Building a Sense-Distinguished Multilingual Lexicon from Monolingual Corpora and Bilingual Lexicons

Marcus Sammer and Stephen Soderland

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

Abstract

Both lexical translation and knowledge-based translation systems require sense-distinguished translation lexicons, yet such lexicons are expensive to create manually. However, the abundance of untagged monolingual corpora and the availability of bilingual, machine-readable dictionaries (MRDs) suggest an opportunity. Our PanLexicon system takes advantage of these resources to automatically construct a sense-distinguished multilingual lexicon. The challenge for PanLexicon is that free, bilingual MRDs do not make sense distinctions, and often have spotty coverage.

PanLexicon uses word contexts from monolingual corpora to guide it in finding translation sets – sets of words that share the same word sense across multiple languages. By maintaining word sense distinctions, PanLexicon finds translations between language pairs that are not supported by any of its bilingual source dictionaries. PanLexicon runs in time linear in the size of its input, and thus scales readily to large numbers of languages.

We built a prototype of PanLexicon with inputs from Spanish-English and Chinese-English dictionaries. Our initial experimental results show that PanLexicon is able to find high-quality translation sets despite the limitations of its inputs.

1. Introduction

Translation lexicons play a vital role in several applications related to machine translation (MT). Such lexicons are used for cross-language information search (Reiter et al., 2007; Gey et al., 2006; Hull and Grefenstette, 1996), for Web-based word translation tools (e.g. Google WordTranslator), and for knowledge-based MT systems (Bond et al., 2005; Carbonell et al., 2006).

To be useful for these tasks, a translation lexicon must distinguish which translations are appropriate to which word sense. Such lexicons exist for only a few language pairs with adequate coverage; creating them manually is a major knowledge acquisition bottleneck. A few multilingual lexical resources are under construction, such as EuroWordNet (Vossen, 1998) and the Wiktionary project (www.wiktionary.org), but scaling up these resources beyond a small number of languages is a laborious process.

On the other hand, some lexical resources are becoming more and more prevalent – vast amounts of monolingual text and bilingual machine readable dictionaries (MRDs) that do not distinguish word senses and often have spotty coverage.

We present the PanLexicon system that takes advantage of these plentiful resources to create a word-sense-distinguished multilingual lexicon in a fully automatic fashion. PanLexicon utilizes a combination of bilingual MRDs to find *translation sets*, where each translation set has one or more words in each of *k* languages that all represent the same word sense.

PanLexicon's source dictionaries alone are insufficient for maintaining word senses. They provide a set of translations in languages $L_2, \ldots L_k$ from a word in language L_1. There is no guarantee that a given translation in L_2 corresponds to a translation in L_3, since they may not each share the same word sense with the word from L_1.

PanLexicon maintains the same word sense across languages by finding word usage contexts from monolingual corpora and computing similarity of contexts across languages as described in Section 3.2. These contexts also serve the place of glosses in the lexicon, giving an example of the intended meaning to either a human reader or an automatic word sense disambiguation (WSD) tool.

We make the following contributions in this paper:

- We present PanLexicon, a scalable automated mechanism to build word-sense-distinguished multilingual lexicons from bilingual MRDs and monolingual corpora.
- We show that PanLexicon is scalable to a large number of languages with time and space linear in the number of languages and corpus size. Additional languages require only a bilingual lexicon with a hub language, a monolingual corpus, a word stemmer, and a stop word list.
- We evaluate performance of PanLexicon on English-Spanish-Chinese translation, and show how parameter settings can effect a tradeoff between coverage and precision.

The paper continues by describing the output of PanLexicon in Section 2 and the mechanics of creating a lexicon in Section 3. Section 4 presents an evaluation of its quality. After reviewing related work in Section 5, we conclude with some thoughts on future research.

2. An Example Lexicon

Before describing the details of PanLexicon, we present an example of the type of lexicon entry it creates. Figure 1 shows a PanLexicon translation set for the factory sense of "plant". A translation set is a multilingual extension of a WordNet synset (Fellbaum, 1998). It contains words that express a given sense in each of the *k* languages that comprise the lexicon, where each language may have multiple synonyms for that sense.

English		Spanish		Chinese	
plant	aluminum smelting **plant** that employs about 930 workers	**planta**	materiales nucleares de las **plantas** de energía para fabricar armas atómicas	厂	工人到**厂**里来, 就是来干活的
				厂房	该厂有8间**厂房**、5间仓库
factory	food warehouses, an insecticide plant and a fertilizers **factory**	**fábrica**	trabajadores de una **fábrica** privada estaban fundiendo pedazos de aluminio	工厂	生产车间作为**工厂**的 "特区"

Figure 1: Example lexicon entry for the concept of "industrial plant." This concept is expressed by two English words, two Spanish words, and three Chinese words, with a usage illustration for each word to indicate its meaning.

