

Comparing Business Intelligence Methods: Textometry and Information Extraction for mining the Enron Crisis

Erin MacMurray¹

¹Université Paris 3 Sorbonne Nouvelle – erin.macmurray@gmail.com

Abstract

Business Intelligence, a field involving information science and economics attempts to collect and analyze pertinent information for developing business strategies. The aim of this paper is to present a Textometric method for identifying events and to compare these results to an NLP approach: pattern-based information extraction. The results of a series of Textometric experiments on a corpus of New York Times articles from the Business/Financial section from November 2001 to March 2002 will be compared to the relationships extracted between Enron and other named entities by an information extraction system. Emerging co-occurrences will be used to chronologically identify the event the Enron crisis.

Résumé

A la frontière des disciplines de l'économie et des sciences de l'information, la veille stratégique se préoccupe de la collecte et de l'analyse d'informations pertinentes pour l'établissement de la stratégie d'une entreprise. Notre objectif est de présenter une méthode textométrique pour détecter des événements à partir du texte brut et de comparer les résultats de cette méthode à ceux d'une approche TAL robuste : extraction de l'information à base de patterns. Dans cette perspective, une série d'expériences textométriques sur le corpus New York Times de Novembre 2001 à Mars 2002, (rubrique Business/Financier) seront comparées aux relations extraites entre Enron et d'autres entités nommées pour la même période par un système d'extraction d'informations. Les cooccurrences évolutives sont utilisées ici comme méthode de fouille dans le but de détecter le déroulement d'un événement économique, la faillite d'Enron.

Keywords: Textometry, Information Extraction, business intelligence, co-occurrences, event identification

1. Introduction

Business Intelligence, a field involving information science, economics and the humanities, attempts to collect and analyze pertinent information for developing business strategies. One of the many tasks in this vast field is extracting information from textual sources such as online news productions. Mining online news media for information is a laborious task that has seen a recent surge in automatic solutions for processing textual data. However, few studies have been done on the information gain actually attained by such solutions, especially for textual data in the economic sector (Alex *et al.*, 2008). This raises the question of how to measure the

information gain actually obtained by these solutions for the specific business intelligence task of current event identification. One angle of attack is to compare different methods of mining textual data. This study attempts to compare the results of an analysis done on the same corpus concerning the same named entity (NE) “Enron” using two different methods of mining textual data: a pattern-based information extraction methodology and the Textometric method of co-occurrences.

There are many natural language mining techniques: machine learning and information extraction (semantic and morpho-syntactic patterns) to name just a couple as discussed during the Message Understanding Conferences (MUC) (Grishman et Sundheim, 1996). Text Mining, generally seen as a subfield of Data Mining, is roughly defined as the processes used to extract and structure unstructured data (Feldman et Sanger, 2006), where ‘unstructured’ usually means in ‘natural language’. In this case, the Text Mining process aims to structure unstructured text in order to apply standard data mining techniques to the newly generated structured output. This process generally uses an Information Extraction (IE) step to identify and organize the textual data according to named entities and the relationships between them (Feldman et Sanger, 2006; Grishman, 2001; Poibeau, 2003). These techniques use pre-defined information models to determine not only the information of interest, but also the possible interactions between the named entities before applying the extraction system to the text, in other words an analytical approach.

Textometry, however, does not use a pre-defined information model and attempts to derive trends across comparable zones of text using statistical and probabilistic calculations on the textual units therein (Lebart et Salem, 1994; Tufféry, 2010). The notion of structure is therefore inherent to the distribution of the words across such zones. In this study, the co-occurrence method (Lafon, 1980; Martinez, 2003) was used to gather information specifically implicating the NE Enron during a given month of the corpus, in other words an empirical approach. The hypothesis is that textual units that are statistically significant for specific month of the corpus will provide similar information to the Relationships extracted for the same month by an IE system.

2. Corpus and Event Identification

The focus for this study was on one type of business intelligence task: identifying information concerning a company NE. The NE Enron and the major economic crisis caused by its collapse in 2001-2002 were selected as the object of analysis for this study. Though this event is not current, the corpus was treated as such. In both analyses, IE and Textometric, the text was separated by month in order to simulate a monthly information flow as in many business intelligence contexts. This analysis was done with little prior knowledge of the Enron crisis and the chronological order of events that made up the crash.

The corpus for this study was taken from the New York Times Annotated Corpus (Sandhaus, 2008). Articles corresponding to the Business/Financial Desk were stripped of their xml metadata and put into txt format for more efficient analysis by the Textometric tool *Le Trameur* (Fleury, 2007). Only the months corresponding to the core of the Enron crisis were considered for this study, from November 2001 to March 2002. This was detected beforehand using the absolute frequency of the NE as measurement. During this five-month period the frequency of

Enron is considerably higher than before or after, which leads to the hypothesis that something is “happening” involving the NE during this time. The final sub-corpus contains only Business/Financial articles that mention Enron during this period. A total of 749 articles, 758,424 tokens for 45,608 types make up the corpus Enron11-03. Lower case letters were forced for all tokens in the corpus Enron11-03 for the Textometric analysis.

