Lexical Translation with Application to Image Search on the Web

Oren Etzioni, Kobi Reiter, Stephen Soderland, and Marcus Sammer

Turing Center
Dept. of Computer Science and Engineering
University of Washington, Seattle, WA 98195 USA
{etzioni, dbgalur, soderlan, sammer} @cs.washington.edu

Abstract

We introduce a novel approach to the task of lexical translation. We utilize the *translation graph*, a massive lexical resource where each node denotes a word in some language and each edge denotes a word sense shared by a pair of words. Our current graph contains 1,267,460 nodes and 2,315,783 edges. The graph is automatically constructed from machine-readable dictionaries and Wiktionaries. Paths through the graph suggest word translations absent from any of the input dictionaries. We define a probabilistic inference procedure that enables us to quantify our confidence in a translation derived from the graph, and thus trade precision against recall.

We demonstrate the graph's utility by employing it in the PANIMAGES cross-lingual image search engine. Google retrieves images based on the words in their "vicinity", which limits the ability of a searcher to retrieve them. Although images are universal, an English searcher will fail to find images tagged in Chinese, and conversely. PANIMAGES addresses this problem by translating and disambiguating queries, using the translation graph, before sending them to Google. Our experiments show that, for queries in "minor" languages, PANIMAGES increases the number of correct images in the first 15 pages of results by 75%.

1 Introduction

Lexical translation is the task of translating individual words or phrases, either on their own (e.g., search-engine queries or meta-data tags) or as part of a knowledge-based Machine Translation (MT) system. In contrast with statistical MT, lexical translation does not require aligned corpora as input. Because large aligned corpora are non-existent for many language pairs, and are very expensive to generate, lexical translation is possible for a much broader set of languages than statistical MT.

While lexical translation has a long history (*cf.* (Helmreich *et al.*, 1993; Copestake *et al.*, 1994; Hull and Grefenstette, 1996)), interest in it peaked in the 1990's. Yet, as this paper shows, the proliferation of Machine-Readable Dictionaries (MRDs) and the rapid growth of multi-lingual Wiktionaries offers the opportunity to scale lexical translation to an unprecedented number of languages. Moreover, the increasing international adoption of the Web yields opportunities for new applications of lexical translation systems.

This paper presents a novel approach to lexical translation based on the *translation graph*. A node in the graph represents a word in a particular language, and an edge denotes a word sense shared between words in a pair of languages. Our TRANSGRAPH system automatically constructs a graph from a collection of independently-authored, machine-readable bilingual dictionaries and multi-lingual Wiktionaries as described in Section 2. Figure 1 shows an example translation graph.

When all the edges along a path in the translation graph share the same word sense, then the path denotes a correct translation between its end points. When word senses come from distinct dictionaries, however, we are uncertain about whether the senses are the same or not. Thus, we define an inference procedure that computes the probability that two edges denote the same word sense and use this probability, coupled with the structure of the graph, to compute the probability that a path denotes a correct translation.

Before we consider lexical translation in more detail, we need to ask: is lexical translation of any practical utility? While it does not solve the full machine-translation problem, lexical translation is valuable for a number of practical tasks including the translation of search queries, meta-tags, and individual words or phrases. For example, Google and other companies have fielded WordTranslator tools that allow the reader of a Web page to view the translation of particular word, which is helpful if you are, say, a Japanese speaker reading an English text and you come across an unfamiliar word.

In the case of image search, the utility of lexical translation is even more readily apparent. Google retrieves images based on the words in their "vicinity", which limits the ability of a searcher to retrieve them. Although images are universal, an English searcher will fail to find images tagged in Chinese, and a Dutch searcher will fail to find images tagged in English. To address this problem, we built the PANIMAGES cross-lingual image search engine.[1] PANIMAGES enables searchers to translate and disambiguate their queries before sending them to Google. PANIMAGES utilizes the translation graph; thus it also enables us to evaluate the quality of translations inferred from the graph in the context of a practical application.

The key contributions of the paper are as follows:

- We introduce the *translation graph*, a unique, automatically-constructed lexical resource, which currently consists of over 1.2 million words with over 2.3 million edges indicating possible translations.

[1] cs.washington.edu/research/panimages

- We formalize the problem of lexical translation as probabilistic inference over the translation graph and quantify the gain of inference over merely looking up translations in the source dictionaries.
- We identify a set of challenges in searching the Web for images, and introduce PANIMAGES, a cross-lingual image search application that is deployed on the Web to address these challenges.
- We report on experiments that show how PANIMAGES substantially increases image precision and recall for queries in "minor" languages, thereby demonstrating the utility of PANIMAGES and the translation graph.

The remainder of the paper is organized as follows. Section 2 introduces the translation graph. Section 3 describes PANIMAGES, our cross-lingual image search application. Section 4 reports statistics on the translation graph and evaluates the utility of the graph by reporting on the precision and recall of the PANIMAGES application. Section 5 discusses related work, followed by conclusions and future work in Section 6.

2 The Translation Graph

This paper introduces a novel lexical resource, which we call the *translation graph*. This section describes how the TRANSGRAPH system constructs a graph from multiple dictionaries, and uses paths in the graph to infer lexical translations.

