Microformats for Innovative Lexicons

Sunny Fugate, Jeffrey Ellen, Dr. Stuart Rubin
SPAWAR Systems Center, Pacific - San Diego, CA
sunny.fugate@navy.mil, jeffrey.ellen@navy.mil, stuart.rubin@navy.mil

Abstract

A microformat is a set of design principles for including semantic information within standard X/HTML markup. Individual microformat entities are distributed yet share a common semantics. Each microformat is a granule of structured information containing a set of attributes. These information granules can be produced, distributed, aggregated, and consumed without reliance on centralized services. We describe the impact on the gathering, distribution, and analysis of concepts within our visual communication research.

In this paper we present solutions for managing lexical consistency when microformat-structured information granules are distributed and maintained independently and asynchronously. Lexical groups and hierarchies leverage the resulting inconsistencies, utilizing term aggregation across visual and linguistic features to dynamically compose lexicons and to perform lexical analysis. We think of this as creating semantic and orthographic ‘projections’ of the lexicon into different feature spaces. We show how to use this approach to construct situated lexicons which derive from shared context and social communities.

1. Introduction

The purpose of our visual communication research is to find new ways of facilitating communication using representational terms which are comprised of various metaphoric, iconic, and indexical forms (e.g. pictures, ideograms, pictograms, emoticons, etc) [7]. As a literal interpretation of a "picture being worth a thousand words", the intent is to accelerate communication speed while decreasing ambiguity through the use of visual elements in addition to plain text. Visual communication strategies often rely strongly on environmental, cultural, and cognitive contexts. Lacking the structure of a natural language or a sufficiently expressive grammar, visual communication strategies rely on usage context to enforce a desired semantics.

Fundamental to our approach was the elimination of ambiguity by enabling a communicator to select specific meanings for ambiguous terms during the inscription process. This resulted in a dramatic decrease in the number of ambiguous terms at the cost of inscription speed. Because each visual term was defined precisely, the recipient could simply select an ambiguous term to query its intended meaning. In essence, we utilized a shared dictionary as fixed context in order to enforce a non-ambiguous semantics.

Figure 1: A noun? A metaphor for the adjective "fast" or the verb "accelerate"? We allow the author to specify during inscription.

One problem with this approach is the potential for a large number of different uses of a term, each requiring a dictionary entry. A conventional lexicographer suffers from this same challenge and resolves it by choosing to define the most common semantics, ignoring uncommon usage and contexts. If terms used for visual communication were as well tracked we may have taken the same approach. Instead, we intend to determine the most common semantics of a visual term through analysis of term usage, context, and term definition overlap.

Figure 2: Rotary-wing and fixed-wing aircraft icons: collocation provides context

Rather than do this via conventional means, we utilize the XHTML-based Microformat concept to facilitate lexicon creation and analysis. Through decentralization, we hope to quickly create a global lexicon with a broad coverage of features and usage contexts. This global lexicon can then be used to build both conventional and unconventional dictionaries. Zesch et al provide a compelling analysis of existing
linguistic and collaborative knowledge bases to justify the loss in quality [17]. We believe the inconsistencies resulting from distributed creation and editing provide important differences in features that can assist in determining the right level of granularity across visual and linguistic features. As a result, our current and proposed approach relate directly to granular computing techniques in respect to creating aggregations at different scales based on shared attributes between individually defined terms.

2. Visual Communication Lexicon

Our research focuses on human cognition and communication efficiency. Microformat markup for each element in our visual lexicon allows us to achieve both usability and transparency for our team while maintaining machine-readability. Using a wiki-hosted microformat, each researcher was able to perform dictionary editing and composition, items such as figure 3, without fear of corrupting dictionary elements.

![Figure 3: Hazmat placards and road trailer identification - extracted from the 2008 Emergency Response Guidebook.](image)

Use of XML or other centralized formats for editing and maintenance of the dictionary is not practical. This is exactly the type of challenge that microformats are intended to address. Whereas many data formats are meant to be primarily read and edited by machine and act as a centralized store of information, microformats are intended to be primarily read and edited by humans, while still allowing machine processing. Later conversion of them to another format such as (Resource Definition Format) RDF is straightforward via an XSLT [11], so it is not ruled out. However, we did not feel that direct use of RDF along with an ontological markup language such as OWL was an appropriate format for our task. RDF and OWL are both intended to provide ontology support for information with a focus on interoperability between machines first, and human readability, second. Application of Microformats to new knowledge areas is a significant area of research due to the ability of combined human and machines management. There are other investigations which discuss conversation between various formats [1, 2].

