Available online at www.sciencedirect.com

SCIENCE DIRECT®

Artificial Intelligence 167 (2005) 13–30

Artificial Intelligence

www.elsevier.com/locate/artint

Word sense disambiguation with pictures

Kobus Barnard [a,*], Matthew Johnson [b]

[a] *Computer Science Department, University of Arizona, USA*
[b] *Department of Engineering, University of Cambridge, USA*

Received 25 July 2004; received in revised form 21 February 2005; accepted 14 April 2005

Available online 11 August 2005

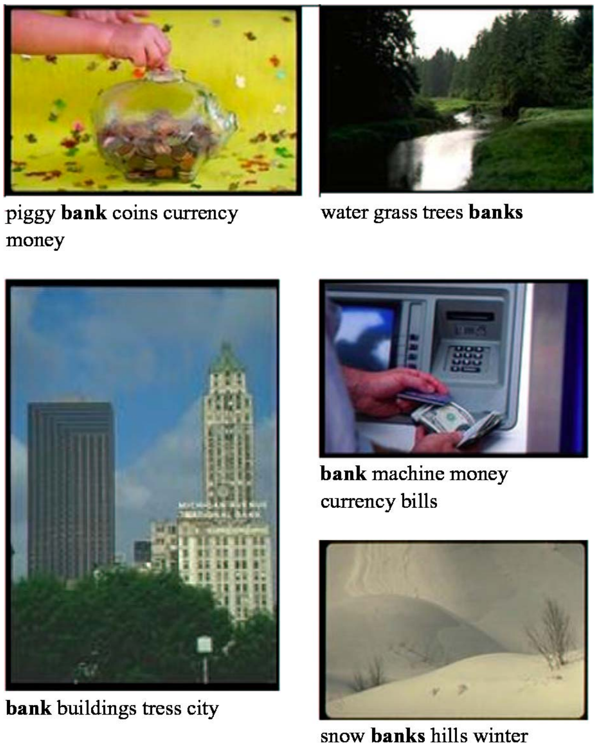
Abstract

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.

Keywords: Word sense disambiguation; Image auto-annotation; Region labeling; Statistical models

[*] Corresponding author.
E-mail address: kobus@cs.arizona.edu (K. Barnard).

0004-3702/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.artint.2005.04.009



piggy **bank** coins currency money

water grass trees **banks**

bank machine money currency bills

bank buildings tress city

snow **banks** hills winter

Fig. 1. Five senses of bank, illustrated 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 Fig. 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 (polysemes) 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 [2–4,27,32,40,41,47,49,50].

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 [47].

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. However, we acknowledge related work using WordNet [42] to propagate sense (and thus semantic) information between feature based classes in the context of multimedia information systems [12,13].

To use image information we exploit a recently developed method for predicting likely words for images [5,9,22]. 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 two state of the art text based word sense disambiguation algorithms.

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 [31].) Our experiments using this corpus suggest that visual information can be very useful for disambiguating word senses.

This work suggests approaches to exploiting multiple data modes to increase our ability to automatically search and browse multi-media information. For example, text data on the web is often augmented with image data. Searches based on text currently do not make use of that information, even though in many cases it would be helpful. While computational methods for effectively understanding arbitrary visual data are still a long way off, using visual features to improve the rankings of query results may not require such a full understanding. For example, if text data can be better sense disambiguated by using image data, then an unambiguous query can be better executed against this data.

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 [16]. 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 [48]. Going further, most word sense disambiguation algorithms use a semantic network such as WordNet [42]. 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 [2,20,35,41].

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 that which it considers to be most commonly used). Going further, researchers have exploited the SemCor sense-attributed corpus [28,41,43,46]. SemCor, short for the WordNet Semantic Concordance [26], consists of 25% of the Brown corpus [25] files which have been fully tagged with part-of-speech and is sense disambiguated.

A number of word sense disambiguation methods have been compared at the three Senseval conferences [1,23,33]. Based on the results from the second Senseval we chose to implement an algorithm based on iterative word sense disambiguation, SMUaw [41]. We were also intrigued by the fact that choosing the most common sense according to WordNet evaluates higher than many of the algorithms currently in use [47]. Thus we also implemented an algorithm which provides a usage distribution over the senses to provide additional evaluation of our algorithm [36].

