Word Sense Disambiguation with Pictures

University of Arizona Computer Science Technical Report, TR-04-12

Kobus Barnard[1] and Matthew Johnson

Abstract

We introduce using images for word sense disambiguation, either alone, or in conjunction with traditional text based methods. The approach is based on a recently developed method for automatically annotating images by using a statistical model for the joint probability for image regions and words. The model itself is learned from a data base of images with associated text. To use the model for word sense disambiguation, we constrain the predicted words to be possible senses for the word under consideration. When word prediction is constrained to a narrow set of choices (such as possible senses), it can be quite reliable. We report on experiments using the resulting sense probabilities as is, as well as augmenting a state of the art text based word sense disambiguation algorithm. In order to evaluate our approach, we developed a new corpus, ImCor, which consists of a substantive portion of the Corel image data set associated with disambiguated text drawn from the SemCor corpus. Our experiments using this corpus suggest that visual information can be very useful in disambiguating word senses. It also illustrates that associated non-textual information such as image data can help ground language meaning.

[1]Computer Science Department, Gould-Simpson Building, #77, 1040 E. 4th Street, P.O. Box 210077, University of Arizona, Tucson, AZ 85721-0077, kobus@cs.arizona.edu

Figure 1: Five senses of bank, illustrated using using images from the Corel dataset.

1 Introduction

A significant portion of words in natural language have a number of possible meanings (senses), depending on context. This is illustrated in Figure 1 with the arguably overused "bank" example. A priori, the word "bank" has a number of meanings including financial institution and a step or edge as in "snow bank" or "river bank". Words which are spelled the same but have different meanings (polysems) confuse attempts to automatically attach meaning to language. As there are many such ambiguous words in natural language texts, word sense disambiguation — determining the exact sense of words — has been identified as an important component of natural language processing, and has been studied by many researchers leading to a large body of literature [3, 27, 41, 40, 22, 2, 1, 31, 32, 38].

Since the words are spelled the same, resolving what they mean requires a consideration of context. A purely natural language based approach considers words near the one in question. Thus in the bank example, words like "financial" or "money" are strong hints that the financial institution sense is meant. Interestingly, despite much work, and a number of innovative ideas, doing significantly better than choosing the most common sense remains difficult [38].

In this paper we develop a method for using image information to disambiguate the senses of words. We posit that image information can be an orthogonal source of information for distinguishing senses. In the extreme case, disambiguation using nearby text alone is impossible as in the sentence: "He ate his lunch down by the bank." In such cases, alternative sources of information offer attractive possibilities for grounding the word meanings. Even when not essential, non-textual information has the capacity to be helpful. Our method for using associated visual information can be used alone, or in conjunction with text based methods. Naturally, when no images are available, the system must fall back on non-image methods. Incorporation of computer vision into the word sense disambiguation process is a novel approach. As far as we know, all other word sense disambiguation methods use document text and/or additional text carrying domain or document context semantic information.

To use image information we exploit a recently developed method for predicting likely words for images [7, 18, 4]. The method is based on a statistical model for the joint probability distribution of words and image region features. The model is learned from a training set of images with associated text. Additional details are provided below (Section 3).

To use the model for word sense disambiguation, we constrain the predicted words to be from the set of senses for the word under consideration. In general, when word prediction is constrained to a narrow set of choices (such as possible senses), it can be quite reliable. We report on experiments using the resulting sense probabilities as is, as well as augmenting a state of the art text based word sense disambiguation algorithm.

In order to evaluate our approach, it was necessary to develop a new corpus, ImCor, which consists of a substantive portion of the Corel image data base associated with disambiguated text drawn from the SemCor corpus. (We have made ImCor available for research purposes [26]). Our experiments using this corpus suggest that visual information can be very useful for disambiguating word senses.