Each node n in the graph is an ordered pair (w, l) where w is a word in a language l. An edge in the graph between (w_1, l_1) and (w_2, l_2) represents the belief that w_2 is a translation into l_2 of a particular sense of the word w_1. The edge is labeled by an integer denoting an ID for that word sense. Paths through the graph represent correct translations so long as all the edges on the path share a single word sense, and thus enable TRANSGRAPH to identify translations that are absent from any of its source dictionaries.

Figure 1 shows a portion of a translation graph for two senses of 'spring' in English. The graph also shows two corresponding French words 'printemps' (spring season) and 'ressort' (flexible spring).

TRANSGRAPH builds the translation graph incrementally on the basis of entries from multiple, independent dictionaries, as described in detail in Section 2.1. As edges are added on the basis of entries from a new dictionary, some of the new word sense IDs are redundant because they are equivalent to word senses already in the graph from another dictionary. For example, TRANSGRAPH assigns one word sense ID to the seasonal sense of 'spring' from an English dictionary, a new word sense ID to the French dictionary entry for 'printemps', and so forth (see labels '1' and '3' in Figure 1). We refer to this phenomenon as *sense inflation*.

Sense inflation would severely limit the utility of the translation graph, so we have developed a mechanism for identifying duplicate word senses automatically. TRANSGRAPH computes the probability $prob(s_i = s_j)$ that a pair of distinct IDs s_i and s_j refer to the same word sense (see Section 2.1 for the details). Thus, TRANSGRAPH determines that word sense ID '3' on edges from 'printemps' has a high probability of being equivalent to ID '1'.

The following section discusses building the graph, focusing on the algorithm for merging word senses originating from different dictionaries.

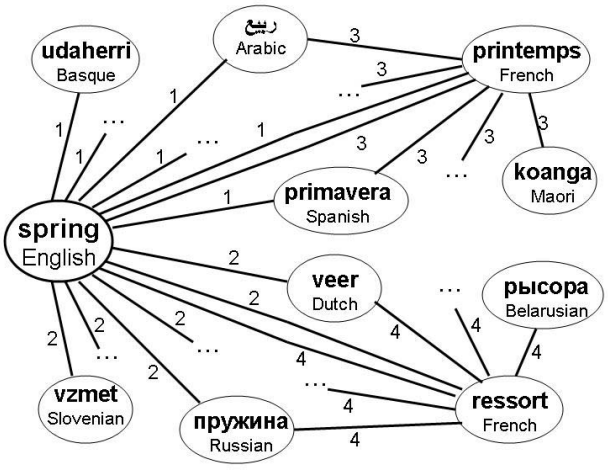


Figure 1: A fragment of a translation graph for two senses of the English word 'spring'. Edges with the label '1' or '3' are for spring in the sense of a season; edges labeled '2' or '4' are for the flexible coil sense. This graph shows translation entries from an English dictionary merged with translation entries from a French dictionary.

2.1 Building the Translation Graph

TRANSGRAPH builds the translation graph from online dictionaries and Wiktionaries of two kinds: bilingual dictionaries that translate words from one language to another, and multilingual dictionaries that translate words in a source language to multiple target languages. Some dictionaries provide separate translations for each distinct word sense, which is particularly helpful for our purposes, but others do not.

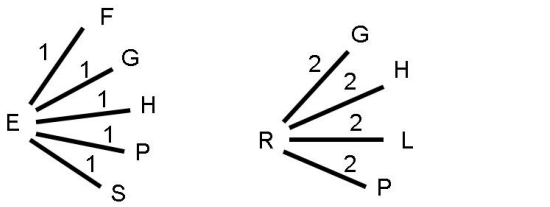
As TRANSGRAPH adds to the graph from each new entry in a dictionary, it assigns a new, unique word sense ID for each word sense in that entry. Thus, edges for translations of the season 'spring' from the English dictionary have one word sense ID, edges for translations of the flexible coil 'spring' have a different word sense ID, and so forth. When the translation in the entry is not word-sense distinguished, TRANSGRAPH makes the conservative assumption that each translation is in a distinct word sense. Section 2.2 explains how we recover from word sense inflation caused by this assumption and from integrating multiple dictionaries.

We implement the translation graph as a relational database. Each row in the *Translation table* represents an edge in the graph, while each row in the *Word sense equivalence table* represents the probability, $prob(s_i = s_j)$, that two word sense IDs s_i and s_j are equivalent.

2.2 Word-Sense Equivalence

As pointed out earlier, accumulating entries from multiple dictionaries results in sense inflation. Below, we explain how TRANSGRAPH addresses this problem by computing word-sense equivalence probabilities of the form $prob(s_i = s_j)$.

Figures 2 and 3 give a schematic illustration of how TRANSGRAPH accumulates entries from multiple dictionaries. Figure 2 shows graph edges from an entry for the word E from an English dictionary that gives translations into French, German, Hungarian, Polish, and Spanish. TRANSGRAPH assigns the word sense ID 1 for these edges. This figure also shows edges from an entry for word



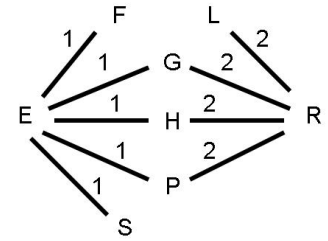
Entry for word E from an English dictionary

Entry for word R from a Russian dictionary

Figure 2: Schematic diagram of edges from an entry for the word E from an English dictionary and edges from an entry for the word R from a Russian dictionary.