There has been some work done incorporating multiple alternative knowledge sources to help disambiguate words in context. In [19], "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 [44], 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 features, so that during testing, a feature set obtained for a word can be compared against many training sets. The proposal is that the similarity so found 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 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 [11,37] 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 [5,9,22]. 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

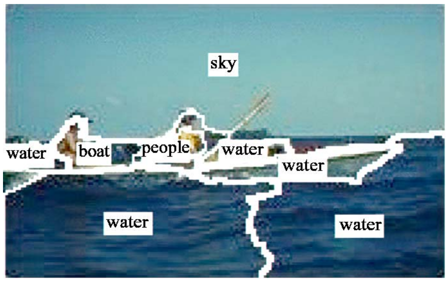


Fig. 2. Illustration of region labeling. Each region is labeled with the maximally probable word, but has a distribution over the entire vocabulary. In the word sense disambiguation task we combine the probability distributions over the regions to provide an "annotation" relevant to the entire an image. We emphasize region based approaches here because we believe that good image annotation requires reasoning about image components.

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 Fig. 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 [29] and statistical machine translation [14,15,38]. A number of additional methods for linking image features to words have been recently proposed [17,24,30,34], and these could also be considered for word sense disambiguation. For this work we use one of the models from [5]. In particular, we use the dependent model, D-2, with linear topology. We do not use the hierarchical clustering version as it 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 [45] for segmentation. For each image region we compute a feature vector representing color, texture, size, position, shape [5], and color context [8]. More specifically,

- Size is represented by the portion of the image covered by the region.
- Position is represented using the coordinates of the region center of mass normalized by the image dimensions.
- Color is represented using the average and standard deviation of $(\mathrm{r} = \mathrm{R}/(\mathrm{R}+\mathrm{G}+\mathrm{B})$, $\mathrm{g} = \mathrm{G}/(\mathrm{R}+\mathrm{G}+\mathrm{B})$, $\mathrm{S} = (\mathrm{R}+\mathrm{G}+\mathrm{B}))$ over the region. We use this color space instead of RGB to reduce correlation among the three bands.
- Texture is represented using the average and variance of 16 filter responses. We use 4 difference of Gaussian filters with different sigmas, and 12 oriented filters, aligned in 30 degree increments. See [45] for additional details and references on this approach to texture.

- Shape is represented by the ratio of the area to the perimeter squared, the moment of inertia (about the center of mass), and the ratio of the region area to that of its convex hull.
- Color context is represented by four colors each one representing the color of adjacent regions, restricted to four 90 degree wedges [8].

A region, together with its feature vector, will be referred to as a "blob" [18].

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 [6]. In this work we use a subset of the SemCor [26] 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—region models can be used in arbitrary configuration to handle images with known regions but in different arrangements—we would need to model all possible combinations of entities. For example, we would have to model tigers on grass, tigers in water, tigers on sand, and so on. Clearly, one tiger model should be reused when possible.

We model the joint probability of a particular blob, b, and a word w, as

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

where l indexes over the concepts, $P(l)$ is the concept prior, $P(w \mid l)$ is a frequency table, and $P(b \mid l)$ is a Gaussian distribution over features. We further assume a diagonal covariance matrix (independent features) because fitting a full covariance is generally too difficult for a large number of features. This independence assumption is less troublesome because we only require conditional independence, given the concept. Intuitively, each concept generates some image regions according to the particular Gaussian distribution for that concept. Similarly, it generates one ore more words for the image according to a learned table of probabilities.

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

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

where by assumption

$$P(l \mid B) \propto \sum_{b} P(l \mid b) \tag{3}$$

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

Some manipulation [7] 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:

$$P(w \mid B) \propto \sum_{b}^{N} P(w \mid b). \tag{4}$$

We limit the sum over blobs to the largest N blobs (in this work N is sixteen). 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:

$$\begin{aligned} P(W \cup B) = {} & \prod_{b \in B} \left(\sum_{l} P(b \mid l) P(l) \right)^{\frac{\max(N_b)}{N_b}} \\ & \times \prod_{w \in W} \left(\sum_{l} P(w \mid l) P(l|B) \right)^{\frac{\max(N_w)}{N_w}} \end{aligned} \tag{5}$$

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 $P(l \mid B)$ 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 ($P(w \mid l)$, $P(b \mid l)$, and $P(l)$) is not tractable. We thus estimate values for the parameters using expectation maximization (EM) [21], treating the hidden factors (concepts) 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 ($P(w \mid l)$, $P(b \mid l)$, and $P(l)$) by maximizing the expected log-likelihood computed during the E-step.