2 Disambiguating Words using Textual Content

Research into automatic methods for disambiguating word senses has resulted in a variety of ways of using the surrounding text, or the "textual context", to infer word sense. Disambiguating sense is a semantic problem, and the underlying assumption is that the word to be disambiguated is semantically linked to the nearby words, as text tends to be semantically

coherent. Co-occurrence statistics will reflect semantic linking, and thus researchers have developed methods based on statistical models for senses [12]. A large number of other methods attempt to quantify this linking using known word semantics. For example, word classes, as defined by a Thesaurus, can be integrated into a combined weight of indicators in the textual context [39]. Going further, most word sense disambiguation algorithms use a semantic network such as WordNet [33]. WordNet is a machine-readable dictionary covering a large proportion of the English language (152,059 words) organized into 115,424 sets of synonyms (synsets). It provides relationships between the sets, the most commonly used one being the hypernym ("is a") relationship. The graph created by hypernym relationships forms a tree in which every node is a hypernym of its children. The path connecting two words can be used to define semantic distances, which has been used in word sense disambiguation algorithms [1, 28, 16, 32].

Usage statistics are also helpful for word sense disambiguation. In WordNet, the "sense number" roughly corresponds to decreasing common usage frequency (the first WordNet sense is most common). Going further, researchers have exploited the SemCor sense-attributed corpus [32, 37, 23, 34]. SemCor, short for the WordNet Semantic Concordance [21], consists of 25% of the Brown corpus [20] files which have been fully tagged with part-of-speech and is sense disambiguated.

A number of word sense disambiguation methods were compared on the same data at the Senseval2 [19] contest for word sense disambiguation methods. The first place algorithm was *SDW* [32] which uses information from both WordNet and SemCor. Thus we use that algorithm in this work. Overall, however, the results from Senseval2 indicate that word sense disambiguation is still very much an open problem [38].

There has been some work done incorporating multiple alternative knowledge sources to help disambiguate words in context. In [15], "world knowledge" derived from alternative synset contexts obtained through WordNet was used to supplement a learning algorithm and showed marked improvement over the unaided version. Another interesting example is found in [35], where, for every word being disambiguated, a feature set is formed based on multiple sources, including the part of speech of neighboring words, morphological form, the unordered set of neighboring words, local collocations and verb-object syntactic relation. During training, disambiguated sentences were mined for training data, such that during testing, a feature set obtained for a word can be compared against many training sets, with the intent that the degree of similarity is directly proportional to the probability that the sense of the word in a training set is the correct sense for the test word. While this system

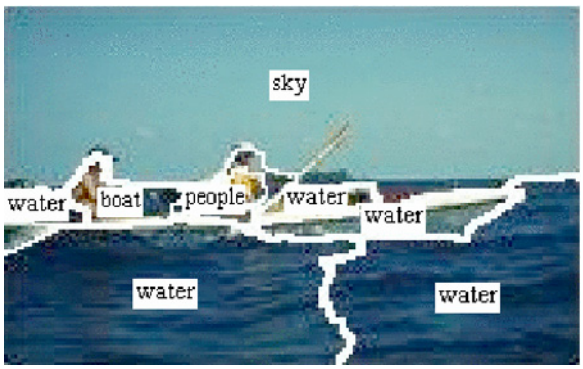

Figure 2: Illustration of region labeling. Each region is labeled with the maximally probable word, but has a distribution over the entire vocabulary.

relied on the surrounding text to obtain the feature set during testing, training data could have potentially come from a number of different sources. This and other similar efforts [29, 9] indicate that intelligent and efficient integration of multiple knowledge sources can result in enhanced performance of a variety of algorithms dealing with textual analysis in general, and word sense disambiguation in particular.

