Demonstration: Using WordNet to Build Hierarchical Facet Categories

Emilia Stoica
School of Information
UC Berkeley
estoica@sims.berkeley.edu

Marti A. Hearst
School of Information
UC Berkeley
hearst@sims.berkeley.edu

ABSTRACT

We have designed an algorithm for automatically generating hierarchical faceted metadata from textual description of items, to be incorporated into browsing and navigation interfaces for large information collections. From an existing lexical database (such as WordNet), the Castanet algorithm carves out a structure that reflects the contents of the target information collection. The algorithm has been successfully applied to collections as diverse as recipes, biomedical journal titles, and art history image descriptions; the resulting category hierarchies require only small adjustments to achieve intuitive results with good coverage.

1. INTRODUCTION

A considerable impediment to the wider adoption of faceted interfaces is the need for creation of the faceted hierarchies and the assignments of terms from the hierarchies to the information items.

In this paper, we describe an algorithm called Castanet that makes considerable progress in automating faceted metadata creation. Castanet creates domain-specific overlays on top of a large general-purpose lexical database (WordNet [3]), producing surprisingly good results in a matter of minutes for a wide range of subject matter. Our approach is relatively simple but at the same time effective at disambiguating and converting the hierarchy into understandable facets.

2. WORDNET CHALLENGES

The main idea behind the Castanet algorithm is to carve out a structure from the hypernym (IS-A) relations within the WordNet [3] lexical database. The primary unit of representation in WordNet is the synset, which is a set of words that are considered synonyms for a particular concept. Each synset is linked to other synsets via several types of lexical relations; we only use hypernymy in this algorithm.

One well-known difficulty with using WordNet for automated analysis tasks is the fine granularity of the word sense distinctions. WordNet distinguishes between, for example, {*newspaper, paper*} as "a daily or weekly publication on folded sheets" and {*newspaper, paper*} as "a newspaper as a physical object." (Mihalcea and Moldovan [5] describe an algorithm that folds together these very fine sense distinctions.)

WordNet also has more easily distinguishable sense distinctions, such as {*course*} as "part of a meal served at one time" and {*course*} as "facility consisting of a circumscribed area of land or water laid out for a sport." (Other senses of this word are described by multiword synsets, such as {*course, course of study, class*}).

This ambiguity can extend up the IS-A hierarchy as well. For example, one sense of the word *tuna* is as a spiky cactus and another is a fish, but the fish sense as well can have two different paths up the IS-A hierarchy, through the {*food fish*} or the {*bony fish*} synsets.

Each sense of a word is assigned a sense number, and the first sense is usually the most common one. A commonly applied and very useful WordNet heuristic is to use the first sense when an algorithm has no way to choose among senses.

Another problem with applying WordNet to a human-readable interface is that the subdivisions within the hypernym hierarchy are sometimes very fine-grained, and paths from root to node can become quite deep (e.g., sense #3 of *tuna* is 14 levels down from its root node).

Another well-known problem with WordNet is its sparse coverage of proper names (including product names) and its spotty coverage of two and three-word compounds. For this kind of application it is important to use a separate mechanism for including proper names and uncommon compounds; we do not address these issues directly here.

3. ALGORITHM OVERVIEW

The Castanet algorithm [7] assumes that there is text associated with each item in the collection, or at least with a representative subset of the items. The textual descriptions are used *both* to build the facet hierarchies and to assign items (documents, images, citations, etc.) to the facets. The text does not need to be particularly coherent for the algorithm to work; we have applied it to fragmented image annotations and short journal titles, but if the text is impoverished, the information items will not be labeled as thoroughly as desirable and additional manual annotation may be needed.

The algorithm has five major steps, each of which is described in more details below:

1. Select target terms from textual descriptions of information items
2. Build the Core Tree:
   - For each term, if the term is unambiguous or a domain term (see below), add its synset's IS-A path to the Core Tree,
   - Increment the counts for each node in the synset's path with the number of documents in which the target term appears.