The model is not particularly sensitive to the number of concepts, and we did not attempt to optimize the number of concepts for this work. In previous studies [5,6,9] we found that 500 concepts has adequate for five to ten thousand images. In this work we used 1000 concepts for the experiments with training sets of the order of 18,000 images, and 100 concepts for the experiment with training sets of the order of 1500 images.

The model generalizes well because it learns about image components. These components can occur in different configurations and still be recognized. For example, it is possible to learn about “sky” regions in images of tigers, and then predict “sky” in elephant images. Of course, predicting the word elephant requires having elephants in the training set.

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*, $\ldots\} \in S$. 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 $P(s \mid B)$. 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 $P(s \mid w, B)$.

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 Fig. 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 $P(s \mid w, B)$ 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 $P(s \mid w, W)$ for the sense, s, for w, based on the observed words, W (the senses are not known a priori). We discuss our choice of $P(s \mid w, W)$ below (Section 4.2).

We assume that these two estimates are relatively independent, which gives the following simple expression for combining them:

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

While the two estimates are likely to have some degree of mutual information, the results below suggest that there is enough independence to be useful. We have considered the possibility that both estimates might embody the empirical sense distribution, and that compensating for this may provide a better strategy, but our most robust results have been with the simple independence assumption above.

4.2. Traditional word sense disambiguation

The probability $P(s \mid w, W)$ in (6) is assumed to come from a traditional text based word sense disambiguation algorithm. In preliminary [6] work we used a naïve algorithm based on distances computed using WordNet [42] among words forming the context and

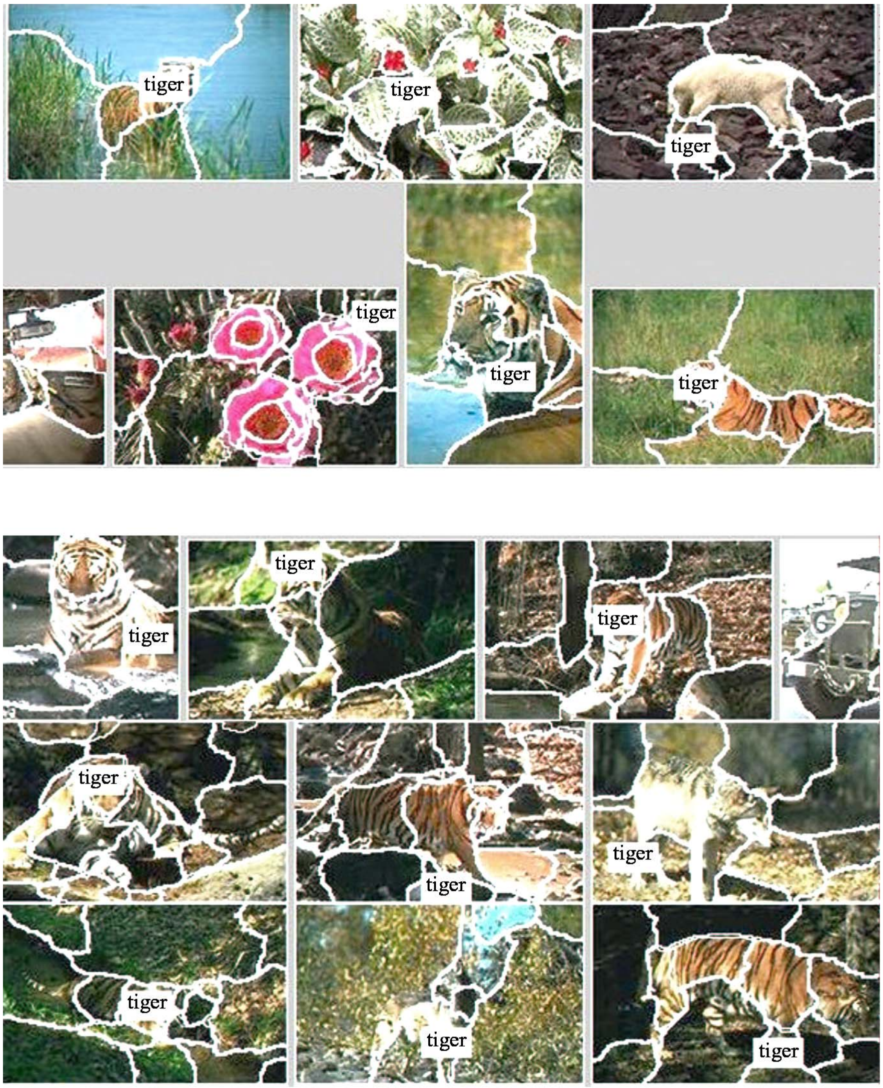


Fig. 3. Illustration of the improvement in region-labeling due to being able to restrict the predicted words to those known to be in the caption. The task here was to find tiger regions in the image data base. The best tiger regions found are shown. The top group was determined only using image data, whereas the bottom group was found using both image data and the five keywords, one of which was tiger. We emphasize that this task is not precisely analogous to word sense disambiguation. The key point is that our difficult prediction problem becomes easier when we can constrain our predictions to a small number of choices.

