In Proceedings of the 11th Annual conference of the European Association
for Machine Translation (EAMT), June 2006, Oslo, Norway.

Leveraging Recurrent Phrase Structure in Large-scale Ontology Translation

G. Craig Murray
Bonnie J. Dorr
Jimmy Lin

Institute for Advanced Computer Studies
University of Maryland
{gcraigm,jimmylin,bdorr}@umd.edu

Jan Hajič
Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University
{hajic,pecina}@ufal.ms.mff.cuni.cz

Abstract. This paper presents a process for leveraging structural relationships and reusable phrases when translating large-scale ontologies. Digital libraries are becoming more and more prevalent. An important step in providing universal access to such material is to provide multi-lingual access to the underlying principles of organization via ontologies, thesauri, and controlled vocabularies. Machine translation of these resources requires high accuracy and a deep vocabulary. Human input is often required, but full manual translation can be slow and expensive. We report on a cost-effective approach to ontology translation. We describe our technique of prioritization, our process of collecting aligned translations and generating a new lexicon, and the resulting improvement to translation system output. Our preliminary evaluation indicates that this technique provides significant cost savings for human-assisted translation. The process we developed can be applied to ontologies in other domains and is easily incorporated into other translation systems.

1 Introduction

Human translation of any text comes at a certain cost. Wherever components of translated text are reused, it is preferable not to incur costs of retranslation. Identifying reusable component elements and prioritizing their translation is essential to maximizing effectiveness and controlling expense. This paper presents a process for identifying the most valuable phrase components in a large thesaurus and leveraging their reusability in machine translation.

Digital collections of any significant size require an organizing principle to provide value to users. Generally, this can be supported through a simple controlled vocabulary of keyword phrases, through hierarchical arrangement of concepts in a detailed ontology, or any number of options in between.

Translation of domain-specific works can require a highly specific dictionary of terms. Such dictionaries are not readily available in all languages for all domains. Some translation efforts will therefore require a certain amount of resource building in support of domain-specific vocabularies. In the case of thesauri for digital archives, accuracy in translation is essential for providing access to contained work. Highly-accurate human translations incur a cost that is generally fixed to the number of words being translated. However, not every term in a thesaurus is equally applied to works in a collection. In many collections, a minority of terms may provide the majority of categorizations. Similarly, a great number of the highly specific terms will describe only one or two items each. Therefore, not every keyword in a thesaurus of categorization terms carries the same value. The cost/benefit ratio for translating infrequent keywords is much higher than for frequent keywords. Moreover, certain components of some phrases will have more value for re-use than others. For manual translation efforts, it is preferable to identify and translate those phrases that have the highest value.

In the work presented here, we leveraged a small set of English-Czech translations to pro-

vide Czech speakers access to 116,000 hours of video testimonies in 32 different languages. Starting from an initial out-of-vocabulary (OOV) rate of 85%, we show that a small set of prioritized translations can be elicited from a human translator, aligned to the original phrases, decomposed and then recombined to cover the majority of terms in a complex ontology. Moreover, we show that prioritization of human translation based on hierarchical position and frequency of use facilitated extremely efficient reuse of human input. The reusable resources we obtained cover 90% of the access value of the thesaurus using human translations of less than 5% of the thesaurus terms.

2 Motivation

Our work is motivated by an extreme case of vocabulary resource sparsity in a highly structured context. The Survivors of the Shoah Visual History Foundation has collected what is presently the world's largest archive of videotaped oral histories (USC, 2006). The archive contains several million segments of video from the testimonies of over 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust. If viewed end to end, the collection amounts to 13 years of continuous video. The MALACH project (Multilingual Access to Large Spoken ArCHives)(Gustman et al., 2002; CLSP, 2005) is currently using this vast resource to research a number of objectives in cross-language information retrieval. Among these objectives is improved support for searching and browsing through collections such as the Survivors of the Shoah Visual History Foundation collection of oral histories.

The Survivors of the Shoah Visual History Foundation (VHF) has been cataloging these video testimonies with a structured thesaurus of keyword phrases representing relevant concepts in the domain. They assign keyword phrases to segments of video as a means of describing the content of the video segment. At present, their thesaurus of English keyword phrases provides the sole means by which users can search the content of the video archive. Czech has been selected as the first language for machine translation research.