R from a Russian dictionary, which in this case has translations into German, Hungarian, Latvian, and Polish. These edges are assigned word sense ID 2.

Figure 3 shows the situation after both sets of edges have been added to the translation graph. There are 6 nodes with edges labeled with word sense ID 1, { E, F, G, H, P, S }; 5 nodes with edges labeled 2, { G, H, L, P, R }; and an intersection of these sets comprising 3 nodes, {G, H, P}. The three nodes in the intersection have two incident edges with distinct sense IDs 1 and 2. The proportion of intersecting nodes provides evidence that these IDs refer to the same word sense.



nodes(1) = { E, F, G, H, P, S }
nodes(2) = {G, H, L, P, R }
nodes(1) ∩ nodes(2) = {G, H, P }

overlap(1,2) = max (3/5 , 3/6) = 0.6

Figure 3: After the entries from Figure 2 have both been added to the graph, the set of nodes with word sense ID 1 overlaps with the set of nodes for word sense ID 2. The proportion of overlapping nodes gives evidence that the two word senses may be equivalent.

TRANSGRAPH determines the probability that two word sense IDs s_i and s_j are equivalent as follows:

- A word sense is equivalent to itself: $prob(s = s) = 1$.
- If s_i and s_j are alternate word senses from the same entry in a sense-distinguished dictionary, then they are assumed to be distinct: $prob(s_i = s_j) = 0$.
- If word senses s_i and s_j have at least k intersecting nodes, then set the probability by equation 1 below.
- In all other cases, the probability is undefined.

TRANSGRAPH estimates the probability that s_i and s_j are equivalent word senses by the following equation.

If $|nodes(s_i) \cap nodes(s_j)| \geq k$, then:

$$prob(s_i = s_j) = max(\frac{|nodes(s_i) \cap nodes(s_j)|}{|nodes(s_i)|}, \frac{|nodes(s_i) \cap nodes(s_j)|}{|nodes(s_j)|}) \quad (1)$$

where nodes(s) is the set of nodes that have edges labeled by word sense ID s, and k is a sense intersection threshold.

As an example of computing the probability of word sense equivalence, our translation graph has 56 translations for the season sense of 'spring' from an English dictionary, and 12 translations for 'printemps' from a French dictionary. 8 of these translations overlap, giving a probability of $\frac{8}{12} = 0.67$ that the two senses are equivalent.

2.3 Computing Translation Probabilities

Given the translation graph coupled with the word sense equivalence probabilities, TRANSGRAPH can compute the probability that a particular word is a translation of another word in a given word sense. First, we show how to compute the probability of a single translation path. Then, we show how we combine evidence across multiple paths.

Consider a single path P that connects node n_1 to n_k, where n_i is the word w_i in language l_i and the ith edge has word sense s_i. Let $pathProb(n_1, n_k, s, P)$ be the probability that (w_1, l_1) is a correct translation of (w_k, l_k) in word sense s, given a path P connecting these nodes.

The simple case is where the path is of length 1. If s is the same sense ID as s_1, then the probability is simply 1.0; otherwise it is the probability that the two senses are equivalent:

$$pathProb(n_1, n_2, s, P) = prob(s = s_1) \quad (2)$$

Where the path P has more than one edge, the path probability is reduced by $prob(s_i = s_{i+1})$ whenever the word sense ID changes along the path. We make the simplifying assumption that sense-equivalence probabilities are mutually independent. Formally, this gives the term

$$\prod_{i=1...|P|-1} prob(s_i = s_{i+1}).$$

If the desired sense s is not found on the path, we also need to factor in the probability that s is equivalent to at least one sense s_i on the path, which we approximate by the maximum of $prob(s = s_i)$ over all s_i. Formally, this gives the term

$$\max_{i=1...|P|}(prob(s = s_i)),$$

which is equal to 1.0 if s is found on path P.

Putting these two terms together, we have the following formula for simple paths of length greater than one (*i.e.*, $|P| > 1$):

$$pathProb(n_1, n_k, s, P) = \max_{i=1...|P|}(prob(s = s_i)) \times \prod_{i=1...|P|-1} prob(s_i = s_{i+1}) \quad (3)$$

Note that we disallow paths that contain non-consecutive repetition of sense IDs (*e.g.*1, 2, 1).

There are typically multiple paths from one node to another in the translation graph. The simplest way to compute $prob(n_1, n_k, s)$ is to take the maximum probability of any path between n_1 and n_k.

$$prob(n_1, n_k, s) = \max_{P \in paths}(pathProb(n_1, n_k, s, P)) \quad (4)$$

We also experimented with another method that gives higher probability if there are multiple, distinct paths between words. We define two paths from n_1 to n_k to be distinct if there is a distinct sequence of unique word sense IDs on each path.

