

Human-Assisted Word Sense Disambiguation

Janara Christensen, Anthony Fader, Thomas Lin

Computer Science & Engineering

University of Washington

Seattle, WA 98105 USA

{janara, afader, tlin}@cs.washington.edu

ABSTRACT

Word Sense Disambiguation is a challenge in Machine Translation. The PanMail project takes the approach that people writing emails to be translated will be willing to help in the disambiguation process because they have an interest in accurate translation of their emails. This not only improves upon regular translation, it also collects valuable data for training better translation techniques.

When asking a user to help disambiguate, PanMail needs a way to present the different senses for the user to choose among. PanMail aims to be able to translate between any two languages in the world. However, only 14 languages have sense-disambiguated dictionaries. Over 500 languages in the world do not have sense-disambiguated dictionaries, and for those languages the system has to use methods like synonyms of the different senses, or language-independent methods like images that represent the different senses.

This study evaluates the use of definitions, synonyms, and images for user disambiguation. We report on lessons learned for how to choose images that reflect different senses, as well as users' impressions of the methods. We find that even though images are not as effective as definitions, they are significantly more effective than the baseline of just choosing the most common word sense. Thus, for the sparse languages where definitions and synonyms are unavailable, images are a good option.

INTRODUCTION

Statistical machine translation is a data-intensive task that requires a large parallel corpus between the source and target languages. A simpler related task is *lexical translation*, where the goal is to translate a single word or phrase. Lexical translation is currently being used in a prototype panlingual email system called *PanMail*, where users can type short emails in their native languages and have the computer give a rough translation to a given target language. The PanMail system leverages machine readable bilingual and multilingual dictionaries to perform lexical translation between language pairs where no parallel corpora are available.

Lexical translation errors are largely due to multiple word senses, or polysemy. For example, Figure 1 illustrates how the French word *avocat* translates to English as both *lawyer* and *avocado*.

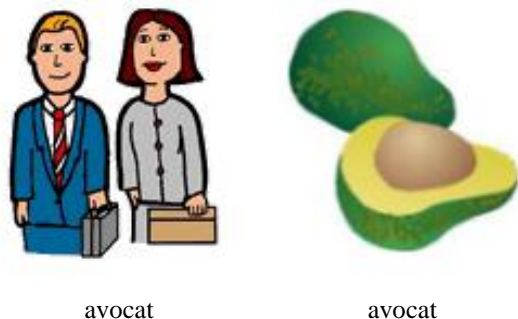


Figure 1. Which sense of *avocat* is the user referring to?

A lexical translation system must make a decision about which of these senses the user had in mind while composing the message. Fortunately, in the PanMail system the user is present at the keyboard and is most likely willing to help the computer make the right decision. If the system can prompt the user to choose the intended sense of a given word, then it can make a more informed translation.

Sense disambiguated dictionaries exist for the 14 most common languages (e.g., English and French), so for those languages the PanMail system could simply display the dictionary entry and have the user choose the intended sense. However, for hundreds of less common languages, these types of dictionaries are not available (as shown in Figure 2). Instead, we would have to rely on disambiguation methods like synonyms or images. When there is still a rich bilingual dictionary, the backtranslation method can offer synonyms for different senses.

When the bilingual data is sparse, synonyms will be less available and we will have to rely on language-independent methods like images. One way to prompt the user to disambiguate a word is to find all candidate target language translations, look each one up in an image database (e.g., Google Image Search), and choose a representative image for each possible sense of the target word. Continuing the above example, *avocat* has the possible translations *lawyer* and *avocado*, so the system would find an image for each of these possible senses. The user could then disambiguate *avocat* by picking which image best represents the sense they had in mind.







Disambiguation Options		Sense-Disambiguated Dictionaries	Synonyms: Backtranslation	Images
14 lang	Major Languages (e.g. English)			
500 languages	Minor Languages (e.g. Thai)			
	“Sparse” Languages (e.g. Swahili)			

Figure 2. Only 14 languages have sense-disambiguated dictionaries.

