A Multimodal Vocabulary for Augmentative and Alternative Communication from Sound/Image Label Datasets

Xiaojuan Ma **Christiane Fellbaum** **Perry R. Cook**
Princeton University
35 Olden St. Princeton, NJ 08544, USA
{xm,fellbaum,prc}@princeton.edu

Abstract

Existing Augmentative and Alternative Communication vocabularies assign multimodal stimuli to words with multiple meanings. The ambiguity hampers the vocabulary effectiveness when used by people with language disabilities. For example, the noun "a missing letter" may refer to a character or a written message, and each corresponds to a different picture. A vocabulary with images and sounds unambiguously linked to words can better eliminate misunderstanding and assist communication for people with language disorders. We explore a new approach of creating such a vocabulary via automatically assigning semantically unambiguous groups of synonyms to sound and image labels. We propose an unsupervised word sense disambiguation (WSD) voting algorithm, which combines different semantic relatedness measures. Our voting algorithm achieved over 80% accuracy with a sound label dataset, which significantly outperforms WSD with individual measures. We also explore the use of human judgments of evocation between members of concept pairs, in the label disambiguation task. Results show that evocation achieves similar performance to most of the existing relatedness measures.

1 Introduction

In natural languages, a word form may refer to different meanings. For instance, the word "fly" means "travel through the air" in context like "fly to New York," while it refers to an insect in the phrase "a fly on the trashcan." Speakers determine the appropriate sense of a polysemous word based on the context. However, people with language disorders and access/retrieval problems, may have great difficulty in understanding words individually or in a context. To overcome such language barriers, visual and auditory representations are introduced to help illustrate concepts (Ma et al., 2009a)(Ma et al., 2010). For example, a person with a language disability can tell the word "fly" refers to "travel through the air" when he sees a plane in the image (rather than an insect); likewise he can distinguish the meaning of "fly" given the plane engine sound vs. the insect buzzing sound. This approach has been employed in Augmentative and Alternative Communication (AAC), in the form of multimodal vocabularies in assistive devices (Steele et al. 1989)(Lingraphica, 2010).

However, current AAC vocabularies assign visual stimuli to words instead of specific meanings, and thus bring in ambiguity when a user with language disability tries to comprehend and communicate a concept. For example, for the word "fly," Lingraphica only has an icon showing a plane and a flock of birds flying. Confusion arises when a sentence like "I want to kill the fly (the insect)" is explained using the airplane/bird icon. Similarly, it will lead to miscommunication if the sound of keys jingling is used to express "a key is missing" when the person intends to refer to a key on the keyboard. People with language impairment are relying on the AAC vocabularies for language access, and any ambiguity may result in communication failure.

To address this problem, we propose building a semantic multimodal AAC vocabulary with visual and auditory representations expressing concepts rather than words (Figure 1), as the backbone of the language assistant system for people with aphasia (Ma et al. 2009b). Our work is exploratory with the following innovations: 1) we target the insufficiency of current assistive vocabularies by resolving ambiguity; 2) we enrich concept inventory and connect concepts through language, environmental sounds, and images (little research has looked into conveying concepts through natural nonspeech sounds); and 3) our vocabulary has a dynamic scalable semantic network structure rather

Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pages 62–70,
Los Angeles, California, June 2010.

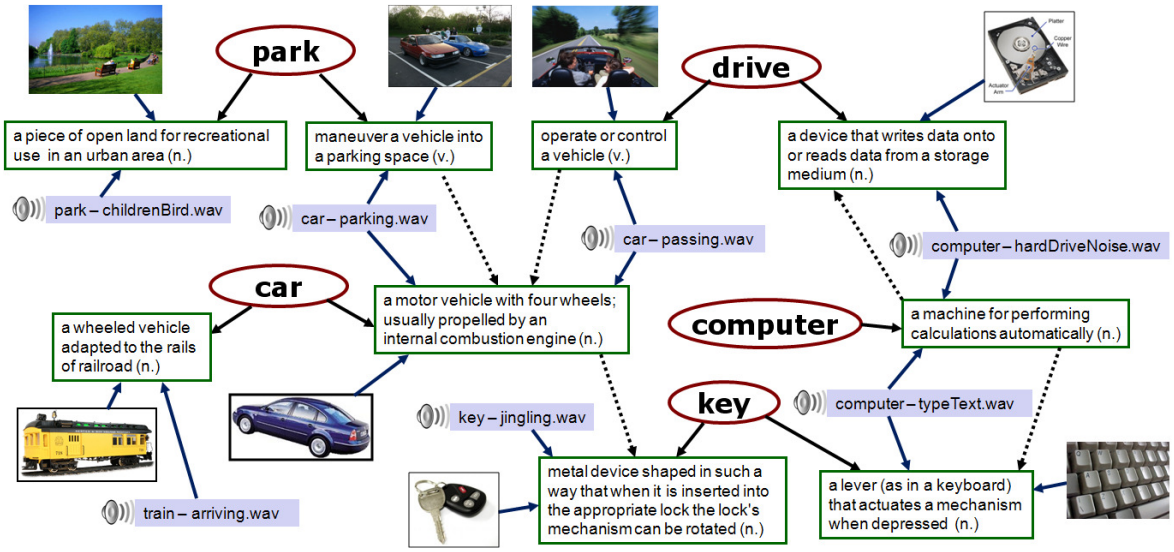

Figure 1. Disambiguated AAC multimedia vocabulary; dash arrows are semantic relations between concepts.

than simply grouping words into categories as conventional assistive devices do.