words related to proposed senses. This algorithm produced a score instead of a true probability, and was calculated using work from [6], which itself was drawn from [3,40].

We found that the performance of this algorithm was poor, leading to the question of whether our original results using image information would be overshadowed by a more sophisticated text based WSD algorithm. Thus we implemented the two methods mentioned above. We describe the first one in detail next. The second algorithm proved to be less interesting in our domain, as it is an attempt to capture usage statistics, which the image based algorithm has access to in training. Thus our need for some independence between the two sources of information breaks down, and the results were not very good.

4.2.1. Iterative word sense disambiguation

The SMUaw algorithm (and a recent derivative, SenseLearner [39]), has been shown to perform very well [1,23]. As such, we based our main text-based algorithms on the technique of iterative word sense disambiguation presented in [41].

This method makes use of both WordNet and the semantically tagged corpus SemCor, and 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 do not make the threshold, we normalize the array of sense frequency counts, and if one of the senses scores above 0.75, 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 0.5.
(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 0.5.
(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 previous work [10] 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.

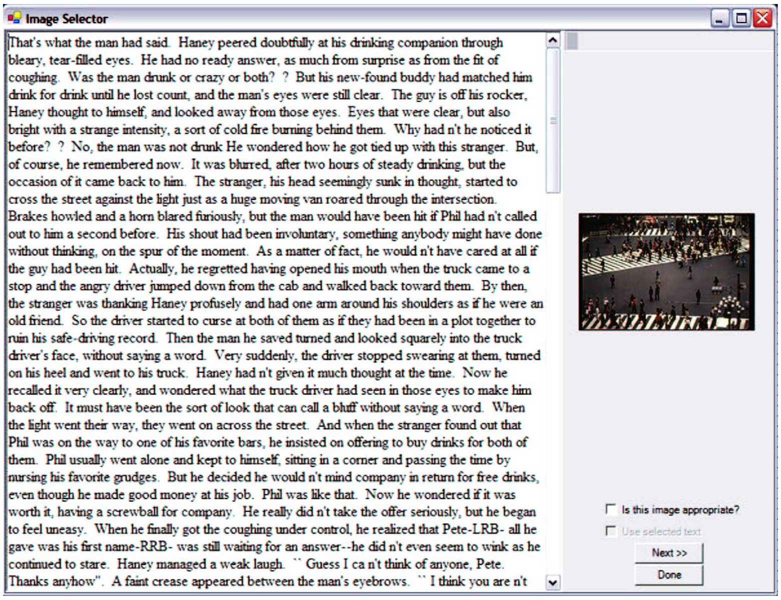
Given the inadequacy of this and all other existing image datasets for this kind of work, we created a new research corpus named ImCor. This corpus links the images from the Corel dataset with the sense disambiguated SemCor corpus to provide a new corpus which links images with semantically tagged text. (We have made ImCor available for research purposes [31].)

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 tool to facilitate human selection of text for the image candidates (Fig. 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. The Kappa statistic for the agreement between the two raters on this subset was 0.575, which is appreciable, but less than hoped for, reflecting the subjective nature of the task.

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 $1/P$, 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 text and 13.17% with random samplings from documents.



