Leximancer White Paper

Leximancer is text mining software that can be used to analyse the content of collections of textual documents and to visually display the extracted information in a browser. The information is displayed by means of a conceptual map that provides an overview of the material, representing the main concepts contained within the text and how they are related.

The Meaning of Meaning

Words and sentences don't always mean what we think they mean, and their meanings can change depending on the time and the situation. Moreover, it is easy to see that the accumulated information from all the nouns, verbs, modifiers, prepositions, modal auxiliaries, clauses, multi-sentence discourse relationships etc. within a conversation or document of even moderate length forms an extremely complex body of meaning.

Because natural language is vital to people in almost any situation, there exist many diverse models and terminologies for describing the meaning of natural language. In the field of Information Technology, a very formal and concrete approach is often taken to models of meaning. This can work adequately for structured relational databases in restricted domains over limited time frames, but does not work well for rich, dynamic, and potentially unreliable natural language in real-life settings. The Cognitive Sciences and the Humanities are much more familiar with the necessity for Ontological Relativity and Dynamics.

Simply put, Ontological Relativity says that a useful mental model of the meaning contained in a piece of information depends to some extent upon the background knowledge, the current context, and the goals of the person. It must be understood that this is not simply saying that different disciplines talk about different things – different disciplines frequently talk about the same things in very different ways. This is not to say that all meaning is infinitely relative – some aspects of meaning can usually be agreed upon by groups of people who have similar backgrounds and goals. Most aeronautical engineers agree on the basic requirements for a plane not to crash. Ontological Dynamics says that a useful mental model of meaning may change over time, because either the external environment is changing in fundamental ways, or the knowledge, context or goal of the person is changing. Leximancer combines the discovery of quantified relational information between concepts with the flexibility and dynamics required to analyse natural language in real-life settings.

Barriers to Meaning

Imagine this situation:

You've just landed a new job fresh out of college. You really need this job, but maybe the course hasn't prepared you enough to get started. Opening the big glass door for the first time, the workplace sounds like a zoo. People are talking across each other, and all talking at once it seems. You desperately filter out most of the noise and just try to get a handle on one

conversation. A bit of it seems to make sense – maybe it sounds like something your last lecturer said. You listen for a while, trying to work out the meaning. But then one of the talkers looks round and wants an answer to a question you barely understand. Do you answer assuming you know what they mean, or do you dumbly shake your head? This could be an example of inappropriate prior knowledge leading to an incorrect interpretation of meaning.

You survive the first week by staying close to the same group of people, and start getting comfortable working with them. Then the CEO asks you for a report by the end of the day on the current morale of the whole organisation. Do you wildly generalise from the five people you know, or do you dumbly shake your head? You may be about to make dangerously misleading conclusions by being forced to rely on a small sample of atypical information.

You stay with the job for three months and decide to be proactive. The marketing manager seems friendly, so you ask her to draw an organisational map to get a handle on the big picture. She draws a map on a big whiteboard with the marketing division as the central node, connected directly to the clients. At the engineering group meeting the next week, you confidently put up this map, but the quality assurance manager dismissively erases it and places QA as the most important link to the clients. The meaning of a situation depends to some extent on the context of the observer and their intentions.

After avoiding the QA section for a while, you establish a good working relationship with the chief scientist. She has been here for years and knows the ropes. You have to test the new production line in a big hurry and she tells you that the only real changes are in the robot welding line. The robots test out fine and time is running out. Do you declare the tests a success? Confident expectations or appeals to authority can lead to flawed appraisal of the actual situation – this is a form of Confirmation Bias.

You work for the chief scientist for several months analysing and summarising factory data sent to you by the production manager's assistant. The production manager position is vacant, and this would be a great promotion. On the day of the presentations to the selection panel, the production manager's assistant gets up before you and presents quite different results from your own. The credibility of information sources can vary widely, and some sources may deliberately lie.

Leximancer is designed to assist people cope with these barriers to meaning by providing a tool that enables the text to tell you what it means, without the biases and other factors that distort meaning.

Operational Capability

Put simply, Leximancer text analysis simultaneously offers maximum automation and maximum control. Our technology improves on previously available text mining and information extraction systems in five main areas: automation, clarity, deep meaning, completeness, and efficiency -

- **Automation:** Given the increasing amount of text requiring examination, and the decreasing amount of time available in most work situations, it is very important that a

text analyst support system does all of the drudge work. Users usually don't have time to hand prepare or code larger amounts of text, and even trying to guess the correct words or topics to query from a retrieval system is time-consuming and error prone. Other systems that are positioned in the text analysis market offer diverse ways of dealing with this task. Some rely on manual insertion and coding of text, while others rely on time-consuming supervised training of each concept, from a set of hand selected exemplar documents. Other systems offer full automation but suffer from a lack of clarity in the analysis - they provide lexical statistics but do not offer a clear conceptual overview. Still other approaches are heavily tailored to specific languages, domains, or styles of document. Leximancer can take the effort and guess work out of this task. Leximancer can process many different formats and styles of text from many languages, and can automatically select and learn a set of concepts that characterise the text. The algorithms used to automatically select important terms, called concept seeds, and the ability to learn a thesaurus of words for each seeded concept are both significant improvements on existing technology. A novel clustering system that allows easy visualisation of the concepts and themes also represents a breakthrough in network clustering. Leximancer algorithms are not strongly language dependent.