One intuitive way to build a disambiguated multimodal vocabulary is to manually assign meanings to each word in the existing vocabulary. However, the task is time consuming with poor scalability – no new multimedia representations are generated for concepts that are missing in the vocabulary. ImageNet (Jia et al., 2009) was constructed by people verifying the assignment of web images to given synonym sets (synsets). ImageNet has over nine million images linked to about 15 thousands noun synsets in WordNet (Fellbaum, 1998). Despite the huge human effort, ImageNet, with the goal of creating a computer vision database, does not yet include all the most commonly used words across different parts of speech. It is not yet suitable for a language support application.

We explore a new approach for generating a vocabulary with concept to sound/image associations, that is, conducting word sense disambiguation (WSD) techniques used in Natural Language Processing on sound/image label datasets. For example, the labels "car, drive, fast" for the sound "car – passing.wav" are assigned to synsets "car: a motor vehicle," "drive: operate or control a vehicle," and "fast: quickly or rapidly" via WSD. It means the sound "car – passing.wav" can be used to depict those concepts. This approach is viable because the words in the sound/image labels were shown to evoke one another based on the auditory/visual content, and their meanings can be identified by considering all the tags generated for a given sound or image as a context. With the availability of large sound/image label datasets, the vocabulary created from WSD can be easily expanded.

A variety of WSD methods (e.g. knowledge-based methods (Lesk, 1986), unsupervised methods (Lin, 1997), semi-supervised methods (Hearst, 1991) (Yarowsky, 1995), and supervised methods (Novischi et al., 2007)) were developed and evaluated with corpus data and other text documents like webpages. Compared to the text data that WSD methods work with, labels for sounds and images have unique characteristics. The labels are a bag of words related to the visual/auditory content; there is no syntactic or part of speech information, nor are the words necessarily contextual neighbors. For example, contexts suggest landscape senses for the word pair "bank" and "water", whereas in an image, a person may drink water inside a bank building. Furthermore, few annotated image or sound label datasets are available, making it hard to apply supervised or semi-supervised WSD methods.

To efficiently and effectively create a disambiguated multimodal vocabulary, we need to achieve two goals. First, optimize the accuracy of the WSD algorithm to minimize the work required for manual checking and correction afterwards. Second, construct a semantic network across different parts of speech, and thus explore linking semantic relatedness measures that can capture aspects different from existing ones. In this paper, we target the first goal by proposing an unsupervised sense disam-

biguation algorithm combining a variety of semantic relatedness measures. We chose an unsupervised method because of the lack of a large manually annotated gold standard. The measure-combined voting algorithm presented here draws advantages from different semantic relatedness measures and has them vote for the best-fitting sense to assign to a label. Evaluation shows that the voting algorithm significantly exceeds WSD with each individual measure.

To approach the second goal, we proposed and tested a semantic relatedness measure called evocation (Boyd-Graber et al., 2006) in disambiguation of sound/image labels. Evocation measures human judgements of relatedness between a directed concepts pair. It provides cross parts of speech evocativeness information which supplements most of the knowledge-based semantic relatedness measures. Evaluation results showed that the performance of WSD with evocation is no worse than most of the relatedness measures that we applied, despite the relatively small size of the current evocation dataset.

2 Dataset: Semantic Labels for Environmental Sounds and Images

Our ultimate goal is to create an AAC vocabulary of associations between environmental sounds and images and groups of synonymous words that are relevant to the content. We are working with two datasets of human labels for multimedia data, SoundNet and the Peekaboom dataset.

2.1 SoundNet Sound Label Dataset

The SoundNet Dataset (Ma, Fellbaum, and Cook, 2009) consists of 327 environmental "soundnails" (5-second audio clips) each with semantic labels collected from participants via a large scale Amazon Mechanical Turk (AMT) study. The soundnails cover a wide range of auditory scenes, from vehicle (e.g. car starting), mechanical tools (e.g. handsaw) and electrical devices (e.g. TV), to natural phenomena (e.g. rain), animals (e.g. a dog barking), and human sounds (e.g. a baby crying). In the AMT study, participants were asked to generate tags for each soundnail labeling its source, possible location, and actions involved in making the sound.