Each word in a translation set includes a number of contexts to illustrate the intended sense. Only the first context for each word is shown in Figure 1. PanLexicon may be applied to a set of *k* languages, where *k* is only limited by available resources, although the figure shows only three languages, English, Spanish, and Chinese.

The underlying meaning of "industrial plant" is expressed as "plant" and "factory" in English, as "planta" and "fábrica" in Spanish**,** and by "厂", "厂房", and "工" in Chinese.

We have three goals for each entry:

- **Intra-language consistency** – Words within a language should be synonyms. In the English portion of the example, the contexts of "plant" and "factory" make it clear that the words are synonyms in the sense of "industrial plant."
- **Inter-language consistency** – Words from different languages should be translations of each other, and their contexts should illustrate the same word sense.
- **Complete sense** – PanLexicon attempts to create complete entries. For example, it could be argued that our example entry is incomplete, because it is missing the English word "mill," which can also be used in the sense of "industrial plant."

3. Building the Lexicon

PanLexicon begins with lexical resources for a set of *k* languages, where one of the languages is designated as the *hub* language and the others as *spoke* languages. There is a monolingual corpus for each language and one or more bilingual dictionaries between the hub language and each of the spoke languages.

A preliminary step is to index each corpus, giving us efficient access to the contexts surrounding each word. PanLexicon iterates through each word *w* in the hub language and uses the bilingual dictionaries to find possible translations $t_1, \ldots t_n$ in the other languages. The bilingual dictionary also assists in translating context words to compute a matching score between each context of *w* and each context of a translation t_i. This score is discussed in more detail in Section 3.2.2. Figure 2 shows an example of this schematically. There is a strong similarity score for the first context of "plant" and the context of "fábrica," indicating that this usage of "plant" and this usage of "fábrica" share the same word sense. The second context of "plant" has a strong similarity to the given context of "planta."

PanLexicon uses these context similarity scores to begin building translation sets that include the hub word *w*. For each context of *w*, the matching process returns one best translation with its best matching context from each of the spoke languages. For example, one context of "plant" may match best with a context of "fábrica" in Spanish and a context of "厂" (chǎng) in Chinese, while another context of "plant" may match best with "planta" in Spanish and "植物" (zhí wù) in Chinese.

Matches that score below a threshold value are discarded as unlikely to be mutual translations across all the languages. The remaining matched contexts of *w* are partitioned into groups that share the same set of best matching translations. These sets of translations form preliminary translation sets, but they contain only one word from each language.

The final step of PanLexicon is to merge together similar translation sets based on the similarity of the contexts in the different translation sets. The algorithm iterates through each of the hub language words. For each word *w*, the system uses the bilingual dictionaries to find a set of potential synonyms of *w*. A linear-time clustering algorithm then merges translations sets of *w* and its synonyms based on the similarity of a vector representation of their matched contexts. Section 3.2.3 gives the details of synonym finding and translation set merging. After merging, PanLexicon creates one entry for each merged translation set. The entry contains the union of the words from the original translation sets, together with their ranked contexts.

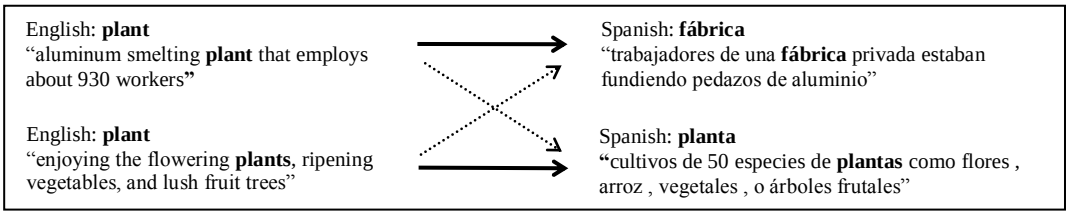

Figure 2: Two contexts of the English word "plant" are matched with contexts of two Spanish translations. The heavy lines indicate strong matches with many context words being translations of each other, while thin lines represent poor matches.

Figure 3 shows an example of three preliminary translation sets containing the hub word "plant." Translation sets 1 and 2 are likely to merge as they share most of their top 10 ranked vector components, while translation set 3 is unlikely to merge with either 1 or 2, as it shares only a single top 10 ranked component with translation set 1 and none with translation set 2.