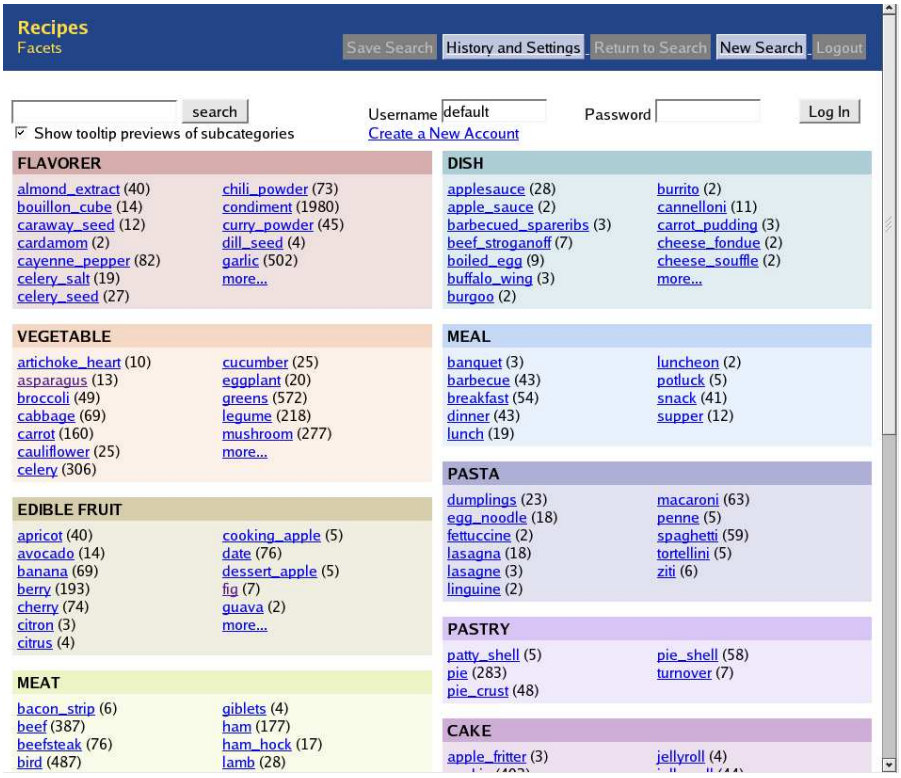

Figure 1: Recipes collection overview.

3. Augment the core tree with the remaining terms' paths:
   - For each candidate IS-A path for the ambiguous term, choose the path for which there is the most document representation in the Core Tree.
4. Compress the augmented tree.
5. Remove top-level categories, producing a faceted set of hierarchies.

3.1 Select Target Terms

Castanet selects only a subset of terms, called *target terms*, that are intended to best reflect the topics in the documents. We use the *term distribution* – defined as the number of item descriptions containing the term – as the selection criterion. The algorithm retains those terms that have a distribution larger than a threshold (currently those terms within 10% of the maximum distribution).

3.2 Build the Core Tree

We build the core tree with the terms that have only one sense within WordNet and the terms that match one of the pre-selected WordNet domains.

For every such target term, the algorithm looks up its synset and its hypernym path in WordNet. (If a term does not have representation in WordNet, then it is not included in the category structure.) To add a path to the Core Tree, its path is merged with those paths that have already been placed in the tree.

In addition to augmenting the nodes in the tree, adding in a new term increases a count associated with each node on its path; this count corresponds to how many documents the term occurs in. Thus the more common a term, the more weight it places on the path it falls within.

3.3 Augment the Core Tree

The Core Tree contains only a subset of terms in the collection. To add in the paths for the remaining target terms, Castanet looks at the number of documents assigned to the deepest node that is held in common between the existing Core Tree and each hypernym path for the ambiguous term. The path with the largest number of documents is selected.