2.1. Pattern-based IE

The pattern-based IE system TEMIS Skill Cartridge Competitive Intelligence™ (CI™) was used to extract relationships involving the NE Enron for this analysis. This IE component is specifically used to extract economic relationships between company or person NE for business intelligence applications (Pauna et Guillemain-Lanne, 2010, Grivel *et al.*, 2002). This system uses lexical patterns to determine the relationships to be extracted between NE. The TEMIS IE apparatus follows these steps:

1. Part of Speech tagging and lemmatization by Xelda™,
2. Information Extraction by CI™ using lexical trigger patterns,
3. Output extractions structured in NE and economic relationships.

The output was manually evaluated for precision; only precise extractions were retained for comparison in this study. The average precision of CI™ relationships relevant to the NE Enron on this corpus was about 42%. This result is low in comparison to internal evaluations done by Temis that showed overall precision between 60% and 70% depending on the corpus. For purposes of this comparison, only relationships with a precision of 60% or higher were retained¹. Sentences can correspond to two different relationships; such examples are therefore extracted twice and evaluated for precision according to the relationship with which they are tagged.

2.2. Emerging Co-occurrences

The Textometric method used for this study follows the chronological specificities methodology (Salem, 1994). The co-occurrences method, as recently defined by W. Martinez, (2003) display what vocabulary is significant to the selected pivot-type for the period studied compared to all previous periods in the corpus. This sheds light on vocabulary that *emerges* as well as vocabulary that *disappears* for the time-period considered. In order to obtain enough comparative data for the Enron crisis, each month was compared to all previous months dating back to January 2001. The co-occurrences therefore *emerge* for the current month being analyzed for the pivot-type *enron*. What results is a co-occurrence network of specific lexical units for each month from November 2001 to March 2002. The resulting networks of co-occurrences can be interpreted according to the following parameters (Lafon, 1980; Martinez, 2003):

Frequency: the total frequency of the type in the corpus;

Co-Frequency: the frequency with which the pivot-type and the co-occurrence simultaneously appear in the defined context;

¹ A precision of 60% or higher corresponds to average scores obtained by Temis on press corpora such as Media Box and Coface, as well as the National Institute of Standards and Technology, information extraction task:http://www-nlpir.nist.gov/related_projects/muc/. Automatic Content Extraction (ACE) evaluation ranges relationship and event precision between 50%-70% (Sarawagi, 2010).

Specificity: index indicating the over-representation (or under-representation) of the co-occurrence in the defined context in relationship to the pivot-type and the co-occurrence in separate contexts;

Number of contexts: the number of contexts in which the pivot-type and the co-occurrence appear simultaneously.

For reasons of maintaining control parameters in this analysis, a co-frequency of 10 and a *specificness* of 6 were applied for each co-occurrence calculation². These figures were chosen after experimenting with various perimeters to find a combination that allowed for the most exhaustive co-occurrence network without making reading the graph impossible.

Due to the difficulty in displaying the numerical figures for the co-occurrence calculations below, figures 3 to 5 show results using color code for *specificness* level and line thickness for the number of contexts. Table 1, here, serves as a guide for reading these figures.





Color	Level of <i>Specificness</i>	Threshold Thickness	Number of Contexts
red	> 50	1 	1-20
orange	13 <= 50	3 	21-40
green	9 <= 12	5 	41-60
blue	6 <= 8	7 	60 and over

Table 1- Cooccurrence information guide for figures 3-5

3. Results on the Enron crisis

The Enron crisis was, and still is, one of the largest bankruptcies in history riddled with overt accounting fraud. This company, “too big to fail”, crashed on December 2, 2001, announcing 2002 as the start of an economic crisis with many other major accounting scandals to follow during that year (Worldcom, Vivendi, ImClone). Table 2 provides selective chronological facts of the Enron crisis to guide the results discussed below.

² The resulting figures will be noted hereafter in the examples: cofrequency(specificness)contexts

October 2001: Destruction of documents, the SEC begins an investigation of Enron and its financial partnerships. A. Fastow, Chief Financial Officer is dismissed from Enron.
November 2001: Enron discovers an over estimation of profit of \$600 million dating back to 1997.
November 2001: Dynegy, energy company, attempts to acquire Enron. After a month Dynegy refuses the merger due to fraudulent accounting practices.
November 2001: The SEC extends its investigation to Enron’s accounting firm Arthur Andersen.
December 2001: Enron files for chapter 11 bankruptcy and terminates 4,000 employees.
December 2001: Shareholders file a lawsuit against Enron executives. Dynegy also files a lawsuit against the company.
January 2002: A congressional committee begins its criminal investigation of Enron. The firm Arthur Andersen admits to the destruction of documents and K. Lay resigns as CEO of Enron.
March 2002: Congressional committee continues its criminal investigation. Arthur Andersen is implicated for obstruction of justice.