Given a set of candidate words and senses, two problems arise. First, the system has to have some way to determine which image best represents the given sense. Second, the user has to be able to easily choose the sense they have in mind from the system's list of images. Our work addresses the second problem with user studies to determine how well people can disambiguate a word using images.

Our driving question is: *how well can users disambiguate a word using images?* To answer this question, we test an image-based disambiguation mechanism with real users. Although an automated system for sense disambiguation images is the end goal, we do not have a working implementation of this kind of system, so in our experiments we manually create a set of images for a small controlled vocabulary. Using a manually created dataset for our experiments might not be representative of the performance of an automated system, but it has two advantages.

First, our goal is to test the performance of the users, not the performance of an image retrieval system. Therefore, using human annotated images will remove errors that would come from an imperfect image retrieval system and will make the results easier to interpret with respect to our question.

Second, humans are much better than computers at finding images representative of a given word sense, so any conclusions drawn from our experiments will act as an upper bound on the quality of an ideal system. Thus our proposed experiments can be thought of as a feasibility study to determine whether further research should go into the image search part of the system.

Our contributions are the following:

1. We designed and conducted a user study to investigate the effectiveness of images for human-assisted word sense disambiguation.

2. We show that humans are able to disambiguate words using hand-picked images with an accuracy of 83%, compared to a baseline of 44%.

The remainder of this paper is organized as follows. The following section provides an overview of related work. Next, the paper describes a study we designed to evaluate the different disambiguation methods, and the test interface we set up. We then review the results of the study, analyze the results, and conclude with a brief discussion and comments on future work.

RELATED WORK

Lexical translation has a long history and has been studied through the lens of multilingual information retrieval [Helmreich 1993, Copestake 1994, Hull 1996], where machine-readable dictionaries (MRDs) are leveraged to perform multilingual search.

Recently, there has been work on performing lexical translation using the *translation graph* constructed from a large collection of multilingual MRDs [Etzioni et al., 2007]. The translation graph has a node for each word in each language and labeled edges between words that share the sense given by the edge label. The main advantage of the translation graph is that it allows lexical translation between language pairs, even if there is no bilingual dictionary linking them.

For example, if a Slovenian-Chinese bilingual dictionary is not available, Slovenian-English and English-Chinese dictionaries can be composed to provide a translation. This composition of MRDs corresponds to finding a sense-preserving path in the translation graph. The PanMail system discussed in the introduction is built on the translation graph introduced by Etzioni *et al.*

In order to find a correct translation of a word in the translation graph, its sense must first be determined.

The problem of word sense disambiguation (WSD) is well-studied in the natural language processing literature [Ide

1998]. However, some of the most successful WSD systems rely on hand-labeled datasets [Yaworsky, 1994] or large monolingual corpora [Yarowsky, 1995], which are not available for less-spoken languages. The goal of this paper is to explore options for language-independent word sense disambiguation, so the work presented here is orthogonal to traditional research in WSD.

Another related line of research is combining information from images and text to perform WSD.

Barnard *et al.* introduce a method for WSD that models the joint distribution of words and related images (e.g., images and their captions) and found that images provide useful information for WSD [Barnard et al., 2003]. This is motivation for the work presented in this paper: if users can assist the computer with WSD, then a joint model of text and images could be used as a method to find images representing a given word sense.

There has been previous work on human-assisted WSD [Sammer et al., 2006] [Colowick and Pool, 2007]. Our work is closest to Sammer *et al.*'s, in which a translation system presents word sense glosses to the user in order to perform WSD. However, for many less-spoken languages these glosses are not available. The work presented in this paper explores the feasibility of alternative methods that do not rely on the availability of glosses.

USER STUDY

To test user performance, we will assume the following model of how the user would interact with an ideal system. A user writes a sentence containing a word w with some sense s of w in mind. The system then prompts the user to identify s in a set of candidate senses s_1, \dots, s_n using some disambiguation method M . M could be a sense disambiguated dictionary entry, our proposed image disambiguation mechanism, or some other way of representing word senses.