The VHF collection thesaurus contains more than 56,000 keyword phrases/concepts with over 50,000 additional synonymous alternate phrases. These phrases have an average length of 6.7 words with a mode of 4 words per phrase. Although the testimonies in the collection represent 32 languages, the controlled vocabulary used to catalogue them is currently only available in English. To make the collection truly accessible this thesaurus must be translated into as many languages as possible. Term specificity has lead to an extreme out-of-vocabulary problem for machine translation.

Ontologies and thesauri are used to describe concepts in detail and to record the relationship between these concepts. As a result, they usually include highly specific terms that may not be available in statistical MT resources such as aligned parallel text or probabilistic dictionaries. This is no less true in the VHF collection thesaurus. In our first pass at machine translation of this thesaurus we found that only 15% of the words in the vocabulary could be found in an available aligned corpus (Čmejrek, Cuřín, Havelka, Hajič & Kubon, 2004). The other 85% were out of the vocabulary (OOV=.85). Lexical information for translating these terms must be acquired from humans.

The following sections describe our general approach to reducing the time and cost of translation by optimizing the reusable value of human input and feeding this input into a machine translation system. We report on our evaluation of the system output with a detailed look at increases in performance from incremental increases in human input. We relate our work to other studies that have taken more costly approaches and we conclude with a summary of our findings.

3 General approach

We describe a prioritized, human-in-the-loop approach to machine translation that is fast, efficient, and cost effective for the task at hand. We collected only 2,500 translations from a human translator and then reused their components to generate a lexicon useful for machine translations of the rest of the thesaurus. We began by selecting the most valuable keyword phrases in terms of access to the VHF video collection and in terms of reusability of their component sub-phases. After collecting translations from one human informant, we had a

second human informant check the translations and align each of their terms to the original English terms using a graphical user interface. From these alignments we then constructed a probabilistic dictionary that maps English words and phrases to Czech words and phrases. To test the validity of this lexical acquisition process we implemented a relatively simple translation system that used this newly generated lexicon and a phrase-based back-off strategy. In Section 4 we report on an evaluation of the translation system's output that quantifies the success of our approach.

3.1 Maximizing Value and Reuse

To reduce overall cost and maximize the value of human translations, we began by defining two values for each keyword phrase in the thesaurus: a *thesaurus value*, representing the importance of the keyword for access to the collection, and a *translation value*, representing the usefulness of having the keyword translated. These are not identical, but the second is related to the first. The concepts we set out to translate are arranged into a poly-hierarchy, with child nodes having multiple parents as "broader terms". The internal nodes generally represent broader concepts or types of concepts. The lower nodes and leaf nodes generally represent very specific concepts. Leaf nodes of the hierarchy are assigned as descriptors of video segments, with parent terms above them. Some of the internal/parent nodes are also assigned as descriptors of video segments; others are not. Therefore, the usefulness of any keyword phrase for providing users access to the digital archive is directly related to its position in the thesaurus hierarchy.

3.1.1 Thesaurus value

Simply providing translations of the leaf nodes of the hierarchy would not provide usable access. Most of the leaf nodes are very specific. Browsing from one concept to another requires translation of internal nodes. Also, many of the internal nodes are assigned to testimonies themselves. For example, various aspects of Auschwitz are described by 35 different keyword phrases. Some of these 35 keyword phrases are broader terms, some are narrower terms. A hierarchical relationship exists between broader and narrower terms. Both the broader terms and the narrower terms may be assigned to segments. Consider the example in Figure 1. The death camp itself, is described by the keyword phrase "Auschwitz II-Birkenau (Poland: Death Camp)". That keyword is assigned to 17,555 video segments in the collection. Some of the broader and narrower terms are also assigned to segments but not all. Notably, "German death camps" is not assigned to any video segments. However, "German death camps" has very important narrower terms including Auschwitz II and others.

Auschwitz II-Birkenau (Poland : Death Camp)
Assigned to 17555 video segments
Has as broader term phrases:
Cracow (Poland : Voivodship)
[534 narrower terms] [204 segments]
German death camps
[6 narrower terms] [0 segments]
Has seven narrower term phrases including:
Block 25 (Auschwitz II-Birkenau)
[leaf node] [35 segments]
Kanada (Auschwitz II-Birkenau)
[leaf node] [378 segments]
...
disinfection chamber (Auschwitz II-Birkenau)
[leaf node] [9 segments]

