Journal : 10579 | Dispatch : 20-9-2008 | Pages : 14
Article No. : 9074 | LE | TYPESET
MS Code : CHUM641R2 | CP | DISK

Lang Resources & Evaluation
DOI 10.1007/s10579-008-9074-8

A cost-effective lexical acquisition process for large-scale thesaurus translation

Jimmy Lin · G. Craig Murray · Bonnie J. Dorr · Jan Hajič · Pavel Pecina

Abstract Thesauri and controlled vocabularies facilitate access to digital collections by explicitly representing the 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 thesauri precludes fully-automatic translation using general-domain lexical resources. In this paper, we present an efficient process for leveraging human translations to construct domain-specific lexical resources. This process is illustrated on a thesaurus of 56,000 concepts used to catalog a large archive of oral histories. We elicited human translations on a small subset of concepts, induced a probabilistic phrase dictionary from these translations, and used the resulting resource to automatically translate the rest of the thesaurus. Two separate evaluations demonstrate the acceptability of the automatic translations and the cost-effectiveness of our approach.

Keywords Thesauri · Controlled vocabularies · Manual translation process

1 Introduction

Providing multilingual access to digital collections is an important challenge in today's increasingly interconnected world. For the most part, research in multilingual access focuses on the content of digital repositories themselves, often neglecting significant knowledge present in the associated metadata. Many

J. Lin (✉) · G. C. Murray · B. J. Dorr
University of Maryland, College Park, MD, USA
e-mail: jimmylin@umd.edu

J. Hajič · P. Pecina
Charles University, Prague, Czech Republic

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

collections employ controlled-vocabulary descriptors, hierarchically arranged in a thesaurus, to characterize the content of items in the collection. Such structures explicitly encode the organizing principles of a collection and facilitate access via searching, browsing, or a combination of both. Multilingual access to such thesauri can enhance content-oriented technologies such as cross-language information retrieval and machine translation in helping users access content in foreign languages.

Building on previous work (Murray et al. 2006a, b), this article tackles the question of how one might, given limited resources, efficiently translate a large thesaurus to facilitate multilingual information access. Due to limited vocabulary coverage, off-the-shelf translation technology provides little help for specialized domains. Instead, we propose a process for lexical acquisition that yields high-value reusable resources for automatic translation. The key to a cost-effective process is to model the utility of each descriptor within the thesaurus, taking into account thesaurus structure and the reusability of component phrases. Guided by such a utility function, we elicited manual translations for a small selection of thesaurus terms, and from these induced lexical resources for translating the rest of the terms automatically. Experiments suggest that our approach yields acceptable translations and provides significant cost-savings compared to an unoptimized translation process.

2 The problem

Our work is situated in the context of MALACH (Multilingual Access to Large Spoken Archives), an effort funded by the U.S. National Science Foundation (Gustman et al. 2002). 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 h of video testimonies of over 52,000 survivors, liberators, rescuers, and witnesses of the Holocaust. The Shoah Foundation uses a hierarchically-arranged thesaurus that contains approximately 56,000 domain-specific concepts represented by keyword phrases (descriptors). These descriptors are assigned to time points in the video testimonies as a means for indexing the video content. Although testimonies are available in other languages, the thesaurus is currently available only in English. Translation of this resource into different languages would greatly enhance multilingual access. As a proof-of-concept, this article focuses on translation into Czech.

Initial attempts to automatically translate the thesaurus revealed that only 15% of the vocabulary could be found in an available aligned corpus, the Prague Czech-English Dependency Treebank (PCEDT) (Čmejrek et al. 2004). Due to the specificity of the domain, translations for the remaining terms could not be found in general electronic resources, including dictionaries at our disposal. Since reliable access requires high accuracy, we found it necessary to acquire lexical information from humans. However, it would be cost prohibitive to manually translate the entire thesaurus. Our solution involves acquiring human translations for a small selection of phrases from the thesaurus and then leveraging this information to automatically translate the remainder.

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

In this work, we propose a human-assisted translation process that takes into account characteristics of the thesaurus. A relatively small number of keyword phrases provides access to a large portion of the video content. Similarly, a large number of highly specific phrases describe only a small fraction of content. Therefore, not every phrase carries the same utility. The hierarchical arrangement of the keyword phrases presents another challenge: some phrases, while not of great value for directly accessing content, may be important for organizing other concepts and for browsing. These factors must be balanced in developing a cost-effective translation process.

3 A proposed solution

This article presents a cost-effective, human-in-the-loop approach to translating large thesauri. Using this approach, we collected 3,000 manual translations of keyword phrases from the Shoah Foundation's thesaurus and reused the translated terms to generate a lexicon, which was then employed to automatically translate the rest of the thesaurus. This section describes our process model in detail.