Each soundnail was labeled by over 100 people. The tags were clustered into meaning units that SoundNet refers to as "sense sets." A sense set includes a set of words with similar meanings. For instance, for the soundnail pre-labeled "bag, zipOpen" which is the sound of opening the zipper of a bag, the following sense sets were generated:

(a) "**zipper**" {zipper, zip up, zip, unzip};
(b) "**bag**" {bag, duffle bag, nylon bag, suitcase, luggage, backpack, purse, pack, briefcase};
(c) "**house**" {house, home, building}, and
(d) "**clothes**" {clothes, jacket, coat, pants, jeans, dress, garment}.

The word in **bold** is was judged by SoundNet to be the best representative of the sense set, and other words, possibly belonging to different parts of speech are included in the curly brackets enclosing the sense sets. SoundNet uses sense sets rather than single words because 1) people may use different words to describe the same underlying concept, (e.g. "baby" and "infant;" "rain" as a noun and as a verb); 2) people cannot draw fine distinctions between objects and events that generate similar sounds, and thus may come up with different but related categories (e.g. "plate," "cup," and "bowl" for the dish clinking sound); and 3) people may perceive objects and events that are not explicitly presented in the sound very differently (e.g. "bag" vs. "clothes" for the sound made by a zipper). In this experiment, only sense sets (labels) that were generated by at least 25% of the labelers were used.

In our disambiguation experiment, two kinds of contexts were explored. In the Context 1 scheme, each label is treated separately: all its members plus the representatives of the other sense sets are considered. Take the soundnail "bag, zipOpen" as an example. The context for disambiguating label (a) "**zipper**" {zipper, zip up, zip, unzip} is:

zipper, zip up, zip, unzip, **bag**, **house**, **clothes**.

The context for label (d) "**clothes**" {clothes, jacket, coat, pants, jeans, dress, garment} is:

clothes, jacket, coat, pants, jeans, dress, garment, **zipper**, **bag**, **house**.

In the Context 1 scheme, all **representative** words will be disambiguated multiple times. The final result will be the synset that gets the most votes. In the Context 2 scheme, as for the image dataset described below, all members from each sense set are put together to create the context, and each word is disambiguated only once.

2.2 Peekaboom Image Label Dataset

The ESP Game Dataset (Von Ahn and Dabbish, 2004) contains a large number of web images and human labels produced via an online game. For example, an image of a glass of hard liquor is labeled "full, shot, alcohol, clear, drink, glass, beverage." The Peekaboom Game (Von Ahn et al., 2006) is the successor of the ESP Game. In our experiment, part of the Peekaboom Dataset (3,086 images) was used. For each image, all the labels together form the context for sense disambiguation.

The Peekaboom labels are noisier than the SoundNet labels for several reasons. First, random objects may appear in a picture and thus be included in the labels. For example, an image is labeled "computer, shark" because there is a shark picture on the computer screen. Second, texts in the images are often included in the labels. For example, the word "green" is one of the labels for an image with a street sign "Green St." Third, the Peekaboom labels are not stemmed, which adds another layer of ambiguity. For example, the labels "bridge, building" could refer to a building event or to a built entity. In the experiment, all labels for an image are used in their unstemmed form to construct the context for WSD.

3 Evocation and Other Semantic Relatedness Measures

A set of measures were selected to assess the relatedness between possible senses of words in the sound/image labels. Apart from existing methods, an additional measure, evocation, is introduced.

3.1 Evocation

Evocation (Boyd-Graber et al., 2006) measures concept similarity based on human judgment. It is a directed measure, with evocation(synset A, synset B) defined as how much synset A brings to mind synset B. The evocation dataset has been extended to scores for 100,000 directed synset pairs (Nikolova et al., 2009).

The evocation data were collected independently of WordNet or corpus data. We propose the use of evocation in WSD for image and sound labels for the following reasons. First, the sound and image labels are generated based on human perception of the content and common knowledge. In SoundNet in particular, many of the evoked labels reflected the most obvious objects or events in a sound scene. For example, "bag" and "coat" were evoked from the zipper soundnail. In this case, the evocation score may be a good evaluation of the relatedness between the labels. Second, evocation assesses relatedness of concepts across different parts of speech, which is suitable for identifying image and sound labels containing nouns, verbs, adjectives, adverbs, etc.

This paper is a first attempt to compare the effectiveness of the use of evocation measure in sense disambiguation to the conventional, relatively better tested similarity measures, in the context of assigning synsets to sound/image labels. Considering that the evocation dataset is small in size and susceptible to noise given the method by which it was collected, we have not yet incorporated evocation into the measure-combined voting algorithm described in the Section 4.

3.2 Semantic Relatedness Measures

Nine measures of semantic relatedness[1] between synsets are used in the experiment, both as contributors to the voting algorithm and as baselines for comparison, including:

1) WordNet path based measures.