3.4 Compress the Augmented Tree

The tree that is obtained in the previous step usually is very deep, which is undesirable from a user interface perspective. Castanet uses two rules for compressing the tree:

1. Starting from the leaves, eliminate a parent that has fewer than k children, unless the parent is the root or has an item count larger than $0.1\times$(maximum term distribution).
2. Eliminate a child whose name appears within the parent's name, unless the child contains a WordNet domain name.

3.5 Remove Top Level Categories

Generally, the resulting tree is several levels deep, but the categories of interest are close to the leaves, which means the user has to unnecessarily descend several levels in the hierarchy to reach them.

To remove top level categories, first Castanet simply eliminates the categories that are very general WordNet synsets (e.g., abstraction, entity). If eliminating these nodes produces the desired hierarchies, then the algorithm is done.

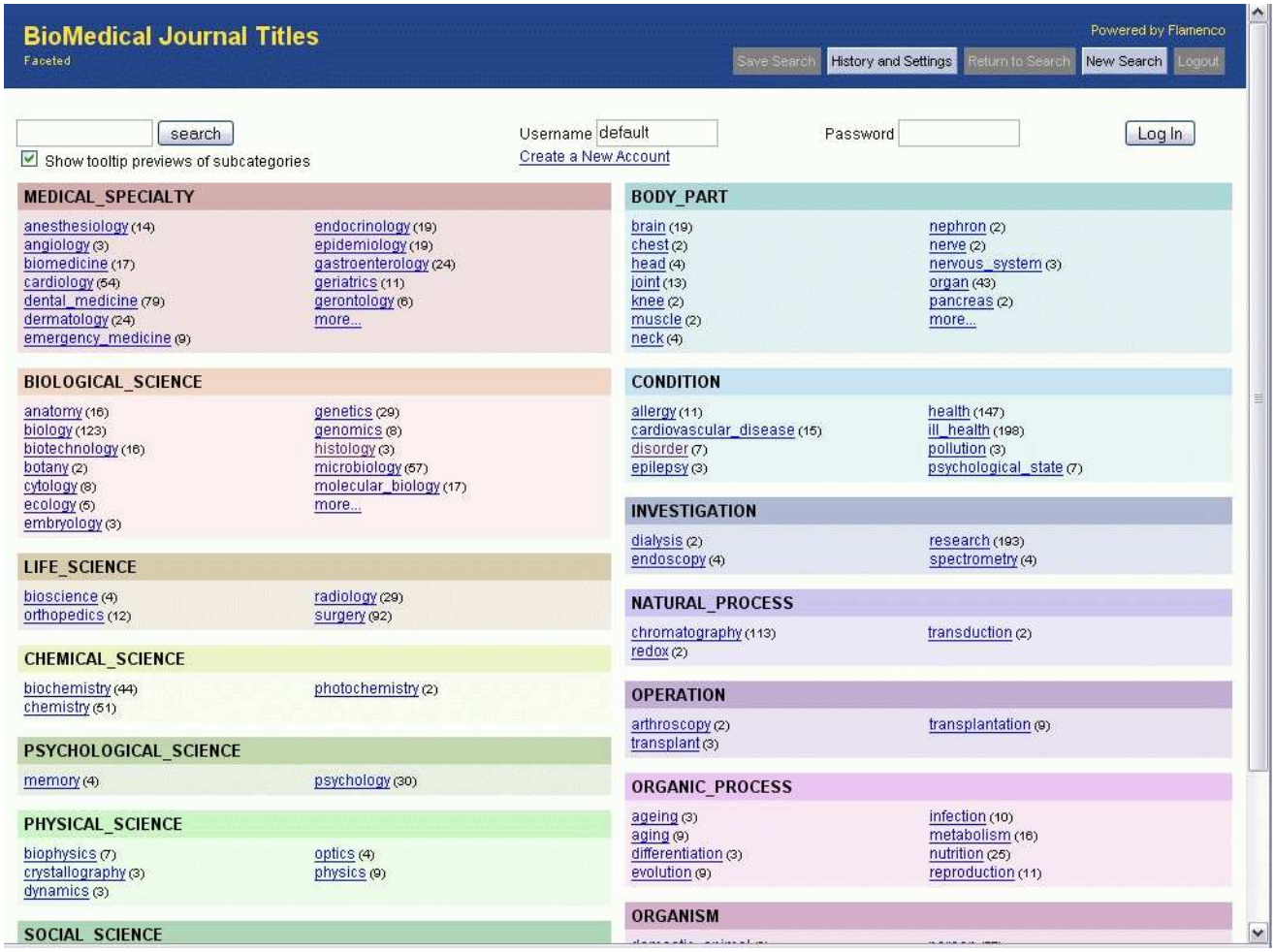

Figure 2: Biomedical journal titles colletion.

Otherwise, it further eliminates the levels that have depth shallower than the depth of the categories of interest.

Space limitations preclude inclusion of the full details of the algorithm. A detailed description appears in [8].

4. EXAMPLE OUTPUT

In this section we show examples of the results Castanet produces on two datasets: recipes and biomedical documents.

4.1 Recipes Dataset

We obtained a small collection of recipes from Southwest website (http://www.southwestcuisine.com), consisting of 3,537 items. A typical entry within the this dataset is:

> *Spaghetti with Beef Casserole - Ground Beef Recipe Favorite: 1 clove garlic, minced; 1 medium onion, chopped; 2 tablespoons olive oil; 1 pound ground chuck, 8 oz thin spaghetti... In a large skillet over medium low heat, saute onion and garlic in olive oil until onion is tender. ... Cook spaghetti according to package instructions ...*

Figure 1 shows part of the results of the CastaNet algorithm on this dataset, displaying the top levels in the Flamenco interface [4]. For this collection the categories are clean and highly understandable, grouping terms especially well into ingredient facets, and doing less well with the messier category "dishes". Many of the categories map directly onto carefully crafted category hierarchies seen at recipes websites such as epicurious.com and marthastewart.com, but the Castanet results also have hierarchy which these sites do not support.

