Leveraging Reusability: Cost-effective Lexical Acquisition for Large-scale Ontology Translation

G. Craig Murray
Bonnie J. Dorr
Jimmy Lin
Institute for Advanced Computer Studies
University of Maryland
{gcraigm,bdorr,jimmylin}@umd.edu

Jan Hajič
Pavel Pecina
Institute for Formal and Applied Linguistics
Charles University
{hajic,pecina}@ufal.mff.cuni.cz

Abstract

Thesauri and ontologies provide important value in facilitating access to digital archives by representing underlying principles of organization. Translation of such resources into multiple languages is an important component for providing multilingual access. However, the specificity of vocabulary terms in most ontologies precludes fully-automated machine translation using general-domain lexical resources. In this paper, we present an efficient process for leveraging human translations when constructing domain-specific lexical resources. We evaluate the effectiveness of this process by producing a probabilistic phrase dictionary and translating a thesaurus of 56,000 concepts used to catalogue a large archive of oral histories. Our experiments demonstrate a cost-effective technique for accurate machine translation of large ontologies.

1 Introduction

Multilingual access to digital collections is an important problem in today's increasingly interconnected world. Although technologies such as cross-language information retrieval and machine translation help humans access information they could not otherwise find or understand, they are often inadequate for highly specific domains.

Most digital collections of any significant size use a system of organization that facilitates easy access to collection contents. Generally, the organizing principles are captured in the form of a controlled vocabulary of keyword phrases (descriptors) representing specific concepts. These descriptors are usually arranged in a hierarchic thesaurus or ontology, and are assigned to collection items as a means of providing access (either via searching for keyword phases, browsing the hierarchy, or a combination both). MeSH (Medical Subject Headings) serves as a good example of such an ontology; it is a hierarchically-arranged collection of controlled vocabulary terms manually assigned to medical abstracts in a number of databases. It provides multilingual access to the contents of these databases, but maintaining translations of such a complex structure is challenging (Nelson, et al, 2004).

For the most part, research in multilingual information access focuses on the content of digital repositories themselves, often neglecting significant knowledge that is explicitly encoded in the associated ontologies. However, information systems cannot utilize such ontologies by simply applying off-the-shelf machine translation. General-purpose translation resources provide insufficient coverage of the vocabulary contained within these domain-specific ontologies.

This paper tackles the question of how one might efficiently translate a large-scale ontology to facilitate multilingual information access. If we need humans to assist in the translation process, how can we maximize access while minimizing cost? Because human translation is associated with a certain cost, it is preferable not to incur costs of retranslation whenever components of translated text are reused. Moreover, when exhaustive human translation is not practical, the most "useful" components should be translated first. Identifying reusable elements and prioritizing their translation based on utility is essential to maximizing effectiveness and reducing cost.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 945–952,
Sydney, July 2006.

We present a process of prioritized translation that balances the issues discussed above. Our work is situated in the context of the MALACH project, an NSF-funded effort to improve multilingual information access to large archives of spoken language (Gustman, et al., 2002). Our process leverages a small set of manually-acquired English-Czech translations to translate a large ontology of keyword phrases, thereby providing Czech speakers access to 116,000 hours of video testimonies in 32 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 human informants, aligned, decomposed and then recombined to cover 90% of the access value in a complex ontology. Moreover, we demonstrate that prioritization based on hierarchical position and frequency of use facilitates extremely efficient reuse of human input. Evaluations show that our technique is able to boost performance of a simple translation system by 65%.

2 The Problem

The USC Shoah Foundation Institute for Visual History and Education manages what is presently the world's largest archive of videotaped oral histories (USC, 2006). The archive contains 116,000 hours of video from the testimonies of over 52,000 survivors, liberators, rescuers and witnesses of the Holocaust. If viewed end to end, the collection amounts to 13 years of continuous video. The Shoah Foundation uses a hierarchically arranged thesaurus of 56,000 keyword phrases representing domain-specific concepts. These are assigned to time-points in the video testimonies as a means of indexing the video content. Although the testimonies in the collection represent 32 different languages, the thesaurus used to catalog them is currently available only in English. Our task was to translate this resource to facilitate multilingual access, with Czech as the first target language.