That's what the man had said. Haney peered doubtfully at his drinking companion through bleary, tear-filled eyes. He had no ready answer, as much from surprise as from the fit of coughing. Was the man drunk or crazy or both? ? But his new-found buddy had matched him drink for drink until he lost count, and the man's eyes were still clear. The guy is off his rocker, Haney thought to himself, and looked away from those eyes. Eyes that were clear, but also bright with a strange intensity, a sort of cold fire burning behind them. Why had n't he noticed it before? ? No, the man was not drunk He wondered how he got tied up with this stranger. But, of course, he remembered now. It was blurred, after two hours of steady drinking, but the occasion of it came back to him. The stranger, his head seemingly sunk in thought, started to cross the street against the light just as a huge moving van roared through the intersection. Brakes howled and a horn blared furiously, but the man would have been hit if Phil had n't called out to him a second before. His shout had been involuntary, something anybody might have done without thinking, on the spur of the moment. As a matter of fact, he would n't have cared at all if the guy had been hit. Actually, he regretted having opened his mouth when the truck came to a stop and the angry driver jumped down from the cab and walked back toward them. By then, the stranger was thanking Haney profusely and had one arm around his shoulders as if he were an old friend. So the driver started to curse at both of them as if they had been in a plot together to ruin his safe-driving record. Then the man he saved turned and looked squarely into the truck driver's face, without saying a word. Very suddenly, the driver stopped swearing at them, turned on his heel and went to his truck. Haney had n't given it much thought at the time. Now he recalled it very clearly, and wondered what the truck driver had seen in those eyes to make him back off. It must have been the sort of look that can call a bluff without saying a word. When the light went their way, they went on across the street. And when the stranger found out that Phil was on the way to one of his favorite bars, he insisted on offering to buy drinks for both of them. Phil usually went alone and kept to himself, sitting in a corner and passing the time by nursing his favorite grudges. But he decided he would n't mind company in return for free drinks, even though he made good money at his job. Phil was like that. Now he wondered if it was worth it, having a screwball for company. He really did n't take the offer seriously, but he began to feel uneasy. When he finally got the coughing under control, he realized that Pete-LRB- all he gave was his first name-RRB- was still waiting for an answer--he did n't even seem to wink as he continued to stare. Haney managed a weak laugh. `` Guess I ca n't think of anyone, Pete. Thanks anyhow''. A faint crease appeared between the man's eyebrows. `` I think you are n't

□ Is this image appropriate?

□ Use selected text

Fig. 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.

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.

6. Experiments

For our experiments we produced twenty different breakdowns of our corpus into training and testing sets (90% training, 10% testing). In our corpus there are a number of images which are used two or more times, and thus we took care to ensure that in these cases the entire group was assigned either to the training or testing sets. For each split, we then determined the vocabulary based on the training set. We first removed stop words from the corpus to reduce computation. We then eliminated all word senses which occurred less than 20 times (50 times in a second experiment). If this produced images without words, then they were removed, and the vocabulary was recomputed, iteratively, if needed. Typical vocabulary sizes were 3800 senses from about 3100 sense blind words (300/2600 for the second experiment). To provide some idea of the vocabulary, we noted 193 senses starting with the letter "m", of which 137 were unambiguous, and 56 had at least two senses. Those 56 senses were:

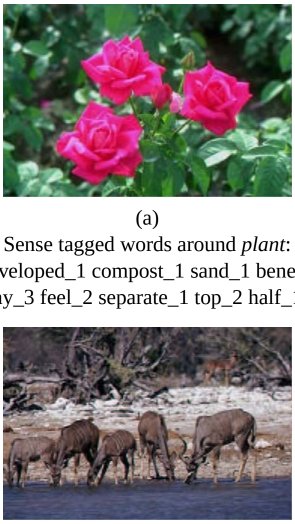
> machinery_1 machinery_2 major_1 major_2 major_3 make_1 make_10 make_12 make_13 make_17 make_2 make_3 make_4 make_6 make_8 man_1 man_2 man_3 man_4 man_5 marvel_1 marvel_2 mass_1 mass_3 mass_4 matter_1 matter_2 maturity_1 maturity_2 mean_1 mean_2 mean_3 measure_2 measure_3 memory_1 memory_2 mention_2 mention_3 mind_1 mind_2 miss_2 miss_6 moment_1 moment_2 monotonous_1 monotonous_2 month_1 month_2 moral_1 moral_2 mortal_1 mortal_2 mouth_2 mouth_3 musician_1 musician_2

Next we trained the word prediction model (Section 3) on the combined image sense data. We used the features described above for the 16 largest image regions, or, if there were fewer than 16, then we used all of them. 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 to get two sets of final results for word sense disambiguation. Fig. 5 shows a few examples where the text based method gives the wrong sense but adding image information leads to the correct sense.

We compute performance using *only* documents which have at least one ambiguous word. By construction, if all the words in a test document have only one sense, then 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. For our baseline we use the performance of the empirical distribution of the training set, which was roughly 60%. This is a harsher baseline than the simple "most common sense" method, which has been found to be surprising effective [47], as the empirical distribution gives the common sense for the particular corpus being investigated. We omit results using a second text WSD method [36] as they were roughly comparable to our base line (a score of zero), which is not surprising after the fact given the nature of that algorithm and our corpus.