We use the standard Noisy-Or model to combine evidence. The basic intuition is that translation is correct unless every one of the translation paths fails to maintain the desired sense s. We multiply the probability of failure for each path. We then subtract that probability from one to get the probability of correct translation. The probability that n_1 is a correct translation of n_k in word sense s is:

$$prob(n_1, n_k, s) = 1 - \prod_{P \in distinctP} (1 - pathProb(n_1, n_k, s, P)) \quad (5)$$

where $distinctP$ is the set of distinct paths from n_1 to n_k.

We found that our current implementation of the Noisy-Or model tends to give inflated probability estimates, so we use the maximum path probability in experiments reported here. Defining distinct paths as those with distinct sense IDs in not sufficient to ensure that paths are based on independent evidence. We are exploring better methods for determining independent paths, and more sophisticated probability models to combine evidence.

2.4 Confidence in Dictionary Entries

Our methods for computing translation probabilities have, thus far, made a strong assumption. We assume that each word sense ID comes from a sense-distinguished dictionary entry. This means that $nodes(s_i)$, the set of nodes with edges to sense s_i, are mutual translations of each other in the same sense.

We found that many of the errors in computing $pathProb(n_1, n_k, s, P)$ are from cases where this assumption is violated by some word sense ID along the path. If all words in the set $nodes(s_i)$ do not share the same sense, any path that passes through sense s_i may result in translation errors.

These "impure" word sense IDs may arise either from errors in a dictionary or from errors parsing the dictionary. As an example, the French Wiktionary has an entry for the word 'boule' with English translations as 'ball', 'boule', 'bowl', 'chunk', 'clod', and 'lump'. These are all good translations of 'boule', but clearly not all in the same sense. An example of a parsing error is the truncation of translation phrases in some dictionary entries, causing bizarre translations.

To compensate for these impure sense IDs, we have begun experimenting with methods to compute $prob(s_i)$, the probability that all words in $nodes(s_i)$ share a common word sense. This adds the term $prob(s_1)$ to Equations 2 and 3, and adjusts Equation 3 to include $prob(s_{i+1})$ for each new sense s_{i+1} along the path.

The *a priori* probability for $prob(s_i)$ is set according to a global confidence in the dictionary. If the dictionary has a high ratio of word senses per entry, the assumption is that the dictionary entries distinguish word senses, and the default $prob(s_i)$ is set to 1.0.

The existence of multiple, possibly non-synonymous translations into the same language lowers our confidence that a dictionary entry is pure. While it is possible to find evidence that two words are synonyms, determining that they are *non-synonymous* is more difficult. We found that even English WordNet is not a strong source of evidence for non-synonymy. Of the cases where $nodes(s_i)$ includes two English translations that are not WordNet synonyms, they were actually synonymous about half the time. Our preliminary experiments indicate that even crude estimation of $prob(s_i)$ can improve the precision of translation graph traversal. The results shown in Section 4 include a early attempt to estimate $prob(s_i)$.

2.5 Bilingual Dictionaries

The method for computing word-sense equivalence discussed in 2.2 relies on having multiple translations for each word sense. Unfortunately, we do not always have this luxury. In response, we have identified cliques in the graph as an additional structure that helps to combat sense inflation.

Consider, for example, the simple clique shown in Figure 4. The figure shows a 3-node clique where each of the edges was derived from a distinct dictionary, and hence has a distinct word sense ID. The edge from (spring, English) to (printemps, French) is labeled '1' and comes from an entry for the season of spring from the English Wiktionary. The edge '2' from (xuân, Vietnamese) to (spring, English) is from a Vietnamese-English dictionary that does not specify which sense of spring is intended. The edge '3' from (xuân, Vietnamese) to (printemps, French) is from a Vietnamese-French dictionary, again without any indication of word sense.

It has long been known that this kind of *triangulation* gives a high probability that all three words share a common word sense (Gollins and Sanderson, 2001). We empirically estimated the probability that all three word sense IDs of a 3-node clique are equivalent to be approximately 0.80 in our current translation graph. The TRANSGRAPH compiler finds all cliques in the graph of size 3 where two word senses are from bilingual dictionaries. It then adds an entry to the word sense equivalence table with probability 0.80 for each pair of sense IDs in the clique. We plan to investigate longer cliques and evidence from other elements of graph structure.

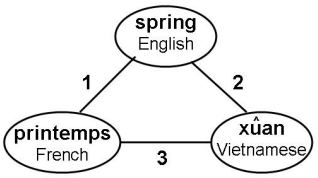


Figure 4: TRANSGRAPH infers that word senses are equivalent with high probability when nodes form a 3-node clique in the graph. In this example the Vietnamese word 'xuân' is translated to English 'spring' and French 'printemps' in the season sense of spring.

3 Image Search with PanImages

We now turn to a discussion of the application of the translation graph to cross-lingual image search.

The Web has emerged as a rich source of images that serve a wide range of purposes from children adorning their homework with pictures to anthropologists studying cultural nuances. Most people find images on the Web by querying an image search engine such as Google's.

Google collects images as part of its crawl of the Web and tags them with the words that appear in their vicinity

on the crawled HTML documents and links. It is not surprising that most of the tags are in "major" languages such as English. So while images are universal, most of them can be found through Google only if you can query in the "right" language.

