AutoMap Introduction

Version 1.0

Francis Nimick

May 27, 2011

1 Introduction: What is AutoMap?

AutoMap is a text mining tool developed by the CASOS (Computational Analysis of Social and Organizational Systems) Group at Carnegie Mellon University. It takes as input any number of unstructured, ordinary text documents, and generates from the raw text within an network map that can be easily imported into *ORA or similar network analysis & visualization software.

AutoMap employs Network Text Analysis methods to extract several types of data from unstructured documents. Data that can be extracted includes content analytic data (words and frequencies), semantic network data (the network of concepts), meta-network data (the cross classification of concepts into their ontological category such as people, places and things and the connections among these classified concepts), and sentiment data (attitudes, beliefs). AutoMap integrates this data seamlessly into *ORA, and the data can be easily exported, in standard DyNetML format, for use in other network analysis software.[1]

1.1 Methods of Text Analysis

AutoMap performs a specific type of Network Text Analysis known as Semantic Network Analysis. Semantic analysis involves the extraction and analysis of links between words to model the author's mental map as a network. AutoMap includes support for computer-assisted coding; the user specifies coding rules used by the software to model the inputted texts as networks of concepts. These coding rules involve text pre-processing and statement formation, which together form the coding scheme. [2]

2 Getting Started

2.1 Introductory Note

This primer was created using *ORA v2.2.9 and AutoMap v3.0.7. It may not be accurate with respect to older or newer versions of the software.

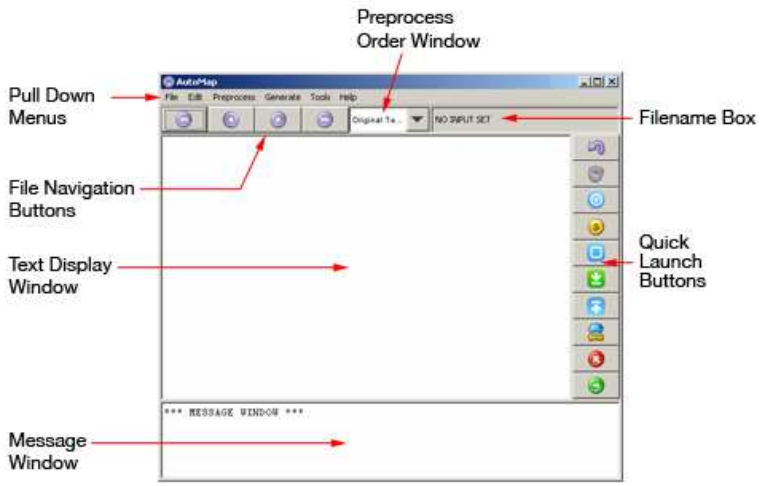
2.2 Downloading & Installing the Software

- *ORA can be downloaded from http://www.casos.cs.cmu.edu/projects/ora/software.html
- AutoMap can be downloaded from http://www.casos.cs.cmu.edu/projects/automap/software.php
- Download the appropriate files for your system & install (using defaults) as directed.

3 Welcome to AutoMap

Open AutoMap, which by default should be installed under Start→Programs→AutoMap. The main display window, as depicted below, will appear.

3.1 The Display Window



The main AutoMap Display Window

This main display window is the screen you encounter when you first launch AutoMap. The elements of this window are as follows:

3.1.1 The Menu Bar (Pull Down Menus)

File This menu contains functions related to loading and saving text files.

Edit This menu contains functions for setting the layout of the AutoMap Display Window.

Preprocess This menu contains functions related to the preprocessing of text files. This preprocessing is done before the generation of any output. These functions alter original text files only.

Generate This menu contains functions for the generation of output from previously preprocessed files.

Tools This menu contains a number of built-in viewers and editors for the support files used in preprocessing.

Help This menu contains a link to the AutoMap help file as well as AutoMap version information.

3.1.2 The File Navigation Buttons

These buttons are used to display the files in the main window.