Our first pass at automating thesaurus translation revealed that only 15% of the words in the vocabulary could be found in an available aligned corpus (Čmejrek, et al., 2004). The rest of the vocabulary was not available from general resources. Lexical information for translating these terms had to be acquired from human input. Reliable access to digital archives requires accuracy. Highly accurate human translations incur a cost that is generally proportional to the number of words being translated. However, the keyword phrases in the Shoah Foundation's archive occur in a Zipfian distribution—a relatively small number of terms provide access to a large portion of the video content. Similarly, a great number of highly specific terms describe only a small fraction of content. Therefore, not every keyword phrase in the thesaurus carries the same value for access to the archive. The hierarchical arrangement of keyword phrases presents another issue: some concepts, while not of great value for access to segments of video, may be important for organizing other concepts and for browsing the hierarchy. These factors must be balanced in developing a cost-effective process that maximizes utility.

3 Our Solution

This paper presents a prioritized human-in-the-loop approach to translating large-scale ontologies that is fast, efficient, and cost effective. Using this approach, we collected 3,000 manual translations of keyword phrases and reused the translated terms to generate a lexicon for automated translation of the rest of the thesaurus. The process begins by prioritizing keyword phrases for manual translation in terms of their value in accessing the collection and the reusability of their component terms. Translations collected from one human informant are then checked and aligned to the original English terms by a second informant. From these alignments we induce a probabilistic English-Czech phrase dictionary.

To test the effectiveness of this process we implemented a simple translation system that utilizes the newly generated lexical resources. Section 4 reports on two evaluations of the translation output that quantify the effectiveness of our human-in-the-loop approach.

3.1 Maximizing Value and Reusability

To quantify their utility, we defined two values for each keyword phrase in the thesaurus: a *thesaurus value*, representing the importance of the keyword phrase for providing access to the collection, and a *translation value*, representing the usefulness of having the keyword phrase translated. These values are not identical, but the second is related to the first.

Thesaurus value: Keyword phrases in the Shoah Foundation's thesaurus are arranged into a poly-hierarchy in which child nodes may have multiple parents. Internal (non-leaf) nodes of the hierarchy are used to organize concepts and support concept browsing. Some internal nodes are also used to index video content. Leaf nodes are

very specific and are only used to index video content. Thus, the usefulness of any keyword phrase for providing access to the digital collection is directly related to the concept's position in the thesaurus hierarchy.

A fragment of the hierarchy is shown in Figure 1. The keyword phrase "Auschwitz II-Birkenau (Poland: Death Camp)", which describes a Nazi death camp, is assigned to 17,555 video segments in the collection. It has broader (parent) terms and narrower (child) terms. 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-Birkenau" and others.

From this example, we can see that an internal node is valuable in providing access to its children, even if the keyword phrase itself is not assigned to any segments. The value we assign to any term must reflect this fact. If we were to reduce cost by translating only the nodes assigned to video segments, we would neglect nodes that are crucial for browsing. However, if we value a node by the sum value of all its children, grandchildren, etc., the resulting calculation would bias the top of the hierarchy. Any prioritization based on this method would lead to translation of the top of the hierarchy first. Given limited resources, leaf nodes might never be translated. Support for searching and browsing calls for different approaches to prioritization.

To strike a balance between these factors, we calculate a *thesaurus value*, which represents the importance of each keyword phrase to the thesaurus as a whole. This value is computed as:

$$h_k = count(s_k) + \frac{\sum_{i \in children(k)} h_i}{|children(k)|}$$

For leaf nodes in our thesaurus, this value is simply the number of video segments to which the concept has been assigned. For parent nodes, the *thesaurus value* is the number of segments (if any) to which the node has been assigned, *plus* the average of the *thesaurus value* of any child nodes.

This recursive calculation yields a micro-averaged value that represents the reachability of segments via downward edge traversals from a given node in the hierarchy. That is, it gives a kind of weighted value for the number of segments described by a given keyword phrase or its narrower-term keyword phrases.