For our user study, we needed to know the intended sense of w in order to evaluate the user. Therefore, we used the following simplification, which preserves the first step of the original model:

1. The user is prompted with a word w and an example sentence containing w used in sense s .
2. The user is prompted to choose s from a set of candidate senses s_1, \dots, s_n using disambiguation method M .

In our user study, we attempted to find the effects of different types of word-senses pairs (w, s) and disambiguation methods M on a user's ability to identify the correct sense.

We considered showing participants both the word within the context of a sentence and the dictionary definition of the word. While showing both sentence and definition would give the user a better idea of the exact sense intended, showing the definition also had a number of disadvantages. First, one of our disambiguation methods is disambiguating

the word by use of a dictionary definition, and showing the definition would give a bias to this method. Furthermore, part of the task is to determine the exact sense of a word within a sentence. Users do not always have exact definitions in mind when writing. By asking the user to determine the definition, the task becomes more realistic. For these reasons, we decided to only show the sentence.

We also considered whether to include the part of speech for each sense the user was to choose from. While part of speech is generally available and might add to the user's ability to disambiguate, English is also quite complex and part of speech rules are not always consistent. For example, nouns can be morphed into adjectives and verbs easily ("I *emailed* John yesterday"). However, we decided that, overall, viewing the part of speech was advantageous and so we included the part of speech for each of the senses the user was to choose from. In a real application we will generally have part of speech information for the majority of languages, and including part of speech information for all methods means they are affected equally.

Disambiguation Methods

In our user study, we asked the participant to pick the correct sense s using disambiguation method M . We used the following five disambiguation methods for M :

M1 Sense definitions. The user is shown a definition for each candidate sense and asked to choose the definition that most closely matches s . This case represents the ideal situation where we have a sense disambiguated dictionary in every source language (which, as discussed above, is not the actual case).

M2 Images. The user is shown an image for each sense and is asked to choose the image that most closely matches s . This case represents our proposed system.

M3 Synonyms. The user is shown several synonyms for each sense and is asked to choose the set of synonyms that most closely matches s . This case represents an ideal version of existing text-based disambiguation methods like backtranslation, where the source word is translated to the target language and then back again.

M4 Synonyms and images. The user is shown one image and one set of synonyms for each sense and will be asked to choose the pair that most closely matches s . We consider this hybrid approach because it is likely to work better than synonyms or images alone, and in theory, images will usually be available too whenever synonyms are available.

M5 Dominant senses. The system automatically chooses the first dictionary sense of w , which is usually the most common usage. This case represents our baseline method.

Query Terms

In our study, we tested each of the disambiguation methods on words evenly distributed in 6 categories: {concrete, abstract} \times {noun, adjective, verb}. For each category, we used 6 words, for a total of 36 words. Four words per

Which sense of **blind** is this email snippet referring to?

... Of **blind** people (which are generally well-explored these days), is pretty skimpy. We're -still- in the exploration/discovery process of "what works", and as we see more solutions like Reef's -- which allow you to build an interface for ONE audience ...

Please click on the sense you choose.




 <p>Sightless, unseeing <i>adjective</i></p>	 <p>Bedazzle <i>verb</i></p>	 <p>Curtain, cover <i>noun</i></p>
---	---	---

Figure 3. The study interface provides the context and asks participants to choose the correct word sense.

category (24 total) were used for testing the disambiguation methods, and 2 words per category (12 total) were used as training examples to familiarize study participants with the interfaces for the different disambiguation methods.

Our rationale for choosing these categories was that certain disambiguation methods might perform better for certain categories of words. For example, we thought our participants might have more trouble disambiguating abstract senses using images, because abstract terms can be difficult to express with images. It's easy to imagine finding good representative images for a concrete term like "apple" but harder to imagine a good representative image for an abstract term like "judgment." Similarly, we thought verbs might prove more difficult to disambiguate than nouns.

