed or opposed at the date indicated by the peak. This is still a heavy task, which explains the low number of tested instances.

The accuracy of strong peaks is 0.90, making them highly reliable. Accuracy of weak peaks is 0.702. Note that very

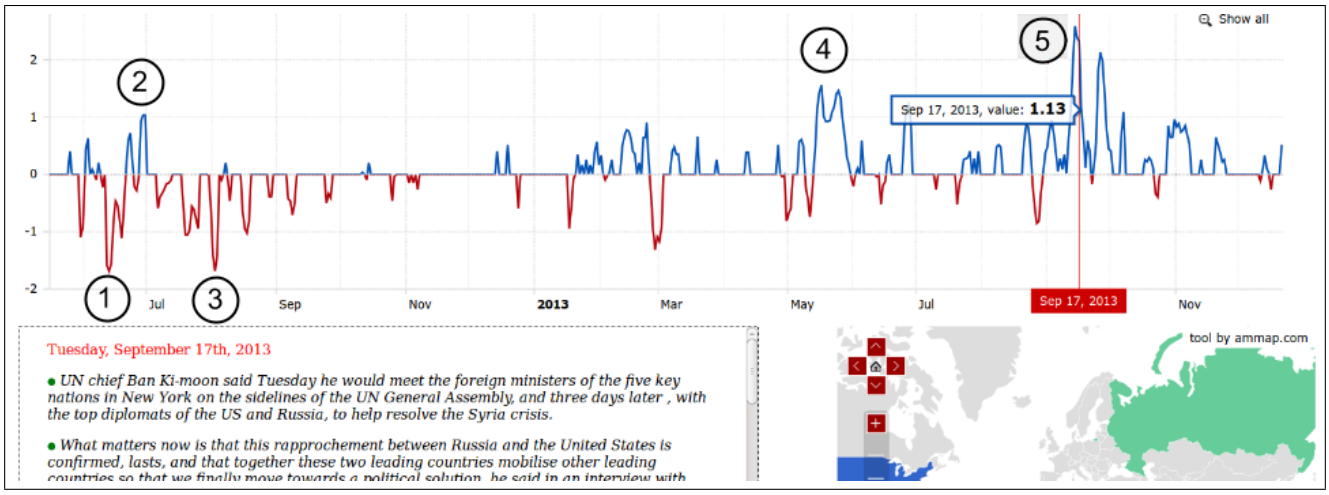

Figure 3: Example of plot produced by the system for bilateral relations between United States and Russia on the query "Syria". The bottom left frame shows sentences corresponding to the user-selected date (Sep. 17, 2013). Circled numbers have been manually added to the screenshot. They correspond to: ① Mutual accusations of supplying arms to Syrian authorities or opposition (bad relation, $sw(d) << 0$); ② Planning of a meeting to discuss the problem (better relation, $sw(d) > 0$); ③ Vetos of China and Russia for United Nations resolutions; ④ Announcement of a peace conference; ⑤ Agreement at this conference.

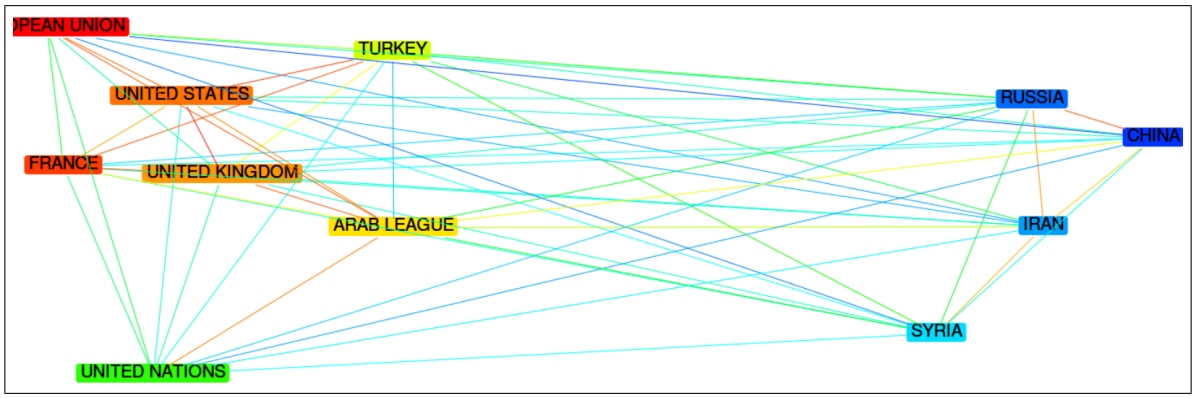

Figure 4: Example of graph produced by the system for relations between different states on the query "syria", for the year 2012. The graph is based on information collected in 18,582 sentences. Edge colors indicate the kind of relation (from dark red for strong alliance to dark blue for strong opposition), and vertice colors reflects proximity of countries with each other.

small peaks (< 0.4) are far less reliable. A better data smoothing could reduce this problem.

6 Discussion and Perspectives

The system described in this paper shows that it is possible to use NLP techniques to aggregate information and provide reliable numerical data that could hardly be obtained in another way. This kind of NLP-driven data journalism can bring significant added value to journalists, as well as end users, in many fields. This opens the way for many applications of the same kind, for studying relations between people, countries, organisations in any domain.

The main guidelines for building such a system are:

- Think about precision first. Do not neglect recall, since it is all about getting data, but the vast amount of information can make up the lack of coverage.
- High variation in the corpus is better than low variation, to have more chance to get data from your accurate classifier, and to give less importance to misclassifications.
- Smoothing the resulting data within temporal windows and correcting them with logistic functions is essential for hiding errors and producing reliable information.
- Make it time-aware. Interests are two-fold: some data are valid only in a limited time range, and the evolution of data can be a main interest of the study.

A wide new range of knowledge can thus become available to data journalists, who generally make use of factual data from structured bases. Such applications would bring high added value to the final users, in terms of aggregation, contextualization and hierarchical organization of information. We need however to reduce drastically the amount of needed supervision in order to make NLP enter the journalism world.

References

[Allan, 2002] J. Allan, editor. *Topic Detection and Tracking*. Springer, 2002.

[Boschee *et al.*, 2005] E. Boschee, R. Weischedel, and A. Zamanian. Automatic information extraction. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[Chambers *et al.*, 2015] N. Chambers, V. Bowen, E. Genco, X. Tian, E. Young, G. Harihara, and E. Yang. Identifying Political Sentiment between Nation States with Social Media. In *Proceedings of EMNLP 2015*.

[Dumais *et al.*, 2002] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web Question Answering: Is More Always Better? In *Proceedings of the 25th Annual International ACM SIGIR Conference*.

[Hoffart *et al.*, 2013] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 2013.

[Kambhatla, 2004] N. Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the 42nd Annual Meeting of the ACL*, 2004.

[Kessler *et al.*, 2012] R/ Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar. Finding Salient Dates for Building Thematic Timelines. In *Proceedings of the 50th Annual Meeting of the ACL*, 2012.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the ACL*, 2003.

[Lehmann *et al.*, 2013] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2013.

[Miller *et al.*, 2000] S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. A novel use of statistical parsing to extract information from text. In *Proceedings of NAACL*, 2000.

[Strötgen and Gertz, 2013] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[Turney, 2002] P. D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *40th Annual Meeting of the ACL*, 2002.

[Zhou *et al.*, 2005] G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, 2005.