More broadly, monolingual image search engines face the following challenges:

- **Limited Resource Languages** - The lower the Web presence of a language, the fewer hits a speaker of that language gets from a query. A query for 'grenivka' (Slovenian for 'grapefruit') produces only 24 results, of which only 9 are images of grapefruits. Yet translating the query into English produces tens of thousands of images with high precision.
- **Cross-Cultural Images** - Results of an image search may vary considerably depending on the language of the query term. Translating the query 'baby' or 'food' into Chinese, Arabic, or Zulu allows an interesting cultural comparison.
- **Cross-Lingual Masking** - A word in one language is often a homonym for an unrelated word in another language. Relevant results can be swamped by results for the unrelated word. The Hungarian word for tooth happens to be 'fog'; the only way to get images of teeth rather than misty weather is to query with a translation that doesn't suffer from cross-lingual masking.
- **Word Sense Ambiguity** - Searching for an image that corresponds to a minor sense of a word is problematic. Most results for the query 'spring' are images of flowers and trees in bloom. If a user wants images of flexible coils or of bubbling fountains, the most effective queries are translations of this sense of 'spring' into languages where that word is not ambiguous.

PANIMAGES, a cross-lingual image-search application deployed on the Web, enables a monolingual user to select from any of 50 input languages, automatically looks up word-sense specific translations into more than 100 languages, and lets the user control which translations are sent to an image search engine. At compile time, PANIMAGES merges information from multiple Wiktionaries and open-source dictionaries into a translation graph as described in Section 2. At run time, PANIMAGES accepts a query from a user, presents the user with possible translations found in the translation graph, then sends the translations selected by the user to Google's image search as described below.

This section describes the implementation and interface of our PANIMAGES system for cross-lingual image search, accessible at `www.cs.washington.edu/research/panimages`. Figure 5 shows the system architecture.

3.1 Interface Design

The PANIMAGES graphical user interface allows a user to enter a search query in any of n source languages ($n = 50$ currently). PANIMAGES presents translations of the query term, presenting multiple sets of translations if the graph has multiple senses of the term. The user selects one or more translations, and PANIMAGES sends this as a query to Google Images.

Finding Translations:
PANIMAGES looks up the node (w_i, l_i) in the translation graph that corresponds to the query word and language, then follows edges in the graph to create one or more sets

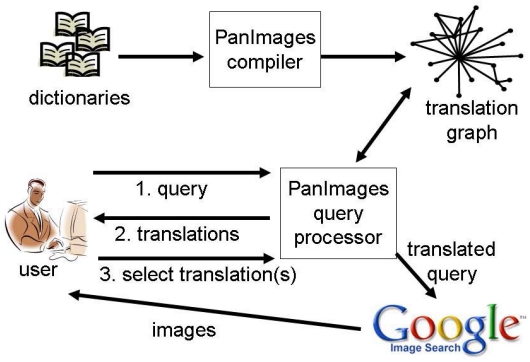


Figure 5: Architecture: The PANIMAGES compiler creates a translation graph from multiple dictionaries. The query processor takes a user query and presents a set of translations. The user selects the desired translation(s), which PANIMAGES sends to Google Image Search.

of nodes (w_j, l_j) where w_j is a translation into l_j for a particular sense of w_i. For each word sense, PANIMAGES follows paths of length up to k in which the probability that the word sense has not changed according to Equation 4 is above a threshold τ. In our experiments we set k to 3 and τ to 0.2.

In the example in Figure 1 for the English word 'spring', translations in sense 1 include nodes reachable from sense 1 and nodes reachable from (printemps, French) along edges for sense 3. Beginning from 'spring' with sense 3 and continuing on paths for sense 1 or 3 produces an identical set of translations that TRANSGRAPH later merges with translations for sense 1.

Presenting Translations to the User:
PANIMAGES presents these sets of translations and allows the user to select one or more translation to be sent to Google Images. As a practical consideration, PANIMAGES defaults to selecting translations in a language with high Web presence: an English translation for all source languages but English, and a French translation for English queries. The user may add or remove any of the translation-language pairs to the query before clicking on *Show Images*. Another option is to click on a single translation to immediately send that translation as a query to Google's image search.

Handling Word Senses:
PANIMAGES lists each distinct word sense along with a gloss if available and the number of translations for this word sense. The user can click on a word sense to see the list of translations for that sense. PANIMAGES presents the word sense with the largest number of translations first, and selects this as the default word sense.

4 Experimental Results

This section presents statistics on our current, automatically constructed translation graph; reports on an evaluation of translation inferences over the graph; and reports on recall and precision results from a sample of image search queries over this translation graph.

4.1 Graph Statistics

The translation graph is composed of 1,267,460 words in more than 100 languages. 3 of the languages have over

100,000 words and 58 of the languages have at least 1,000 words. The words were extracted from 3 multilingual dictionaries (English and French Wiktionaries, and an Esperanto dictionary) and 14 bilingual dictionaries, giving a total of 2,315,783 direct translations or edges in the graph. Further translations can be found from graph paths with length greater than one edge.