Table 2 – Facts on the Enron Crisis³

3.1. Relationships extracted from November 2001 to March 2002

A total of 638 correct extractions containing the NE Enron were collected in the corpus. A total of 8 different relationships⁴ describe Enron’s actions from November to March with a precision of at least 60%. The different relationships extracted are provided in table 3 below. Only 3 relationships of the 8 were consistently extracted from November 2001 to March 2002 ([*Bankruptcy*], [*Court case*], [*Stock Information*]) the remaining relationships appear and disappear unevenly over this period as will be discussed below.

Relationship CI TM	Extractions	Precision	Exemple Extraction
Acquisition	32	62%	<i>The acquisition of Enron by Dynegy.</i>
Bankruptcy	261	88%	<i>Enron filed for chapter 11 bankruptcy last week.</i>
Court case	136	84%	<i>Dynergy faces a possible lawsuit from Enron</i>
Financial information	6	75%	<i>Enron reported losses for this month.</i>
Financial reporting	39	88%	<i>Enron overstated its earnings by \$6 billion.</i>
Merger	11	92%	<i>Enron was formed by the merger of Houston Natural Gas and InterNorth</i>
Management changes	16	94%	<i>Kenneth Lay has resigned as chief executive of Enron.</i>
Stock information	49	74%	<i>Enron’s stock fell as low as 7 cents a share.</i>

Table 3- Number of CITM extractions, precision, and example sentences in Enron11-03 corpus

Figure 1 shows the fluctuations of the five most precise relationships extracted in the corpus according to month. It is interesting to note that, with the exception of the [*Bankruptcy*] and

3 For this chronology we compiled facts analyzed in the following sources: (Andersen, 2002 ; Lowensetein, 2004), Enron Timeline- documentary (Gibney, 2005, from the book by McLind et Elkind, 2004) : *The Smartest Guys in the Room* <http://www.pbs.org/independentlens/enron/timeline.html> (consulted 10/2011)

4 Relationships will be designated [*Relationship*] in this paper.

[*Financial Reporting*] relationships, the others follow the chronological unfolding of events that make up the Enron crisis. The [*Acquisition*] relationship has a greater number of extractions for the month of November and decreases thereafter. The extractions as in the example [extraction-acquisition NOV] correspond to the attempt by Enron to be bought by the smaller energy company Dynegy during this month.

[extraction-acquisition NOV] **Dynegy Is Said to Be Near to Acquiring Enron for \$8 Billion**

The same analysis can be done on [*Stock Information*] relationship, which extracts information on the fluctuations of shares. Due to Enron's uncertain future, it is consistent that its shares would shift depending on news of its acquisition.

[extraction-stock information NOV] **Enron's shares have fallen by more than half** in the last two weeks because of the S.E.C. investigation and worries about off-balance-sheet debts and transactions with investment partnerships involving the company's former chief financial officer, Andrew S. Fastow, who was ousted last week.

The [*Court case*] relationship, which steadily increases from November to March, is in keeping with the lawsuits filed against Enron by shareholders, the company Dynegy and also by the American government. These lawsuits often followed investigations opened in January. Finally, the [*Financial Reporting*] relationship, which extracts information on numerically reported profits or losses, increases in a manner similar to the [*Bankruptcy*] relationship, both reaching a peak for the month of January. Enron officially declared its bankruptcy in December, yet the number of extractions for this relationship increases during this period. From these results it would seem that this kind of information, along with Enron's profit misstatements ([*Financial Reporting*] extractions), was discussed by the media more frequently during January than the actual month of occurrence.

[extraction-bankruptcy JAN] Investors have largely shrugged off the collapse of **Enron, the Houston energy trader that filed for bankruptcy** after admitting that it had for years overstated its profits and understated its debts.

[extraction-financial reporting JAN] Investors have largely shrugged off the collapse of **Enron, the Houston energy trader that filed for bankruptcy** after admitting that it had for years overstated its profits and understated its debts.

[*Financial Information*] focuses on information such as profits, losses without associated numerical figures. This relationship only has a few extractions for the months of December 2001, January and February 2002. It is difficult to accurately evaluate this relationship due to the low number of extractions; 3 of the 6 appear in January following the trend observed for [*Financial Reporting*].