Figure 1: Sample keyword phrase with broader and narrower terms

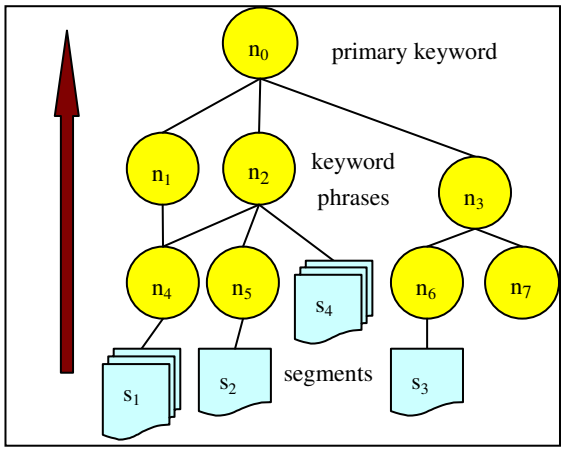

Figure 2: Bottom-up micro-averaged expectation value

Parent nodes in the hierarchy need to be given some of the value of their children in order to reflect their importance in providing access to materials in the archive. However, we cannot simply take a sum of the value of all children, grandchildren, etc. under a parent node—that would unevenly bias the top of the

hierarchy. A prioritization of terms based on total value of narrower terms (i.e. child nodes) would lead to translation of the top of the hierarchy first. If only part of the thesaurus were being translated, this would eliminate most of the lower nodes, including leaf nodes, from the translated set of keyword phrases. This is the exact opposite of the problem with translating only leaf nodes.

In our measure of keyword phrases' *thesaurus value*, we strike a balance between these two extremes. We calculate an importance value for each node as the sum of the number of segments assigned to that keyword phrase, and the average value of all of its children.

$$value_k = count(segments_k) + \frac{\sum_{children(k)}^{i} value_i}{count(children(k))}$$

Because they have no children, the *thesaurus value* of each leaf nodes is simply the number of video segments to which the keyword has been assigned. For the parents of the leaf nodes, the *thesaurus value* is the number of segments (if any) to which the parent node has been assigned, *plus* the average of the *thesaurus value* of its child nodes (if any). This recursive calculation yields a micro-averaged value representative of the expected number of segments one could reach within one or two downward edge traversals in the hierarchy. Put in another way, it gives a kind of weighted value for the number of segments that might be described by a given keyword phrase or its immediate narrower-term keyword phrases. For example, in Figure 2 each of the leaf nodes n_4, n_5, n_6 and n_7 has a value based solely on the number of segments that have been assigned that keyword phrase (in some cases zero). The node at n_2 has value both as an access point to the segments at s_4, and as an access point to the keyword phrases at nodes n_4 and n_5. Other internal nodes, such as n_3 will have value only as access to other nodes/keyword phrases (e.g. nodes n_6 and n_7), and not all keyword phrases will be of value for access to anything (e.g. n_7).

Working from the bottom of the hierarchy up to the primary node (n_0) we calculated this *thesaurus value* for each node in the hierarchy. In our example, we would start with nodes n_4 through n_7, counting the number of the segments that have been assigned each keyword phrase. Then we would move up to nodes n_1, n_2 and n_3. At n_2 we count the number of segments s_4 and add that count to the average of the *thesaurus values* for n_4 and n_5. At n_3 we simply average the *thesaurus values* for n_6 and n_7. The end result values are then used to calculate an estimate of how valuable the translation of any given keyword phrase would be.

3.1.2 Translation value

After obtaining the *thesaurus value* for each node, we calculated a *translation value* for each word in the vocabulary as the *sum* of the *thesaurus value* for every keyword phrase that contains that word. For example, Auschwitz occurs in 35 important keyword phrases. As a candidate for translation, it carries a high impact value for the translation of the whole thesaurus. We estimate the utility value of translating Auschwitz into another language by summing the micro-averaged *thesaurus value* of all of the keyword phrases in which it is used. The end result of this step is a vocabulary of words and the impact that each correctly translated word would have on the overall value of the translated thesaurus.

However, to translate individual words, rather than individual phrases would be an inefficient use of our translator's time. Therefore, we elicited aligned translations of entire keyword phrases, and prioritized their order of translation based on the *translation value* of the words they contain. The value that any keyword phrase holds for translation is only indirectly related to its own value as a point of access to the collection (i.e. its *thesaurus value*). It is the case that some keyword phrases contain words with high *translation value*, but the keyword phrase itself has low *thesaurus value*. The value gained by translating any given phrase is more accurately estimated by the total value of any untranslated words it contains. Therefore, our next step was to iterate through the thesaurus keyword phrases, prioritizing their translation based on the assumption that any words contained in a keyword phrase of higher priority would already be translated.