In our language research a distributed dictionary is desirable for a number of reasons. Decentralization allows dictionary entries or entire catalogs to exist anywhere and to be created by anyone without reliance on or the restrictions of a central authority [4]. Nonetheless, this approach doesn’t preclude large, centralized dictionaries. In fact, distribution allows the dictionary to grow quickly as terms and term usages are added while search engine indexers (spiders) can compile a global dictionary which reflects the content of the distributed lexicon. Distribution allows dictionaries which may contain very similar elements to be framed by their domain contexts, but provides sufficient structure to allow these to be combined for use within cross-domain and global dictionaries.

The primary source of confusion dealing with homographs (words that share the same spelling or visual form but have different meaning) is ambiguity. To address this problem we used a universally unique identifier to reference each particular symbol use (rather than just referencing each particular symbol). In this way the specific definition for the representation can be referenced unambiguously.

When a dictionary is first constructed and published each element is given a globally unique ID. In our implementation, future edits create an entirely new entry containing the original content, subsequent modifications, and a new unique ID. This forces even subtle differences between dictionary entries to be retained over time. In many cases the changes were irrelevant to any linguistic insights, but in others the modifications would expose generalizations of the concept, its usage, or its properties. The use of a straight-forward aggregation scheme across feature subsets allows many different analyses to be performed. Many of these analyses consist of the construction of a dendogram across a subset of features ranked by information content or by some insight or query of the researcher. Each tree shows how the elements of different dictionary collections are collocated and demonstrates relationships and generalizations in an intuitive manner.

In addition to analysis of lexical, linguistic, and visual features, the global lexicon can be compacted in order to derive a dictionary with a concise, conventional set of representative terms. The process of refining and extracting a concise lexicon can utilize the compactification procedure described in other investigations[14], utilizing both textual and numeric features in which specialized distance metrics and fuzzy constraints may need to be explicitly defined[16].
2.1 Granules in Visual Languages

The use of glyphs, signs, and otherwise icons is synonymous with granulization. For example, a Stop sign is an information granule, which carries both explicit meaning — informing one as to the point at which to bring one’s vehicle to a complete stop, as well as implicit meaning — informing one to remove one’s foot from the accelerator and to depress the brake pedal. We see that such granules serve to summarize a more verbose representation using natural language. This principle is termed *randomization* out of respect for the definition of random numbers; namely, a sequence of numbers, which effectively has no more compact generator than the sequence itself (i.e., a mathematical fixed point).

What is needed is a way to randomize terms in a visual language so that arbitrary statements may be properly interpreted. While pictograms are useful such usage is necessarily limited. While arrows are readily interpreted as directional indicators, and maybe a man standing on his head could be interpreted to mean the inverse of a matrix (i.e., given a mathematical context), how does one say differentiate a positive definite matrix from a diagonally dominant one as a matter of education? The answer is that you need natural language as augmented by mathematical formalisms.

The extraction of meaning is properly a machine learning task involving complex pattern matching, which is beyond the scope of this paper. Nevertheless, to summarize, information granules are not necessarily static. They can be dynamic (i.e., changing in definition over time and in context) and fuzzy (i.e., as in shades of meaning). The development of a system, which uses language users and usage to capture such granular semantics, serves to underscore the importance of developing a science of granular computing.

2.2 Microformat Specification

Two initial microformats were created. The first is a markup which contains the URI of an image or Unicode glyph representation, a definition, as well as linguistic and visual attributes. A representative example provides semantic markup for a 2-color pictogram and refers via reference to an antonym, simple linguistic references, several visual features, and URI references to data analytical transforms.

![Figure 4: Definition microformat example.](image)

Microformats are meant to be simple, human readable and human editable. In our implementation we chose to represent properties that we were immediately useful to us. The specification can be easily expanded as we determine new requirements and refine our specification. The potential uses of a stand-alone linguistic microformat are notably interesting, but beyond the scope of our original implementation.

The second microformat is intended to be embedded during usage. This microformat is used to specify a particular representation when a language element (a word, or pictogram) was used on a webpage. This microformat solely contains a reference to an artifact’s universally unique identifier (UUID) [10].

![Figure 5: Embedded microformat example.](image)

2.3 Wiki

To make our microformats easily editable and available to linguistic researchers we chose to use a web wiki as a primary editing format. This allows microformats and their associated images to be placed directly on a wiki page. The wiki software provides version control and editing management as well as a simple and accessible user interface.
We have one wiki page for each uniquely defined term. Each wiki page contains one or more microformat lexicon entries, each of which directly specifies the URI of the image or the Unicode string that is being defined. For the purposes of our research an automated Perl script runs on the wiki periodically to create an updated XML dictionary containing all of the updated dictionary entries and image URIs. The Perl script uses an extension to the Text-Microformat CPAN module to extract microformat data [8].