Building a translation graph from a combination of these dictionaries provides more translations than any of these dictionaries alone. The English Wiktionary had translations for 19,500 words – after adding the other dictionaries, the graph has translations for over 255,000 English words and phrases, the bulk of them from bilingual dictionaries. Similarly, coverage of French went from 12,700 words in the French Wiktionary to 32,800 in the graph.

4.2 Evaluating Inferred Translations

We evaluated the precision and recall gain from inference using Equations 1 through 4 as follows. We took a random set of 1,000 English words and found Hebrew or Russian translations using the translation graph. We also took a random set of 1,000 Turkish words and found Russian translations.[2] The set of random words was not weighted by word frequency, thus they contained many relatively obscure words (*e.g.*, abashment, abjectly, Acrididae, 'add up') for which no translation was found in the target language.

The baseline is the number of words in the source language that can be translated using only direct edges in the graph. We then added inferred translations that can be made from a single application of the word sense equivalence equation (Equation 1) with k set to 2 at a probability threshold of 0.2. Finally, we found all inferred translations using Equations 1 - 4 and using graph paths from all 17 source dictionaries with path length up to 3 word sense IDs at a probability threshold of 0.2.

Figures 6 through 8 compare the number of words translated and the proportion of correct translations. The total height of each bar represents the number of source language words that have at least one translation. We measure precision as the number of correct translation pairs divided by the number of translation pairs that the system outputs. Note that precision is computed over *all* translations for a given word, some of which may be correct and others may be erroneous.

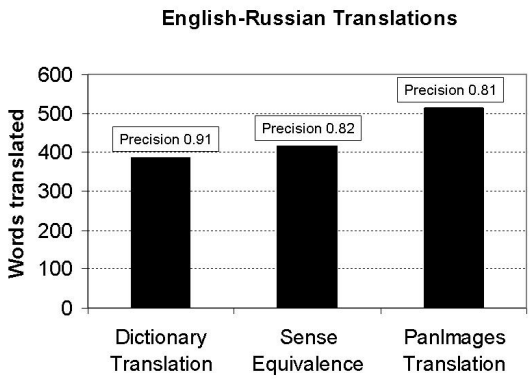


Figure 6: A comparison of direct vs. inferred translations from English to Russian. Inference from graph traversal boosted the number of translated words by 33% with a modest drop in precision.

[2]We chose these language pairs because we were able to find bilingual speakers who would evaluate the results for us.

The language pair English-Russian is an interesting test of the translation graph, because we had neither a bilingual nor a multilingual Russian dictionary. The only edges to Russian words came from one of the multilingual dictionaries. The baseline accuracy of the source dictionaries is 91%. Adding inferences from Equation 1 gave an 8% increase in translated words with a drop in precision to 82%. There was a greater gain from combining all translation paths in the graph – 33% more translated words than the baseline with negligible further drop in precision.

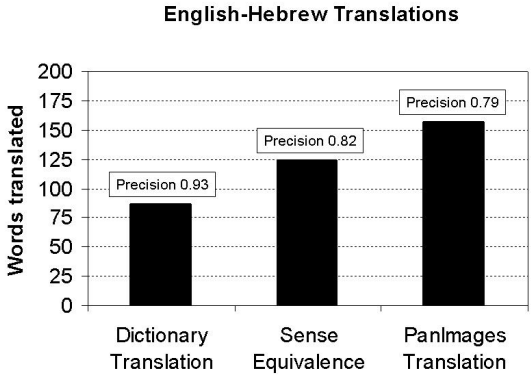


Figure 7: Graph traversal increased the number of translations from English to Hebrew by 80%, again with a modest drop in precision.

Like Russian, there are no bilingual dictionaries for Hebrew and no Hebrew multilingual dictionary. Inference based on Equation 1 boosts translated words by 43% and using all translation paths gives a gain of 80% over the baseline. The baseline precision drops from 93% to 79%.

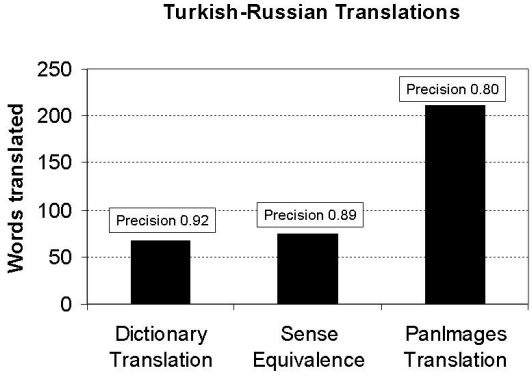


Figure 8: Translation from Turkish to Russian benefited from interaction between several bilingual dictionaries, resulting in 3.15 times as many translated words as the baseline.

Translations from Turkish to Russian showed a large gain from inferences based on bilingual dictionaries. While direct edges came only from the three multilingual dictionaries, there were also three bilingual dictionaries between Turkish and English, German, or Kurdish. In turn, these dictionaries interacted with other bilingual dictionaries for English, German, and Kurdish. Inference from all paths resulted in a three-fold increase in translated words, while maintaining high precision (80%).

In summary, we see that inference over the translation graph yields a tradeoff between translation coverage and precision. We can control the tradeoff using the probability threshold—lowering the threshold increases coverage

but reduces precision. In the Web image retrieval context, where precision is already far-from-perfect, the tradeoff seems like a good one, particularly for the numerous "minor" languages where few images are returned in response to many queries.[3] Finally, we anticipate that as we add dictionaries to TRANSGRAPH, and as Wiktionaries grow in size, both coverage and precision will increase in tandem.