- "path" – shortest path length between synsets, inversely proportional to the number of nodes on the path.
- "wup" (Wu and Palmer, 1994) – ratio of the depth of the Least Common Subsumer (LCS) to the depths of two synsets in the Wordnet taxonomy.
- "lch" (Leacock and Chodorow, 1998) – considering the length of the shortest path between two synsets to the depth of the WordNet taxonomy.

2) Information and content based measures.

- "res" (Resnik, 1995) – the informational content (IC) of a given corpus of the LCS between two synsets.
- "lin" (Lin, 1997) – the ratio of the IC of the LCS to the IC of the two synsets.
- "jcn" (Jiang and Conrath, 1997) – inversely proportional to the difference between the IC of the two synsets and the IC of the LCS.

[1] "hso" (Hirst and St-Onge, 1998) extensively slows down the WSD process with over five context words, and thus, is not included in the experiment.

3) WordNet definition based measures.

- "lesk" (Banerjee and Pedersen, 2002) – overlaps in the definitions of two synsets.
- "vector" (Patwardhan and Pedersen, 2006) – cosine of the angle between the co-occurrence vector computed from the definitions around the two synsets.
- "vector_pairs" – co-occurrence vectors are computed from definition pairs separately.

The computation of the relatedness scores using measures listed above were carried out by codes from the WordNet::Similarity (Pedersen et al., 2004) and WordNet::SenseRelate projects (Pedersen and Kolhatkar, 2009). In contrast to WordNet::SenseRelated, which employs only one similarity measure in the WSD process, this paper proposes a strategy of having several semantic relatedness measures vote for the best synset for each word. The voting algorithm intends to improve WSD performance by combining conclusions from various measures to eliminate a false result. Since there is no syntax among the words generated for a sound/image, they should all be considered for WSD. Thus, the width of the context window is the total number of words in the context.

4 Label Sense Disambiguation Algorithm

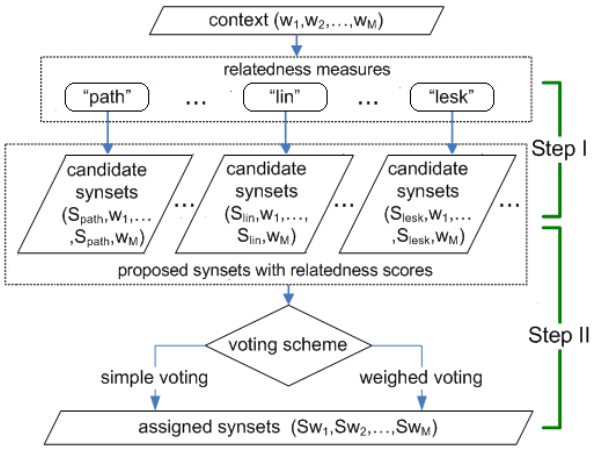

Figure 2. Measure-Combined Voting Algorithm.

Figure 2 shows the overall process of the measure-combined voting algorithm for disambiguating sound/image labels. After the context for WSD is generated, the process is divided into two steps. In Step I, the relatedness scores of each sense of a word based on the context is computed by each measure separately. Step II combines results from all measures and generates the disambiguated synsets for all words in the sound/image labels. Evocation did not participate in Step II.

4.1 Step I: Generate Candidate Synsets Based on Individual Measures

Given the context of M words (w_1, …, w_M), and K relatedness measures (k = 1, …, K), the task is to assign each word w_j (j = 1, …, M) to the synset $s_{x,wj}$ that is the most appropriate within the context. Here, the word w_j has N_j synsets, denoted as $s_{n,wj}$ (n = 1, …, N_j). Step I is to calculate the relatedness score for each synset of each word in the context.

$$score_k(s_{i,w_j}) = \sum_{m=1,\ldots,M}^{m\neq j} \max_{n=1,\ldots,N_m} (measure_k(s_{i,w_j}, s_{n,w_m}))$$

The evocation score between two sysnets s_a, s_b is the maximum of the directed evocation ratings.

$$\underset{evocation}{score}(s_a, s_b) = \max(evocation(s_a, s_b), evocation(s_b, s_a))$$

$$\underset{evocation}{score}(s_{i,w_j}) = \sum_{m=1,\ldots,M}^{m\neq j} \max_{n=1,\ldots,N_m} (\underset{evocation}{score}(s_{i,w_j}, s_{n,w_m}))$$

The synset that evocation assigns to word j is the one with the highest score.

$$s_{w_j} = s_{x,w_j}, if \quad \underset{evocation}{score}(s_{x,w_j}) = \max_{i=1,\ldots,N_j} (\underset{evocation}{score}(s_{i,w_j}))$$

4.2 Step II: Vote for the Best Candidate

Three voting schemes were tested, including unweighted simple votes, weighted votes among top candidates, and weighted votes among all synsets.

