Strategic Intelligence Analysis: From Information Processing to Meaning-Making

Yair Neuman, Liran Elihay, Meni Adler,
Yoav Goldberg, and Amir Viner

The Laboratory for the Interdisciplinary Study of Symbolic Process
Ben-Gurion University of the Negev, Israel
yneuman@bgu.ac.il

Abstract. Strategic intelligence involves the efforts to understand the "Big Picture" emerging from data sources. Concerning textual data, this process involves the extraction of meaning from textual information. In this paper we present a new methodology for meaning extraction from news articles. This methodology is based on the gradual construction of visual maps from processed textual information. The methodology is not a substitute for meaning-making by human agents but combines computational power with human evaluation and provides a tool for identifying emerging patterns of meaning.

1 Introduction

Information science and information technology have been shown to be indispensable for national security and intelligence analysis. However, one should realize that there are two different and complementary axes of information: "The syntactic axis concerns the formal representation of information in a digital code. The semiotic axis concerns the *meaning* of information for interpreting agents as this information appears in the analogical mode" ([6]). While the first axis has been intensively studied and formalized in information sciences, the second axis has been almost exclusively studied by other fields such as Semiotics ([10]), the study of signs and signification, and Pragmatics ([5]), the branch of linguistics dealing with language use in context. The need for a meaning-based information analysis is evident in strategic intelligence that aims to provide officials with the "*big picture*". Our aim is to present a novel methodology for meaning-based analysis of texts for strategic intelligence. The methodology leans on insights gained from studying meaning-making in various domains such as political rhetoric ([4], [7], [9]), and computational linguistics ([6]), and from the interdisciplinary background of the authors. Due to space limits, we present only the general outlines of the methodology without dwelling on technical details.

When trying to understand a text, whether a political speech or a newspaper article the reader faces a cognitive problem: The text is *linearly* organized as a string of linguistic signs. However, in order to understand the text he has to map it into an *associative* network in his mind. If we adopt the theoretical position that meaning emerges as a global property from an interconnected network of signs/concepts ([10], [11]), then the analyzer of a given text should cope with the cognitive burden of

extracting an associative network from a linear text. This task is cognitively demanding and the logic governing the process is not self-evident. It was argued by Neuman and Muchnick ([8]) that this problem is to a certain extent solvable if the linear text is mapped into a network of signs characterized by scale-free topology ([1], [2]) in which few signs have the highest connectivity to other signs. By analyzing political speeches and news articles Neuman and Muchnick ([8]) exposed a scale-free topology of the texts. A text organized as a scale-free network of signs has 3 main benefits that may guide its representation in the human mind. These benefits will be exploited in the suggested methodology. First, the text is organized around just a few hubs (i.e., its organizing concepts) that function as coordinates for understanding the whole text. Second, scale-free networks are constructed through the dynamics of *preferential attachment* ([2]). The dynamics of preferential attachment is such that a new concept/sign/node entering the text/network is more likely to attach to one of the hubs/organizing concepts. Therefore, it is the topology that directs the linkage of new entering concepts. These dynamics load the new entering concepts with meaning by associating them with few existing hubs. This is an important characteristic in the cognitive economy of mapping a linear text into an associative semiotic network. Third, the hubs do not only get their meaning from the context of their associated nodes. By playing the role of a "magnet" for other concepts, through the dynamics of preferential attachment, they also create the context for their own interpretation. In other words, the hubs are interpreted by their surrounding context and recursively provide the context for the interpretation of their associative nodes.

2   Extracting Meaning from Texts

2.1   The Data

As data we used all the news articles (N = 800) of the online monthly English-language Palestinian newspaper: *Palestine Times*. The newspaper appears in 104 volumes from 1997 to 2006. The newspaper is targeted to a Western and educated English speaking audience. As any communicative text, this political newspaper aims to provide its readers with a certain representation of the world. In this context, uncovering hidden patterns of meaning in the newspaper's articles may be of great relevance for intelligence analysis seeking to understand the way Palestinians appeal to the Western audience with regard to key issues such as the American involvement in Iraq.

2.2   The Methodology

The suggested methodology is built around several phases. At the first step a news article is segmented to sentences. The components of a sentence are classified with Part-of-Speech tagging. Only nouns, corresponding to concepts, were included in this phase. The nouns were cleaned for singular/plural and for noun-compounds. For each article we created a network. The nouns were represented as nodes and co-occurrence of nouns in a given sentence was used to establish a link/edge between the nodes. We found that the networks have a scale-free topology with only a few nouns having the greatest degree of connection. For each news article we ranked the nouns according to