We start by assuming the entire vocabulary is untranslated. Then we select the one keyword phrase that contains the most valuable untranslated words—we simply add up the *trans-*

lation value of all the words in each keyword, and select the keyword with the highest value. We add this keyword to a prioritized list of phrases to be translated and we remove it from the list of untranslated phrases. We update our vocabulary list and assume all the words in the prior keyword phrase to now be translated. Working from this assumption, we again select the one keyword phrase that contains the most valuable untranslated words. As our prioritized list of keyword phrases grows, we obtain more and more words for our vocabulary and the *translation value* for untranslated keyword phrases drops toward zero. Ultimately, we wind up with a distilled list of keyword phrases that should be translated, in a prioritized order, to maximize coverage of the vocabulary with minimized translation effort.

Naturally, there will be some words that appear more than once for translation because they appear in more than one keyword phrase with high *translation value*. This is desirable. To build an accurate dictionary, we want to have more than one translation of important words. This is particularly desirable for morphologically rich languages such as Czech. However, we do not want to exhaustively translate all the morphological variants of a word at the cost of leaving entire sections of the vocabulary untranslated. Therefore, it is also natural that some errors will result from translating only some variants of a word and not others. We address the acceptability of these errors in our results section below. Stop words are not prevalent in this thesaurus and were not removed from the calculation of *translation value*.

Following this scheme, the most important parts of the thesaurus will be translated first, and the most important vocabulary terms will quickly become available for machine translation of those keyword phrases with high *thesaurus value* that did not make it onto the prioritized list (i.e. low *translation value*). We find that the gain rate of *thesaurus value* rises very quickly after the first few translations. With each subsequent translation of keyword phrases on the prioritized list we gain tremendous value in terms of providing access to the collection of video testimonies. The rate of gain is shown in Figure 3. In this figure it can be seen that prioritization based on *translation value* gives much higher yield of total access than prioritization based on *thesaurus value*.

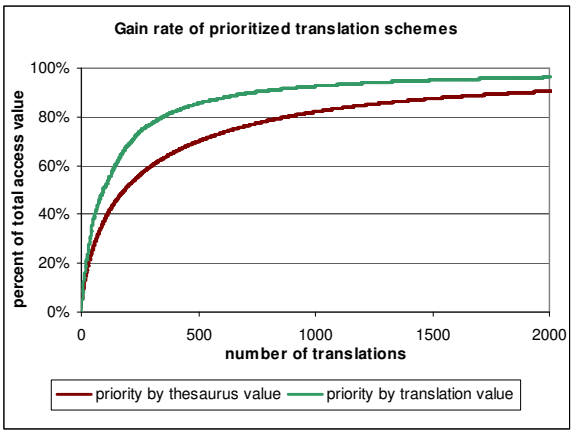

Figure 3: Gain rate of access value based on number of human translations

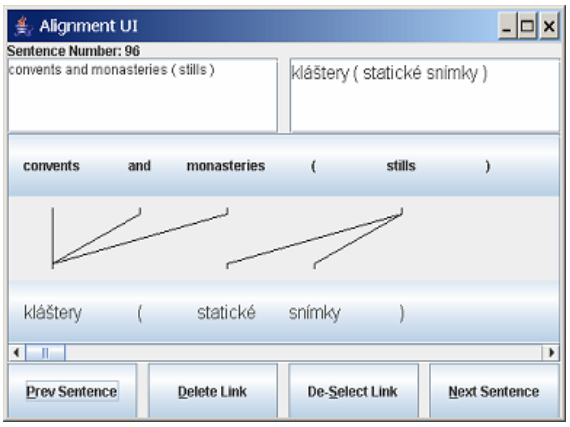

Figure 4: Alignment tool user interface

3.2 Alignment and Decomposition of Human Translations

Following the prioritization scheme above, we obtained professional translations for the top 2500 English keyword phrases. If multiple translations were possible for a phrase, we encouraged our translator to provide more than one. We tokenized the translations that were provided and presented them to another bilingual native Czech speaker for verification and for alignment to the original English text. Using a tool developed at the University of Maryland (Madnani, 2004), this second human informant would mark each Czech word in a translated keyword phrase with a link to the equivalent English word(s). Figure 4 shows an example of the human input. Multiple links are used to convey the relationship between a single word in one language and a string of words

in another. During the alignment, informants could also give feedback on errors in translation or tokenization.