3.1.3 Preprocess Order Window

This window contains a list of the preprocesses performed on the files. These can be undone, one process at a time, with the Undo command (Preprocess→Undo Last Step).

3.1.4 Filename Box

This area displays the file name of the currently selected text file.

3.1.5 Text Display Window

This displays the text contained in the currently selected text file.

3.1.6 Message Window

This area contains information about the actions AutoMap is currently performing, as well as information about any errors that have occurred.

3.2 A Brief Tutorial

The details of the user interface will be covered in a later section. We will begin, first, with a brief tutorial introduction:

3.2.1 Retrieving the Example

On the AutoMap sample data site, http://www.casos.cs.cmu.edu/projects/automap/sample.php, download the input texts and pre-processing material for the "Power Outage" sample data. Save and extract these files to separate folders in a known location.

Note The examples on the AutoMap website are currently broken. Please download, in addition to the examples, the corrected files from the directory in which this PDF was located.

3.2.2 Loading Sample Texts

- In the main AutoMap display window, choose File→Import Text Files.
- Navigate to the location of the saved input data and choose Select.
- Verify that the option for text encoding is set to "Let AutoMap Detect", and the text direction is set to "Left→Right, Top→Bottom." Choose Confirm.
- The texts are now loaded into AutoMap. Verify that all three files are loaded with the file navigation buttons.

3.2.3 Initial Pre-Processing (Delete List)

- Select Preprocess→Text Refinement→Apply Delete List.
- In the window that appears, choose Customized Delete List. Select the file from our downloaded example. Confirm.
- Choose Rhetorical delete processing and choose OK.
- Observe the changes to the text files. Words have been replaced with "xxx" separators.

3.2.4 Stemming

- Select Preprocess→Text Refinement→Apply Stemming.
- In the window that appears, choose KStemmer.
- When asked whether to stem capitalization, choose "No" and continue.
- Observe the changes to the text. The endings of many uncapitalized words (e.g. plurals, past tense) have been deleted, and the words have been reverted to their simplest forms.

3.2.5 The Generalization Thesaurus

- Select Preprocess→Text Refinement→Apply Generalization Thesauri
- Select the generalization thesaurus downloaded for pre-processing and confirm.
- Choose to sort the thesaurus.
- Choose to *not* use thesaurus content only.

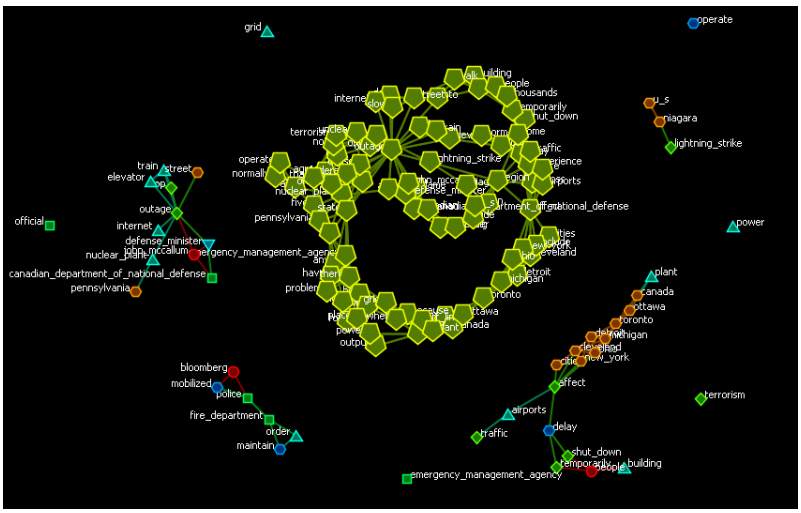
- Observe the effects. Multiple-word items, such as “North America,” have been shortened to single items such as “North_America” for analysis.

3.2.6 Meta-Network Generation

- Meta-Networks generated by AutoMap contain various node classes. These classes include Concept, agent, event, knowledge, location, organization, resource, role, and task.
- To generate meta-networks for the sample data, select Generate→MetaNetwork→MetaNetwork DyNetML (Per Text).
- Set window size to 5, and select the downloaded meta-network thesaurus. Confirm.
- Choose an appropriate output directory for the MetaNetwork data and select. AutoMap will generate the MetaNetwork data in the folder specified.