2.4 Universally Unique Identifier

As is true with most data structures, elements must be able to be referenced uniquely. Instead of a centrally managed system (such as ranges of 32-bit integers like Unicode), we use UUIDs which can be generated asynchronously with good guarantees that no collisions will occur. By using a distributed data source for maintaining dictionary entries we will have to entertain the possibility of two microformat containers being referenced by the same UUID but containing different information. There are several published ways of solving this problem, and are outside the scope of this paper[13].

3. Lexicon Analysis

3.1 Resolving Ambiguities

One of the side effects of the distributed lexicon generation process is the potential for ambiguity and repetition. It is important to have a disambiguation process to discern between separate concepts vice separate entries which are intended to refer to the same concept. Examined individually, each entry is composed of a set $S$ of attribute-value pairs, $(a_i, v_i) \in S$. Entries which share all attribute-value pairs, with the exception of a unique ID, are understood to be identical with the possible exception of usage context. When terms are potentially identical, several possible hypotheses can be made concerning usage context: a) the usage contexts are actually the same and the terms are truly identical, b) the existing attribute-value pairs are insufficient to differentiate usage contexts, and c) the existing attribute-value pairs are sufficient, but inconsistent due to error or omission.

In the first case we assume that the terms are identical and ignore characteristics of the usage contexts. We can combine the terms by annotating the equivalence class as an attribute within both terms.

In the second case we can assume that there exists some hypothetical attribute-value pair $(h, v)$, which differentiates the terms. We can then assume that $(h, v)$ is a member of the set, $E$, of all known $(h, v)$, which extend $S$. If such pairs exist in the global lexicon, $|E| = 0$, then we can examine these pairs for suitability in extending $S$. An automated mechanism could rank the pairs in $E$ based on whether the attribute-value pairs can be derived from usage context (e.g. part-of-speech, term collocation, N-gram occurrences, namespace). If no such pair exists, $|E| = 0$, and then we need to innovate a set of new pairs, $F$, by examining usage context using the common arsenal of linguistic tools. This process is labor intensive, but some automation could be achieved depending upon the nature of the usage context (e.g. whether it is part of a tagged corpora, is structured or semi-structured text, or has other lexicon elements which can be used to determine collocations).

The third case is the most interesting when considering community-constructed lexicons and has applications to other domains of analysis outside our research. If all terms were precisely specified in all usage contexts there would be a fixed hierarchy of equivalence classes. Since there is good agreement between lexicographers for most natural languages this is essentially the structure of existing lexicons. In our application area, little is known about which visual communication terms are actually in use or their properties. There is no common visual language, but there are numerous visual and graphical symbols which are used within various knowledge domains for communication. Thus, each domain may define similar symbols for widely different uses. Because the context fixes the meaning, there is generally no need for disambiguation or lexical analysis. However, although cataloged visual symbol sets exist, these catalogs are underspecified, descriptive of only a few features and not intended for linguistic or lexical analysis.

We believe that the same issues occur within other application areas in which a significant amount of latent error and omission occurs due to community driven efforts. However, this is not a deficit. It is these very differences which can provide crucial information about the set of possible generalizations and assist in feature selection for the purposes of term disambiguation.

3.2 Deriving Equivalence Classes using Feature Aggregation

There is likely latent error within individual entries. The number of equivalence classes is exponential in the number of features and therefore exponentially larger by a factor of the error and omission rates. This not computationally pleasant, however, using aggregation across different feature-sets it can provide the necessary insights useful for deriving the common
lexicon. Those features that are persistent in the face of error and omission can be used to determine equivalence classes which are generalizations of the terms. The expected result is the ability to extract a generalization ontology by directly leveraging errors in term specification. In order to determine general meaning of aggregates of terms we can create equivalence classes based on shared portions of the attribute-value system. Naively, we may do so in an exhaustive fashion and look at all possible sets $S'$ of length $|S|-j$ for $j=1..|S|$.

The most granular view of the lexical entries distinguishes each individual entry uniquely by the complete set of characteristics, $S$, and usage contexts $C$. At this level there is no analysis of shared features. Each term is treated as unique if it has a unique UUID, even if all other characteristics are identical. This is sensible since its usage contexts may differ in such a way that the term should be treated as unique. This allows terms to be underspecified. In many cases, there may not exist a linguistic model for the differences between usage contexts. These cases are interesting for performing analysis of usage context in order to model new linguistic features. The lexicographer can use the hypotheses set $E|F$ to explore new linguistic features and subtleties in meaning.