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

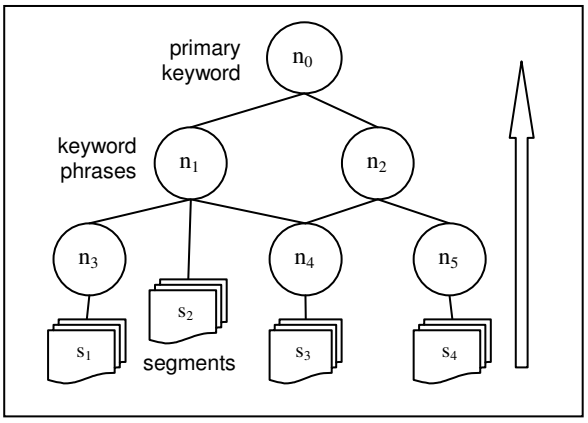

Figure 2. Bottom-up micro-averaging

For example, in Figure 2 each of the leaf nodes n_3, n_4, and n_5 have values based solely on the number of segments to which they are assigned. Node n_1 has value both as an access point to the segments at s_2 and as an access point to the keyword phrases at nodes n_3 and n_4. Other internal nodes, such as n_2 have value only in providing access to other nodes/keyword phrases. Working from the bottom of the hierarchy up to the primary node (n_0) we can compute the *thesaurus value* for each node in the hierarchy. In our example, we start with nodes n_3 through n_5, counting the number of the segments that have been assigned each keyword phrase. Then we move up to nodes n_1 and n_2. At n_1 we count the number of segments s_2 to which n_1 was assigned and add that count to the average of the *thesaurus values* for n_3, and n_4. At n_2 we simply average the *thesaurus values* for n_4 and n_5. The final values quantify how valuable the translation of any given keyword phrase would be in providing access to video segments.

Translation value: After obtaining the *thesaurus value* for each node, we can compute the *translation value* for each word in the vocabulary

as the sum of the thesaurus value for every keyword phrase that contains that word:

$$t_w = \sum_{k \in K_w} h_k \quad \text{where } K_w = \{x \mid \text{phrase } x \text{ contains } w\}$$

For example, the word "Auschwitz" occurs in 35 concepts. As a candidate for translation, it carries a large impact, both in terms of the number of keyword phrases that contains this word, and the potential value of those keyword phrases (once they are translated) in providing access to segments in the archive. The end result is a list of vocabulary words and the impact that correct translation of each word would have on the overall value of the translated thesaurus.

We elicited human translations of entire keyword phrases rather than individual vocabulary terms. Having humans translate individual words without their surrounding context would have been less efficient. Also, the value 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*). Some keyword phrases contain words with high *translation value*, but the keyword phrase itself has low *thesaurus value*. Thus, the value gained by translating any given phrase is more accurately estimated by the total value of any untranslated words it contains. Therefore, we prioritized the order of keyword phrase translations based on the *translation value* of the untranslated words in each keyword phrase.

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 have been translated. Starting from the assumption that the entire thesaurus is untranslated, we select the one keyword phrase that contains the most valuable un-translated words—we simply add up the *translation value* of all the untranslated words in each keyword phrase, and select the keyword phrase with the highest value. We add this keyword phrase to a prioritized list of items to be manually translated and we remove it from the list of untranslated phrases. We update our vocabulary list and, assuming translations of all the words in the prior keyword phrase to now be translated (neglecting issues such as morphology), we again select the keyword phrase that contains the most valuable untranslated words. We iterate the process until all vocabulary terms have been included at least one keyword phrases on the prioritized list. Ultimately we end up with an ordered list of the keyword phrases that should be translated to cover the entire vocabulary, with the most important words being covered first.

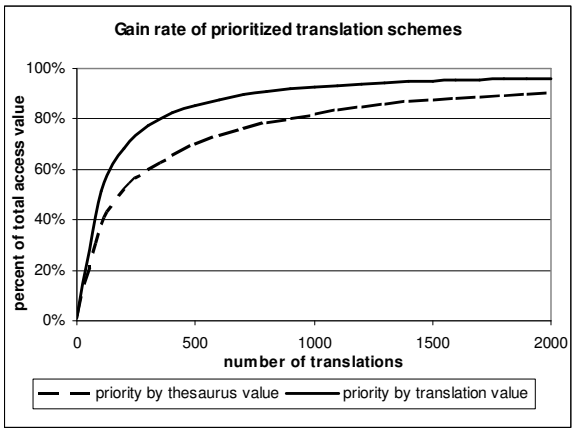

Figure 3. Gain rate of access value based on number of human translations

A few words about additional characteristics of this approach: note that it is greedy and biased toward longer keyword phrases. As a result, some words may be translated more than once because they appear in more than one keyword phrase with high *translation value*. This side effect is actually desirable. To build an accurate translation dictionary, it is helpful to have more than one translation of frequently occuring words, especially for morphologically rich languages such as Czech. Our technique makes the operational assumption that translations of a word gathered in one context can be reused in another context. Obviously this is not always true, but contexts of use are relatively stable in controlled vocabularies. Our evaluations address the acceptability of this operational assumption and demonstrate that the technique yields acceptable translations.