3.1 Required Resources

Because our goal for PanLexicon is to scale it to a very large number of languages, we designed it to require a minimal set of resources. The first required resource is a set of bilingual lexicons between each of the *k-1* spoke languages and a single hub language. Secondly, the system requires a monolingual corpus in each language. The corpora do not need to be aligned, but they must have some overlap on topic matter. Because we use the bilingual lexicons to match contexts between the corpora, PanLexicon requires word segmenters and morphological tools as necessary to convert text as it appears in the corpora into words as they appear in the bilingual lexicons. Finally, a stop word list is also required for each language.

3.2 System Details

We now discuss some of the details concerning indexing, matching, and merging in the PanLexicon system.

3.2.1 Preprocessing

We build the corpora for each language, indexing each sentence as a separate document. The index stores each sentence as a string of space separated tokens. Additionally each sentence is indexed on lowercased, stemmed versions (if applicable) of each token. The system tabulates co-occurrence counts between the lowercased, stemmed tokens, where two tokens are considered to co-occur if they appear in the same sentence together.

Next, the system collects and stores the contexts for each word *w* in another inverted index. We are currently using language dependent, heuristic methods for determining the context around each word, but these are based only on stop word lists and punctuation. Each context is stored in tokenized form and is indexed by the lowercased, stemmed versions of each non-stop-word token that contains no punctuation or numeric digits.

3.2.2 Finding the Best Matching Contexts

To find the context of a translation *t* that best matches context *c* of hub word *w*, the system queries the context index with the bag of words consisting of all of the translations of all of the non-punctuation, non-stop-word tokens from *c*. The original words from *c* are also included in the query bag, in order to facilitate matching on named entities. The contexts are returned from the index sorted by Lucene's internal sorting algorithm.[1] The top *n* contexts returned by the index are then re-ranked using a scoring metric discussed next. The highest ranked context is returned as the best match.

To produce the similarity score, the context words of *t* are weighted by their pointwise mutual information (PMI) with respect to *t* and the context words of *w* are weighted by their PMI with respect to *w*. The PMI score between words *x* and *y* is defined as:

$$\log\left(\frac{P(x \wedge y)}{P(x)P(y)}\right)$$

where $P(x)$ is the probability that *x* appears in a given sentence, and $P(x \wedge y)$ is the probability that both *x* and *y* appear in the sentence. PMI is used to weight the context words because it helps select contexts that are highly predictive of the given word.

Context words of *w* with a translation appearing in the context of *t* are called matching context words of *w*. Similarly, matching contexts words of *t* are those with a translation appearing in the context of *w*. We obtain a score for each context by squaring the sum of the weights of the matching context words and then dividing by the length of the context. This scoring method provides a good tradeoff between favoring short and long contexts.

[1] See http://lucene.apache.org/java/docs/scoring.html

	1	2	3
Translation Set	{fábrica, **plant**, 厂}	{planta, **plant**, 厂}	{planta, **plant**, 植物}
Top Scoring Hub Language Context	aluminum smelting **plant** that employs about 930 workers	factory that manufactures power **plant** equipment in Brno	world ' s seed – producing **plants** and ferns _ nearly 34 , 000 species
Hub Vector	112 **produce** 86 **build** 70 **worker** 55 **manufacture** 48 ethylene 42 000 36 **assembly** 35 **car** 33 **smelt** 32 ton	100 **manufacture** 94 power 85 **build** 80 factory 55 **smelt** 46 **produce** 39 **assembly** 31 **worker** 30 **car** 24 equipment	702 animal 624 specie 104 insect 94 endanger 90 rare 70 wild 65 thousand 63 extinction 59 000 51 world

Figure 3: Three translation sets containing the hub word "plant". The context vectors for translation sets 1 and 2 have a high similarity, as indicated by the intersection of the top 10 context words. This is evidence that the translation sets {fábrica, plant, 厂} and {planta, plant, 厂} should be merged, while the dissimilar context for {planta, plant, 植物} indicates that it should not merge with either of the other translation sets.

We then use the harmonic mean to combine these non-symmetric scores for the contexts of *t* and *w* into a single match score.

There is also an optional verification procedure for determining whether the highest ranking context *d* in a spoke language is indeed strongly predictive of the desired word sense. The matching procedure can be applied in reverse to find the best matching translation and context for *d* back in the hub language. If the best matching translation for *d* is a word different from the starting word *w*, the original match may be spurious. Matches failing this back verification procedure can be eliminated at this point.

3.2.3 Merging

The goal of translation set merging is to combine translation sets representing the same sense. To do this we start with the translation sets from all of the potential synonyms of a given hub word *w*. We define the set of potential synonyms of *w* as the intersection of the sets of back translations of *w* through each of the spoke languages. With English as our hub language for example, if every spoke language contains some word that translates as both "big" and "large," then "big" and "large" would be considered potential synonyms.