The output of the alignment process was then used to build a probabilistic dictionary. To do this we simply decomposed the translations based on the links that were assigned to pairs of words by the human informant. In the case where a single word in one language was aligned to multiple words in the other, we consider the string of multiple words in the second language as a phrasal unit and include the phrase in our lexicon. In the example shown in Figure 4, "stills" is recorded as translatable with "statické_snímky" and "kláštery" is recorded as a translation for "convents_and_monasteries." In the case where entire sub-phrases are directly aligned, we preserve the sub-phrase alignment as well as the individual tokens. (We count the number of occurrences of each alignment in all of the translations and calculate probabilities for each direction. For example, in the top 2500 keyword phrases "stills" appears 28 times. It was translated and aligned with "statické_snímky" 27 times, and only once with "statické_záběry." By this count, the probability that "stills" should be translated "statické_snímky" is 27/28=0.9643.

The translation of the English keyword phrases into Czech took approximately 60 hours, and the alignments took 45 hours. The overall cost of human input was less than 900 €. The projected cost of fully translating the entire thesaurus would have been close to 20000 €. Moreover, it would have involved repeated effort for frequently recurring terms and would not have produced any reusable resources. Our alignment system was pre-existing and is freely available. Naturally, costs for building resources with our approach will vary with language and other factors, but in our case the cost savings for human input is approximately 20-fold.

3.3 Machine Translation

To verify the validity of our approach, we need to show that a lexicon acquired in this way is beneficial to a translation system. To do this we implemented a very simple translation system, although more sophisticated systems would benefit from the exact same resources. In our system, we take English input and first look for phrasal matches in the dictionary we created, then for individual words. Using a greedy coverage algorithm, we look for longest matching strings of tokens as a means of finding phrases. Building on the example above, our system looks for "monasteries and convents stills", finds nothing in the dictionary, then looks for "monasteries and convents" and finds "kláštery". If no match were found for the phrase "monasteries and convents" the system would have backed-off to look for matches on individual tokens. If no match is found in the dictionary we created, we back-off to the Prague Czech-English Treebank dictionary, a much larger dictionary but with broader scope. If no match is found in either dictionary for the full token, we look for matches based on the stem. Tokens that are unmatched even after back-off and stemming are simply passed through un-translated.

A minimal set of heuristic rules were applied to reordering the Czech tokens but the output is primarily word by word translation. Because our focus here was on building lexical resources, not on building a translation system, we have not yet added morphological adjustment of Czech words or sophisticated word order correction. Our evaluation scores below will partially reflect the simplicity of our system. Our system is simple by design. Any improvement or degradation to the *input* of our system has direct influence on the *output*. Thus, measures of accuracy and translation error from our system can be directly interpreted as measures of the quality of the lexical resources we used, and in turn, of the process by which those resources were developed.

4 Evaluation and Results

We performed two different kinds of evaluation to validate our process. (1) We compared the output of our system to a prior set of human generated translations. We had human informants provide translations of a set of keyword phrases prior to the prioritized translation described above. A minimum of two independent translations were obtained for each keyword phrase. There is no overlap between these phrases and the aligned input phrases that were translated from the prioritized list. (2) After

collecting aligned translations we generated a probabilistic dictionary and translated a random sample of 100 keyword phrases. These translations were then corrected by native Czech speakers (i.e. word order, word choice, and morphology were adjusted to be correct translations of the English keyword phrases.) We compared our machine translations to both the human generated translations, and the human corrected translations. Below, we report accuracy using two different measures (TER and Bleu).

As a baseline, we first generated translations using only the dictionary available in the Prague Czech-English Dependency Tree Bank. We randomly selected keyword phrases from a pool of keyword phrases that had been translated by human informants prior to our prioritized translation process. We obtained two or three reference translations for each of these keyword phrases.

We used the Bleu metric (Papineni, Roukos, Ward, & Zhu, 2002) to measure the difference between our generated translations and human reference translations. We also measured the difference between our generated translations, and human *corrections* of our generated translations. We take this second measure to be an upper bound on realistic performance improvement. If the human corrected translation was acceptable to the human translators, it is hard to imagine our system surpassing that level of accuracy and fluency. The results are shown in Figure 5. Each bar in this graphs shows performance after adding a different number of aligned translations into the lexicon (i.e. performance after 0 aligned translations, 500, 1000, 1500, 2000, and 2500 aligned translations.)