3.2.7 Viewing the Data

- Open *ORA.
- Select File→Open Meta-Network. Navigate to the AutoMap output directory, open the MetaNetwork1 folder, and select all of the .xml files within. Open the files.
- You can now visualize and analyze the networks produced by AutoMap.



A typical AutoMap generated network, produced in this case from the USA sample text

4 Working with User-Provided Text

For working with arbitrary, user-provided input text, AutoMap provides a number of features:

4.1 Preparing Text for Pre-Processing

4.1.1 Importing Text

Text can be imported into AutoMap in multiple ways. Users can:

- Import a folder of text files, as described in Section 3.2.2 on page 4.
- Create & save text files into which content is manually entered (File→Create New Text File).
- Scrape web-sites for raw text (Extractors→Extract Web Pages to Text Files).

4.1.2 Cleaning Text

Text that has been imported into AutoMap can be cleaned & prepared for pre-processing in a variety of ways.

- To clean text, AutoMap includes a variety of tools, located under Preprocess→Text Cleaning. For most text, all cleaners can be run with little or no repercussions.
- To prepare text for pre-processing, AutoMap includes tools, located under Preprocess→Text Preparation. For most text, these preparation tools can all be run, but care must be taken with more complex inputs.

4.2 Text Pre-Processing

AutoMap mainly uses delete lists and generalization thesauri to pre-process texts for analysis. Multiple delete lists and thesauri can be applied in any order, though it usually advisable to apply thesauri first to avoid potential loss of meaning through deletion.

4.2.1 Delete Lists

Delete lists allow the user to remove non-content words from text, such conjunctions, articles, and other noise words. The format of the list is a plain text file, containing a list of words to be deleted, one on each line. AutoMap includes delete lists for common words, but you may wish to create your own to match a particular input file.

Creating a Delete List A delete list can be created manually, using a text editor, though it is often faster to create one within AutoMap. After opening the text file(s) in AutoMap, create and save a Concept List (Generate→Concept List). When the list is opened in the Concept List viewer, the user can check any number of terms and save them to be deleted (File→Save As Delete List).

Applying a Delete List To apply a delete list from within AutoMap, choose Preprocess→Text Refinement→Apply Delete List. Either a built-in or custom list can be selected. Delete processing can be either Rhetorical or Direct:

Rhetorical For each word deleted, inserts a meaningless filler word, typically "xxx." This enables the original spacing of the words in the text to be preserved.

Direct Simply removes the words to be deleted. This does not preserve the original spacing of words in the text.

In most cases, the default option (Rhetorical processing) should be sufficient.

4.2.2 Generalization Thesauri

AutoMap includes support for generalization thesauri, which link multiple disparate representations of the same concept. While AutoMap includes generalization thesauri that cover common cases, it is likely that analysis will require the creation and application of a custom generalization thesaurus for the given input text.

Creating a Generalization Thesaurus A generalization thesaurus is stored in CSV format, which can be created with any spreadsheet editor. The first column contains the concept to be replaced, and the second column contains the key_concept with which to replace it. The concept to be replaced can be any number of words, while the key_concept is usually a single word or a set of words separated by underscores.

Important note The order of entries in a generalization thesaurus is important. AutoMap will replace a phrase according to its first match in the generalization thesaurus. For example, if you had the following generalization thesaurus:

United States	united_states
United States of America	united_states

The sentence "United States of America" would be processed to "united_states of America," as the first two words would be replaced according to the first thesaurus entry. Therefore, the longest and most specific entries in a generalization thesaurus should always be listed first. To verify a generalization thesaurus and ensure that it is properly ordered, use the Thesaurus Sort tool, located at Procedures→Thesaurus Procedures→Sort Thesaurus.