Because our target application is email, we chose to sample our words from an email corpus. For our email corpus we used the British Columbia Conversation Corpus (BC3)¹, which consists of 40 email threads and 3222 sentences. We obtained our query words by randomly sampling words from the corpus until we had enough words for each

category. Note that this method ensures that more common words were weighted higher than less common words.

We determined the correct sense of each word by looking at the list of possible senses in Wiktionary. We then voted to confirm the correct sense. To prevent users from being overwhelmed by the number of possible senses, we limited the number of senses shown to just the most common five senses for each word. In the uncommon cases where the correct sense was not in these top five, we dropped the word and sampled an additional one from the same category. Wiktionary senses tended to be at a more appropriate level for translation disambiguation than WordNet or the Merriam Webster dictionary. WordNet was often too fine-grained. For example, it has 52 different senses for the word "play," many of which would translate into the same target words in most languages.

Images

We decided to choose the images ourselves rather than by use of an image retrieval system. Our rationale for this decision is as follows. Our goal is to test the performance of the users, not the performance of an image retrieval system. Therefore, using human annotated images will remove errors that would come from an imperfect image retrieval

¹ Available for download at <http://www.cs.ubc.ca/labs/lci/bc3.html>

system and will make the results easier to interpret with respect to our question. Second, humans are likely much better than computers at finding images representative of a given sense, so any conclusions drawn from our experiments will act as an upper bound on the quality of an ideal system. Thus our proposed experiments can be thought of as a feasibility study to determine whether further research should go into the image search part of the system. Our eventual goal is to use this study to better inform the design of a potential system.

Each of the three authors found a potential image for each sense of each word by querying Google Images and Flickr. Google Images generally had more appropriate images for this task than Flickr. The images on Flickr were often larger and tended to emphasize artistic attributes that are less important for disambiguation. Querying Google Images for the source word and the synonyms or related words that only match one potential sense was the best method we found for finding images for specific word senses. One possible reason for this is that Google licensed the ESP Game [von Ahn and Dabbish, 2004] technology, in which humans provide as many labels as possible for images. This means images will often be tagged with synonyms of their primary descriptions too.

Even with this technique, however, there tended to be a good amount of noise (undesirable images) in the search results and many senses of words that were difficult to find good images for. Some senses did not naturally lend themselves to images. For example, “approach – to take approaches to,” “clear – bright, not dark, or obscured,” and “find – to decide that, to form the opinion that.” Additionally, sometimes the differences between senses were too small to easily distinguish between with images, for example “offer – to propose something.” and “offer – to proffer.”

For each word sense, the authors voted on which of the three potential images (one found by each author) was most representative of that sense, and we went with the images with the most votes. The challenges for finding good images of particularly difficult senses still meant that even in this idealized case of image selection, some of the more difficult senses did not end up with good images.

Synonyms

We also handpicked the synonyms. Again, this allows us to test the ideal case and gain an upper bound. Generally synonyms were chosen from <http://thesaurus.reference.com/>.

Participants

We recruited 12 study participants. Participants ranged in age from 20 to 60, were of approximately balanced gender (5 female), and were all fluent in English. Participants were generally recruited from our classmates, friends, and families. There was a small bias toward computer science graduate students, who in general may be more comfortable with computer usage than the general population, but there

is no initial reason to believe that they would have a different behavior in response to using images for word sense disambiguation tasks. Participants generally completed the study within half an hour.

Web-based Study Interface

In our user study, we asked the participant to pick the correct sense s using disambiguation

The study was run via a website interface. The website began by explaining how the goal of the study was to evaluate different approaches toward user-assisted word sense disambiguation. After this brief introduction, participants began the disambiguation evaluations. At the beginning of each of the four disambiguation methods, participants were first presented with three examples (that did not count toward the results) in order to practice with the interface. As shown in Figure 3, the interface first provided the word in the context of an email, and then asked the participant to pick between the possible word senses.