4.2 BioMedical Journal Titles

We obtained a list of 3,275 journal titles from those available in the MEDLINE (http://www.medline.com) citation database. Sample titles are *Archives of otolaryngology–head & neck surgery, Journal of women & aging*, and *The Journal of bone and joint surgery. British volume.* To our knowledge there exists no hierarchical organization for the full set of journals.

Figure 2 shows the top level organization of the results, again using the Flamenco software to display them. These facets were built from the journal titles only and so most journals have only one or at most two labels assigned to them. In future we plan to augment the journal names with text from the articles within the journals in order to make

the automated item assignment more thorough.

The difficulty of evaluating ontologies is well-established [2]. We are planning a two-pronged approach to evaluation via usability studies and comparison with other state-of-the-art algorithms, in particular, LDA [1] and Subsumption [6].

5. DEMO DESCRIPTION

Our demo will show the facets that Castanet builds for the two collections: biomedical journal titles and recipes. As user interface we will use Flamenco image browser.

We will perform various searches and show how Castanet facilitates the user in finding the desired information.

Acknowledgements: Portions of this research were funded by NSF IIS-9984741 and by NSF DBI-0317510.

6. REFERENCES

[1] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS),16*, Cambridge, MA, 2004.

[2] P. Buitelaar, S. Handschuh, and B. Magnini, editors. *ECAI-2004 Workshop: Towards Evaluation of Text-based Methods in the Semantic Web and Kn owledge Discovery Life Cycle*, August 2004.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] M. A. Hearst, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9), September 2002.

[5] R. Mihalcea and D. I. Moldovan. Ez.wordnet: Principles for automatic generation of a coarse grained wordnet. In *Proceedings of FLAIRS Conference 2001*, May 2001.

[6] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of SIGIR '99*, 1999.

[7] E. Stoica and M. Hearst. Nearly-automated metadata hierarchy creation. In *HLT-NAACL'04, Companion Volume*, 2004.

[8] E. Stoica and M. A. Hearst. Automating creation of hierarchical faceted metadata. In preparation., 2006.