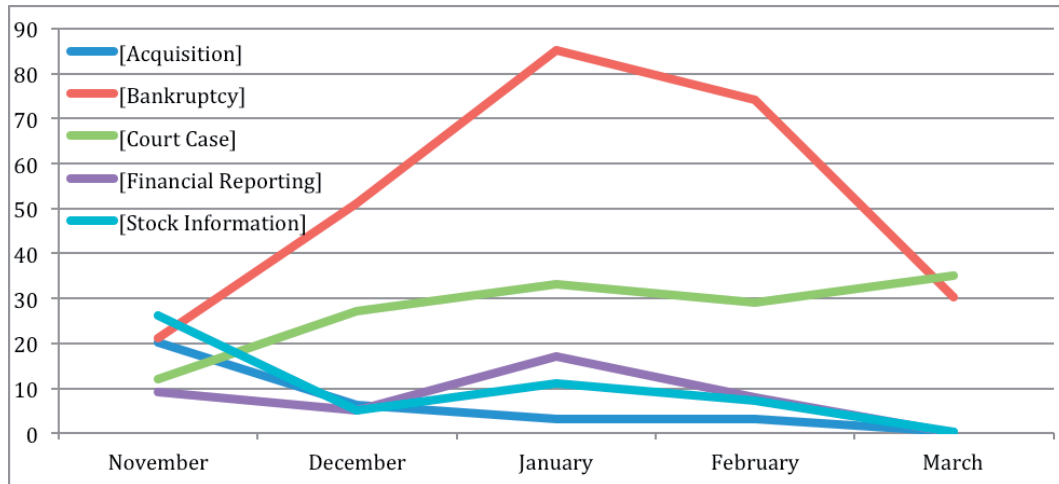


Figure 1- extracted CI™ relationships from November 2001 to March 2002

Other CI™ relationships that are less present in the corpus than the five mentioned above, follow trends that are consistent with emerging events of the Enron crisis. [*Merger*] extracts information similar to that of [*Acquisition*], meaning sentences that correspond to a company merging or purchasing a company. This relationship also reaches its peak for the month of November during the merger discussions between Enron and Dynegy. The [*Merger*] relationship is extracted during the month of January during the investigations of the merger contract between the two companies as well as for background information on Enron as in the following example.

[extraction-merger JAN] Enron had been formed in mid-1985 by the merger of Houston Natural Gas and InterNorth.

Likewise, the peak for the relationship [*Management Changes*] corresponds to the resignation of K. Lay as CEO of Enron, which also occurs during this month. Surprisingly, the lexical unit *lay* does not appear in the co-occurrence network for this month (figure 5).

[extraction-management-changes JAN] By stepping down as Enron's chairman, Kenneth L. Lay, begins his new career, that of a defendant in lawsuits and a witness before congress.

This relationship also extracts background information for the months of November and February. The extracted content gives information on previous position changes, such as the dismissal of A. Fastow and the resignation of J. Skilling as CEO of the company. These actors (units *fastow* and *skilling*) do appear in the co-occurrence networks for February, however, not necessarily in the same contexts as the relationship.

3.2. Emerging co-occurrences from November 2001 to March 2002

When applying the co-occurrence calculation to the pivot-type *enron* on a monthly basis, similar information is obtained to that of the relationships in the extractions above.

November 2001

For November 2001, the co-occurrence network displays lexical units pertaining to the *acquisition* and *merger* of Enron by the company Dynegy, as also extracted by the IE system. Further information is shown on the financial health of Enron through the units, *debt*, *debts*, in the same way the relationship [Stock information] extracted sentences on the fluctuation of Enron shares.

[art 58, 2001-11] but while jp morgan chase is proud of serving alongside citigroup as both lead lender and adviser to **enron** on its *acquisition* by *dynegey*, the dual role it has worked to achieve sometimes proves complicated for the bank. (*acquisition*: 18(7.7)18; *dynegey*: 228(52.1)207)

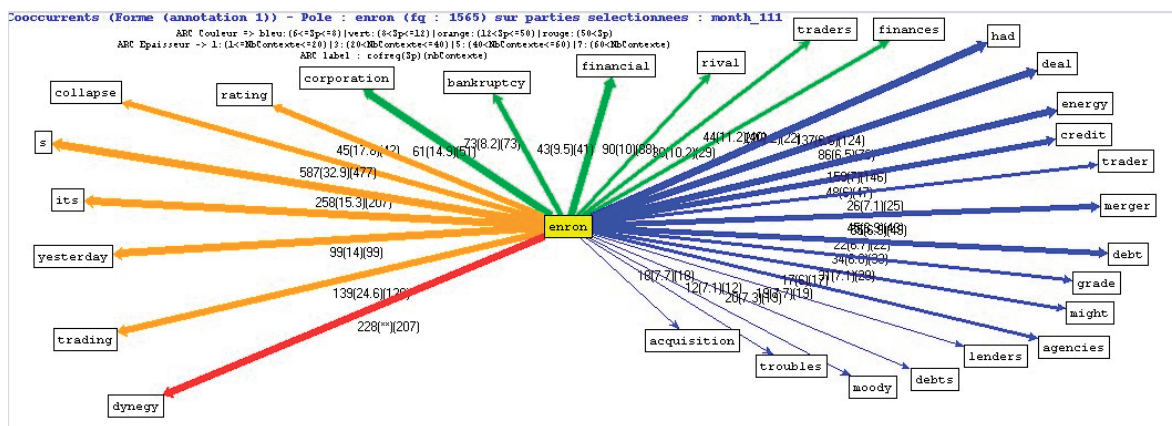


Figure 3- “enron” co-occurrences for November 2001

December 2001

The co-occurrences for the month of December display, as expected, Enron’s filing for bankruptcy. This month also displays the emerging lexical units *lawsuit*, *court case*, and *sued* that Enron is involved in, similar to the increase noticed in the [Court case] relationship for the month of December, figure 2.