There is a notable jump after the very first translations are added into our probabilistic dictionary. Without any elicitation and alignment we get an initial Bleu score of 0.35 compared to human reference. After we elicited only 500 translations and added the aligned terms to our dictionary, our Bleu score rose to 0.45. After 2500 elicited and aligned translations, we achieve 0.47. Comparison to human corrected machine translations is even more impressive, jumping from 0.35 to 0.84. There is a consistent rise as more and more translations are added.

The fact that the score continues to rise indicates that the approach is successful in quickly expanding the lexicon without introducing much noise. However, we should be careful to draw conclusions from the trend and not the specific values. The Bleu score of 0.84 after 2000 aligned translations (using human corrected references) is only meaningful in comparison to the scores of 0.80 at 1500 aligned translations and 0.35 at 0 aligned translations.

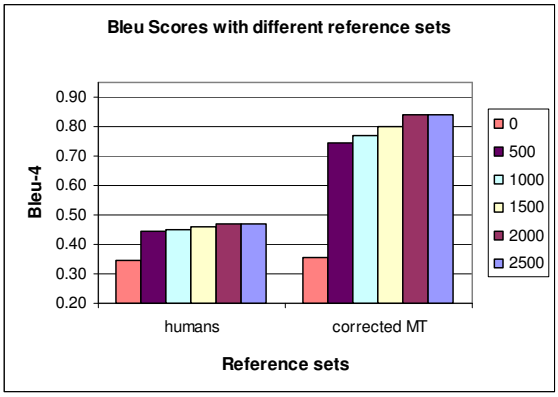

Figure 5: Blue scores against different reference sets

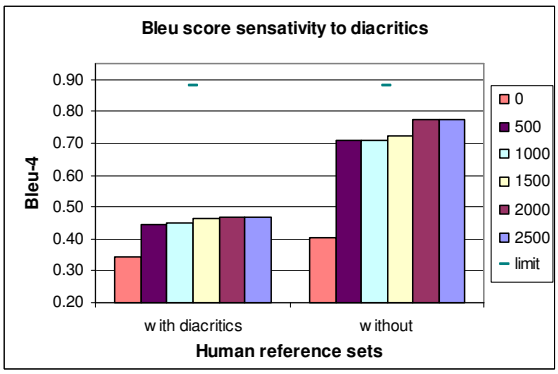

Figure 6: Diacritics have significant impact on performance measures

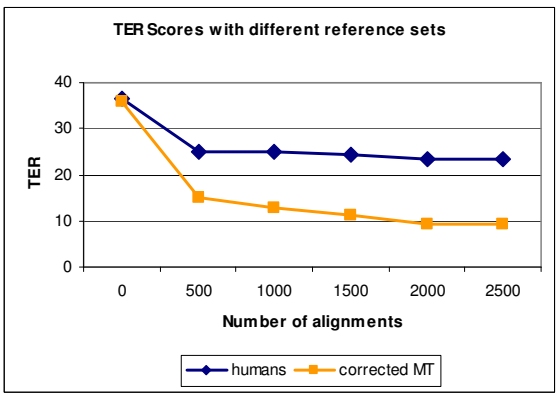

Figure 7: TER score against different reference sets

On closer inspection we found that some of the errors when compared to human references were due to inconsistencies in use of diacritics by human informants. For example, some translators occasionally left the caron off of the character "Č" and used "C" in their translations instead. As a result, our initial human translations occasionally had different diacritic marks than later translations from other human translators. To gauge the impact this inconsistency had on our translation metrics, we also ran evaluation of Bleu scores with diacritic marks removed. Figure 6 shows this comparison. By removing consideration of diacritic marks we get a 53% average improvement in scores. We wish to emphasize again that the scores are only meaningful in comparison between different levels of treatment—in our case, different numbers of translations aligned and decomposed into our lexicon. Figure 6 also includes an indication of where the upper limit on performance without diacritics should be. We compared our human corrected translation output (diacritics removed) to the initial human reference translations (diacritics removed.) We found that human corrected machine translations achieve a Bleu score of 0.880 when compared to human reference translations collected before machine translation. This gives an indication of the upper limit of the scores obtainable for "correct" machine translations.