Applying a Generalization Thesaurus To apply a generalization thesaurus to input text, choose Preprocess→Text Refinement→Apply Generalization Thesauri. Either a default or user-created thesaurus can be used. "Use Thesaurus Content Only," if selected, will preserve *only* the content that was replaced by the thesaurus; all other text will be removed. This option is not recommended except in cases of extremely thorough thesauri.

4.2.3 Stemming

Stemming is the process by which words in the input text are *normalized*, i.e. derivations of words are detected and are replaced with the normal form. AutoMap includes two well-known stemming algorithms, the Porter and KSTEM algorithms. Here is an example output:

Original Hospitals switched to using emergency generators.

Porter hospit switch to us emerg gener.

KSTEM Hospitals switch to USE emergency generator.

Stemming should be done with care, as it can sometimes distort the meaning of phrases in the input text.

4.2.4 Other Pre-Processing Tools

4.3 Text Processing

Once the text has been cleaned and pre-processed, it can be analyzed. The simplest type of analysis to perform is a semantic analysis.

4.3.1 Semantic Network Generation

To generate a semantic network from a given input text, select Generate→Semantic Network. Per Text will generate one network per text file entered, while Union will generate a single network encompassing the concepts in all currently active files. The user will then be presented with a number of options:

Directionality This determines the directionality of the created network. "Bidirectional" creates a network with directed links, and "Unidirectional" creates a network with symmetric links.

Window Size This sets the maximum distance between concepts to be connected. Higher numbers result in the generation of a more connected, less specific network. Values from 2-5 are appropriate for typical analysis.

Stop Unit This sets the window delimiter, and should be set according to the nature of the input text.

Stop Unit Value This maximum number of stop units that can be included with the Window Size parameter. ([2], p. 74)

Upon completion, the generated semantic network(s) will be saved in the specified folder, where they can be opened with *ORA. This process is described in Section 3.2.7 on page 5.

4.3.2 Meta-Network Analysis

As seen in the example, AutoMap can also perform Meta-Network Analysis, in which varying textual elements are identified and linked as network nodes of varying types. This requires the creation of a Meta-Network Thesaurus.

The Meta-Network Thesaurus A Meta-Network Thesaurus consists of a thesaurus, similar to a generalized thesaurus, but linking concepts not to other concepts, but to node categories. These Meta-Network categories include Agent, Knowledge, Resource, Task/Event, Organization, Action, Role, Attribute, and any other user-defined categories. An initial Meta-Network thesaurus can be generated from *preprocessed text* by selecting Generate→MetaNetwork→Suggested MetaNetwork Thesauri. The generated thesauri should then be modified to rectify any errors and better classify the elements of the Meta-Network.

Running a Meta-Network Analysis Select Generate→MetaNetwork→MetaNetwork DyNetML. As before, Per Text will generate a separate network for each text file being processed, while Union will aggregate and link the concepts in all files. A previously created Meta-Network Thesaurus must be selected, and the other options are the same as they were for the Semantic Network, as described in Section 4.3.1 on the preceding page.

Note The other capabilities of Meta-Network analysis, such as the use of Submatrix Selection and Speech Tagging, will be described in a later revision of this guide.

After the Meta-Network(s) have been generated, they can be imported into *ORA as described in Section 3.2.7 on page 5. They can then be visualized and analyzed.

4.4 A Final Note

For a more detailed tutorial on the creation of concept lists, deletion lists, generalization thesauri, and semantic networks, please see pages 11-21 of the 2010 AutoMap User's Manual [3].

References

[1] AutoMap Project, CASOS. http://www.casos.cs.cmu.edu/projects/automap/

[2] Carley, K. M., Columbus, D., Bigrigg, M., Diesner, J., & Kunkel, F. (2009) *AutoMap User's Guide 2009* (CMU-ISR-09-114). Carnegie Mellon University. http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-09-114.pdf

[3] Carley, K. M., Columbus, D., Bigrigg, M., Diesner, J., & Kunkel, F. (2010) *AutoMap User's Guide 2010* (CMU-ISR-10). Carnegie Mellon University. http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-10-121.pdf