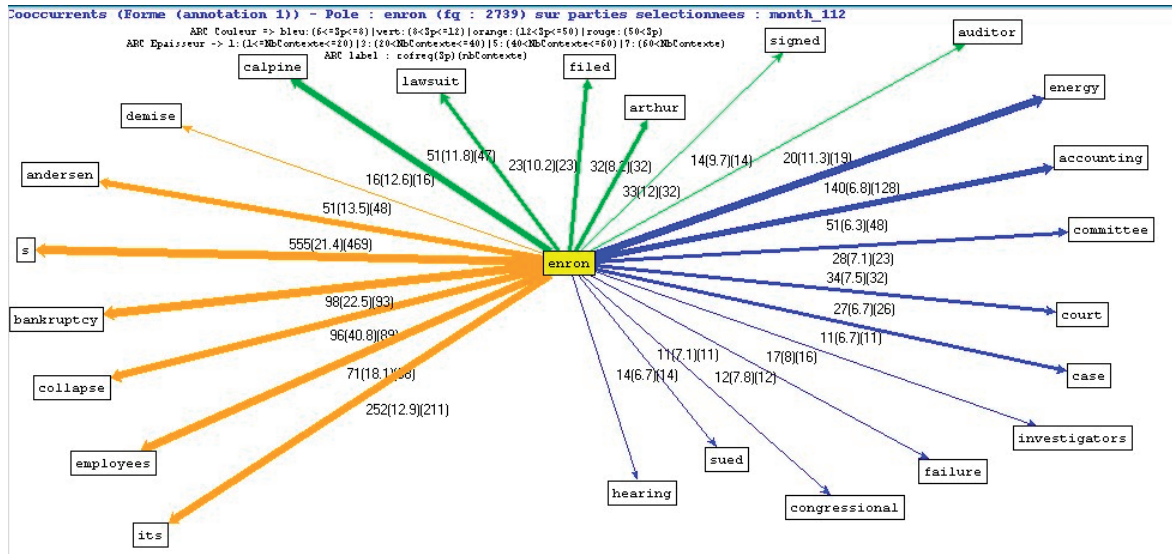


Figure 4- "enron" co-occurrences for December 2001

Contrary to the results found by the IE system, the co-occurrences of *enron* detect all possible expressions of the company's fall, as seen in the examples (art 96, 2001-12 and art 96, 2001-12). The lexical unit *collapse*, is not one of the many patterns pre-defined by the information for the [Bankruptcy] relationship in the IE system studied. In this case, such information would go undetected by the system.

[art 96, 2001-12] the fallout from *enron*'s *collapse* continued on Friday as the company struggled to line up financing. (*collapse*: 96(40.8)89)

[art 96, 2001-12] ripples spreading from *enron*'s expected *bankruptcy*. (*bankruptcy*: 98(22.5)93)

January 2002

The month of January produces the largest co-occurrence network of the five months considered, figure 5. This can be explained in part by the increase in the number of tokens for this month compared to the other months, a difference of 147,717 tokens between December and January. In contrast to the increased [Bankruptcy] and [Financial reporting] relationships observed for this month, the co-occurrences display lexical units relevant to the congressional investigations of Enron and the revelations of corrupt activities during these hearings:

- destruction of documents: *shredding* (28(7.9)27), *destruction* (60(13.1)59), *documents* (145(26.8)131),
- fraudulent partnerships: *partnerships* (123(19.1)121), *ljm2* (31(13.0)25), *raptor* (30(10.3)23),

Certain lexical units, other than the unit *bankruptcy* (144(6.9)137) which appears for this month, describe the crash of Enron, *collapse* (320(111.4)307), *debacle* (40(19.7)40), *fall* (66(11.8)65), and *scandal* (43(17.1)43). These units are not included in the patterns defined for CI™, and therefore, this kind of information was not extracted for the month of January.