3 Predicting Words from Images

To integrate image information with text data we exploit recent work on linking images and words [7, 18, 4]. The general approach is to build statistical models for the co-occurrence of image regions and words. A key assumption is that words are linked to images via regions. These models can be used to predict words for image regions (region-labeling) as well as entire images (auto- annotation). Region labeling is illustrated in Figure 2. To label regions, probabilistic inference using these models provides a posterior probability distribution over the vocabulary for each region, and we label the region with the one which has maximal probability. We fit the models using large image data sets with associated text. Critically, we do not require that words in the training data be identified as belonging to particular image regions, as such data is rare.

These models owe much to previous work in the text domain [24] and statistical machine translation [10, 11, 30]. A number of additional methods for linking image features to words have been recently proposed [13, 25], and these could also be considered for word sense disambiguation. For this work we use one of the models from [4]. In particular, we use the dependent model, D-2, with linear topology. We do not use the hierarchical clustering

version as that is better suited characterizing a known data set, and less suited for predicting words for novel images.

We first segment images into regions which have coherent color and texture. This simplification is essentially a data reduction step allowing semantic analysis to be done on groups of pixels. In this work we use a modified version of Normalized Cuts [36] for segmentation. For each image region we compute a feature vector representing color, texture, size, position, and shape [4]. A region, together with its feature vector, will be referred to as a "blob" [14].

Our language model is the commonly used "bag of words" where word order is not used. Various pre-processing strategies can be used to increase the likelihood that words can be connected to visual attributes of image regions [5]. In this work we use a subset of the SemCor [21] vocabulary as described further below (Section 6).

To statistically link blobs with words we assume that there are hidden factors (concepts) which are each responsible for generating *both* the words and blobs associated with that factor. This binding of their generation leads to the capacity to link words and blobs. We further assume that the observations (image and associated text) are generated from multiple draws from the hidden factors or nodes. Without modeling image generation as being compositional — the same model of a tiger region can be used for all images with such regions — our models would need to model all possible combinations of entities. Furthermore, modeling specific combinations in the training data would lead to poor generalization on novel combinations in new images. We model the joint probability of a particular blob, b, and a word w, as

$$P(w,b) = \sum_{l} P(w|l)P(b|l)P(l) \tag{1}$$

where l indexes over the concepts, $P(l)$ is the concept prior, $P(w|l)$ is a frequency table, and $P(b|l)$ is a Gaussian distribution over features. We further assume a diagonal covariance matrix because fitting a full covariance is generally too difficult for a large number of features.

To go from the blob oriented expression (1) to ones for an entire image, we assume that the observed blobs, B, yield a posterior probability, $P(l|B)$, which is proportional to the sum of $P(l|b)$. Words are then generated conditioned on the blobs from:

$$P(w|B) \propto \sum_{l} P(w|l)P(l|B) \tag{2}$$

where by assumption

$$P(l|B) \propto \sum_b P(l|b) \tag{3}$$

and Bayes rule is used to compute $P(l|b) \propto P(b|l)P(l)$.

Some manipulation [6] shows that this is equivalent to assuming that the word posterior for the image is proportional to the sum of the word posteriors for the regions:

$$P(w|B) \propto \sum_b^N P(w|b) \tag{4}$$

We limit the sum over blobs to the largest N blobs (typically N is eight). While training, we also normalize the contributions of blobs and words to mitigate the effects of differing numbers of blobs and words in the various training images. The probability of the observed data, $W \cup B$, given the model, is thus:

$$P(W \cup B) = \prod_{b \in B} \left(\sum_l P(b|l)P(l) \right)^{\frac{max(N_b)}{N_b}} \prod_{w \in W} \left(\sum_l P(w|l)P(l|B) \right)^{\frac{max(N_w)}{N_w}} \tag{5}$$

where $max(N_b)$ (similarly $max(N_w)$) is the maximum number of blobs (words) for any training set image, N_b (similarly N_w)is the number of blobs (words) for the particular image, and $P(l|B)$ is computed from (3).