4.3 Image Retrieval Performance

We also evaluated coverage and precision of PANIMAGES image search for non-English queries, comparing the results of sending the non-English query directly to Google Image search with the results of sending the default PANIMAGES translation instead. We chose a limited test set of languages and words to limit the manual tagging effort necessary for the experiment.

To generate our test set of words, we selected 10 arbitrary concepts that are associated with distinctive images, 6 nouns (ant, clown, fig, lake, sky, train), 2 verbs (eat, run), and two adjectives (happy, tired). Next, we selected 32 languages with a limited Web presence ranging from Danish and Dutch to Telugu and Lithuanian.[4] Now, for each concept, we chose 1/4 of the languages at random, and recorded the word for the concept in the language. These 80 words became our set of non-English queries. We then compared precision and recall of Google's image search for these 80 words "as is" with the precision and recall of Google's image search for these words translated by PANIMAGES into English.

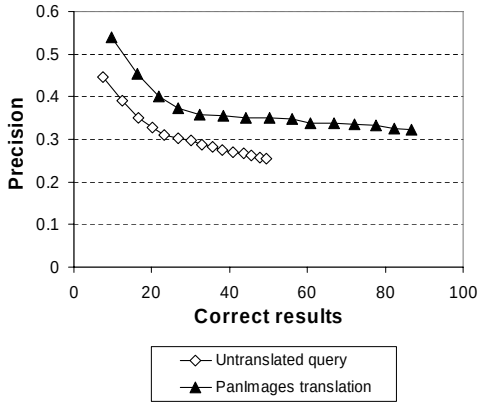


Figure 9: Image search results for random words in 32 languages with a limited Web presence. The PANIMAGES translation into English increases correct results by 75% from an average of 49.6 correct results on the first 270 images (the rightmost white diamond in the graph) to 86.8 for PANIMAGES (the rightmost black triangle in the graph). In addition, PANIMAGES boosts precision by approximately 27% throughout the graph.

Translating the queries from minor languages into a major language gives a large boost in recall. The average number of results as estimated by Google was 33,000 for minor language queries and 1,856,000 for the queries translated by PANIMAGES, a 57-fold increase. For 10% of the minor language queries Google failed to return *any* images.

[3] An experiment in Section 4.3 shows that 10 % of the "minor' language queries in our experiments returned no results whatsoever!

[4] We selected languages that had words for these concepts present in our translation graph.

More significant is the number of correct results found in the first 15 pages (containing 270 images). Here the PANIMAGES translation resulted in a 75% gain over the untranslated query, from an average of 49.6 correct results to an average of 86.8. Average precision also rose 27% from 0.25 to 0.32. The main cause of low precision for the minor language queries was cross-lingual masking. The query term was a homonym of a completely unrelated word in a major language.

5 Related Work

There was considerable research in the 1990's on methods to acquire translation lexicons for knowledge-based MT (Neff and McCord, 1990; Helmreich *et al.*, 1993; Copestake *et al.*, 1994). Many of these systems used MRDs to assist manual creation of lexicons, or used automated acquisition with post editing. Despite the shift in emphasis towards statistical MT, research on knowledge-based MT has continued, with its need for lexicon acquisition.

Translation lexicons are also a vital resource for cross-lingual information retrieval (CLIR), a subfield prompted in part by the TREC conferences (Harman, 1996) and a series of SIGIR CLIR workshops (Gey *et al.*, 2006). Surveys of CLIR research may be found in (Oard, 1997) and (Kishida, 2005). Much of the CLIR research, in contrast to PANIMAGES, has focused on a small number of language pairs, much of it building systems that must be adapted to one language pair at a time.

While early CLIR systems typically relied on bilingual dictionaries (Hull and Grefenstette, 1996), corpus-based methods or hybrid methods soon outstripped purely dictionary-based systems (Yang *et al.*, 1998). Methods that derive word-translations from parallel text include (Gale and Church, 1991; Fung, 1995; Melamed, 1997; Franz *et al.*, 2001). There are also hybrid systems (Ballesteros and Croft, 1998; Sharoff *et al.*, 2006) that use corpus-based techniques to disambiguate translations provided by bilingual dictionaries.

The main drawback of using bilingual dictionaries, in past work, has been word-sense ambiguity. A single term in the source language is typically translated into multiple terms in the target language, mixing different word senses. Combining information from multiple bilingual dictionaries only exacerbated this problem: translating from language l_1 into l_2 and then translating each of the possible l_2 translations into a third language l_3, quickly leads to an explosion of translations. But this is exactly the problem that the translation graph inference mechanism solves. This use of inference to reduce ambiguity is a key capability of PANIMAGES that makes it more powerful than any collection of dictionaries and Wiktionaries.

The ability of TRANSGRAPH to automatically determine distinct senses of a word is similar to the work on "semantic mirrors" (Dyvik, 2004). Dyvik clusters translations between Norwegian and English to find alternate senses of words in either language.