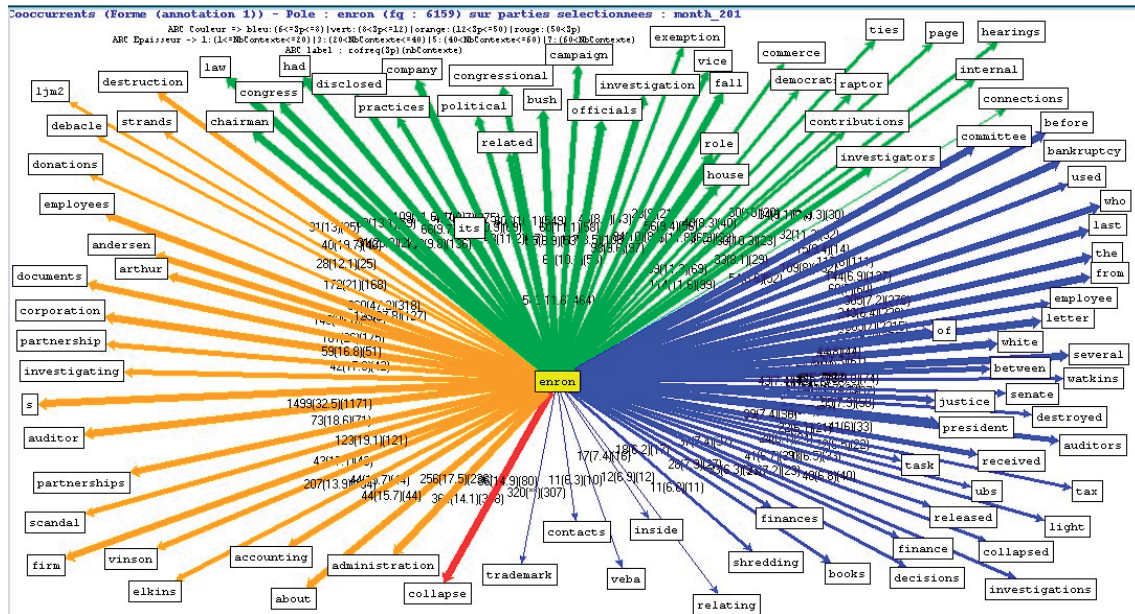


Figure 5- “enron” co-occurrences for January 2002

February-March 2002

Both February and March produce dense co-occurrence networks compared to the first two months studied. However, these months do not generate a great deal of new lexical units compared to those that emerged for the month of January. Vocabulary associated with the collapse of Enron, corrupt accounting practices, and their investigation, remain pertinent during these two months (art 476, 2002-02). Furthermore, this type of detail is not be picked up on by the IE system.

[art 476, 2002-02] investigators picking throught the wreckage of **enron**, seeking to understand what caused its collapse in decembre, have explored its **byzantine partnerships** and financial strategies. (*byzantine*: 11(10.5)11; *partnerships*: 192(38.5)185)

The example (art 487 2002-02) shows the co-occurrence of *enron* and the *401k*⁵ plan. Here, again this lexical unit is not part of the pre-defined information model, yet the co-occurrence method shows that it is statistically significant in the context of the pivot-type *enron* for the month of February.

[art 487, 2002-02] **enron** executives sold large amounts of stock, the company barred **employees** from selling their shares in their **401(k)** plans last fall as the price plummeted. (*employees*: 201(25.4)183; *401*: 58(6.5)57; *k*: 88(7.4)87)

The example (art 719, 2002-03) shows the co-occurrence of *enron* and *wessex water*, which is statistically significant for the month of March due to Enron’s shedding of most of its assets. This information was also found by the IE system with the *Selling* relationship that spiked for the month of March, figure 2.

5 401k is used to describe the type of retirement savings account in the United States that allows employers to help employees save for retirement while reducing taxable income under a specific provision of the US tax code.

[art 719, 2002-03] the azurix corporation of houston, which was formed as holding company for enron's water assets when it bought wessex water in 1998 for \$1.9 billion ... (houston: 61(8.5)60; water: 22(10.5)15; wessex: 12(10.2)11)

4. Comparing IE output to Co-occurrence Networks

4.1. Comparable content

As seen above, the relationships extracted are comparable to the lexical units in co-occurrence with the type *enron*. Namely, the [Bankruptcy] relationship can be compared to the lexical unit *bankruptcy* that appears from November to February. The chronology of both the relationship (in red below) and the co-frequency of *enron* and *bankruptcy* (in blue below) are very similar, peaking in the month of January. In this case, the pattern used to extract the relationship corresponds in most cases to the actual co-occurrence of *enron* and *bankruptcy* in the sentence.