Since we do not know which concept is responsible for which observed blobs and words in the training data, determining the maximum likelihood values for the model parameters ($P(w|l)$, $P(b|l)$, and $P(l)$) is not tractable. We thus estimate values for the parameters using expectation maximization (EM) [17], treating the hidden factors responsible for the blobs and words as missing data. In the EM computation we alternate between the following two steps:

Expectation(E) Estimate the expectations of the unobserved data from the previous estimates of the parameters. In particular, for each blob and word in the training data, we estimate the probability that it comes from each of the hidden factors (concepts).

Maximization(M) Estimate the model parameters ($P(w|l)$, $P(b|l)$, and $P(l)$) by maximizing the expected log-likelihood computed during the E-step.

4 Using Word Prediction for Sense Disambiguation

In the context of word sense disambiguation, our vocabulary is assumed to be sense disambiguated. Formally, we use an extended vocabulary S, which contains the senses of the words in a vocabulary W. Notationally, if the word *bank* $\in W$ then {*bank_1, bank_2,* ...} $\in S$. Thus, every sense $s \in S$ is the sense of only one word $w \in W$. Once a model has been trained on S, we can use the annotation process to compute $P(s|B)$. Different than annotation, word sense disambiguation has the additional characteristic that we are trying to *only* distinguish between the senses, s, for a particular word, w, rather than produce a number of good choices from all of S, which is clearly more difficult.

Given a word, w, under consideration, we assume that senses for all other words should not be predicted. Operationally we simply take the posterior probability over all the senses in our vocabulary, and set those not corresponding to w to zero. We then rescale the posterior so that it sums to one. This computation yields the probability of a word sense, s, given w, and the visual context, B, which we denote as $P(s|w, B)$.

Being able to constrain the word prediction domain makes the process more accurate and thus more useful. Linking words — which carry semantics — to images, is a difficult task, and limiting the choices the system has to make is generally helpful. For example, as shown in Figure 3, if we know the words in a caption, and thus can constrain region labeling to those words, then labeling performance increases substantively.

4.1 Combining Word Prediction and Traditional Word Sense Disambiguation

The quantity $P(s|w, B)$ can be used as is for word sense disambiguation, and we provide results for this strategy. It is also natural to combine it with text based methods, as it seems to provide an orthogonal source of information. Here we assume that a text based method can provide a second estimate of the probability $P(s|w, W)$ for the sense, s, for w, based on the observed words, W (the senses are not known a priori). We discuss our choice of $P(s|w, W)$ below (Section 4.2).

In preliminary work [8] we proposed that these two estimates were relatively independent, giving the following simple expression for combining them:

$$P(s|w, B, W) \propto P(s|w, B)P(s|w, W). \tag{6}$$

We provide results for this approach for combining the two kinds of information. However, preliminary work suggested that these two quantities are less independent than we originally

Figure 3: Illustration of the improvement in region-labeling due to being able to restrict the predicted words to those known to be in the caption. The task here was to find tiger regions in the image data base. (This task is not precisely analogous to word sense disambiguation). The best tiger regions found are shown. The top group was determined only using image data, whereas the bottom group was found using both image data and the five keywords, one of which was tiger.

assumed. In particular, both approaches typically embody the likelihood of common senses. Thus we propose here a alternative heuristic for combining the two quantities:

$$P(s|w, B, W, \propto max(\mathbf{0}, P(s|w, B) + P(s|w, W) - P(s|w)) \tag{7}$$

Intuitively, we assume that both $P(s|w, B)$ and $P(s|w, W)$ embody $P(s|w)$ as a major component, making them not particularly independent. Thus we approximate them as each providing two sources for the sense, $P(s|w)$ and the difference, either $P(s|w, B)$-$P(s|w)$ or $P(s|w, W)$-$P(s|w)$. If we wish to treat these equally important but alternative hypothesis, the we need to subtract off the doubly counted $P(s|w)$. We then treat any negatives as zero, and renormalize.

