

Whitepaper on a 360 Degree Strategy for Text Analysis

Cohan Sujay Carlos

Researcher, Aiaioo Labs
Benson Town, Bangalore, India
<http://www.aiaioo.com>
cohan@aiaioo.com

Abstract

We propose a strategy for converting pretty much any kind of unstructured text into structured forms that are compatible with relational database management systems. The benefits of such a conversion would be that it makes it possible to further analyse the information using dimension tables, OLAP cubes, and other data analysis techniques that are normally used with structured data. We also propose a novel analysis technique for unstructured text that we call intention analysis, and explain how it subsumes sentiment analysis. We also argue that intention analysis, when integrated with event and fact analysis, might produce a 360 degree view of the meaning of any text. We conclude with some thoughts on transforming images into structured information.

1 Introduction

Tools for Business Intelligence tend to offer users a fairly standard set of analytics capabilities that include, among others, classification, clustering, association rule mining, OLAP cubes and query languages for the same. All of these capabilities are focused on structured text that usually originates in what is known as an OLTP system. Thus there is a need for a strategy for integration of unstructured data in the BI pipeline. This is particularly relevant in the light of a Gartner report that claims that unstructured information comprises no less than 80% of the useful information available to a company.

In this whitepaper, we propose methods for analysing text that convert text (unstructured information) into structured tables that can be stored in an RDBMS (also known as the OLTP system). This structured information can be analysed using any of the available BI analysis tools - using

OLAP cubes for instance. The conversion of the unstructured information into structured form also makes it possible to use the standard reporting and visualization tools of BI products to display results of analyses to users (in the form of standard graphs and charts).

2 Background

BI products typically consist of a stack of tools build on top of an OLTP system which is typically an RDBMS. The OLTP system consists of relational tables containing fields (entities). An instantiation of the entities or fields constitutes a relation or a row in the table.

So, to convert unstructured textual information into a form that a standard BI product can use, you need to map portions of text into rows in a table. These rows can be stored in the OLTP system, and then processed using the layers above it. The results of analysis using OLAP cubes, etc., can then be displayed using the BI Dashboards that typically come with the BI product.

3 Conversion Process

In this section, we examine how portions of text can be converted into rows in a table.

In a paper in 1987, titled A Language/Action Perspective on the Design of Cooperative Work, Winograd proposed the concept of a "Conversation for Action (CfA)". Essentially, he proposed that all speech might belong to one of two categories:

- Speech that is intended to communicate information, and
- Speech that is intended to cause the listener to act in some manner (this class of speech is called a speech act).

The first category (i.e., speech that is meant to convey information, be it a fact or event) can be

Category of Intention	Subtype of Intention
Purchase	
Sell	
Inquire	
Direct	
Compare	
Suggest	
Opine	
Praise	Opine
Criticise	Opine
Complain	
Accuse	
Quit	
Express	
Thank	Express
Apologize	Express
Empathise	Express

Table 1: List of Intention Types.

analyzed using semantic web tools to extract entities, relations and events.

The second category (i.e., speech that is meant to cause the listener to act) can be analyzed using a novel type of analysis called intention analysis, for which Aiaioo Labs has published the first commercial API. The result of intention analysis is a set of entities, namely an actor, object and optionally an alternative (apart from the category of intention).

Events, facts or intentions can be represented in tabular form (where each entity in the event corresponds to a field in the table, and each instance of an event corresponds to a row in the table). Essentially, text analysis converts text into tuples, and tuples are nothing but the rows of a database table.

So, every sentence uttered by humans can be construed either as communicating information (be it event or fact), or as communicating an intention. The set of intentions that Aiaioo Labs has used is given in Table 1.

3.1 Intention Analysis

Intention Analysis is the identification of intentions from text, be it the intention to purchase or the intention to sell or to complain, accuse or to inquire, in incoming customer messages or in call center transcripts.

The set of intentions that Aiaioo Labs has used is given in Table 1.

3.1.1 Use Cases

In July 2011, we used intention analysis to study the GooglePlus launch. We especially looked at quit intentions to see how frequently people were threatening to quit FB over time and saw how the number dropped sharply once people got to try GooglePlus (once the by-invite-only period ended).

This was a powerful observation, because in just four days, we could tell that GooglePlus could not replace Facebook, at least not yet. Details of the study can be found at <http://www.aiaioo.com/cami>

3.1.2 Background

The work that intention analysis is based on goes as far back as 1962 when J. L. Austin noted that not all utterances are statements whose truth and falsity are at stake, and that there was a class of utterances like “I pronounce you man and wife” that are actions [taken from Winograd, 1987].