3.1 Prioritization

Given unlimited financial resources, one could simply elicit manual translations for all concepts in a thesaurus. However, since most projects face resource constraints, one must devise a prioritization scheme for manual translation, placing "more useful" terms before "less useful" ones. We define two measures to quantify the utility of a keyword phrase in our thesaurus: *thesaurus value*, which represents the importance of a particular keyword phrase for providing access to the collection, and *translation value*, which quantifies the usefulness of having the keyword phrase translated. We describe these measures in detail.

Keyword phrases in the Shoah Foundation's thesaurus are arranged in a poly-hierarchy where nodes can have multiple parents. Internal (non-leaf) nodes of the hierarchy are primarily used to organize concepts and support browsing, although some of these nodes are also used to index video content. Leaf nodes represent specific concepts and are only used for indexing. Thus, the utility of a keyword phrase for providing access to the collection is directly related to the concept's position in the thesaurus hierarchy.

A concrete example will help to make this clear. Consider the fragment of the thesaurus hierarchy shown in Fig. 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, but not all, of the broader and narrower terms are also assigned to segments. Notably, "German death camps" is not assigned to any video segments, although it is important because it facilitates access to six frequently assigned narrower terms. This example demonstrates the value of internal nodes in providing access to the structure of the thesaurus, even when those concepts are not directly

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

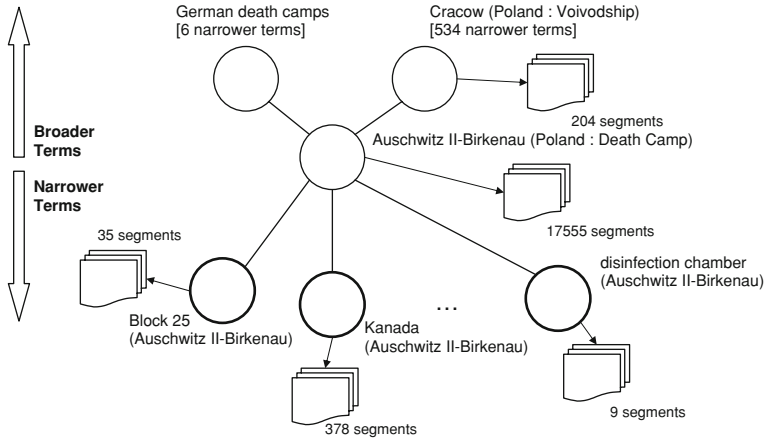

Fig. 1 A fragment from the thesaurus showing broader and narrower terms. Nodes with thick borders indicate leaf nodes

assigned to any content segments. As such, translation of these nodes provides multi-lingual access to the arrangement concepts within the thesaurus.

We interpret the number of video segments under any given node in the hierarchy (directly or via child nodes) as an indication of that node's potential importance for accessing collection content. We have no principled reason to assume that any particular video segment is more important than any other. Therefore, we treat each as equally important. We use these counts to estimate the importance of each keyword phrase's inclusion in the thesaurus and, by extension, of the utility gained from translating that keyword phrase.

Parent nodes high in the hierarchy help users manage concepts, but nodes low in the hierarchy are closer to the content. The hierarchy organizes content, but navigating the nodes comes at some cost to the user (e.g., cognitive load, physical interaction, etc.) Thus, our utility function must balance the value of direct content access with the value supplied by internal nodes that provide structure and facilitate browsing. To strike this balance, we introduce *thesaurus value*, which quantifies the importance of each keyword phrase with respect to the thesaurus:

$$h_k = \text{count}(s_k) + \frac{\sum_{i \in \text{children(k)}} h_i}{|\text{children}(k)|} \quad (1)$$

For leaf nodes in our thesaurus, this value is simply the number of video segments s to which the concept k has been assigned. For non-leaf nodes, the thesaurus value is the number of segments (if any) to which the concept has been assigned, plus the average of the thesaurus values of 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, the thesaurus value captures the number of segments described by a given keyword phrase and the average number of segments described by its children in the hierarchy (i.e. narrower terms).