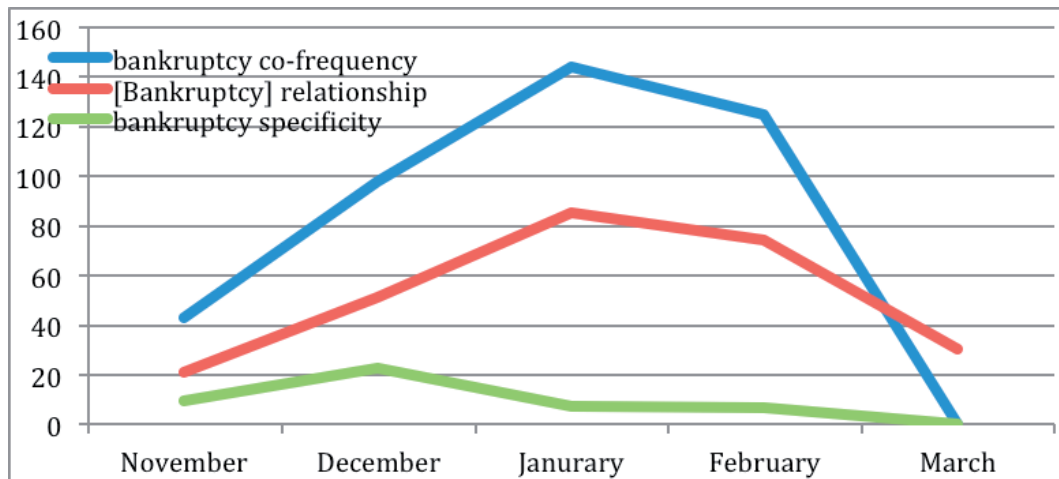


Figure 6- chronology of the relationship [Bankruptcy], co-frequency and specificity for “enron” and “bankruptcy”

Similar chronologies between the number of extracted relationships and the co-frequency of corresponding lexical units were noted for [Acquisition] and [Merger] as well as [Court Case]. Both [Acquisition] and [Merger] have co-occurring units *acquisition* and *merger* for the month of November which is in keeping with the sequence of events of the Enron crisis. Likewise, [Court Case] can be compared to such units as *lawsuit*, *court*, *case*, *sued*, *investigations*, *hearings*, etc. that appear from January to March.

4.2. *Emerging characteristics*

When observing the chronology of the specificity for the co-occurring unit *bankruptcy* (in green figure 6), the result is slightly different. This unit reaches its high point in December, the actual month Enron declared bankruptcy. Here, the specificity is the result of the co-occurrence calculation. For the month of December the units *enron* and *bankruptcy* display a greater attraction than that of the preceding months, thus showing how the units *emerge* on a temporal axis. This characteristic was also seen when looking at the average specificity for *court case* vocabulary, *lawsuit*, *court*, *case*, *sued*, *investigations*, *hearings*, for example. This result seems to correspond to the *real-world* unfolding of events. Relationships and the co-frequency of corresponding terms follow repetition in the discourse of the press.

Lexical units that co-occur with the type *enron* also provide more detail of what is salient for a particular month rather than an overview of normalized relationship categories. The co-occurrence methodology is not based on a pre-defined information model and can therefore follow the natural trends of the chronological text dynamic. In other words, co-occurring units that described the fraudulent accounting practices and the collapse of Enron emerged clearly over the five month period; however, no equivalent relationships were available to pick up on this information. Furthermore, precoded patterns in the IE system are also relatively costly to develop and maintain as they need to be updated regularly to remain current. Any new relationship must first be defined, modeled, and then coded before the extraction process can take place. A Textometric method works on directly on the textual units thus having no maintenance cost issues.

4.3. *Single extractions and normalized view of information*

The co-occurrence network often missed background information on the Enron crisis that was no longer current for a given month. Background information often corresponds to few or single extractions in the month. In the example [extraction-acquisition FEB] this February article recalls the attempted merger between Enron and Dynegy in November. The co-occurrence between the forms *enron* and *dynegy* were no longer sufficiently salient for February to detect this type of content. This was particularly the case for the [Acquisition] and [Merger] relationships.

[extraction-acquisition FEB] When **Dynegy canceled its proposed acquisition of Enron** -- a deal that came together in less than 10 days in November before Enron filed for bankruptcy ...

Other relationships showed no equivalents in the co-occurrence networks. This was especially the case for [Financial Reporting], [Financial Information] and [Management Changes]. The articles that make up the corpus correspond to the Business/Financial column; therefore, information on earnings, profits, losses as well as position title changes are quite common in the press discourse. Such information can be found in a variety of contexts, not only co-occurring with *enron*.

The IE system also normalized the extractions providing a more analytical and categorical representation of the content. For example, a business analyst would interpret Arthur Andersen as a single segment, yet it appears as two lexical units in the co-occurrence network. Likewise, the [Court Case] relationship corresponds to a number of co-occurring units described above.

Normalizing information can thus help target certain categories that pertain to a NE, even if the information is not new.

5. Conclusion

This study presented a comparison of two methods, analytical and empirical, for mining information on the NE Enron during its collapse from November 2001 to March 2002. Both methods were useful in pinpointing events as they unfolded chronologically and gave similar results. The IE system provided a more homogenous display of information categories whereas the Textometric method gave greater detail as to the emerging actions the company was involved in. Information was missed by both methodologies and actual interpretation of the results was always left to the analyst.