All other levels of granularity rely on combining terms into equivalence classes based on sets of shared characteristics. Because we deal with visual forms our set of characteristics needs to include both graphical and linguistic attributes.

In describing visual features of the digital images in our lexicon, the lowest level is a pixel-by-pixel or a vector-by-vector description. This is essentially an atomic description for digital imaging, further decomposition is not possible. The image can also be described using different attributes such as edge, spline, and vertex locations and characteristics. Higher-level features aggregate large sets of granular features, such as contiguous regions (measured at different thresholds of color, hue, and saturation), object models, object relationship models, frequency domain representations, etc. Note that these descriptions will each have different computational efficiencies depending upon the task. Atomic descriptions have the least redundancy, and are computationally efficient for mapping transforms across pixels, but are not efficient for transforms which operate over aggregates. There exist a number of computer vision/imagery analysis algorithms that can derive these metrics.

Linguistic models may need to incorporate conventional attributes (character level model, phone model, N-gram models, sentence models, etc) as well as unconventional attributes that relate to the usage in a visual communication context (e.g. adjacency, subsumption, the use as an affix or containing affixes, indexicality, etc.) [5, 15]. We note that proximity between the orthographic, visual, and phonological forms of language terms will imbue shared semantics and induce problems in the use and interpretation of the terms. A word which sounds similar but means something different can cause problems. Contextual clues are very important in all languages. We believe that our distributed lexicon can assist in the analysis of these and other interesting features of emerging and specialized lexicons.

4. Conclusions: Folksonomy and the Living Lexicon

A lexicon of a language is its vocabulary, including its words and expressions and their characteristics. The intent of our graphical lexicon is to formally define the use of various icons and symbols such that such forms can be more effectively used in human and machine-mediated communication. We illustrated how granular computing concepts can be applied to help with the lexicon tractability and with issues related to the use of community efforts in constructing and sharing terms.

As linguists, we want to analyze the usage and lexical similarities between terms. By allowing different views of the same collection at different granularities we can determine the semantic coverage of an individual term with respect to its usage contexts. We can determine the definition of a term by performing an analysis of attributes shared across variations.

As language innovators we seek to determine if sufficient terminology exist for a given concept. Performing such a reverse lookup in a conventional lexicon is not practical. In general, our search becomes one in which we abstract gradually until we find terminology general enough to cover the potentially new term. The messiness of a distributed lexicon can assist us in our search and our process of abstraction. As such we can much more easily determine whether a term with our desired semantics exists.

The approach of using social networks and citizen science to create and manage semantic information is not new [9, 12]. Using a semantic markup for linguistic information can directly improve the machine processing capabilities for existing online dictionaries such as Wiktionary [6, 17]. By using a microformat reference to a specific definition we can specify which definition is meant instead of solely relying on context. By allowing the lexicon to be distributed we allow innovative uses of a term to be specified and shared. A centralized lexicon does not immediately allow for this type of community-driven lexicography, although community managed centralized lexicons are fairly
common. We believe that there is significant value in community managed lexicons which are distributed, especially in regard to the use of visual communication strategies due to the lack of formal lexicography.

Whereas a conventional dictionary requires centralization and the burden of centralized maintenance, the distributed dictionary approach leverages a folksonomy-like use of very simple microformats which can be compiled into centralized repositories as needed. Using our simple approach we have created a way to share lexical information in a distributed manner and leverage the differences between uses to expand that dictionary in a structured way. If there are many different uses of a particular term, each one of these uses has some validity within the scope of human language. This is true even if such usage is very limited. If a described term is similar to one in widespread use we can use a coarse grained view of the global lexicon to create a lexicon which aggregates the two terms. However, the underlying differentiation remains within the global lexicon, preserving the rich tapestry of terms, uses, and the many dialects of the contributors.

Tracking microformat references results in a real-time measure of usage as well as the usage context. Such information would greatly assist in content analysis of images just as our research benefits by the use of latent semantic analysis of text and image collocations for image characterization [3]. Using our approach we can automatically compile a dictionary that tracks the usage and currency of both new and antiquated visual and textual terms. Also, by looking at higher levels of granularity, relationships and aggregates can be analyzed to discern overall language trends. If the lexicon is distributed, then that meta-information can be included for analysis as well. For example, information regarding language users can be used to contrast between different community usages or contexts. Most intriguing is the possibility of facilitating automated language recognition by using a usage contexts and microformat-enabled terms as training sets. We only applied the microformat lexicon to our visual communication research, but the same technique could be easily used within the context of natural language.

5. References