Bleu presents one way to measure translations, based on accuracy of n-grams when compared to a pool of human references. A different way to measure the improvement of quality in the translations we generated is to examine the rate of error. Translation Error Rate (TER) is an alternative measure that counts the number of insertions, deletions, and other changes required to correct a translation (Snover et al., 2005). Like Bleu, TER compares the machine translation to human references. Unlike Bleu, it measures the required changes to reach the closest reference translation. We ran TER evaluations for each of the test conditions reported above. The results were consistent. Where Bleu scores went up TER scores go down—where translation quality goes up, error rates go down.

Figure 7 shows the translation error rate (TER) against different reference sets. These plots show the average error rate at different levels of aligned translations added to the lexicon. At zero alignments, the error rate is quite high; then it drops quickly at 500 and has an overall downward trend.

We experimented with the effect of removing the back-off dictionary. This simulates applying our approach in the absence of pre-existing resources. We also presented the output of our system to native Czech speakers and collected measures of acceptability and fluency. The details of these analyses are forthcoming (Murray, et al., in press). Only a small percentage of the translations had meanings that were far from the intended meaning. Disfluencies were manageable and were restricted to occasional errors in morphology and word order.

In summary, our results show an overall improvement in translation quality and an overall reduction in translation error after only a few hundred translations. We then see a continued drop in error rate and a continued rise in performance as more and more translations are added. While our evaluation used only 100 samples, we fully expect the trends observed in our experiments to hold across larger test sets.

5 Related work

Several studies have taken a knowledge-acquisition approach to collecting multilingual word pairs. For example, Sadat et al. (2003) automatically extracted bilingual word pairs from comparable corpora based on the assumption that if two words are mutual translations, then their most frequent collocates are likely to be mutual translations as well. Others have made similar mutual-translation assumptions for lexical acquisition (Echizen-ya, Araki, & Momouchi, 2005; Déjean, Gaussier, & Sadat, 2002; Kaji & Aizono, 1996; Rapp, 1999; Tanaka & Iwasaki, 1996). Most of these studies make use of either parallel corpora or a bilingual dictionary for the task of bilingual term extraction. While Echizen-ya et al. (2005) avoided using a bilingual dictionary, they required a parallel corpus to achieve their goal, whereas Fung (2000) and others have relied on a bilingual dictionary. In either case, the bilingual resources had to be somewhat large. In addition, these approaches focused on the ex-

traction of single-word pairs, not more complex phrasal units.

Several dictionary and thesaurus translation initiatives have adopted a term-by-term labor-intensive approach to the translation task. For example, in "Webster's Online Dictionary with Multilingual Thesaurus Translation" (Parker, 2006), the project goal is to create the largest dictionary of modern language usage by collecting human translations on the web over a period of years. However, the task is so unconstrained, it is difficult to determine when it is complete—and there is little room for taking advantage of regular structures or domain-specificity across the resource, as we have done in our thesaurus translation effort.

Many recent approaches to dictionary and thesaurus translation are geared toward domain-specificity. These approaches are motivated by a need to provide domain-specific thesauri to specialists in a particular field, e.g., medical terminology (Déjean, Gaussier, Renders, & Sadat, 2005) and agricultural terminology (Chun & Wenlin, 2002). Researchers working on these projects are faced with either finding human translators who are specialized enough to manage the domain-particular translations—or applying automatic techniques to large-scale parallel corpora where data sparsity makes it difficult to translate low-frequency terms. Data sparsity is also an issue for more general state-of-the-art bilingual alignment approaches (Brown, Della-Pietra, Della-Pietra, & Mercer, 1993; Melamed, 2000; Och & Ney, 2003; Wantanabe & Sumita, 2003).

6 Conclusion

We achieved acceptable machine translation of a thesaurus of 56,000 keyword phrases using information gathered from human translation of only 2,500 keyword phrases. For the lexicon we developed, our overall cost of human input was less than 900 €. Had we paid for human translation of the entire thesaurus it would have cost close to 20000 €. By prioritizing translations based on the *translation value* of the translated words and the *thesaurus value* of the keyword phrases in which they appear, we were able to optimize the rate of return on investment and choose an acceptable trade-off point. For this project we chose a point where less than 0.01% of the value of the thesaurus would be gained from each additional human translation. Our evaluations demonstrate that this choice produces a lexicon with significant positive impact on translation quality. For other applications a different trade-off point will be warranted depending on the initial OOV rate and the importance of detailed coverage.