Although this comparison yields interesting results, there are a few limits to this research. First, the amount of noise produced by IE system was not taken into account in these figures, in other words, only precise extractions were considered here. Wrongly extracted relationships by an IE system can greatly increase the amount of time it takes to analyze information, almost defeating the purpose of any automatic interpretation of the data thereafter. This comparison was done after the evaluation process and only takes into account accurate extractions. In contrast, Textometry can work on raw textual data and therefore has the advantage of not producing incorrect annotations. However, a Textometric method almost systematically demands manual interpretation of the data. Though the co-occurrence network appears to be an accurate *summary* of the data, the analyst must refer back to the text in order to correctly understand the presence of a lexical unit.

Second, the IE system had the advantage of normalizing and homogenizing fixed linguistic expressions. It is possible to imagine a Textometric methodology combining both repeated segments (Salem, 1987) and co-occurrences that would resolve this issue. This, however, is neither implemented, nor tested at the moment.

Finally, the comparison introduced by this paper presents a certain number of difficulties in precisely measuring the information output by both methodologies. A sentence-by-sentence, precision-like comparison would not give a true representation of the contributions of either approach. Both methodologies are sufficiently distinct, each catering to a different vision of how information is to be extracted and stored. Determining the frequency of certain types of information pertinent to an event, such as background information vs. emerging actions, as discussed in part 4 may help give a clearer measure of the advantages and disadvantages of both approaches.

Future research should also be extended to how the two approaches can be used in combination. Such methods are already being employed for research in Consumer Relationship Management or opinion mining (Feldman et al, 2010). Textometry can also be used as a means for evaluating IE output results by aiding recall measures through the comparison of the co-frequency of lexical units to the number of relationships extracted. The pertinence of extracted relationships as they represent current events, can be compared to the statistically emerging units for a given month.

References

- Alex B., Grover C., Haddow B., Kabadjov E., Matthews M, Roebuck S., Tobin R., Wang X. (2008). *Assisted Curation: Does Text Mining Really Help?* Pacific Symposium on Biocomputing.
- Anderson J.M. (2002). Enron: a Select Chronology of Congressional Corporate, and Government Activities. CRS Report for Congress.
- Feldman R., Goldenberg J. Netzer O. (2010). *Mine Your Own Business: Market Structure Surveillance Through Text Mining*. Columbia University, Working Paper.
- Feldman R. et Sanger J., (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, p. 422
- Fleury, S. (2007). *Le Métier Textométrique: Le Trameur, Manuel d'utilisation*. University Paris 3 Centre de Textométrie. <http://tal.univ-paris3.fr/trameur/>
- Grishman, R. and Sundheim, B. (1996). *Message Understanding Conference- 6: A Brief History*. Proceedings of the 16th International Conference on Computational Linguistics (COLING), I. Kopenhagen, p.466–471
- Grishman, R. (2003). Information Extraction, in R. Mitkov, *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, p. 545-559.
- Grivel L., Guillemin-Lanne, S., Coupet, P. Huot, C. (2001). *Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance*. Proceedings Veille Stratégique Scientifique and Technologique, Toulouse.
- Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, 1 octobre p. 127-165.
- Lebart, L. and Salem, A. (1994). *Statistique textuelle*. Paris, Dunod, versions auteur disponibles en ligne : à l'ENST <http://ses.telecom-paristech.fr/lebart/ST.html> et à Paris 3 <http://www.cavi.univ-paris3.fr/lexicométrica/livre/st94/st94-tdm.html>
- Lowenstein R. (2004). *The origins of the crash: The Great Bubble and its Undoing*. Penguin Books, NY, USA.
- Martinez, W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3.
- McLean B. and Elkind P. (2003). *The Smartest Guys in the Room, The Amazing Rise and Scandalous Fall of Enron*. Penguin Books, NY, USA.
- Pauna R., Guillemin-Lanne, S. (2010) *Comment le text mining peut-il aider à gérer le risque militaire et stratégique ?* Proceedings Veille Stratégique Scientifique and Technologique, Toulouse.
- Poibeau T. (2003). *Extraction automatique d'information. Du texte brut au web sémantique*. Paris : Hermès Sciences.
- Salem A. (1987). *Pratique des segments répétés*, Paris, Klincksieck.
- Salem A. (1994). *La lexicométrie chronologique*, Actes du colloque de lexicologie politique « Langages de la Révolution », collection « St. Cloud » Paris, Klincksieck.
- Sandhaus, E. (2008). *The New York Times Annotated Corpus*. Philadelphia Linguistic Data Consortium. http://www ldc.upenn.edu/Catalog/docs/LDC2008T19/new_york_times_annotated_corpus.pdf
- Sarawagi, S. (2010), Information Extraction, in *Foundations and Trends in Databases*, vol. 1, no. 3.
- Tufféry S. (2010). *Data mining et statistique décisionnelle: l'intelligence des données*. Paris : Editions Technip.