Before the study, we randomly assigned a word ordering such that for each participant, each disambiguation method would be evaluated on one word from each category (from $\{\text{concrete, abstract}\} \times \{\text{noun, adjective, verb}\}$). We kept the ordering constant among the different participants. For example, if the ordering was $w_1 = \text{call}$, $w_2 = \text{work}$, \dots , $w_{36} = \text{firm}$, then every participant will see the words in that order. This way, word ordering had a minimal effect on the results.

With four disambiguation methods and 12 study participants, we used the following 4×4 Latin squares to determine the order of methods to present each participant in order to remove confounding factors due to ordering:

$$\begin{pmatrix} \text{A} & 1 & 2 & 3 & 4 \\ \text{B} & 2 & 1 & 4 & 3 \\ \text{C} & 3 & 4 & 1 & 2 \\ \text{D} & 4 & 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} \text{E} & 1 & 3 & 4 & 2 \\ \text{F} & 2 & 4 & 3 & 1 \\ \text{G} & 3 & 1 & 2 & 4 \\ \text{H} & 4 & 2 & 1 & 3 \end{pmatrix} \begin{pmatrix} \text{I} & 1 & 4 & 2 & 3 \\ \text{J} & 2 & 3 & 1 & 4 \\ \text{K} & 3 & 2 & 4 & 1 \\ \text{L} & 4 & 1 & 3 & 2 \end{pmatrix}$$

This means that, for example, participant A will be prompted with method M1 first, then M2, M3, and M4.

After the study, participants were also given a brief post-study survey where they were asked “How difficult was it to use each method for the task of disambiguation?” (radio button choices: Very Hard, Hard, Medium, Easy, and Very Easy for each method) and “How fun/enjoyable was each method?” (radio button choices: Not at all, Not much, Medium Usually, and Very Much for each method). The radio buttons started out unselected to avoid biasing the results toward any default selections.

RESULTS AND ANALYSIS

In this section we will first analyze correctness and time spent to disambiguate with the different methods. Our comparisons are biased toward an even distribution of concrete/abstract and nouns/adjectives/verbs for query terms, so we next examine how the methods performed

within the different query term categories. Last, we will examine the post-study survey results.

As the baseline for comparison, we use the most common word sense of the test queries. This baseline will perform better than the even more naïve method of picking a random word sense each time. A more complicated baseline could have been to use existing WSD methods based on word co-occurrence frequencies (e.g. *avocat* as *lawyer* would tend to occur more frequently around words like *court* and *avocat* as *avocado* would occur more frequently around words like *ate*), but that method has its own set of limitations, would not work for short emails without much context, and we did not have the resources to try it.

Correctness

The first measure we look at is whether the different disambiguation methods have any effects on the correctness of user disambiguations.

A binary measure for correctness may not be the most effective because some senses will translate correctly for some languages but not for others. For example, two of the senses of the word *green* include “the color green (noun)” and “having the color green (adjective).” If the intended sense is the first one but the study participants picks the second one, then should that count as a correct answer because for some languages that will lead to the correct translation, or should it count as an incorrect answer because for other languages it will lead to an incorrect translation?

We settled with using a 3 point scale for correctness. The 3 possible ratings include:

1. **Correct** – when a sense fits exactly, and will always translate as intended. (e.g., “the color green” in the example above). For significance testing, we count this as 100% correct.
2. **Almost correct** – the more ambiguous case above when a sense is close and will sometimes translate as intended. (e.g., “having the color green”). For significance testing, we count this as 50% correct.
3. **Incorrect** – when the sense is wrong. (e.g., green as in “environmental”). We count this as 0% correct.

Figure 4 shows the distribution of *correct*, *almost correct*, and *incorrect* responses for the *definitions*, *synonyms*, *images*, and *baseline* disambiguation methods. The mean correctness (on the 0%, 50%, 100% scale) was: *definitions* 96%, *images + synonyms* 89%, *synonyms* 88%, *images* 83% and *baseline* 44%.