To represent each translation set during merging, PanLexicon forms a context vector from the hub language contexts from each translation set. Each non-punctuation, non-stop-word, lowercased stemmed word forms a dimension of the vector. The value of the component of the vector in dimension *u* is taken to be the sum of the scores of the matched contexts in the translation set whose hub language context contains the token *u*. The cosine similarity metric between context vectors is then used as the distance function between the translation sets.

3.3 Scalability

PanLexicon is designed to scale to a large number of languages, so we have carefully considered the scalability of each step of the algorithm. The PanLexicon system runs in time $O(k \cdot n)$, where k is the number of input languages and n is the length of the largest monolingual corpus.

Collecting contexts from each corpus and running the context matching algorithm are linear in time and space in the number of languages. The hub-and-spoke design means that the context matching is done only $k - 1$ times. Finding matching contexts for each hub word context uses an inverted index, so context matching scales linearly with the hub corpus size and number of spoke languages.

The final step that merges translation sets considers a bounded number of potential synonyms for each hub language word. The number of translation sets for each synonym is also bounded by the number of appearances of the word in the corpus. So a linear-time clustering algorithm keeps the entire merging procedure linear in the size of the hub language corpus. In practice, the number of translation sets for each word will be far fewer than the worse case scenario, even with a large corpus size.

4. Experimental Results

We conducted tests of PanLexicon for three languages: Spanish, Chinese, and English as the hub language. Teams of bilingual speakers judged the correctness of output at various stages of the algorithm, two bilingual evaluators for Spanish-English and two for Chinese-English. Even this three language scenario provides valuable data points for the components of our system, and has the ability to produce a word-sense-distinguished Spanish-English-Chinese lexicon. This is a potentially valuable resource as we were unable to find a freely available machine readable Chinese-Spanish dictionary on the Web.

4.1 Resources

For translations between Spanish and English, we used a dictionary from Ultralingua.[1] We also used an Ultralingua tool for converting Spanish and English words in the corpora into dictionary forms. For translations between Chinese and English we merged dictionaries from the Linguistic Data Consortium[2] (LDC), the English Wiktionary,[3] and the CEDICT[4] dictionary. We used a MaxEnt segmenter similar to Xue and Shen (2000) for tokenizing the Chinese corpus. For corpora, we used the English, Spanish, and Chinese gigaword corpora from the LDC,[5] although for the Chinese corpus we only used the portion in Simplified Chinese.

We created four test sets of 50 words each, one set from the source language for each of the following directions: Spanish to English, English to Spanish, Chinese to English, and English to Chinese. The test set words were randomly selected from the most frequently occurring 10,000 words in each language's corpus, subject to the following criteria:

- The word has at least two translations into the target language according to our dictionaries with distinct senses as verified by a bilingual informant.
- The two translations each appear at least 1000 times in their respective corpora.
- For English and Spanish translations, the two words do not have the same stem.
- For Chinese translations, the two words do not share a common character.

To ensure that this procedure did not bias us too much toward high or low frequency words, we then further required that the test sets contained an equal number of words from each of 5 equal size bins, when ranked by word frequency.

4.2 The Experiments

We performed three experiments on PanLexicon. The first experiment tested the quality of the context matching and ranking portion of the algorithm, and ran bilingually in both directions between English and Spanish and between English and Chinese. The second experiment tested the quality of the preliminary Spanish-English-Chinese translation sets with one translation in each language, and the third experiment tested the quality of

[1] http://www.ultralingua.com

[2] http://www.ldc.upenn.edu, catalog numbers LDC2002L27 and LDC2003E01

[3] http://en.wiktionary.org/wiki/Main_Page

[4] http://www.mandarintools.com/cedict.html

[5] http://www.ldc.upenn.edu, catalog numbers: LDC2005T12, LDC2006T12, and LDC2005T14

the final merged lexicon entries that could have multiple synonyms per language in a translation set.

4.2.1 Matching and Ranking

In the PanLexicon system, contexts of words are used to identify word senses to human consumers, and internally to maintain word senses across all the languages. The first experiment tested whether our matching and ranking algorithm is capable of selecting high quality contexts. For this test we considered the context of a word to be good, if it was clear and distinctive enough that at least one of our bilingual informants could select appropriate translations of the word given the context, and those translations included the translation to which our system matched the context.

To this end, we presented just one side of the matched contexts our system produced to our bilingual informants. The informants were given the list of possible translations of the word and asked to choose one or more translations from the list that were most appropriate for the given word as it was used in the context.