Following this process model, the most important elements of the thesaurus will be translated first, and the most important vocabulary terms will quickly become available for automated translation of keyword phrases with high *thesaurus value* that do not make it onto the prioritized list for manual translation (i.e., low *translation value*). The overall access value of the thesaurus rises very quickly after initial translations. With each subsequent human translation of keyword phrases on the prioritized list, we gain tremendous value in terms of providing non-English access to the collection of video testimonies. Figure 3 shows this rate of gain. It can be seen that prioritization based on *translation value* gives a much higher yield of total access than prioritization based on *thesaurus value*.

3.2 Alignment and Decomposition

Following the prioritization scheme above, we obtained professional translations for the top 3000 English keyword phrases. We tokenized these translations and presented them to another bilingual Czech speaker for verification and alignment. This second informant marked each Czech word in a translated keyword phrase with a link to the equivalent English word(s). Multiple links were used to convey the relationship between a single word in one language and a string of words in another. The output of the alignment process was then used to build a probabilistic dictionary of words and phrases.

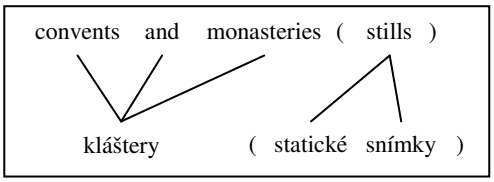

Figure 4. Sample alignment

Figure 4 shows an example of an aligned tranlsation. The word "stills" is recorded as a translation for "statické snímky" and "kláštery" is recorded as a translation for "convents and monasteries." We count the number of occurrences of each alignment in all of the translations and calculate probabilities for each Czech word or phrase given an English word or phrase. For example, in the top 3000 keyword phrases "stills" appears 29 times. It was aligned with "statické snímky" 28 times and only once with "statické záběry", giving us a translation probability of 28/29=0.9655 for "statické snímky".

Human translation of the 3000 English keyword phrases into Czech took approximately 70 hours, and the alignments took 55 hours. The overall cost of human input (translation and alignment) was less than 1000 €. The projected cost of full translation for the entire thesaurus would have been close to 20000 € and would not have produced any reusable resources. Naturally, costs for building resources in this manner will vary, but in our case the cost savings is approximately twenty fold.

3.3 Machine Translation

To demonstrate the effectiveness of our approach, we show that a probabilistic dictionary, induced through the process we just described, facilitates high quality machine translation of the rest of the thesaurus. We evaluated translation quality using a relatively simple translation system. However, more sophisticated systems can draw equal benefit from the same lexical resources.

Our translation system implemented a greedy coverage algorithm with a simple back-off strategy. It first scans the English input to find the longest matching substring in our dictionary, and replaces it with the most likely Czech translation. Building on the example above, the system looks up "monasteries and convents stills" in the dictionary, finds no translation, and backs off to "monasteries and convents", which is translated to "kláštery". Had this phrase translation not been found, the system would have attempted to find a match for the individual tokens. Failing a match in our dictionary, the system then backs off to the Prague Czech-English Dependency Treebank dictionary, a much larger dictionary with broader scope. If no match is found in either dictionary for the full token, we stem the token and look for matches based on the stem. Finally, tokens whose translations can not be found are simply passed through untranslated.

A minimal set of heuristic rules was applied to reordering the Czech tokens but the output is primarily phrase by phrase/word by word translation. 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 translation accuracy for our system can be directly interpreted as quality measures for the lexical resources used and the process by which they were developed.

4 Evaluation

We performed two different types of evaluation to validate our process. First, we compared our system output to human reference translations using Bleu (Papineni, et al., 2002), a widely-accepted objective metric for evaluation of machine translations. Second, we showed corrected and uncorrected machine translations to Czech speakers and collected subjective judgments of fluency and accuracy.

For evaluation purposes, we selected 418 keyword phrases to be used as target translations. These phrases were selected using a stratified sampling technique so that different levels of *thesaurus value* would be represented. There was no overlap between these keyword phrases and the 3000 prioritized keyword phrases used to build our lexicon. Prior to machine translation we obtained at least two independent human-generated reference translations for each of the 418 keyword phrases.

After collecting the first 2500 prioritized translations, we induced a probabilistic dictionary and generated machine translations of the 418 target keyword phrases. These were then corrected by native Czech speakers, who adjusted word order, word choice, and morphology. We use this set of human-corrected machine translations as a second reference for evaluation.

Measuring the difference between our uncorrected machine translations (MT) and the human-generated reference establishes how accurate our translations are compared to an independently established target. Measuring the difference between our MT and the human-corrected machine translations (corrected MT) establishes how acceptable our translations are. We also measured the difference between corrected MT and the human-generated translations. We take this to be an upper bound on realistic system performance.