their connectivity and the 5 highest ranking nouns were identified as hubs. The criterion of 5 hubs is arbitrary, but justified by the lower limit of the working memory. We considered these hubs as the basic organizing concepts of the text. It was found that at least one of the hubs appeared in the title of the article or in the opening paragraph. This finding is not surprising because it is known that in news articles most newsworthy information comes first. In the next step, we used a heuristic we call "hubs first." For each article we identified the first sentence in which one (or more) of the hubs appeared, removed this hub/s from our list, and continued searching for the next sentence in which one (or more) of the remaining hubs appeared. This process was repeated until the list was exhausted. The sentences we extracted are *context sentences* for the interpretation of a hub and were stored in a data bank for queries with the hub's name as an index. The "hubs first" heuristic is at the core of our methodology and produces valuable information for diagnosis and prediction that cannot be discussed here. We will illustrate the methodology with regard to the hub "IRAQ". Let's assume that we would like to understand the way in which the issue of Iraq is represented by the Palestinian newspaper to its audience. As we will show, an interesting pattern emerges from the analysis. Iraq appears as a hub in 8 news articles dated from March 2000 to October 2004. We organized these articles chronologically and for each article extracted (1) the hubs and (2) the context-sentences. For example, consider the first news article in which the hub "IRAQ" appears. The hubs are: Iraq, Palestinians, Government, Land and Plan. According to the "hubs first" heuristics the context-sentences are (the hubs are underlined):

S1: America presses for resettling refugees in Iraq

S2: An Iraqi opposition leader has disclosed that the Iraqi government, under pressure from the United States, was considering a plan whereby a million Palestinian refugees would be settled in Western Iraq.

S3: According to an article written by Sa'ad Farage, President of "The Iraqi Activists," a Washington based organization, the government of Saddam Hussein was giving out signals suggesting that it was receptive and open to the idea of settling Palestinians in Iraq.

S4: Among the signals, he wrote, was a recent Iraqi decree allowing Palestinians living in Iraq—around 62,000—to own land and property in Iraq.

Next we represent each sentence as a list that includes the hub(s) that appears in that sentence. In addition, drawing on Hoey's theory of lexical cohesion ([3]), if in a sentence there is a lexical item that establishes a link to a lexical item from the previous sentence (as a repetition, synonym etc.) then it is added to the list and semantically condensed. For example, the first sentence includes the noun "America" and the second includes "United States". Therefore, "America" is added to the list. The nouns are organized in the list in the order of presentation in the sentence as follows: L1: America, Palestinians, Iraq; L2: government, America, Plan, Palestinians, Iraq; L3: government, Palestinians, Iraq; L4: signals, Palestinians, Iraq, land, Iraq. By Using a Graphical User Interface, the next step is to map the lists onto a single network in which each noun is mapped to a node. A relation between two nodes is established if the nouns co-occur in a sentence. The relations are labeled
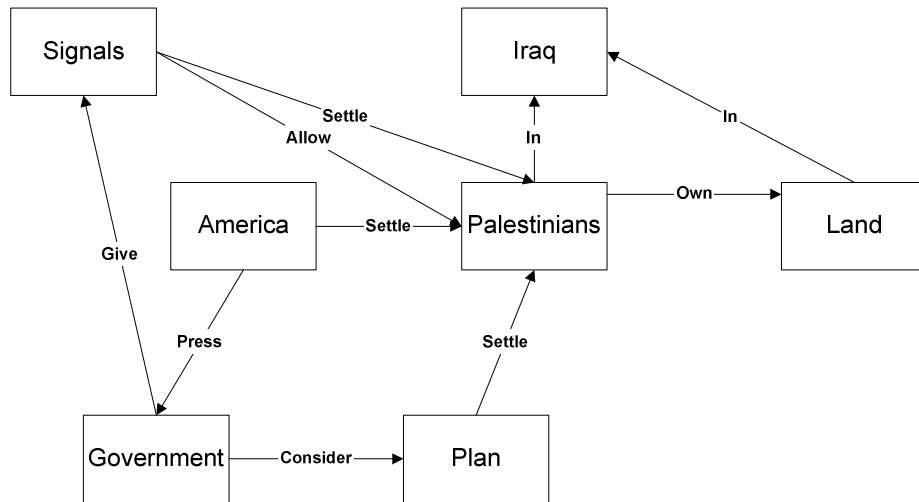
Fig. 1. The network of the first article

according to well-defined instructions. The direction of the link is determined by the question whether the second or the first nouns are the subject of the relation. The network of the first article is as follows:
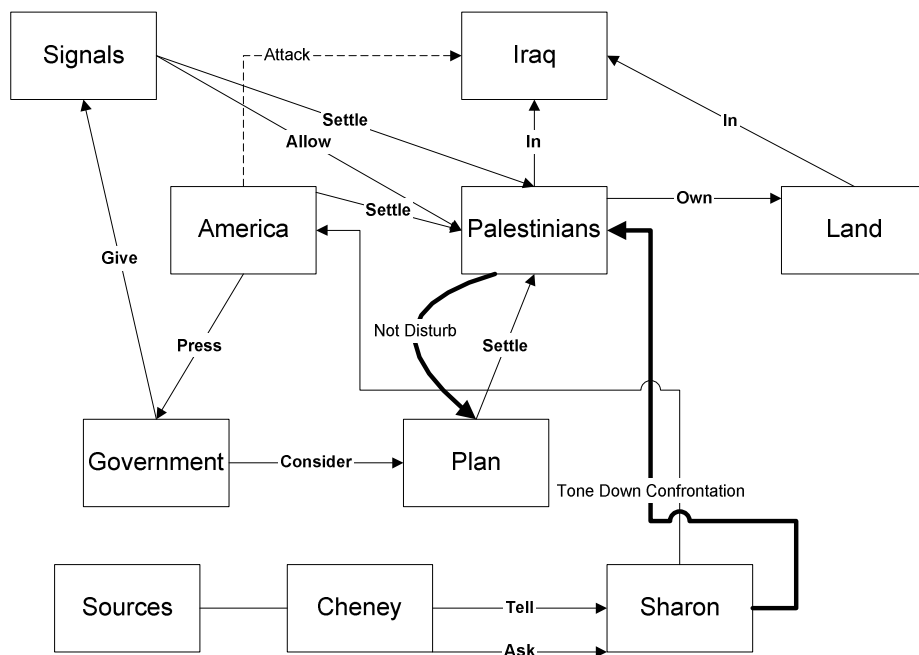