We then analyze disambiguation *correctness* using a mixed-model analysis of variance. We model our variable of interest, *method* (values *definitions*, *images+syn*, *synonyms*, *images* and *baseline*), as a fixed effect. To account for learning or fatigue effects, we model *trial number* as a fixed effect. To examine whether the *correctness* was influenced by whether the query word was

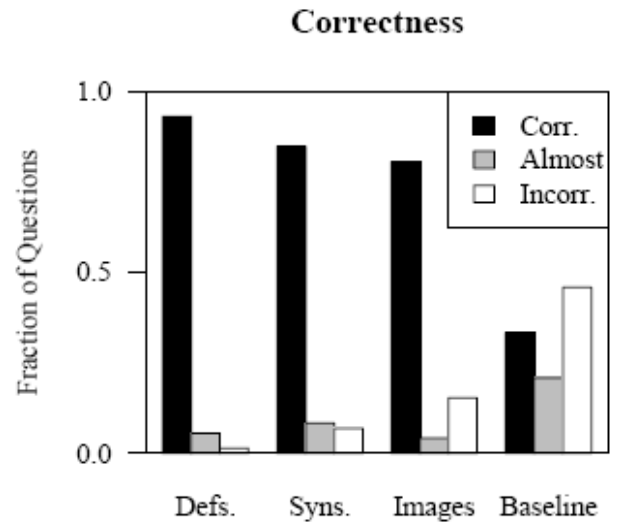


Figure 4. Study Results - Correctness

abstract/concrete or a *noun/verb/adjective*, we model those two variables as fixed effects. To examine whether *correctness* was influenced by the number of word senses to choose from, we modeled *senses* as a fixed effect. Finally, we account for variations in the disambiguation difficulty of the query words (e.g., *make sense* may be more difficult to disambiguate than *joint* because the different senses of *joint* are much more distinct) and variations in the performance of different people by modeling both *query* and *participant* as random effects.

We found no significant effect of *trial number*, *abstract/concrete*, *noun/verb/adjective*, or *senses*, and so we remove them from the remainder of our analyses. The omnibus test reveals a significant main effect of *method* ($F(4, 71) = 14.2, p < .0001$), leading us to investigate pairwise differences. We use Tukey’s Honestly Significant Difference procedure to account for increased Type I error of unplanned comparisons. This shows *definitions* yielded significantly higher *correctness* than *images* ($F(1,274) = 8.22, p \approx .005$) and *baseline* ($F(1,15) = 56, p < .0001$). We did not find significantly higher correctness between *definitions* and (*images+syn* or *synonyms*), nor between (*images+syn* or *synonyms*) and *images*. We found that *images+syn* ($F(1,15) = 42, p < .0001$), *synonyms* ($F(1,15) = 41, p < .0001$), and *images* ($F(1,15) = 31, p < .0001$) all yielded significantly higher correctness than the *baseline*.

Time Spent

Our study interface collected the number of seconds that users took to disambiguate each word. We told users they could be free to take breaks as needed during the study, and some of them did. As a result, there were a few instances where people took breaks and the captured value for time spent was on the order of 5 minutes (while the average is closer to 30 seconds). To account for people taking breaks, we did not use any timing data more than 3 standard deviations away from than the mean. This was less than 2%

of the data, and was not biased toward any particular disambiguation method.

The *baseline* of automatically choosing the most common word sense will never require any of the user's time to pick an answer, so in that sense it is the winner in terms of *time spent*.

We analyze *time spent* (to disambiguate) using a mixed model analysis of variance. We model our variable of interest, *method* (values *definitions*, *images+syn*, *synonyms* and *images*) as a fixed effect. We also model *trial number*, *abstract/concrete*, *noun/verb/adjective*, and *senses* as a fixed effects and *query* and *participant* as random effects. Surprisingly, we did not find any significant effect of *method*. The least squares means of *time spent* for our human-assisted disambiguation methods all fell within the standard errors of each other.