Figure 4 gives an example of one of these questions going from English into Spanish. PanLexicon had matched this context of "fire" with the Spanish context "asentamiento judío de Neveh Dekalim dispararan un proyectil de tanque" of the word "disparar".

Word:	**fire**
Context:	Jewish settlement of Neveh Dekalim **fired** a tank shell
Translations:	descargar despedir disparar fuego incendiar lumbre quema tirar
Other options:	Word is used as a proper noun not listed in the translations. Correct translation is not listed. Correct translation cannot be determined from the context (context is confusing or ambiguous)

Figure 4: Spanish-English evaluators were shown the context for the English word "fire" and asked to select one or more Spanish translations for the word sense indicated by the context. Our system and both evaluators agreed on "disparar" for this example.

For each test word, we created one question for each translation of the word for which our system found a best matching context. Our bilingual dictionaries contained an average of 6.3 translations per word, and we were able to find a best matching context for an average of 5.5 translations per word. Most of the translations for which our system did not find a best matching context were words that appeared in our corpus fewer than 50 times.

The results on each test set for this experiment are shown in Figure 5. A high threshold for context matching

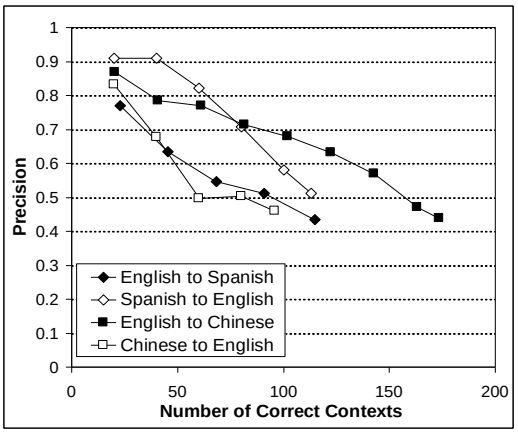

Figure 5: A context threshold on context matching allows PanLexicon to trade off lexicon coverage and precision. A correct context is one in which a bilingual speaker can determine the target language translation from the source language context alone.

scores gives high precision, from 0.77 to 0.92, but the total number of correct contexts at this threshold is relatively low: a total of 25 contexts for the 50 test words. Lower thresholds increase the number of correct contexts, but reduce precision. The graph for English to Chinese extends further to the right, because our English to Chinese dictionary contained several more translations per word on average than the other dictionaries.

On the average across the test sets, PanLexicon finds 50 correct contexts at precision 0.74, and 100 correct contexts at precision 0.55 from an initial set of 50 ambiguous words. These are the cases where PanLexicon can find a context that is strongly predictive of a word sense that is associated with a target translation.

PanLexicon is most successful on senses of words that are well represented in its newswire corpora, particularly if there are news stories in both corpora with the same proper names or technical terms such as in Figure 4's context. Our use of PMI to weight context words tends to bias our system towards choosing contexts with proper nouns and specialized topics. For example, PanLexicon selected the following context for the word "credit": "listing their forests for carbon dioxide credits".

Low context matching scores often occur for a minor word sense, relative to the news corpus, or where the matching context words are ambiguous common words. As an example of ambiguous context words, an English context for the legal sense of "case" matched on a translation of the word "hearing" in a Spanish context about a "trombone case." The Spanish translation of "hearing" was used in the sense of "listening to music."

For those questions where both informants selected one or more translations, there were an average of 9.6 translations presented to them from which they each selected an average of 1.4, agreeing on an average of 0.67. Combining their choices gives an average of 2.2 translations selected by at least one of the two evaluators per question. If we randomly selected translations for the contexts provided by our system we would expect to have 2.2 out of 9.6 judged correct, a precision of 0.226. Our system's precision keeps well above this random baseline.

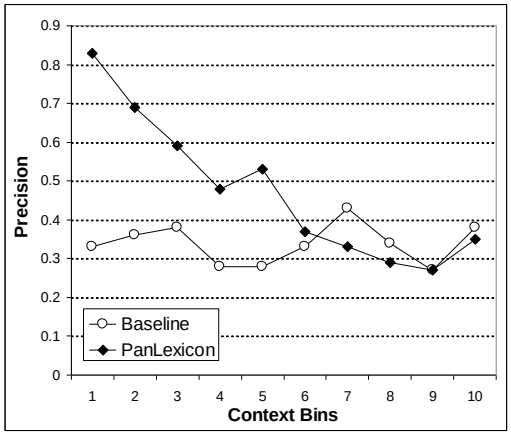

Figure 6: Average precision across the 4 test sets begins at 0.83 for the contexts with highest matching score and dips below precision 0.30 for contexts with low score. Precision for the upper half of the contexts stays well above the baseline choice of always selecting the translation that occurs most frequently in the target language corpus.