4.2 Traditional Word Sense Disambiguation

The probability $P(s|w, W)$ in (6) is assumed to come from a traditional text based word sense disambiguation algorithm. In preliminary [5] work we used a naïve algorithm based on distances computed using WordNet [33] among words forming the context and words related to proposed senses. This algorithm produced a score instead of a true probability, and was calculated using work from [5], which itself was drawn from [2, 31].

We found that the performance of this algorithm was poor, leading to the question of whether our originally preliminary results using image information would be overshadowed by a more sophisticated text based WSD algorithm. Thus we looked at the results of Senseval 2, a competition held between various WSD algorithms in 2001 [19]. Interestingly, the simple algorithm of choosing the most common meaning according to WordNet outperforms all algorithms but the five best [38]. Accordingly, we chose the best of these, SMUaw, based on earlier work by the same team [32].

This algorithm makes use of both WordNet and the semantically tagged corpus SemCor. The Mihalcea iterative approach consists of 10 algorithms which act as filters on the input data. Each algorithm in the pipeline uses a different heuristic to disambiguate a word and moves it from the set of ambiguous words, SAW, into the set of disambiguated words SDW (a process referred to here as "marking"). These procedures range from removing proper nouns and monosemous words to connecting words which have certain semantic distances. The original algorithm gave words a definite sense based on computational heuristics associated with each filter. As the approach described above requires softer output, we modified the algorithm so that information that would otherwise be lost at each filtration step contributes to the score of the sense. Each of the procedures was altered in

the following ways (original procedure in italics):

1. *Mark all proper nouns with a WordNet sense of 1.* No change.

2. *Mark all words with one sense as having that sense.* No change.

3. *Examine the usage of the word and its neighbors in SemCor. If the count of one sense is a certain threshold above the remainder of the senses, remove and mark the word with that highest sense.* Instead of dropping the counts for the senses which don't make the threshold, we normalize the array of sense frequency counts, and if one of the senses scores above .75, we mark the word with that sense but retain the distribution data.

4. *For every sense of every noun in SAW, find all nouns which occur within a window of 10 words from that sense usage and compile them together to create "noun contexts" for each. The sense whose noun-context has the greatest overlap with the textual context of the word (defined as the cardinality of the intersection of the noun context with the words in the document), if it is greater than the next highest sense by a threshold, should be marked.* Again, instead of throwing away the overlap data we instead store the entire array of cardinalities, normalize, and mark the word if the highest is above a threshold, in this case .5.

5. *For every word in SAW, if one of its senses is within a semantic distance of 0 (same synset) from a word in SDW, mark it with that sense.* Instead of throwing away data, a count for each word which was a semantic distance of 0 from a given sense was tabulated, and then these counts were normalized and used as substitute probabilities. Again, we mark a word if it is above the likelihood threshold of .5.

6. *Same as above, but was performed within SAW (i.e. two words in SAW which have senses with a semantic distance of 0 are marked with that sense).* Change is same as above.

7. *Same as fifth procedure, but with a distance of 1 (hypernym/holonym relationship).* Change is same as in 5.

8. *Same as sixth procedure, but with a distance of 1.* Change is same as in 6.

All those words not disambiguated by the process were given a default distribution which favored the most common sense. The end result is that the last 6 of the 8 procedures now produce softer distributions which are more useful as part of (6).

5 ImCor

In preliminary work [8] we used the Corel image data set which has four or five keywords per image. We labeled the senses of these keywords for 16,000 images, and identified a subset of 1,800 images with potential sense problems using heuristics to bias the set towards ambiguous keywords. Nevertheless, the amount of ambiguity across the dataset was not sufficient to provide for realistic testing. For example, while a word such as *head* is usually ambiguous, in the Corel dataset it overwhelmingly tends to be used in one way.

Given the inadequacy of this and all other existing image datasets for this kind of work, we created a new research corpus named ImCor. This corpus links the images from the Corel dataset with the sense disambiguated SemCor corpus to provide a new corpus which links images with semantically tagged text. (We have made ImCor available for research purposes [26]).