We provide results in two forms. In Table 1 we report the average absolute scores over the 20 samples. In Table 2 we report the amount by which the performance of each method exceeds that of the baseline, averaged over the 20 samples. This controls somewhat for subset difficulty, and makes it easy to identify non-trivial performance since doing so results in positive values.

The results of combining the two sources of information are very promising, as the performance went beyond that of either method alone, which was exactly what we were trying to achieve. On the large data set (extended ImCor) we were able to increase performance over the baseline by nearly 20% yielding nearly 80% absolute performance. In the small (seed) data set, the performance increase was more modest, yielding 5% improvement. In all three cases, the results are statistically significant. Specifically, we performed a paired t test for the performance with images and text exceeding that of text alone over the 20 samples with 9 degrees of freedom, reflecting the fact that we have roughly 10 in-



(a)

Sense tagged words around *plant*:
rooting_1 developed_1 compost_1 sand_1 benefit_1 good_1
find_1 day_3 feel_2 separate_1 top_2 half_1 plant_2

(b)

Sense tagged words around *water*:
reach_1 location_1 sundown_1 herd_1 water_2 and_then_1 broad_1 grass_1 flat_1

Fig. 5. Two cases where image information proved to be helpful. In (a), text based word sense disambiguation gives the canonical, abstract meaning of “water”, water_1 (substance). Adding image information gives the correct sense, water_2 (body of water). In (b), using text alone gives the incorrect sense for “plant”, plant_1 (factory). Adding image information gives the correct sense, plant_2 (botanical). In both cases the more visual, but less common, sense was promoted by the statistical model linking image features to words. However, we caution the reader that most words used in this study are not particularly visual, and most examples are not this clear cut. Nonetheless, correlates between visual features and word senses which are consistent between training data and testing data can help disambiguate senses as demonstrated in the quantitative results.

Table 1

Restricted word prediction results for the word sense disambiguation experiments. The first two rows are for the extended ImCor data set (20,153 text passages paired with images) at two different 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. For completeness, the third row is the result using the manually produced seed data set (1,633 pairs), even though the data is a bit sparse for our learning method. The numbers tabulated are the fraction of times the sense was correctly chosen. Every document processed has at least one ambiguous word. Some words are unambiguous, and all algorithms score correctly on those words by construction. The results shown are the average of 20 different breakdowns of training and testing. The error, as estimated from the variance over the 20 test/training splits, is about 0.003 for the first two rows, and about 0.01 for the third row. Incorporating image information is statistically significant at $p = 0.01$ in all three cases, using paired t tests

Data set	Minimum sense count	Baseline	Text only using [41]	Image only	Combined (using (6))
Full	20	0.615	0.683	0.791	0.817
Full	50	0.606	0.674	0.781	0.814
Seed	20	0.571	0.693	0.687	0.741

Table 2
Analogous results for that in Table 1, but here we show the performance *increase* of each method over the empirical distribution baseline, averaged over the samples. Comparisons based on these numbers are more accurate than comparing the overall performances reported in Table 1 because the results for the empirical distribution controls somewhat for sample difficulty. The estimated errors in the numbers are 0.003 for the first two rows, and about 0.01 for the third row

Data set	Minimum sense count	Text only using [41]	Image only	Combined (using (6))
Full	20	0.069	0.177	0.202
Full	50	0.068	0.175	0.208
Seed	20	0.125	0.116	0.173

Table 3
Average counts for the number of senses correctly identified over the 20 samples for each of the three experiments. The total number of ambiguous words is provided in the first column. These results are consistent with those in the previous two tables, but they do not map exactly onto those numbers because here only words which are ambiguous with respect to the vocabulary are counted. The errors in the first two rows are roughly 11, and the errors in the third row are roughly 4. All differences between WSD with text only and WSD with text with images are significant at $p = 0.0005$

Average number of ambiguous words	Baseline	Text only using [41]	Image only	Combined (using (6))
6975	4506	4935	5082	5361
6204	3986	4390	4515	4803
697	411	477	454	498

dependent samples in the 20 sets due to sampling 10% of the data at a time. For the three experiments we have: (1) (M = 0.133, SE = 0.0030) with $t(9) = 44.8$, $p < 0.0005$; (2) (M = 0.140, SE = 0.003) with $t(9) = 49.6$, $p < 0.0005$; and (3) (M = 0.048, SE = 0.011) with $t(9) = 4.5$, $p < 0.001$.