Figure 6 shows the average precision over all 4 test sets for contexts grouped into equal size bins after sorting by descending matching score. We compare this to a baseline that always selects the translation that appears most frequently in the target language corpus. Precision for the highest confidence contexts is 0.83. For the first 5 bins the system's precision is well above the baseline precision. For bins of low scoring matches, the baseline choice performs better than selecting the word from PanLexicon's match.

4.2.2 Translation Sets

The second experiment tested the ability of our system to maintain coherent senses across all three languages at the same time. To this end we tested the quality of the translation sets produced by PanLexicon before they entered the merging process. Each translation set consisted of an English, Spanish, and Chinese word, together with a collection of contexts for each word. We considered the translation set to be correct if the top scoring context for each word in the translation set illustrated the same word sense. We were unable to find an informant who was fluent in Spanish, English, and Chinese, so we tested using our bilingual Spanish-English and Chinese-English informants. For each translation set, we presented the English word and top scoring context together with the Spanish word and top scoring context to the Spanish-English informant. We presented the same English word and context together with the Chinese word and top scoring context to the Chinese-English informant. Each informant was allowed to choose one of three options. They could specify that the senses of both words as used in their respective contexts were the same, different, or that one or more of the contexts was unclear.

To combine the judgments of all four evaluators, we used a simple probabilistic model. We defined the probability that a translation set was correct to be the probability that the English and Spanish portions shared the same word sense and that the English and Chinese portions also shared the same word sense. These two events were assumed to be independent. The probabilities of correctness for the English-Spanish portions and for the English-Chinese portions were defined to be either one or zero when the evaluators agreed on correctness, and one half when the valuators gave mixed judgments.

The translation sets for the second experiment were obtained by running PanLexicon on all of the English translations of both the Spanish and Chinese test sets. For this experiment we produced more translation sets than our informants could assess, so we chose a random selection of them from four different tiers based on the translation sets' combined matching scores. From this we estimated precision scores for four threshold values, presented in Figure 7. Precision begins at 0.73 and declines to 0.53 at lower matching scores. The first data point represents approximately 22 correct and 8 incorrect translation sets for the test set of 100 words; the third data point represents approximately 80 correct and 46 incorrect translation sets. Nearly half the errors have semantically related, but not synonymous words.

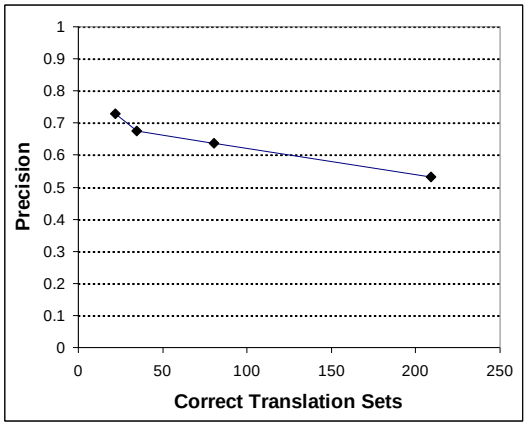

Figure 7: Precision for the Chinese-English-Spanish translation sets begins at 0.73 with highest context matching scores, then declines to precision 0.53 at lower scores. Translation sets were judged correct if all three words in context indicated the same meaning.

We analyzed the sources of errors from the English and Spanish portions of the translation sets from one of the test sets. Most of the errors fell into four categories:

- **Semantically Related Translations (26%):** Although the contexts had plentiful content word matches, the content words alone were insufficient to disambiguate between semantically related, but non-synonymous, translations.
- **Mismatched Translations (20%):** The contexts were strong matches, but one of the context words, instead of the source word itself, was the true match to the target word. An example is shown in Figure 8. This error is related to the previous one, but can be more easily identified automatically.
- **Multiword Expressions (17%):** Our system currently handles only single word translations.
- **Not Enough Matched Context (14%):** The matches were not strong enough. Either the contexts were not long enough, or not enough of the content words were matched between the contexts.

	word	context
English	**bank**	money is *deposited* in Swiss **bank** accounts
Spanish	**depositar**	dinero estaría **depositado** en cuentas en Suiza

Figure 8: Contexts of "bank" and "depositar" are strong matches, but the correct match is between the English "deposit" and the Spanish "depositar."

4.2.3 Merged Entries

The third experiment gives preliminary results for the final output of our system, merging translation sets that represent the same word sense. For this experiment, we produced all of the final PanLexicon entries that contained one of the Spanish or Chinese words from the Spanish and Chinese test sets. Because we were dealing with a small number of translation sets, we used an O(n^2) greedy merge algorithm. We presented the entries that resulted from merging two or more of the original translation sets to our bilingual informants. These are the lexicon entries that contain more than one word in at least one of the three languages.