5.1 Building Imcor

The task at hand was to link images with text passages from SemCor to provide images linked to text more along the lines as one would find in a newspaper or magazine setting. The Corel keywords were used to determine an initial set of 30 candidate images for each of the SemCor articles. We developed a program to facilitate human selection of text for the image candidates (Figure 4). The rater would then be asked to first choose whether the image was appropriate for the text, and, if so, the rater further selected the text passage within the article that was most appropriate.

The magnitude of the task meant that two raters were required to build the corpus. We divided the data between them so that there was an overlap of one article in six. We calculated consistency as the quotient of matching choices over total choices. This enabled us to verify that their results were relatively consistent at 0.74 by this measure.

The end result was a list of documents with associated images marked either as "inappropriate", "no text" (for images which illustrated the article as a whole but no specific part), or "appropriate" with paragraph text from the article. We then gathered the appropriate images into a single corpus with the disambiguated text becoming the captions. We incorporated images which were associated with the article as a whole but no specific text segment by assigning them a random sampling of words from the article with a selection factor of $\frac{1}{P}$, where P is the number of paragraphs in the article. The end result was a corpus of 1633 image/text pairings, in which 86.83% were tagged with specific paragraph

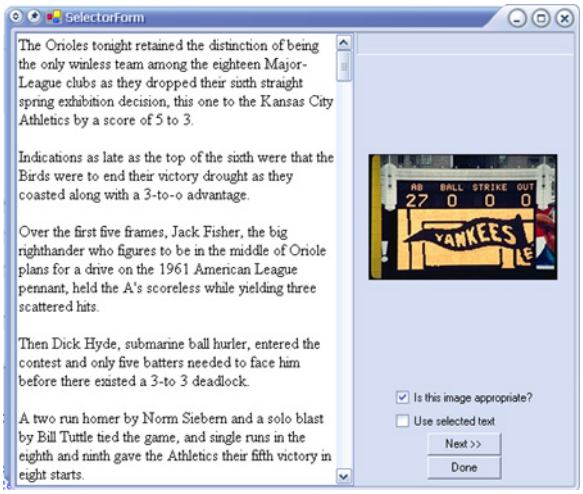

Figure 4: A screen-shot of the program used to select text passages from SemCor semantically linked to images. The rater reads the article on the left and then looks at a picture. If that picture is appropriate, they click the box in the lower right. At that point the rater has the opportunity to select any text which is appropriate, indicate that they have done so, and then move on to the next image.

text and 13.17% with random samplings from documents.

5.2 Expanding ImCor

While a carefully sense disambiguated annotated corpus of 1633 images goes far beyond what is available, it is still relatively small for our purposes. Therefore we exploit the fact that there is much semantic redundancy in the Corel image data (e.g., there are at least 50 images of planes/jets with very similar keywords), to find additional images which are appropriate for the captions found in the first step. Any image which was not already used that shared two or more keywords with an image which had been paired with SemCor text was added to the corpus with that text. This operation produced a new version of the corpus with 20,153 image/text pairings.

Minimum sense count in training data vocabulary	Text only	Image only	Combined (using (6))	Combined (using (7))
20	0.009 (0.004)	0.210 (0.002)	0.178 (0.004)	0.207 (0.002)
50	0.027 (0.002)	0.208 (0.002)	0.200 (0.002)	0.200 (0.002)

Table 1: Results of word sense disambiguation experiments with two values for the minimum number of times that a word sense needs to be used in the training data in order to be considered part of the vocabulary. The numbers tabulated are the fraction of times the sense was correctly chosen *relative* to the performance of choosing the most common sense in the training data. This "empirical" performance is 58.5% for the first row, and 57% for the second row. Thus the absolute performance is about 60% for the text based method, and 80% when images are used. All numbers are positive which means that all algorithms exhibit non-trivial performance. The results are the average of 8 different breakdowns of training and testing, and, in the image case, we further averaged over two different initialization strategies which gave similar results. The numbers in parenthesis are estimates of the error due to test/training sampling and initialization.