1) Unweighted Simple Votes

Synset $s_{n,wj}$ of word w_j gets a vote from relatedness measure k if its $score_k$ is the maximum among all the synsets for w_j, and it becomes the candidate synset for w_j elected by measure k ($C_{k,wj}$):

$$vote_k(s_{x,w_j}) = \begin{cases} 1, if \quad score_k(s_{x,w_j}) = \max_{i=1,\ldots,N_j} (score_k(s_{i,w_j})) \\ 0, else \end{cases}$$

$$candidate_k(s_{w_j}) = s_{x,w_j}, \qquad if\ vote_k(s_{x,w_j}) = 1$$

The candidate list for word w_j (candidates(Sw_j)) is the union of all candidate synsets elected by individual relatedness measures.

$$candidates(s_{w_j}) = \underset{k=1,\ldots,K}{union}(candidate_k(s_{w_j}))$$

For each candidate in the list, the votes from all measures are calculated. The one receiving the most votes becomes the proposed synset for w_j.

$$voteCount(s_{i,w_j}) = \sum_{k=1}^{K} vote_k(s_{i,w_j})$$

$$s_{w_j} = s_{x,w_j}, if$$

$$voteCount(s_{x,w_j}) = \max_{s_{i,w_j} \in candidates(s_{w_j})} (voteCount(s_{i,w_j}))$$

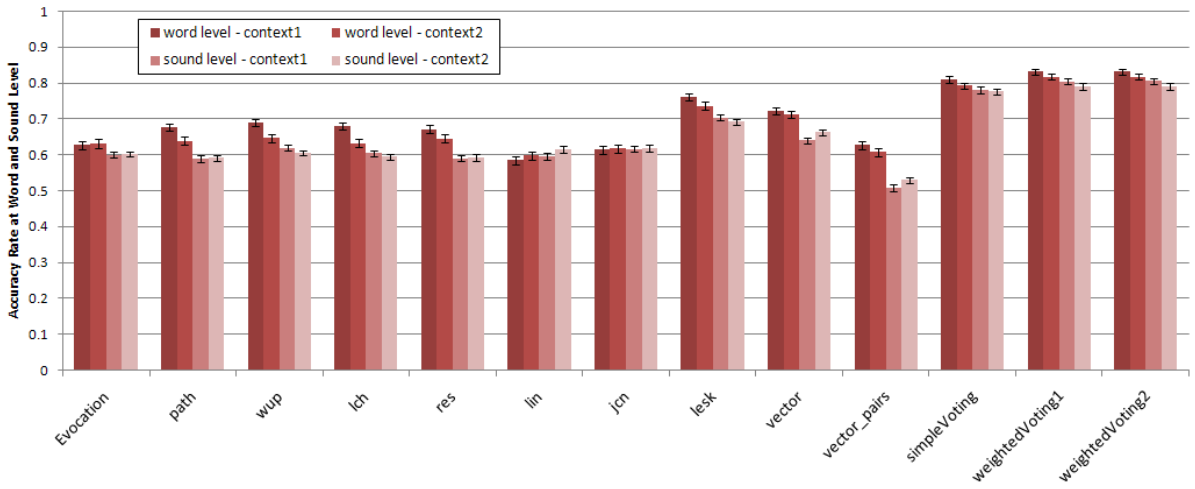

Figure 3. Accuracy rate at word and sound level in comparison among evocation, voting, and nine individual sense similarity measures.

2) Weighted Votes among Top Candidates

The weighted voting scheme avoids a situation where the false results win by a very small margin. The weight under relatedness measure k for $s_{i,wj}$ is calculated as the relative score to the maximum $score_k$ among all synsets for word w_j. It suggests how big of a difference in relatedness score of any given synset is to the highest score among all the possible synsets for the target word.

$$weight_k(s_{x,w_j}) = score_k(s_{x,w_j}) / \max_{i=1,...,N_j} (score_k(s_{i,w_j}))$$

The weighted votes synset $s_{i,wj}$ receives over all measures is the sum of its weight under individual measure. In voting scheme 2, the synset from the candidate list which gets the highest weighted votes becomes the winner.

$$weightedVote(s_{i,w_j}) = \sum_{k=1}^{K} weight_k(s_{i,w_j})$$

$$s_{w_j} = s_{x,w_j}, if$$

$$weightedVote(s_{x,w_j}) = \max_{s_{i,w_j} \in candidates(s_{w_j})} (weightedVote(s_{i,w_j}))$$

3) Weighted Votes among All Synsets

Voting scheme 3 differs from 2 in that the synset from all synsets for word w_j which gets the highest weighted votes is the proposed synset for w_j.

$$s_{w_j} = s_{x,w_j}, if$$

$$weightedVote(s_{x,w_j}) = \max_{i=1,...,N_j} (weightedVote(s_{i,w_j}))$$

5 Evaluation

The evaluation of WSD with evocation and the measure-combined voting algorithm was carried out primarily on the SoundNet label dataset because of the availability of ground truth data. SoundNet provides manual annotation for 1,553 different words for 327 soundnails (e.g. the word "road" appears in 41 sounds).