- **Clarity:** Leximancer is designed to provide a clear and transparent conceptual analysis of text. Other systems in this market claim to learn concepts, but it is difficult for a user to clearly perceive the meaning of these 'black box' concepts. Leximancer maintains a symbolic representation of each concept in the thesaurus, similar to the headwords in a traditional thesaurus. The way in which our system can distil the information contained in text composed of many tens of thousands of vocabulary words into a small and comprehensible set of concept dimensions is novel in the field. Leximancer is also designed to be transparent: a user can inspect a ranked list of words which make up each concept, and can easily drill down to the text to inspect the validity and nature of the induced abstract relationships. The Leximancer user interface uses qualitative and quantitative visualisation techniques taken from Information Science to assist in the rapid comprehension of complex information spaces. The clustering algorithm uses a complex-systems approach to display emergent themes among the conceptual relationships.

- **Control:** The *Compound Concepts* facility allows the user to create their desired custom classifiers by simply constructing Boolean combinations of discovered or selected concepts. These logical combinations of concepts offer a unique combination of machine adaptation and detailed control.

- **Deep Meaning:** Leximancer generates similarly structured conceptual maps from texts which have similar meanings, even when the documents use different styles, formats, or even languages. This gives the user confidence that the concept maps have real meaning, and are not chance artefacts of the document formats. This also shows that it is not easy to conceal the patterns of meaning from Leximancer by using veiled speech, dialect, or non-standard grammar.

- **Completeness:** Some users of Leximancer have commented that this product combines all the most useful features into one offering. We offer a one-stop automatic system to take the user from a large collection of raw text to a concept map, while

allowing inspection and validation of all the intermediate steps. Development and validation has involved the analysis of many different types of text and the refinement of many analytic strategies.

- **Adaptability to Situation:** Leximancer offers the capability for adapting entity or thematic classifiers to changing circumstances as reflected in text data. For example, if a theme such as favorable sentiment is to be watched over time, Leximancer can adapt the classification pattern to each successive time period from a small stable nucleus, or seed word set - the *Sentiment Lens* function will first remove any seed terms which are used out of the favorable context, and then the thesaurus learner will discover other favourable terms peculiar to the data. This gives the discovery power of a concept query while automatically tracking terminology changes. When applied to an entity, such as an organisation or individual, this process can discover and track the set of active participants and aliases. Leximancer also provides a natural extension to this technique, whereby entities and concepts can be discovered which are directly and indirectly associated with prior concepts of interest. This allows automatic extraction of the social network surrounding an entity or entities of interest, or automatic generation of a concept map surrounding any theme of interest.

- **Data Exploration:** Leximancer can independently discover the set of concepts and entities which characterise a set of text – this is important for discovering what is in the data that we were not aware of.

- **Robustness:** Leximancer is largely language independent, and is much more robust than grammatical or rules-based Information Extraction systems at handling poor quality text, such as highly informal spoken language, dialect, veiled speech, inaccurate automatic transcription, or degraded OCR. An additional advantage is that Leximancer will discover new concepts, relationships, and themes in the data, without the operators having to become aware of the change and construct new extraction rules.

- **Efficiency:** The Leximancer system was designed from the outset to be simple and fast. A normal modern PC or notebook computer is quite adequate for analysing 1 GB of text or more. This contrasts to other text mining solutions that require substantial server infrastructure or high performance computation. The Leximancer algorithms are scalable and so much larger amounts of text could be processed using more memory and multi-threaded processing.

Leximancer's Technical Approach

Leximancer draws on several disciplines:

- **Corpus and Computational Linguistics:** simply speaking, a word can be characterised by the words that tend to appear near it, and not apart from it. It is known that the appearance of a word is correlated with the appearance of certain other words, and this correlation has a range in terms of separation before or after the target word in the stream of text. This is a form of short-range order and we exploit this to find

a measure for how relevant one word is to a set of other words. The relevancy measure, or metric, is actually taken from Bayesian theory and the field of Information Retrieval. The *Sentiment Lens* function, discussed below, is a powerful and novel application of the principles of Corpus Linguistics.

- **Machine Learning:** Leximancer uses a machine learning optimisation approach for iteratively growing a thesaurus of words around a set of initial seed words. The actual algorithm is a heavily modified version of one found in the field of Computational Linguistics, and was used originally for word sense disambiguation.

- **Complex Networks Theory:** The cluster map presentation is heavily influenced by the area of Complex Networks, and we are looking for emergent behaviour among the simpler concepts to reveal more abstract concepts. We call these emergent themes and our validation efforts are demonstrating that these emergent themes provide a measure of the meaning of the text.