6 Experiments

In preparation for our experiments, we produced eight different breakdowns of our corpus into training and testing sets (90% training, 10% testing). We then removed stop words from the corpus to reduce computation. For each training set we eliminated all word senses which occurred less than 20 times (50 times in a second experiment). Typical vocabulary sizes were 4400 words, of which about 1600 were ambiguous (3500 / 600 for the second experiment).

We then applied the text based algorithm detailed above (Section 4.2) to the test data to get an estimate of the probability of each sense for each ambiguous word. Next we trained the word prediction model (Section 3) on the combined image sense data. We used the features for the 8 (all if there were fewer) largest image regions. We used exactly the same features as in previous work [4] to represent region size, location, rough shape, color, color variance, and texture. We then applied the model to the test data to predict senses according to (4) restricted to the senses for each word under consideration as described in Section 4. We also combined the image and text results as described in Section 4.1.

The results are shown in Table 1. As in previous work [4] we find it useful to focus on the amount by which the performance exceeds what is possible using the empirical

distribution of the training set, which was roughly 60%. This controls somewhat for subset difficulty, and makes it easy to identify non-trivial performance since it results in positive values. Exceeding the empirical distribution is even more difficult than the simple "most common sense" method, which has been found to be surprising effective [38], as the empirical distribution gives the common sense for the particular corpus being investigated.

We compute performance using *only* words which are ambiguous in a particular image. By construction, if a test image only has one sense of a word, our measurement process would score all algorithms as giving the correct sense, which would inflate performance figures, and dilute the effects that we are investigating.

We found that the text based algorithm barely exceeded the performance of the empirical distribution, suggesting that in this corpus very little text information is available for the kind of processing used by that algorithm. By contrast, our results using image information are very promising, increasing performance over the empirical by roughly 20% yielding around 80% absolute performance. This confirms our main thrust — that image information can help disambiguate senses, and that this ability can go beyond that easily available using text alone.

Unfortunately, combining the two methods to achieve even better performance proved difficult. Using both proposed approaches for doing so gave results that were a little worse than using only image information, although the second method was almost on par. Although disappointing, this result is not overly surprising given that the text results are not that different from that for the empirical distribution. This is because our word prediction model can (and does) implicitly encode some text occurrence statistics, and thus of the information available in the empirical distribution is already likely being used, rendering the text algorithms efforts somewhat redundant.

7 Conclusion

The main conclusion from this work is that visual information can help disambiguate senses, and thus help ground language meaning. In fact, we found that our method for using visual information performs substantively better than a state of the art text based methods on our corpus. Of course, alternative information such as image data is not always available, and here we must fall back on text based methods.

A second important contribution of this work is the development of a new corpus, ImCor, which links images with sense disambiguated text. As linking images with text is an important emerging research area, this data set will help researchers in this area evaluate the extent to which various approaches capture the semantics of the visual data.

We were unable to further improve word sense disambiguation performance by combining image and text based methods. We suspect that this is because the image based word prediction is able to capture some of the statistics of the sense occurrences, and it is difficult to find a word sense disambiguation method which performs beyond this, at least on a restricted corpus like ours. We are thus compelled to investigate a tighter integration of the two processes. In particular, we would like to develop a text based word sense disambiguation component which focuses on information which is orthogonal to that embodied in our statistical models for the co-occurrence data of words and blobs in images.

References

[1] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density, 1996.

[2] Eneko Agirre and German Rigau. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing*, 1995.

[3] Y. Bar-Hillel. Automatic translation of languages. In Donald Booth and R.E. Meagher, editors, *Advances in Computers*, New York, 1960. Academic Press.