The accuracy rate (precision) was computed for each WSD method. The sound level accuracy of a WSD_k is the average percentage of correct sense assignments over the 327 sounds. The word level accuracy is the mean over 1553 distinctive words. Accuracy rates of different measures at both level accepted the null hypothesis in homogeneity test.

$$\underset{sound-level}{accuracy}(WSD_k) = (\sum_{i=1}^{327} (\%correctness)_i) / 327$$

$$\underset{word-level}{accuracy}(WSD_k) = (\sum_{w=1}^{1553} (\%correctness)_w) / 1553$$

Due to the lack of ground truth in the Peekaboom dataset, we only computed the overlap between the WSD result of 3,086 images from the voting algorithm, evocation and each relatedness measures.

5.1 Overall Comparison across WSD methods with Various Relatedness Measures

Figures 3 show the overall comparison among different methods at both sound level and word level. It suggests that the performance of the evocation measure in sense disambiguation is as good as the path-based and context-based measures. The definition-based measures ("lesk" and "vector") are significantly better than other measures if used individually (similar to (Patwardhan et al.2003)).

However, the voting algorithms proposed in this work significantly outperformed each individual measure based on ANOVA results. At sound level, Context 1: ($F(12, 20176) = 102.92$, $p < 0.001$); Context 2: ($F(12, 4238) = 89.42$, $p < 0.001$). At word level, Context 1: ($F(12, 20176) = 68.78$, $p < 0.001$); Context 2: ($F(12, 4238) = 60.72$, $p < 0.001$).

The scheme of composing context (Section 2.1) has significant impact on the accuracy, with Context 1 (taking all members in the related sense set and representatives from the others) outperforming Context 2 (taking all words in all sense sets) at the word level ($F(1, 40352) = 20.19$, $p < 0.001$). The influence of context scheme is not significant at the sound level ($F(1, 8476) = 0.35$, $p = 0.5546$). The interaction between measures and context schemes is not significant, indicating that accuracy differences are similar regardless of context construction.

5.2 Performance of the Voting Algorithm

Figure 4 shows the histogram (distribution) for the accuracy rate at sound and word levels. We see that for the voting algorithm, the accuracy rates are greater than 0.7 for most of the sounds, and greater than 0.9 for majority of the words to disambiguate.

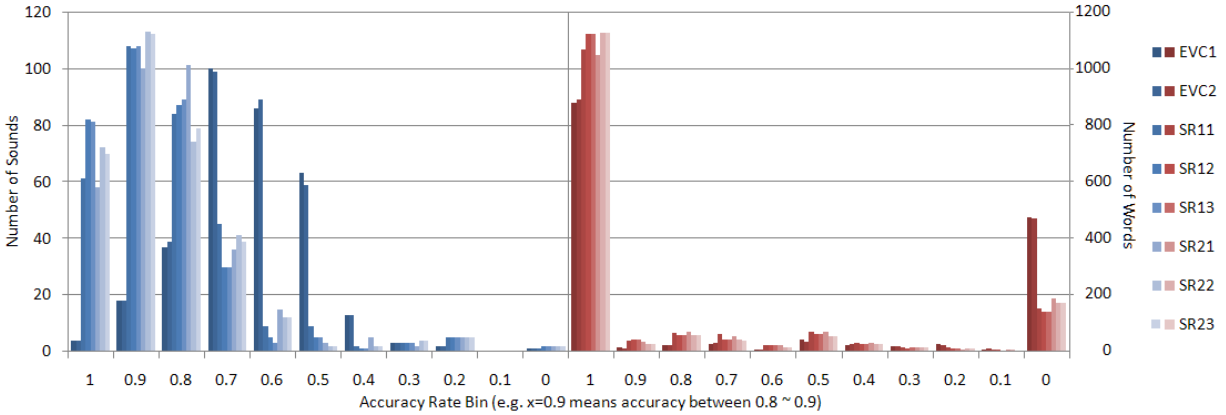

Figure 4. Histogram of accuracy rate at sound (327, left) and word level (1553, right) among different measures, contexts, and voting schemes. EVC1 = Evocation (Context 1); SR11 = Voting (Context 1, voting scheme 1).