On the Web, commercial search engines such as Google, French Yahoo (`http://fr.yahoo.com`) and German Yahoo (`http://de.yahoo.com`), offer query translation capability for only a handful of languages. For example, French and German Yahoo automatically translate query terms into any of several major languages using Systran (`http://www.systran.com`) and translate the resulting Web pages.

In contrast, PANIMAGES translates between a large number of languages, and infers word-sense preserving translations that are not found in any single dictionary. PANIMAGES's translation graph is also a platform for plugging in more and more dictionaries, increasingly comprehensive Wiktionaries, and corpus-based translations, all of which will lead directly to improved cross-lingual image search over time.

6 Conclusions and Future Work

The recent proliferation of bilingual MRDs and multilingual Wiktionaries is a valuable resource for acquisition of translation lexicons. Yet, difficulties arise in making use of these lexical resources. Most bilingual dictionaries do not distinguish between word senses, giving instead a list of translations of all senses of a source word. Many MRDs have only spotty coverage.

To address these issues, we introduced the *translation graph*, which combines multiple independently-authored MRDs and Wiktionaries. Our TRANSGRAPH system has built a graph with over 1.2 million words in more than 100 languages and over 2.3 million edges that represent translations between words. TRANSGRAPH provides a probabilistic inferencing mechanism over the graph that can infer translation pairs that are not found in any of the source dictionaries. Our experiments found that using the graph to infer translations not found in any single dictionary can more than triple the number of translated words, while maintaining high precision (0.80).

This paper also introduces PANIMAGES, a fully-implemented cross-lingual image search system for the Web based on the translation graph. Our experiments show that, for queries in languages with a limited Web footprint, PANIMAGES increases the total number of results 57-fold (from 33,000 to 1,856,000). PANIMAGES increases the number of correct images by 75% on the first 15 pages (containing 270 images), while increasing precision by 27%.

Our future work includes expanding the coverage of the translation graph by increasing the number of source dictionaries, and re-parsing Wiktionaries that have grown in coverage. We have recently added 10 more Wiktionaries and are in the process of adding 20 more bilingual dictionaries to the translation graph.

We are also exploring ways to improve translation precision with better estimates of $prob(s_i)$ and $prob(s_i = s_j)$ and exploring probabilistic models of how to combine evidence from multiple graph paths. Informal evaluation on our latest translation graph shows higher precision than the results presented here.

In future work, we plan to apply the translation graph to tasks other than image search, including the translation of tags in social tagging systems such as *del.icio.us* and in on-line games such as von Ahn's "ESP game"

Acknowledgments

This research was supported by a gift from the Utilika Foundation to the University of Washington's Turing Center. We thank Ethan Phelps-Goodman, Doug Downey, and Jonathan Pool for helpful comments and thank Jonathan Pool and Julia Schwarz for help with evaluating translation accuracy.

References

(Ballesteros and Croft, 1998) Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *ACM SIGIR*, 1998.

(Copestake *et al.*, 1994) A. Copestake, T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodriquez, and A. Samiotou. Acquisition of lexical translation relations from MRDs. *Machine Translation*, 3(3–4):183–219, 1994.

(Dyvik, 2004) H. Dyvik. Translation as semantic mirrors: from parallel corpus to WordNet. *Language and Computers*, 49(1):311–326, 2004.

(Franz *et al.*, 2001) M. Franz, S. McCarly, and W. Zhu. English-Chinese information retrieval at IBM. In *TREC 2001*, 2001.

(Fung, 1995) P. Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *ACL-1995*, 1995.

(Gale and Church, 1991) W. Gale and K.W. Church. A Program for Aligning Sentences in Bilingual Corpora. In *ACL-1991*, 1991.

(Gey *et al.*, 2006) F.C. Gey, N. Kando, C-Y. Lin, and C. Peters. New directions in multilingual information access: Introduction to the workshop at SIGIR 2006. In *Workshop on New Directions in Multilingual Information Access at SIGIR 2006*, 2006.

(Gollins and Sanderson, 2001) T. Gollins and M. Sanderson. Improving cross language retrieval with triangulated translation. In *SIGIR*, 2001.

(Harman, 1996) D. Harman. Overview of the Fourth Text Retrieval Conference (TREC-4). In *TREC-4*, 1996.

(Helmreich *et al.*, 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 and Grefenstette, 1996) D.A. Hull and G. Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *ACM SIGIR 1996*, pages 49–57, 1996.

(Kishida, 2005) K. Kishida. Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41:433–455, 2005.

(Melamed, 1997) I.D. Melamed. A Word-to-Word Model of Translational Equivalence. In *ACL-1997 and EACL-1997*, pages 490–497, 1997.

(Neff and McCord, 1990) M. Neff and M. McCord. Acquiring lexical data from machine-readable dictionary resources for machine translation. In *3rd Intl Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, 1990.

(Oard, 1997) D. Oard. Cross-language text retrieval research in the USA. In *3rd DELOS Workshop*, 1997.

(Sharoff *et al.*, 2006) S. Sharoff, B. Babych, and T. Hartley. Using comparable corpora to solve problems difficult for human translators. In *ACL/HLT*, 2006.

(Yang *et al.*, 1998) Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking. Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence*, 103(1-2):323–345, 1998.