[4] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[5] Kobus Barnard, Pinar Duygulu, and David Forsyth. Clustering art. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 434–441, 2001.

[6] Kobus Barnard, Pinar Duygulu, and David Forsyth. Exploiting text and image feature co-occurrence statistics in large datasets. In Remco Veltkamp, editor, *Trends and Advances in Content-Based Image and Video Retrieval*. Springer, 2004 (to appear).

[7] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *Proceedings of the International Conference on Computer Vision*, volume II, pages 408–415, 2001.

[8] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.

[9] John Bear, John Dowding, and Elizabeth Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Meeting of the Association for Computational Linguistics*, pages 56–63, 1992.

[10] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fedrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85, 1990.

[11] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of machine translation: parameter estimation. *Computational Linguistics*, 19(10):263–311, 1993.

[12] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense disambiguation using statistical methods. In *Meeting of the Association for Computational Linguistics*, pages 264–270, 1991.

[13] Peter Carbonetto, Nando de Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. In *European Conference on Computer Vision*, 2004.

[14] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer, 1999.

[15] Massimiliano Ciaramita, Thomas Hofmann, and Mark Johnson. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003.

[16] C. de Loupy and M. El-Bze. Using few clues can compensate the small amount of resources available for wsd. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 219–223, 2000.

[17] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[18] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of The Seventh European Conference on Computer Vision*, volume IV, pages 97–112, 2002.

[19] Phil Edmonds and Adam Kilgarriff, editors. *Journal of Natural Language Engineering*, volume 9, January 2003.

[20] W. Nelson Francis and Henry Kučera. *Frequency Analysis of English Usage. Lexicon and Grammar*. Houghton Mifflin, 1981.

[21] Miller G., Leacock C., Randee T., and Bunker R. A semantic concordance. In *Procedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, 1993.

[22] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. In *DARPA Workshop on Speech and Natural Language*, pages 233–237, 1992.

[23] G. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL '98 Workship on Usage of WordNet for NLP*, 1998.

[24] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. Technical report, Massachusetts Institute of Technology, 1998.

[25] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, 2003.

[26] Matthew Johnson and Kobus Barnard. *ImCor: A linking of SemCor sense disambiguated text to corel image data.* http://kobus.ca/research/data/index.html, 2004.

[27] A. Kaplan. An experimental study of ambiguity in context, 1950.

[28] Xiaobin Li, Stan Szpakowicz, and Stan Matwin. A wordnet-based algorithm for word sense disambiguation. In *IJCAI*, pages 1368–1374, 1995.

[29] Susan Weber McRoy. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30, 1992.

[30] Dan Melamed. *Empirical methods for exploiting parallel texts.* MIT Press, Cambridge, Massachusetts, 2001.

[31] Rada Mihalcea and Dan Moldovan. Word sense disambiguation based on semantic density. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[32] Rada Mihalcea and Dan Moldovan. An iterative approach to word sense disambiguation. In *Proceedings of Flairs 2000*, 2000.

[33] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[34] Palomar M. Montoyo A. and Rigau G. Wordnet enrichment with classification systems. In *Proceedings of NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, 2001.

[35] Hwee Tou Ng and Hian Beng Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Arivind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 40–47, San Francisco, 1996. Morgan Kaufmann Publishers.

[36] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[37] Jiri Stetina, Sadao Kurohashi, and Makoto Nagao. General word sense disambiguation method based on A full sentential context. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 1–8. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[38] Jonathan Traupman and Robert Wilensky. Experiments in improving unsupervised word sense disambiguation. Technical report, University of California at Berkeley, 2003.

[39] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July 1992.

[40] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Conference on Applied Natural Language Processing*. ACL, 1995.

[41] V. Yngve. Syntax and the problem of multiple meaning. In W. Locke and D. Booth, editors, *Machine Translation of Languages*, New York, 1955. Wiley.