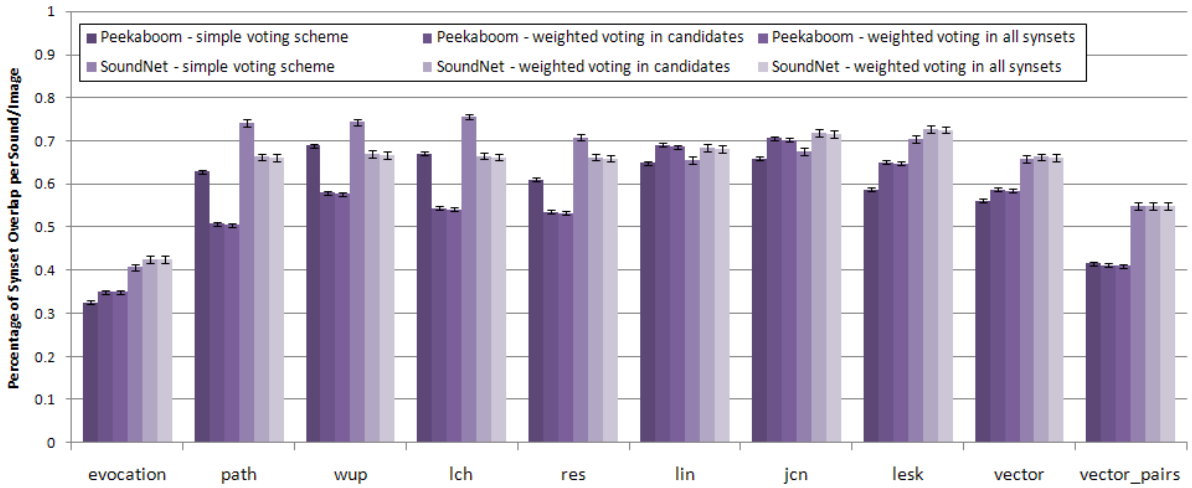

Figure 5. Percentage of sense disambiguation results overlap between voting algorithm, evocation, and individual sense relatedness measures at image (3,086 images) and sound (327 sounds) level.

Figure 5 show the percentage of sense disambiguation results overlapping between voting algorithm and individual relatedness measures. Note that any two methods may come up with different correct results (e.g. "lesk" assigned "chirp" as "a sharp sound" while the voting algorithm assigned "chirp" as "making a sharp sound"). This indicates the change of the contribution of each relatedness measures in different voting schemes. In the simple voting scheme, more disambiguation results came from the "path," "wup," and "lch" (the WordNet path based measures), while the weighted voting

scheme took more of the recommendations from "lesk," "lin," and "jcn" (context and definition based measures) into consideration. At the sound level, there is no significant accuracy difference among the three voting schemes, and the influence of the context composition is similar. However, at the word level (Figure 3), the weighted voting schemes significantly outperformed the simple voting scheme ($F(2, 9312) = 5.20$, $p = 0.0055$), and all of them have significantly better accuracy when the context contains mainly members from the same sense set ($F(1, 9312) = 4.79$, $p = 0.0287$).

5.3 Performance of WSD with Evocation

As shown in Figures 3, the performance of the evocation measure is not significantly different from path-based and some context-based measures at sound level, including "path," "wup," "lch," "res," "lin," and "jcn" (for Context 1, $F(6, 2282) = 2.0582$, $p = 0.0551$; for Context 2, $F(6, 2282) = 1.6679$, $p = 0.1249$); and is significantly better than the vector_pairs measure (for Context 1, $F(1, 652) = 61.37$, $p < 0.001$; for Context 2, $F(1, 652) = 36.47$, $p < 0.001$). At the word level, the performance of the evocation measure is not significantly different from that of measures including "path," "wup," "lch," "res" ($F(4, 7760) = 0.39$, $p = 0.8135$), and "lin," "jcn," and "vector_pairs" ($F(3, 6208) = 1.52$, $p = 0.2077$). Figure 8 (SoundNet) and Figure 9 (Peekaboom) show the percentage of synset assignment overlap between evocation and the other nine relatedness measures. The overlap with "lesk" and "vector" are significantly higher than that with the other measures ($F(8, 5877) = 34.67$, $p < 0.001$). It suggests that evocation as a semantic relatedness measure may be closer to the definition-based measures than path and content based measures.

For the SoundNet dataset, 34% to 44% of evocation WSD results overlap with that of other measures; for the Peekaboom dataset, the overlap is 25% to 35% (Figure 6). Given that evocation performed similarly in accuracy to most of other measures with relatively low overlap in WSD results, evocation may capture different aspects of semantic relatedness from existing measures.