The true value in our approach comes from establishing an operational value for the items to be translated and then optimizing the value gained from each human translation. In our case the items were keyword phrases arranged in a hierarchical thesaurus that describes an ontology of concepts. The operational value of these keyword phrases was based on their facility in providing individuals access to Holocaust survivors' testimonies. Many words in the thesaurus will still be outside the vocabulary of our newly generated lexicon. There are so many place names in the thesaurus that full coverage of the vocabulary would require many more translations than we have obtained so far. The salient point in our work is that we addressed the most important deficiencies in the vocabulary first, and that the gain from more translations will be smaller and smaller following a prioritized scheme. Choosing how far into the prioritized list to go is merely a function of the financial resources available for the task. We have shown that careful prioritization of translations, combined with an operational definition of *translation value* based on hierarchical arrangement of terms and reusability of components, can facilitate cost effective thesaurus translation with minimal human input.

7 Acknowledgments

We would like to thank Doug Oard for his invaluable contribution. We also wish to acknowledge the assistance of our Czech informants, in particular Robert Fischmann and Alena Prunerova. This research also benefited from programming efforts of Soumya Bhat, Nitin Madnani and Rebecca Hwa.

This work was supported in part by NSF IIS Award 0122466 and NSF CISE RI Award EIA0130422. Additional support also came from a grant of the MSMT CR #1P05ME786 and #MSM0021620838, and the Grant Agency of the CR #GA405/06/0589.

References

Brown, P. F., Della-Pietra, V. J., Della-Pietra, S. A., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263-311.

Chun, C., & Wenlin, L. (2002). The translation of agricultural multilingual thesaurus. In *Proceedings of the Third Asian Conference for Information Technology in Agriculture* Beijing, China.

CLSP - Center for Language and Speech Processing. (2005) *Multilingual Access to Large spoken ArCHives* http://www.clsp.jhu.edu/research/malach.

Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., & Kubon, V. (2004). Prague Czech-English dependecy treebank: Syntactically annotated resources for machine translation. In *4th International Conference on Language Resources and Evaluation* Lisbon, Portugal.

Déjean, H., Gaussier, E., Renders, J.-M., & Sadat, F. (2005). Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine, 33*(2), 111-124.

Déjean, H., Gaussier, E., & Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of COLING '02*.

Echizen-ya, H., Araki, K., & Momouchi, Y. (2005). Automatic acquisition of bilingual rules for extraction of bilingual word pairs from parallel corpora. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition* (pp. 87-96).

Fung, P. (2000). A statistical view of bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In Jean Veronis (ed.), *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers.

Gustman, Soergel, Oard, Byrne, Picheny, Ramabhadran, & Greenberg. (2002). Supporting Access to Large Digital Oral History Archives. In *Proceedings of JCDL* (pp. 18-27).

Kaji, H., & Aizono, T. (1996). Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of COLING '96* (pp. 23-28).

Madnani, N. (2004) *UMIACS Word Alignment Interface*. http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm

Melamed, I. D. (2000). Models of translation equivalence among words. *Computational Linguistics, 26*(2), 221-249.

Murray, G. C., Dorr, B., Lin, J., Pecina, P., & Hajič, J. (in press). Leveraging reusability: Cost-effective lexical acquisition for large-scale ontology translation. To appear in *Proceedings of COLIING/ACL '06*.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ALC* (pp. 331-318). 2002.

Parker, P. M. (2006) *Webster's Online Dictionary with Multilingual Thesaurus Translation* http://www.websters-online-dictionary.org

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the ACL* (pp. 519-526).

Sadat, F., Yoshikawa, M., & Uemura, S. (2003). Enhancing cross-language information retrieval by an automatic acquisition of bilingual terminology from comparable corpora . In *Proceedings of the 26th Annual ACM SIGIR* (pp. 397-398).

Snover, M., Dorr, B. J., Schwartz, R., Makhoul, J., Micciulla, L., & Weischedel, R. (2005). A study of translation error rate with targeted human annotation. Technical Reports *LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58* College Park: University of Maryland.

Tanaka, K., & Iwasaki, H. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of COLING '96* (pp. 580-585).

USC. (2006) *USC Shoah Foundation Institute for Visual History and Education* http://www.usc.edu/schools/college/vhi

Wantanabe, T., & Sumita, E. (2003). Example-based decoding for statistical machine translation. In *Proceedings of MT Summit IX* (pp. 410-417).