In 1975, Searle identified the following broad categories of illocutionary (causing an action to happen) speech acts [from Winograd, 1987]:

- Assertive: Committing the speaker to the truth of a proposition
- Directive: Attempting to get the listener to do something
- Commissive: Committing the speaker to a course of action
- Declaration: Bringing about something (eg., pronouncing someone married)
- Expressive: Expressing a psychological state

Interestingly, the expressives include “expression of opinion” which corresponds to the modern day concept of sentiment analysis.

3.1.3 Structured Results

The result of intention analysis is, for each intention detected, the following set of labelled entities.

- Actor: The person holding the intention
- Actee: The other person(s) participating in the transaction desired
- Object: The object in the intention
- Alternative: Optionally the other object in the intention

Category of Event
Acquisition
Merger
Spin Off
Sale
Partnership Formation
Declaration of Bankruptcy
Renaming

Table 2: Partial List of Event Types.

3.2 Sentiment Analysis

Sentiment Analysis is the identification of sentiment from text, be it positive sentiment or negative sentiment.

3.2.1 Use Cases

Sentiment analysis has been around for a long time and is a very popular text analytics operation. However, it is only one of the categories of intention (the intention to opine) that it is possible to detect.

It is dangerous to perform sentiment analysis without first performing intention analysis. For example, the sentence “Is the Sony S4 a good camera?” can be easily misconstrued as a case of positive sentiment when in reality it is a case of intention to inquire (not an intention to opine at all).

3.2.2 Structured Results

The result of sentiment analysis is, for each sentiment detected, the following set of labelled entities.

- Holder: The person holding the sentiment
- Entity: The object about which the sentiment is felt
- Polarity: The polarity of the sentiment

3.3 Event Analysis

Event Analysis is the identification of events from text, be it an acquisition of one company by another, or the merger of two companies, or the hiring of a person by a company.

The set of events that Aiaioo Labs has used is given in Table 2.

3.3.1 Use Cases

Events are mostly useful for knowledge management of news articles. They help in identification of happenings around the world related to certain

Category of Fact
Part of
Feature of
CEO of
Born on Day
Died on Day

Table 3: Partial List of Fact Types.

entities. Events mostly deal with current events and changes.

3.4 Fact Analysis

Fact Analysis is the identification of facts (or relations) from text, be it who is the CEO of a company, or its earnings.

The set of facts that Aiaioo Labs has used is given in Table 3.

3.4.1 Use Cases

Facts are mostly useful for mining knowledge from encyclopedias or static web pages. They help in the identification of relations between certain entities.

3.4.2 Fine Distinctions

The distinction between events and facts is a fine one (sentences describing either are uttered when there is an intent to communicate information). The distinction is that facts describes the state of a system, whereas events describe changes of state. For example, “London is in England” is a fact, whereas “London Bridge is falling down” is an event. “London” and “England” are entities (the first example indicates that there is a relation between them). “London Bridge” is the key entity in the second example.

3.4.3 Background

A lot of the work in the semantic web space, including work on RDF extraction and named entity recognition is related to event extraction. The identification of events is often preceded by the identification of named entities or noun chunks and facts (relations between them). Events can be inferred from the presence of certain entities and the existence of certain relations between them and also the presence of certain trigger words.

3.4.4 Structured Results

The result of event analysis is, for each event detected, the following set of labelled entities.

Intention	Actor	Actee	Object
Purchase	John	Mike	the green car
Quit	John		Z Bank

Table 4: Intention Examples.

Polarity	Entity	Holder	Sentiment
Positive	Jane	John	pretty
Neutral	Mary	John	

Table 5: Sentiment Examples.

- Actor: The main player in the event (usually a person or organization)
- Experiencer: The other persons or organizations participating in the event
- Object: The object involved in the event
- Price: Optionally the monetary transaction involved in the event
- Location: Optionally the location of the event
- Time: Optionally the time of the event

4 Examples

The following examples illustrate how intention analysis (including sentiment analysis) and event analysis can convert unstructured text into structured text.

4.1 Intention Analysis Examples

Table 4 contains the result of intention analysis performed on the sentence “John wants to buy the purple car from Mike” and on the sentence “John said he would close his account with Z Bank”.

4.2 Sentiment Analysis Examples

Table 5 contains the result of sentiment analysis performed on the sentence “John thinks that Jane and not Mary is pretty.”