A hypothesis one might have is that *images* take less *time spent* for the cases where there are more senses to disambiguate between. In these cases, it may be easier for a participant to just look over the images and quickly identify the correct one, instead of having to read through lengthy sense-disambiguated definitions. However, we were also unable to find any patterns or significance when separating out *time spent* per *method* by *senses*. For the query terms with 5 senses to disambiguate between (the maximum we allowed), *definitions* even had a slightly (not significantly) lower mean *time spent* than *images*.

So, it turned out that number of senses did not necessarily lead to images having an advantage. One possible explanation here is that when the disambiguation options are clear and visually distinct (e.g. *lawyer* versus *avocado* for *avocat*), images would be noticeably faster. However, often the disambiguation options are not clear and visually distinct, and in these cases participants may actually need extra time to examine the images carefully, think about what they represent, and think about what the differences are.

Abstract/Concrete, Nouns/Verbs/Adjectives

Were abstract senses harder to disambiguate using images? The mean *correctness* for *concrete* query terms (84.3%) was higher than the mean *correctness* for *abstract* query terms (75.6%). However, there were not enough data points for this to be a statistically significant difference. A future study could be run with more participants to try to draw this distinction. Every disambiguation method we tested (including *baseline*) had a higher mean *correctness* for concrete queries than abstract queries. Images for *concrete* queries had a higher mean (84.9%) than images for *abstract* queries (81%), and definitions for *concrete* queries also had a higher mean (97.2%) than definitions for *abstract* queries (94.4%).

Were verbs more difficult to disambiguate than nouns? After separating by *noun/verb/adjective*, there was not enough data to draw statistically significant conclusions for

method, other than that our human-assisted disambiguation methods led to higher *correctness* than the baseline. However, examining the least squares means suggest that there may be patterns to find in a more extensive study. *Images* for *noun* queries had a mean correctness of 88%, while *images* for *verb* queries had a mean correctness of 71%. *Images* for *adjective* queries had a mean correctness of 89%. On the other hand, *definitions* for *noun* queries, *verb* queries, and *adjective* queries all had means of 95.8%. *Synonyms* for *noun* queries had a mean correctness of 91%, and *synonyms* for *adjective* and *verb* queries had means around 87%. This suggests that disambiguating with definitions may perform similarly for different parts of speech, while image disambiguation may be more effective on nouns and adjectives than for verbs. A possible explanation could be that verbs are harder to express using images, especially abstract verbs.

When were Images Most/Least Effective?

We also examined the query terms that images performed best and worst on, to see if there were any discernable patterns about which images worked best or which types of words image disambiguation was especially good or bad for.

One case where images did extremely poorly is the term *make sense*, one of our abstract verbs. There were two main senses for this term: one meaning "be coherent, be reasonable" and the other meaning "decipher, understand." In this case, the poor performance was probably due to lack of good possible images. Google image search for "coherent, reasonable" and "decipher, understand" both give almost no images that express the desired meanings.

Another interesting case for image was the term *free*. Two of the relevant senses for *free* include "unconfined, free from jail" and "loose, unconstrained." The provided context was "please feel *free* to ..." and matches the second sense. However, we had chosen an image of someone leaving jail to represent the first sense and an image of horses running free for the second sense. We suspect that study participants may have interpreted the first image as a human being free to do something and the second image as a horse being free to do something, and picked the first image for that reason. Even when the images chosen seem to reflect their senses perfectly well, people may notice and act on unintended distinctions between images.

A query term where images did especially well was the word *joint*. We suspect images worked well here because the different meanings of *joint* are all very visually distinct: a cigarette, bones, a hinge, people cooperating, and jail.

How could lessons learned here help inform the algorithms for a future human-assisted word sense disambiguation system? The results suggest that for languages where a sense-disambiguated dictionary is available, that should be used. For languages where only synonyms and images are available, images are probably more effective on *nouns* and *adjectives* than on *verbs*. Also, images tend to be more

effective when expressing very distinct senses (perhaps a metric like semantic distance in WordNet could be used to measure this). When translating to a spare language where images are the only available disambiguation option, it is better to use them than to automatically pick the most common word sense.