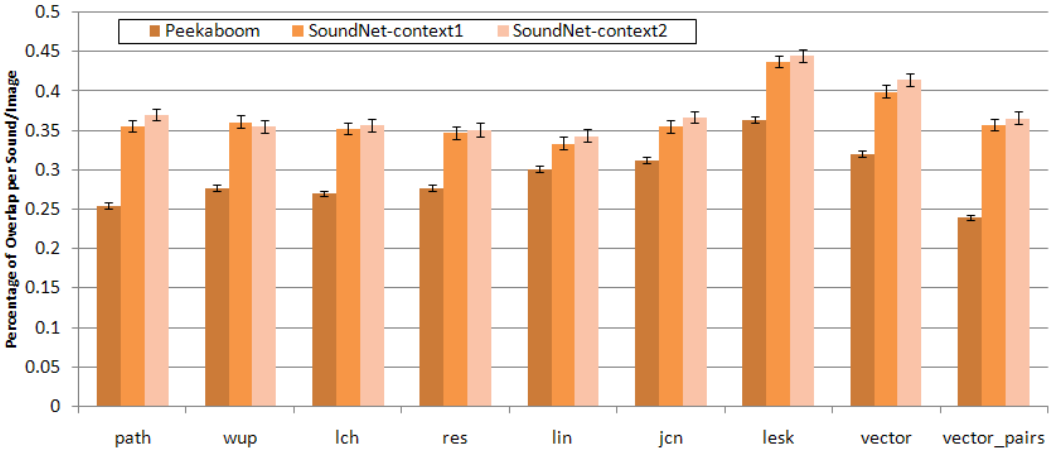

Figure 6. Percentage of WSD results overlap between evocation and various relatedness measures.

6 Conclusion and Future Work

We explored the construction of a sense disambiguated semantic AAC multimodal vocabulary from sound/image label datasets. Two WSD approaches are introduced to assign specific meanings to environmental sound and image labels, and further create concept-sound/image associations. The measure-combined voting algorithm targets the accuracy of WSD and achieves significantly better performance than each relatedness measure individually. Our second approach applies a new relatedness measure, evocation. Evocation achieves similar performance to most of the existing relatedness measures with sound labels. Results suggest that evocation provides different semantic information from current measures.

Future work includes: 1) expanding the evocation dataset and investigating the potential improvement in its WSD accuracy; 2) incorporating the extended evocation dataset into the voting algorithm; 3) exploring additional information such as image and sound similarity to help with WSD.

Acknowledgments

We thank the Kimberley and Frank H. Moss '71 Princeton SEAS Research Fund for supporting our project.

References

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*.

Jordan Boyd-Graber, Christaine Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. *Proceedings of the Thirds International WordNet Conference*.

Jia Deng, Wei Dong, Richard Socher, Li -J. Li, Kai Li and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marti Hearst. 1991. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. *Proc. of the 7th Annual Conference of the University of Waterloo Center for the New OED and Text Research*.

Graeme Hirst and David St. Onge. 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*.

Jay Jiang and David Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings on International Conference on Research in Computational Linguistics*.

Claudia Leacock and Martin Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of SIGDOC'86*.

Dekang Lin. 1997. Using Syntactic Dependency as a Local Context to Resolve Word Sense Ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 64-71.

Lingraphica. http://www.aphasia.com/. 2010.

Xiaojuan Ma, Christiane Fellbaum. and Perry Cook. 2010. SoundNet: Investigating a Language Composed of Environmental Sounds. In *Proc. CHI 2010*.

Xiaojuan Ma, Jordan Boy-Graber, Sonya Nikolova, and Perry Cook. 2009a. Speaking Through Pictures: Images vs. Icons. *Proceedings of ASSETS09*.

Xiaojuan Ma, Sonya Nikolova and Perry Cook. 2009b. W2ANE: When Words Are Not Enough - Online Multimedia Language Assistant for People with Aphasia. *Proceedings of ACM Multimedia 2009*.

Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum. 2009. Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools (in press). *Modelling, Learning and Processing of Text-Technological Data Structures, Springer Studies in Computational Intelligence*.

Adrian Novischi, Muirathnam Srikanth, and Andrew Bennett. 2007. Lcc-wsd: System Description for English Coarse Grained All Words Task at SemEval 2007. *Proceedings of the 4th International Workshop on Semantic Evaluations(SemEval-2007*), pp 223-226.

Siddharth Patwardhan, Satanjeev Benerjee and Ted Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. 2003. *Proceeding of CICLing2003*, pp. 241-257.

Siddharth Patwardhan and Ted Pedersen Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. 2006. *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pp. 1-8

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WorNet::Similarity – Measuring the Relatedness of Concepts. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations*, pp. 38-41.

Ted Pedersen and Varada Kolhatkar. 2009. WordNet::SenseRelate::AllWords - A Broad Coverage Word Sense Tagger that Maximimizes Semantic Relatedness. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations*, pp. 17-20.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Richard Steele, Michael Weinrich, Robert Wertz, Gloria Carlson, and Maria Kleczewska. Computer-based visual communication in aphasia. *Neuropsychologia*. 27(4): pp 409-26. 1989.

Luis von Ahn, Laura Dabbish. 2004. Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, p.319-326.

Luis von Ahn, Ruoran Liu, Manuel Blum. 2006 Peekaboom: a game for locating objects in images. *Proceedings of the SIGCHI conference on Human Factors in computing systems*.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. *Proc. of ACL*, pp 133-138.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting on Association For Computational Linguistics*.