4.3 Event Analysis Examples

Table 6 contains the result of event analysis performed on the sentence “Z Bank acquired Y Bank on Thursday for 20 billion dollars” and on the sentence “Z Bank announced its merger with X Bank on Friday”.

Event	Actor	Experiencer	Price
Acquisition	Z Bank	Y Bank	\$20bn
Merger	Z Bank	X Bank	

Table 6: Event Examples.

Fact	Person	Organization	Date
CEO of	Z	Y	2011

Table 7: Fact Examples.

4.4 Fact Analysis Examples

Table 7 contains the result of fact analysis performed on the sentence “Z is the CEO of Y as of 2011”.

5 Image Processing

Images are a form of unstructured information. Image processing algorithms typically recognize an object in the image (to which a label is assigned, indicating the type of object recognized), and the location and size of the object (the x coordinate, the y coordinate, the width and the height).

5.1 Structured Results

So, image processing results in the following set of entities:

- Object: The type of object recognized in the image
- x: The left edge of the bounding box
- y: The top of the bounding box
- width: The width of the bounding box
- height: The height of the bounding box

5.2 Image Analysis Example

Table 8 contains the result of image processing performed on an image containing a shoe and a face.

Object	x	y	height	width
shoe	10	10	16	44
face	76	10	43	37

Table 8: Image Example.

6 Integration

In our experience, the best way to analyze text is to start with intention analysis. Intention analysis will identify the speech act if the utterance represents a speech act.

If the intention of the utterance is to opine, one additional step that needs to be performed is sentiment analysis to determine if the user is praising something or criticizing something. Both praise and criticism may be present in a single subjective sentence.

If there is no basic speech act identified by the system, then the utterance is possibly informative. So, we then perform event analysis to identify what event the utterance is about. The space of events is large, so typical event analysis systems are built for certain domains and for a closed set of events.

Fact analysis can be performed can also be performed (before or after or concurrently with event analysis) to see if any facts of interest may be gleaned from the utterance.

7 Caveat

The distinction between intentions, events and facts is not watertight. There are times when utterances can cross the boundaries and fall into more than one of these categories. It's just not something that happens very often.

8 Uses

It seems as if each of the types of text analysis would be best applied to a specific type of text source. These would be:

- Event Analysis: News articles, because news reports are always about important happenings or changes in the state of the world, and hence are rich with events and also with facts.
- Fact Analysis: Wikipedia, other Encyclopedias and Knowledge Bases are full of facts, but don't necessarily report current events. They may contain information on events that took place in another age.
- Intention Analysis: Emails, Feedback, Social Media Messages

9 Enterprise Tools

Similarly, each of the types of text analysis seem suitable for a class of enterprise tool. These are as follows:

- Event Analysis: Media Monitoring Tools, Opportunity Identification Tools, Conformance and Discovery Tools
- Fact Analysis: Enterprise Search, Semantic Web, Logic and Inference Engines
- Intention Analysis: CRM Tools, Collaboration Tools, Task Management Tools, Communication Devices

10 Conclusion

As we have seen above, the goal of assimilating unstructured information is best performed by the conversion of unstructured information into tabular form (structured form). We have shown above that it is possible to transform all forms of text into a structured form (when the text is informative, we use event analysis, and when the text is a speech act, we use intention analysis). Between intention analysis and event analysis, it seems we can deal with most human utterances of value, and therefore, we have a complete solution for dealing with unstructured information.

We also introduce intention analysis as a novel method, offered commercially as an API by Aiaioo Labs, and list the different types of intentions that can be identified. We suggest that sentiment analysis is subsumed by intention analysis, that expression of sentiment is nothing but the intention to opine (that is, nothing but a type of intention). We support this argument by pointing out that sentiment analysis is bound to fail if there is no distinction made between a question asked by a user about whether a product is good and an assertion made by a user that a product is good.

We then provide examples of conversion of different types of text into rows in tables through intention analysis and event analysis and list the dimensions (entity labels) that may be used as the fields (columns) of the tables.

Finally, we propose a method of converting images into structured information that can be handled by a typical BI product.

Information about Aiaioo Labs

Demonstrations of the Aiaioo Labs APIs for intention analysis, sentiment analysis and event analysis can be accessed from the Aiaioo Labs home page.

Aiaioo Labs (<http://www.aiaioo.com>) is a research SME working on Applied AI, and on text analytics in particular. We not only develop algorithms for text analytics that we publish as research, but we also develop and evaluate business use cases (we create innovations around how customers can use text analytics).