Post-Study Survey Results

In terms of difficulty (1 = very hard, 5 = very easy), the average ratings were *images* 2.42, *synonyms* 3.08, *definitions* 3.25, and *images + synonyms* 3.67. Several study participants commented that for a number of the *image-only* questions, some of the images were ambiguous and it was difficult to understand what they were supposed to mean. Adding synonyms to the images helped resolve this problem.

In terms of enjoyability (1 = not enjoyable, 5 = very enjoyable), the average ratings were *definitions* 2.75, *synonyms* 2.75, *images* 3.5, *images + synonyms* 3.5. One possible explanation is that being able to use images often made the overall disambiguation task more enjoyable and less rote.

CONCLUSIONS AND FUTURE WORK

Machine translation is a difficult problem, in part because of the Word Sense Disambiguation problem. By enlisting the help of the user, PanMail hopes to address the WSD problem for lexical translation. However, an interesting question here is how well different techniques for human-assisted word sense disambiguation work.

This paper conducts a study comparing several human-assisted disambiguation methods including sense definitions, synonyms, images, synonyms + images, and a baseline of just using the most frequently used sense. For our test query set, sense definitions achieved a precision of 96%, images achieved 83%, and the baseline achieved 44%. This suggests that sense definitions are the best option when they are available, but when they are not, images are good to use too. Images were also given a higher mean enjoyability rating than definitions in post-study survey questions.

Automatically selecting good images that reflect different word senses is still an open problem. As part of this study, we also report on how we found images from image search engines using synonyms and related words, and the lessons learned from the experience. This can help inform future work in development of human-assisted image-based disambiguation setups for systems like PanMail that translate to the hundreds of more minor languages that normally do not receive as much attention.

ACKNOWLEDGEMENTS

We thank James Fogarty, Mausam and Stephen Soderland for helpful comments and suggestions.

REFERENCES

[Alm et al., 2006] C.O. Alm, N Loeff and D. Forsyth. Challenges for Annotating Images for Sense

Disambiguation. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, 2006.

[Barnard et al., 2003] K. Barnard, M. Johnson and D. Forsyth. Word sense disambiguation with pictures. In *Proceedings of HLT-NAACL*, 2003.

[Colowick and Pool, 2007] S.M. Colowick and J. Pool, Disambiguating for the web: a test of two methods, In *Proceedings of the 4th International Conference on Knowledge Capture (KCAP)*, 2007.

[Copestake et al., 1994] A. Copestake, T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodriguez and A. Samiotou. Acquisition of lexical translation relations from MRDs. In *Machine Translation* vol 3 no 3-4, pg. 183-219, 1994.

[Etzioni et al., 2007] O. Etzioni, K. Reiter, S. Soderland and M. Sammer. Lexical Translation with Application to Image Search on the Web. In *Proceedings of Machine Translation Summit XI.*, 2007.

[Helmreich, 1993]. S. Helmreich, L. Guthrie and Y. Wilks. The Use of Machine Readable Dictionaries in the Pangloss Project. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*, 1993.

[Hull, 1996] D. A. Hull and G. Grenfenstette, Querying across languages: a dictionary-based approach to multilingual information retrieval. In *ACM SIGIR*. pg. 49-57, 1996.

[Ide, 1998] N. Ide and J. V'eronis. Introduction to the special issue on Word Sense Disambiguation: The state of the art. In *Computational Linguistics* vol. 24 no. 1 pg. 2-40. MIT Press, Cambridge, MA, March 1998.

[Sammer et al., 2007] M. Sammer, K. Reiter, S. Soderland, K. Kirchhoff and O. Etzioni. Ambiguity Reduction for Machine Translation: Human-Computer Collaboration. *Assoc. for Machine Translation in the Americas (AMTA)*, 2006.

[von Ahn and Dabbish, 1994] L. von Ahn and L. Dabbish, Labeling Images with a Computer Game. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2004.

[Yarowsky, 1994] D. Yarowsky. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of ACL*, 1994.

[Yarowsky, 1995] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of ACL*, 1995.