These merged entries were considered correct only if all of the words in all three languages shared the same word sense as illustrated by their contexts. As in the previous experiment we had to test each entry bilingually in two parts. We used the same probabilistic model to merge the judgments of the four evaluators.

Five words from each of the two 50 word test sets were set aside as a tuning set to determine a threshold merging distance. The selected threshold resulted in PanLexicon creating 41 entries with more than one word in at least one of the languages, from the remaining 90 test words.

On average the merged entries contained 1.1 English words, 2.0 Chinese words, and 1.8 Spanish words. The combined votes of all evaluators judged 43.9% of these entries to be correct, although we noticed wide variation in our informants' judgments. If both evaluators for each language pair were required to agree that an entry was correct in order to count it as correct, then only 19.5% of the 41 entries were considered correct. If only one evaluator from each language pair had to judge the entry as correct, then 75.6% were considered correct.

Many merge errors were based on semantically related words that were not synonyms. For example PanLexicon merged translation sets containing the English words "bullet" and "shot". It is clear from the context vectors in Figure 9, why these semantically related concepts merged, although they are not synonyms.

English context vector for translation set containing the word "bullet"	English context vector for translation set containing the word "shot"
1926 **fire**	2090 **fire**
566 **gun**	450 **gun**
542 **police**	419 **rifle**
517 **disperse**	413 soldier
515 **rifle**	401 **disperse**
478 **wound**	385 **wound**
444 chest	382 **police**
440 plastic	306 warn
387 crowd	284 dead
377 head	270 gunman

Figure 9: English context vector for translation sets containing the words "bullet" and "shoot". Such semantically related words are likely to merge even though they are not synonyms.

5. Related Work

Several researchers in the 1990's developed methods to use bilingual MRDs to assist in manual creation of translation lexicons (Neff and McCord, 1990; Helmreich et al., 1993; Copestake et al., 1994). A notable example of more recent work is the EuroWordNet project[1] based on the Princeton WordNet.[2]

EuroWordNet takes the approach of using the Princeton WordNet as an interlingual standard and linking lexical entries from other languages to this standard. This approach has the advantage that it leverages the considerable amount of human effort that has already gone into creating the WordNet, which is a de facto standard in the research community. However, the Princeton WordNet standard is English centric, making it difficult to incorporate concepts from other languages. It is also built by lexicographers making fine grained sense distinctions, which means the task of linking multilingual entries automatically is extremely challenging – while linking them manually is very expensive. PanLexicon is fully automated, and has the potential to scale to far more languages than EuroWordNet.

Another body of research is the massively collaborative effort to create multilingual lexicons such as Wiktionary[3] and the related OmegaWiki[4] projects. The Wiktionary approach of letting ordinary people add and modify entries has proven to be extremely powerful. The English Wiktionary is currently estimated to have 348,419 entries in 389 languages, although only a handful of languages have high coverage. Since anyone is allowed to edit a Wiktionary, standardization of entry formatting is difficult, and the quality of the entries can vary greatly. The OmegaWiki project attempts to solve some of the standardization problems, but is only in its infancy.

Even with the large number of people editing and adding to the Wiktionaries, they still suffer from a scarcity of data. For example, the English Wiktionary contains only approximately 6,000 translations into Chinese. This can be compared to the over 30,000 translations between Chinese and English freely available for download from the CEDICT dictionary mentioned earlier.

The PanImages system (Etzioni et al., 2007) combines multilingual and bilingual dictionaries into a translation graph, which produces multilingual sense-distinguished translations. The translation graph uses a probabilistic inference mechanism to reason about sense equivalence across the source dictionaries without using corpora. PanImages can only provide glosses for a word when a sense-distinguished dictionary provides that gloss. Thus, PanImages could benefit from the capabilities provided by PanLexicon and the two systems could be combined in future work.

Other corpus-based techniques to build multilingual lexicons have mostly focused on using parallel bilingual

[1] http://www.illc.uva.nl/EuroWordNet/#EuroWordnet

[2] http://wordnet.princeton.edu

[3] http://en.wiktionary.org/wiki/Wiktionary:Main_Page

[4] http://www.omegawiki.org/Main_Page

corpora (Dyvik 2002). Pair-wise approaches like this do not scale to large numbers of languages.

6. Conclusions and Future Work

PanLexicon is a fully automatic and scalable tool to build word multilingual translation lexicons. Each lexicon entry is a *translation set*, a set of words across multiple languages that all share the same implicit word sense, along with illustrative contexts for each word. These contexts provide a cue to the meaning for human readers and can guide automatic WSD tools, as well.