Fig. 2. The third network

The same process is applied to the next articles in the chronological order of their appearance and maps are added to the original map. For example, the third news article was published in April 2002 and it includes the following hubs: Iraq, Sharon, Cheney, Palestinians, and Sources. From this article we extracted the following context sentences: S1: Cheney tells Sharon U.S. wants to attack Iraq for Israel's sake. S2: According to Israeli sources quoted by the Israeli state-run radio on 20 March, Cheney asked Sharon to "tone down" the confrontation with the Palestinians so as not to disrupt or disturb American plans *vis-à-vis* Iraq. We created the lists and mapped them onto the network as follows, where the bold/dashed lines indicate the new important paths:

What is interesting to notice is that in this phase of the evolving network we can observe an *emerging pattern of meaning*: America not only plans to settle Palestinian refugees in Iraq (as we noticed in network 1) or to attack Iraq for "Israel's sake." An interesting link is established between Cheney's request from Sharon to "tone down the confrontation with the Palestinians" and the US plan. Cheney asks Sharon not to disturb the US from actualizing its plan. What is the plan? The plan, as indicated in the first network, is to resettle Palestinian refugees in Iraq. This interpretation makes sense in the context of the first sentence saying that "U.S. wants to attack Iraq for Israel's sake." The attack is for "Israel's sake" because it will solve the Palestinian problem through transfer of Palestinians to Iraq! This emerging *conspiracy theory* is further supported by the analysis of the other articles.

3   Conclusions and Further Work

The methodology we present is preliminary, and we are working to maximize its computational power. These efforts take into account the difficulty of accommodating noise as the analysis unfolds and some other inevitable difficulties. The system is currently being refined and tested on other corpuses. Nevertheless, one should remember that the difficulty of extracting the "Big Picture" is not a technical one. The flood of information facing the analyzer is the major barrier to seeing the "forest." In this context, the extraction of meaning from information involves a process of abstraction in which information is necessarily lost ([12]) in favor of the big picture. We may conclude that a computational tool for meaning extraction should be guided by an "oblivion algorithm" that will filter the irrelevant information in the system. Such an algorithm is currently not at hand and computational theories of abstraction and meaning-making are currently far from being applicable to strategic intelligence analysis. We may conclude by suggesting that the great challenge of intelligence analysis and informatics is exactly in developing tools that support meaning-based information analysis by scaffolding a movement from the micro level of data to the macro level of meaning.

References

1. Barabasi, A, L. *Linked*, Perseus, 2002.
2. Barabasi, A. L, Albert, R., *Emergence of Scaling in Random Networks*, Science, Vol. 286, 1999, pp. 509-512.

3. Hoey, M. *Patterns of Lexis in Text*, Oxford University Press, 1991.
4. Liebersohn, Y., Neuman, Y., Bekerman, Z., *Oh Baby, it's Hard for Me to Say I'm Sorry: Public Apologetic Speech and Cultural Rhetorical Resources*, Journal of Pragmatics, Vol. 36, 2004, pp. 921-944.
5. Mey, J. L. *Pragmatics: An Introduction*, Blackwell, 2002.
6. Neuman, Y. *A Theory of Meaning*, Information Sciences, Vol. 176, 2006, pp. 1435-1449.
7. Neuman, Y., Levi, M., *Blood and Chocolate: A Rhetorical Approach to Fear Appeal*. Journal of Language and Social Psychology*, Vol. 29, 2003, pp. 29-46.
8. Neuman, Y., Muchnick, L. *On the Small-World of Political Rhetoric*, Submitted.
9. Neuman, Y., Tabak, I. *Inconsistency as an Interactional Problem: A lesson from Political Rhetoric*, Journal of Psycholinguistic Research*, Vol. 32, 2003, pp. 251-267.
10. Sebeok, T. A., and Danesi, M. *Forms of Meaning*, Mouton de Gruyter, 2000.
11. Sigman, M., Cecci, G. A. *Global Organization of the Wordnet Lexicon*, PNAS, Vol. 99, 2002, pp. 1742-1747.
12. Zucker, J-D. *A Grounded Theory of Abstraction in Artificial Intelligence*, Phil. Trans. R. Soc. Lond. B, Vol. 358, 2003, pp. 1293-1309.