The results from our objective evaluation are shown in Figure 5. Each set of bars in the graph shows performance after adding a different number of aligned translations into the lexicon (i.e., performance after adding 500, 1000, ..., 3000 aligned translations.) The zero condition is our baseline: translations generated using only the dictionary available in the Prague Czech-English Dependency Treebank. Three different reference sets are shown: human-generated, corrected MT, and a combination of the two.

There is a notable jump in Bleu score after the very first translations are added into our probabilistic dictionary. Without any elicitation and alignment we got a baseline score of 0.46 (against the human-generated reference translations). After the aligned terms from only 500 translations were added to our dictionary, our Bleu score rose to 0.66. After aligned terms from 3000 translations were added, we achieved 0.69. Using corrected MT as the reference our Bleu scores improve from 0.48 to 0.79. If human-generated and human-corrected references are both considered to be correct translations, the improvement goes from .49 to .80. Regardless of the reference set, there is a consistent performance improvement as more and more translations are added. We found the same trend using the TER metric on a smaller data set (Murray, et al., 2006). The fact that the Bleu scores continue to rise indicates that our approach is successful in quickly expanding the lexicon with accurate translations. It is important to point out that Bleu scores are not meaningful in an absolute sense; the scores here should be interpreted with respect to each other. The trend

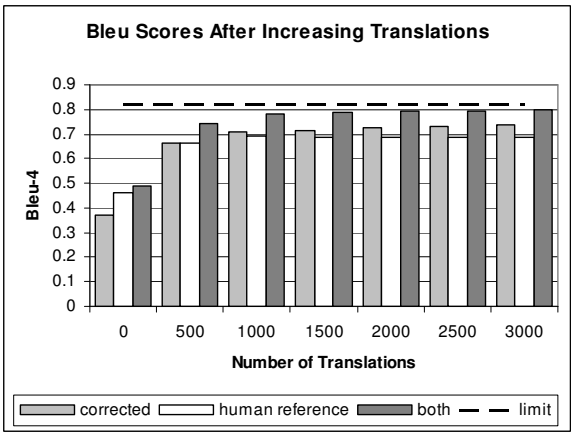

Figure 5. Objective evaluation results

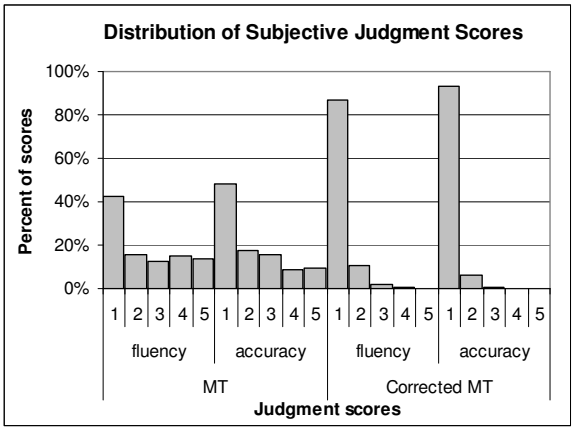

Figure 6. Subjective evaluation results

in scores strongly indicates that our prioritization scheme is effective for generating a high-quality translation lexicon at relatively low cost.

To determine an upper bound on machine performance, we compared our corrected MT output to the initial human-generated reference translations, which were collected prior to machine translation. Corrected MT achieved a Bleu score of 0.82 when compared to the human-generated reference translations. This upper bound is the "limit" indicated in Figure 5.

To determine the impact of external resources, we removed the Prague Czech-English Dependency Treebank dictionary as a back-off resource and retranslated keyword phrases using only the lexicons induced from our aligned translations. The results of this experiment showed only marginal degradation of the output. Even when as few as 500 aligned translations were used for our dictionary, we still achieved a Bleu score of 0.65 against the human reference translations. This means that even for languages where prior resources are not available our prioritization scheme successfully addresses the OOV problem.

In our subjective evaluation, we presented a random sample of our system output to seven

native Czech speakers and collected judgments of accuracy and fluency using a 5-point Likert scale (1=good, 3=neutral, 5=bad). An overview of the results is presented in Figure 6. Scores are shown for corrected and uncorrected MT. In all cases, the mode is 1 (i.e., good fluency and good accuracy). 59% of the machine translated phrases were rated 2 or better for fluency. 66% were rated 2 or better for accuracy. Only a small percentage of the translations had meanings that were far from the intended meaning. Disfluencies were primarily due to errors in morphology and word order.