We can further interpret the results in Table 1 by noting that each run attempts to find senses for about 7,000 words distributed over about 800 documents. The 7,000 words have about 20,000 senses among them, relative to our vocabulary. Thus our baseline method, performing at about 60% specifies the correct sense for about 4,200 words, and misses 2,800. The combined method, performing at about 80%, misses about half that amount (1,400).

Finally, in Table 3, we provide the average counts of correct sense identification, restricted to words which are ambiguous. Again, in all three experiments, there is a significant performance increase due to adding image data. Specifically, in a paired t test for the results using images and text being greater than that using text alone over the 20 samples with 9 degrees of freedom we have, for the three experiments: (1) (M = 426, SE = 10.4) with $t(9) = 41.0$, $p < 0.0005$; (2) (M = 413, SE = 11.4) with $t(9) = 36.2$, $p < 0.0005$; and (3) (M = 21.2, SE = 4.0) with $t(9) = 5.3$, $p < 0.0005$.

We emphasize that our domain was constructed somewhat artificially to test our ideas, and that some of the improvement going from the small (seed) data set to the larger one is likely due to the system taking advantage of the structure of the Corel data. However, even in the seed data case, where there was only limited image data to train on but the corpus was more pure, we found a statistically significant improvement in word sense disambiguation performance when image data was included.

7. Conclusion

The main conclusion from this work is that visual information can help disambiguate senses, and thus help determine language meaning. In fact, on a small, relatively friendly domain, we were able to exceed the performance of two text based methods. We were further able to improve performance by combining text and imaged based information. Our experiments thus suggest that image information as captured by our approach can be sufficiently independent from textual based cues that combining the two sources of information can prove fruitful.

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.

References

[1] Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004.

[2] E. Agirre, G. Rigau, Word sense disambiguation using conceptual density, in: Proceedings of COLING’96, Copenhagen, Denmark, 1996, pp. 16–22.

[3] E. Agirre, G, 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.

[4] Y. Bar-Hillel, The present status of automatic translation of languages, in: D. Booth, R.E. Meagher (Eds.), Advances in Computers, Academic Press, New York, 1960, pp. 91–163.

[5] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, J. Machine Learning Res. 3 (2003) 1107–1135.

[6] K. Barnard, P. Duygulu, D. Forsyth, Clustering art, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. II, 2001, pp. 434–441.

[7] K. Barnard, P. Duygulu, D. Forsyth, Exploiting text and image feature co-occurrence statistics in large datasets, in: R. Veltkamp (Ed.), Trends and Advances in Content-Based Image and Video Retrieval, Springer, Berlin, submitted for publication.

[8] K. Barnard, P. Duygulu, K.G. Raghavendra, P. Gabbur, D. Forsyth, The effects of segmentation and feature choice in a translation model of object recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, II, 2003, pp. 675–682.

[9] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, in: Proceedings of the International Conference on Computer Vision, vol. II, 2001, pp. 408–415.

[10] K. Barnard, M. Johnson, Word sense disambiguation with pictures, in: Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data, 2003, pp. 1–5.

[11] J. Bear, J. Dowding, E. Shriberg, Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog, in: Meeting of the Association for Computational Linguistics, 1992, pp. 56–63.

[12] A.B. Benitez, S.-F. Chang, Automatic multimedia knowledge discovery, summarization and evaluation, 2003.

[13] A.B. Benitez, S. Chang, Image classification using multimedia knowledge networks, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Barcelona, Spain, September 2003.

[14] P.F. Brown, J. Cocke, S.A.D. Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin, A statistical approach to machine translation, Computational Linguistics 16 (1990) 79–85.

[15] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer, The mathematics of machine translation: Parameter estimation, Computational Linguistics 19 (10) (1993) 263–311.

[16] P.F. Brown, S. Della Pietra, V.J. Della Pietra, R.L. Mercer, Word-sense disambiguation using statistical methods, in: Meeting of the Association for Computational Linguistics, 1991, pp. 264–270.

[17] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general contextual object recognition, in: Proceedings of the European Conference on Computer Vision, I, 2004, pp. 350–362.

[18] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, J. Malik, Blobworld: A system for region-based image indexing and retrieval, in: Proceedings of the Third International Conference on Visual Information Systems, Springer, Berlin, 1999.

[19] M. Ciaramita, T. Hofmann, M. Johnson, Hierarchical semantic classification: Word sense disambiguation with world knowledge, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003.

[20] C. de Loupy, M. El-Bèze, 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, 2000, pp. 219–223.