PanLexicon is designed to scale to large numbers of languages with minimal resource requirements. It uses bilingual MRDs that do not make word-sense distinctions, unaligned corpora for each language, and sufficient morphological tools to map surface forms into entries in the MRDs. Because of its hub-and-spoke architecture, the time to create a lexicon scales linearly with the number of languages.

We evaluated a prototype of PanLexicon with three languages: Spanish, Chinese, and English as the hub language. PanLexicon has parameters that allow it to trade off between coverage and precision on each aspect of the system that we evaluated. The highest confidence matches between Spanish and English contexts or between Chinese and English contexts produce contexts that were judged to be predictive of the translation word with precision ranging from 0.77 to 0.92. Evaluation of translation sets across the three languages showed the same coverage-precision trade off, with precision of 0.73 gracefully declining to precision of 0.52.

The final step of PanLexicon merges translation sets that represent the same implicit word sense, producing translation sets that may have sets of synonyms for each language. This was the most difficult task for PanLexicon; only 44% of its merging decisions were entirely correct according to our evaluators, although 75% were considered correct by at least one evaluator. This quantitative evaluation together with our analysis of the errors gives us a strong position from which to continue to improve system performance.

PanLexicon has potential to help overcome the knowledge-acquisition bottleneck that has plagued lexical translation and knowledge-based MT. Most of these applications would benefit from large, sense-distinguished translation lexicons. We are exploring methods to turn PanLexicon's contexts into seeds for automatic WSD. This will increase the utility of the lexicon.

Future work also includes integrating PanLexicon into an actual machine translation system. We see potential in incorporating PanLexicon into an MT system based on the DELPH-IN machinery (Bond et al, 2005) for semantic transfer and on the Lingo Grammar Matrix (Bender et al., 2005) for creating the grammars needed for parsing and generation. The DELPH-IN machinery is scalable to many languages, but in doing so, it requires the creation of lexical transfer rules for each language pair. PanLexicon produces a multilingual lexicon that can help in this transfer step for all the language pairs.

Acknowledgements

This research was carried out at the University of Washington's Turing Center, which is supported in part by a generous gift from the Utilika Foundation. Additionally we would like to thank Emily Bender and Jonathan Pool for their extensive comments and Fei Xia for providing the Chinese word segmentation tool.

References

Bender, E. and Flickinger, D. 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. Proceedings of IJCNLP-05 (Posters/Demos), Jeju Island, Korea.

Bond, F., Oepen, S., Siegel, M., Copestake, A., and Flickinger, D. (2005). Open source machine translation with DELPH-IN. In Open-Source Machine Translation Workshop at MT Summit X.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T. and Frey, J. (2006). Context-Based Machine Translation. In AMTA.

Etzioni, O., Reiter, K., Soderland, S., and Sammer, M. (2007). Lexical Translation with Application to Image Search on the Web. In the proceedings of MT Summit XI.

Helge Dyvik. (2002). Translations as semantic mirrors: from parallel corpus to wordnet. Available at http://www.hf.uib.no/i/LiLi/SLF/Dyvik/

Fellbaum, C. (ed.) (1998). WordNet. An Electronic Lexical Database.Cambridge: The MIT Press.

Gey, F., Kando, N., Lin, C-Y. and Peters, C. (2006). 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.

Helmreich, S., Guthrie, L., and Wilks, Y. (1993) The use of machine readable dictionaries in the Pangloss project. AAAI Spring Symposium on Building Lexicons for Machine Translation.

Lee, C., Lee, G., and Yun, S.J. (2000). Automatic WordNet mapping using word sense disambiguation. 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Levin, L., A. Lavie, M. Woszczyna, and A. Waibel, (2000). "The Janus III translation system," Machine Translation, vol. 15, no. 1-2, Special Issue on Spoken Language Translation.

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

Oepen, S., H. Dyvik, J. T. Lønning, E. Velldal, D. Beermann, J. Carroll, D. Flickinger, L. Hellan, J. B. Johannessen, P. Meurer, T. Nordgård and V. Rosén (2004) `Som å kapp-ete med trollet? Towards MRS-based Norwegian-English machine translation'. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, MD.

Vossen, P. (ed.). (1998). EuroWordNet: A multilingual database with lexical semantic networds. Kluwer Academic Publishers.

Wang, C. and Seneff, S. (2006). High-quality speech translation in the flight domain. Proc. Interspeech.

Xue, N. and Shen, L. (2003). Chinese word segmentation as LMR Tagging. In Proc. of SIGHAN Workshop.