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. This approach is based on the simple assumption that if two words are mutual translations, then their most frequent collocates are likely to be mutual translations as well. However, the approach requires large comparable corpora, the collection of which presents non-trivial challenges. Others have made similar mutual-translation assumptions for lexical acquisition (Echizen-ya, et al., 2005; Kaji & Aizono, 1996; Rapp, 1999; Tanaka & Iwasaki, 1996). Most make use of either parallel corpora or a bilingual dictionary for the task of bilingual term extraction. Echizen-ya, et al. (2005) avoided using a bilingual dictionary, but required a parallel corpus to achieve their goal; whereas Fung (2000) and others have relied on pre-existing bilingual dictionaries. In either case, large bilingual resources of some kind are required. In addition, these approaches focused on the extraction of single-word pairs, not phrasal units.

Many recent approaches to dictionary and thesaurus translation are geared toward providing domain-specific thesauri to specialists in a particular field, e.g., medical terminology (Déjean, et al., 2005) and agricultural terminology (Chun & Wenlin, 2002). Researchers 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 poses a problem for low-frequency terms. Data sparsity is also an issue for more general state-of-the-art bilingual alignment approaches (Brown, et al., 2000; Och & Ney, 2003; Wantanabe & Sumita, 2003).

6 Conclusion

The task of translating large ontologies can be recast as a problem of implementing fast and efficient processes for acquiring task-specific lexical resources. We developed a method for prioritizing keyword phrases from an English thesaurus of concepts and elicited Czech translations for a subset of the keyword phrases. From these, we decomposed phrase elements for reuse in an English-Czech probabilistic dictionary. We then applied the dictionary in machine translation of the rest of the thesaurus.

Our results show an overall improvement in machine translation quality after collecting only a few hundred human translations. Translation quality continued to rise as more and more human translations were added. The test data used in our evaluations are small relative to the overall task. However, we fully expect these results to hold across larger samples and for more sophisticated translation systems.

We leveraged the reusability of translated words to translate a thesaurus of 56,000 keyword phrases using information gathered from only 3000 manual translations. Our probabilistic dictionary was acquired at a fraction of the cost of manually translating the entire thesaurus. By prioritizing human translations based on the *translation value* of the words and the *thesaurus value* of the keyword phrases in which they appear, we optimized the rate of return on investment. This allowed us to choose a trade-off point between cost and utility. For this project we chose to stop human translation at a point where less than 0.01% of the value of the thesaurus would be gained from each additional human translation. This choice produced a high-quality lexicon with significant positive impact on machine translation systems. For other applications, a different trade-off point will be appropriate, depending on the initial OOV rate and the importance of detailed coverage.

The value of our work lies in the process model we developed for cost-effective elicitation of lexical resources. The metrics we established for assessing the impact of each translation item are key to our approach. We use these to optimize 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 determined by the access they provide to video segments in a large archive of oral histories. However, our technique is not limited to this application.

We have shown that careful prioritization of elicited human translations facilitates cost-effective thesaurus translation with minimal human input. Our use of a prioritization scheme addresses the most important deficiencies in the vocabulary first. We induced a framework where the utility of lexical resources gained from each additional human translation becomes smaller and smaller. Under such a framework, choosing the number of human translation to elicit becomes merely a function of the financial resources available for the task.

Acknowledgments

Our thanks to Doug Oard for his contribution to this work. Thanks also to our Czech informants: Robert Fischmann, Eliska Kozakova, Alena Prunerova and Martin Smok; and to Soumya Bhat for her programming efforts.

This work was supported in part by NSF IIS Award 0122466 and NSF CISE RI Award EIA0130422. Additional support also came from grants 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: Chinese Academy of Agricultural Sciences (CAAS) and Asian Federation for Information Technology in Agriculture (AFITA).

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.

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 the Joint Conference on Digital Libraries*. Portland, Oregon. (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).

Murray, G. C., Dorr, B., Lin, J., Hajič, J., & Pecina, P. (2006). Leveraging recurrent phrase structure in large-scale ontology translation. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*. Oslo, Norway.

Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J.-L., & Arluk, N. (2004). The MeSH translation maintenance system: Structure, interface design, and implementation. In *Proceedings of the 11th World Congress on Medical Informatics*. (pp. 67-69). Amsterdam: IOS Press.

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 Association for Computational Linguistics* (pp. 331-318).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. (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 International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 397-398).

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, [online] 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).