[21] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. Roy. Statist. Soc. Ser. B 39 (1) (1977) 1–38.

[22] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in: Proceedings of The Seventh European Conference on Computer Vision, vol. IV, 2002, pp. 97–112.

[23] P. Edmonds, A. Kilgarriff (Eds.), J. Natural Language Engineering 9 (January 2003).

[24] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, in: Proceedings of CVPR'04, vol. 2, 2004, pp. 1002–1009.

[25] W.N. Francis, H. Kučera, Frequency Analysis of English Usage. Lexicon and Grammar, Houghton Mifflin, 1981.

[26] G. Miller, C. Leacock, T. Randee, R. Bunker, A semantic concordance, in: Proceedings of the 3rd DARPA Workshop on Human Language Technology, 1993, pp. 303–308.

[27] W. Gale, K. Church, D. Yarowsky, One sense per discourse, in: Proceedings of the DARPA Workshop on Speech and Natural Language, 1992, pp. 233–237.

[28] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran, Indexing with wordnet synsets can improve text retrieval, in: Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal, Canada, 1998, pp. 38–44.

[29] T. Hofmann, J. Puzicha, Statistical models for co-occurrence data, Technical Report, Massachusetts Institute of Technology, 1998.

[30] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: Proceedings of SIGIR, 2003, pp. 119–126.

[31] M. Johnson, K. Barnard, ImCor: A linking of SemCor sense disambiguated text to corel image data, http://kobus.ca/research/data/index.html, 2004.

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

[33] A. Kilgarriff, Senseval: An exercise in evaluating word sense disambiguation programs, in: Proceedings of LREC, Granada, May 1998, pp. 581–588.

[34] V. Lavrenko, S.L. Feng, R. Manmatha, Statistical models for automatic video annotation and retrieval, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, May 2004.

[35] X. Li, S. Szpakowicz, S. Matwin, A wordnet-based algorithm for word sense disambiguation, in: Proceedings of the IJCAI-95, Montreal, Quebec, 1995, pp. 1368–1374.

[36] D. McCarthy, R. Koeling, J. Weeds, J. Carroll, Finding predominant senses in untagged text, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004, pp. 280–287.

[37] S.W. McRoy, Using multiple knowledge sources for word sense discrimination, Computational Linguistics 18 (1) (1992) 1–30.

[38] D. Melamed, Empirical Methods for Exploiting Parallel Texts, MIT Press, Cambridge, MA, 2001.

[39] R. Mihalcea, E. Faruque, Senselearner: Minimally supervised word sense disambiguation for all words in open text, in: Proceedings of ACL/SIGLEX Senseval-3, Barcelona, Spain, July 2004.

[40] R. Mihalcea, D. Moldovan, Word sense disambiguation based on semantic density, in: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, August 1998.

[41] R. Mihalcea, D. Moldovan, An iterative approach to word sense disambiguation, in: Proceedings of Florida Artificial Intelligence Research Society Conference (FLAIRS 2000), Orlando, FL, May 2000, pp. 219–223.

[42] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller, Introduction to wordnet: An online lexical database, Internat. J. Lexicography 3 (4) (1990) 235–244.

[43] A. Montoyo, M. Palomar, G. Rigau, Wordnet enrichment with classification systems, in: Proceedings of NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations', Carnegie Mellon University, Pittsburgh, USA, 2001, pp. 101–106.

[44] Hwee Tou Ng, Hian Beng Lee, Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach, in: A. Joshi, M. Palmer (Eds.), Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, Morgan Kaufmann, San Francisco, 1996, pp. 40–47.

[45] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Machine Intell. 22 (9) (2000) 888–905.

[46] J. Stetina, S. Kurohashi, M. Nagao, General word sense disambiguation method based on a full sentential context, in: Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference, Association for Computational Linguistics, Somerset, NJ, 1998, pp. 1–8.

[47] J. Traupman, R. Wilensky, Experiments in improving unsupervised word sense disambiguation, Technical Report, University of California at Berkeley, 2003.

[48] D. Yarowsky, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, in: Proceedings of COLING-92, Nantes, France, July 1992, pp. 454–460.

[49] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: Proceedings of the 33rd Conference on Applied Natural Language Processing, ACL, 1995, pp. 189–196.

[50] V. Yngve, Syntax and the problem of multiple meaning, in: W. Locke, D. Booth (Eds.), Machine Translation of Languages, Wiley, New York, 1955, pp. 208–226.