- **Physics:** The idea of a measurable short-range order between words was influenced by solid-state physics. More directly, the actual algorithm used for clustering the concepts is derived from physical force laws and numerical methods used in many-body problem simulations.

- **Content Analysis:** Overall, the strategy used by Leximancer follows the model set down by Content Analysis. This can be briefly described as the quantification of knowledge within text by means of coding or tagging of text segments using a set of concepts, each of which is defined by a set of relevant words.

- **Information Science:** The principles and guidelines for indexing and navigating large amounts of information in a complex concept space came from Information Science.

The Sentiment Lens

The Sentiment Lens function is a powerful application of the principles of Corpus Linguistics. Say we have specified a list of seed words for a concept, such as a sentiment (or *affect*, as they are called technically), prior to inspecting a new text data set. Let us make the reasonable assumption that the majority of these seed words are in fact used in the way we expect within the data. However, it is a common occurrence that some of the terms will be used in ways that are unrelated to the sentiment we are looking for.

A simple example might be a set of customer comments about the performance of Help Desk staff. Since these customers have all contacted the Help Desk, most of them had some sort of *problem*. Their comments will often mention the problem or issue which led them to making the call, but the words *problem* or *issue* do not by themselves indicate that they are unhappy with the way their problem was dealt with. Further, the word *problem* is likely to be frequent in the data. This will result in a substantial **systematic error** in the identified sentiment if such terms remain in the sentiment definition. These errors are not random errors which average away.

The Leximancer Sentiment Lens takes each seed word for a given sentiment and measures its usage within the data in question, by finding the other words which tend to associate with it. This allows the system to identify how much each seed word's usage has in common with the other seed words. Assuming that the majority of seeds form a consensus around the desired sentiment, the Sentiment Lens function can identify and remove the outlier seed words, such as the word *problem* in the above example. In that example, the word *problem* would typically be used in quite a different context from the genuinely negative seed words.

Once the Sentiment Lens has removed the outlier seed words from a concept's starting definition, the thesaurus learner can confidently take the set of core words and discover other relevant but unexpected sentiment words. Note that the Sentiment Lens function can be applied to *any* user seeded concept to automatically refine and focus the seed list. This automatic quality control over user-selected concept terms can be a huge time saver.

Conclusion

Previous statistical text analysis techniques such as Latent Semantic Analysis (LSA), Hyperspace Analog to Language (HAL), and Latent Dirichlet Allocation (LDA) have demonstrated that highly useful and reliable information can be extracted from the word co-occurrence information in text. Leximancer extends and reworks this approach with two stages of non-linear machine learning to provide a statistical means of extracting clear semantic patterns from text.

Scenarios for using Leximancer

Survey Analysis

Survey data which includes text responses can be easily and effectively analysed with Leximancer. Categorical and demographic columns are extracted as variables automatically. The topics which are important to the respondents, along with the vocabulary which they actually use, are identified automatically. A single button click will add in favorable and unfavorable dimensions of sentiment to the analysis. Leximancer's Sentiment Lens will remove the sentiment seed words which are used out of context in this data, and the Thesaurus Learner will then identify other relevant sentiment terms which are particular to this text.

The Insight Dashboard report is a useful way of encapsulating the quantitative correlations between your selected dependent segments and the observed text concepts, along with the supporting textual evidence for better understanding. The Dashboard report is a stand-alone hyper-linked document, available in PDF or HTML format.

Conversation Analysis

Conversation and interview transcripts can be analysed easily. As long as speaker labels are formatted correctly, Leximancer will automatically extract the speaker identifiers as variables, and associate these labels with their utterances. This allows content from selected speakers to

be filtered in or out, and allows comparative analysis between speakers, normally using the discovered concepts as independent variables.

Conversation analysis can be extended to incorporate:

- Blogging/forums etc. on the internet – brand/product analysis etc.
- Email dialogues – litigation eDiscovery
- Voice-to-text translations - call centre dialogues, meetings, scenario training/simulations

Profiling

The Leximancer system can discover concepts and entities which are directly or indirectly related to one or more target concepts (which you specify with a few seed terms). This can be configured to discover a social network of related named entities, or a concept map profiling the related issues. Additional settings allow concepts to be discovered from the intersection of multiple target concepts. This allows you to profile complex themes and scenarios. More importantly, it allows you to discover unstated links between targets. For example, we specified two target concepts – Raynaud's Phenomenon (a medical condition) and Dietary Supplement, and asked Leximancer to profile the intersection between these two target concepts from a collection of Medline abstracts. Even though no abstract in the collection mentions Omega 3 fatty acids (fish oil) alongside Raynaud's Phenomenon, Leximancer prominently ranked the concepts oil and omega in the intersection since Omega 3 fatty acids treat many of the symptoms present in Raynaud's Phenomenon.