Journal : 10579 | Dispatch : 20-9-2008 | Pages : 14
Article No. : 9074 | LE | TYPESET
MS Code : CHUM641R2 | CP | DISK

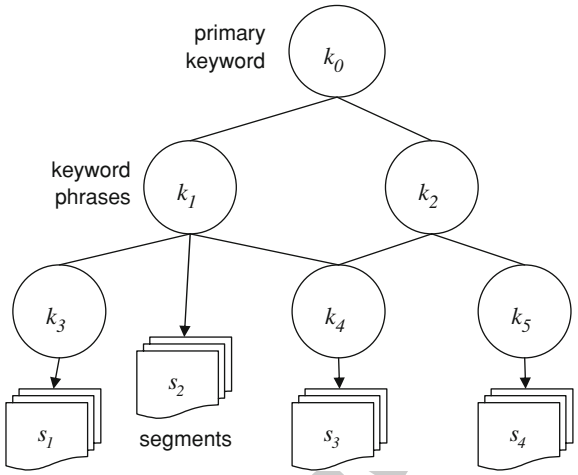

Fig. 2 Thesaurus fragment illustrating the computation of thesaurus values

For example, in Fig. 2, each of the leaf nodes (k_3, k_4, and k_5) has value only as a means for directly accessing content (s_1, s_3, and s_4). Node k_1 has value both as a direct access point to segments s_2 and indirectly to segments s_1 and s_3 (via k_3 and k_4). Other internal nodes, such as k_2, have value only in providing access to other keyword phrases (k_4 and k_5). Working our way up from the bottom of the hierarchy, we can compute the thesaurus value for each node in this simple example as follows: For nodes k_3 through k_5, we simply count the number of segments that have been assigned each keyword phrase. Then we move up to nodes k_1 and k_2. At k_1 we count the number of segments s_2 to which k_1 was assigned and add that count to the average of the thesaurus values for k_3 and k_4. At k_2 we simply average the thesaurus values for k_4 and k_5. And so on up the hierarchy. The final values quantify the utility of keyword phrases in providing access to video content. Although it would make some sense to prioritize human translations based simply on these thesaurus values, we can gain even more efficiency by taking into account the utility of individual lexical components within the keyword phrase.

Our example in Fig. 1 also illustrates the recurrent nature of the individual words that make up keyword phrases. Note that the term "Auschwitz" appears in four of the keyword phrases shown. In fact, the term "Auschwitz" occurs in 35 keyword phrases in the English thesaurus, and these are used as content descriptors for a significant portion of the archive. Thus, the impact of translating any individual term (i.e., word) is a function of the cumulative thesaurus value of all the keyword phrases in which it occurs. As a candidate for translation, "Auschwitz" has high potential impact, both in the number of keyword phrases that contain this term, and the value of those keyword phrases (once translated) in providing multi-lingual access to video segments in the archive.

Here, we introduce a measure of the *translation value* for each term (i.e., word) in the vocabulary. After obtaining the thesaurus values for each keyword phrase, we compute the translation value as the sum of the thesaurus value for every keyword phrase in which the term appears:

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

$$t_w = \sum_{k \in K_w} h_k \quad \text{where } K_w = \{x | \text{phrase } x \text{ contains } w\} \tag{2}$$

The end result of computing translation values is a list of terms and the impact that the correct translation of each term will have on the overall value of the translated thesaurus.

Accurate translation of individual terms requires context. Therefore, we elicited human translations of entire keyword phrases rather than individual terms. But how best to prioritize these translations? The value that any keyword phrase has for translation is only indirectly related to its own value as a point of access to the collection. Some keyword phrases have low thesaurus value but contain terms with high translation value. The impact of translating those keywords is not directly reflected by their use in describing the collection (i.e., their thesaurus value). Thus, the value gained by translating any given keyword phrase is more accurately estimated by the total value of any untranslated terms it contains. Therefore, we prioritized keyword phrases based on the translation value of the untranslated terms in each keyword phrase.

This process is implemented as follows: we iterate through the thesaurus keyword phrases, prioritizing their translation based on the assumption that any terms contained in a keyword phrase of higher priority would already have been translated. Starting from the assumption that the entire thesaurus is untranslated, we choose the keyword phrase that contains the most valuable untranslated terms. This is done by adding up the translation value of all the untranslated terms in each keyword phrase and selecting the keyword phrase with the highest sum. We add this keyword phrase to the prioritized list of items to be manually translated and remove it from the list of untranslated phrases. We update our vocabulary list, assuming that all the terms in the keyword phrase are now translated (neglecting issues such as morphology). Then, we again select the keyword phrase that contains the most valuable untranslated terms. This process iterates until all terms have been added to the prioritized list.

Note that this prioritization scheme is greedy and biased toward longer keyword phrases. In addition, some terms may be translated more than once because they appear in more than one keyword phrase with high (total) translation value.[1] This side effect is actually desirable. To build an accurate translation dictionary, it is helpful to have more than one translation of frequently occurring terms, especially for morphologically rich languages such as Czech. Our approach assumes that translations of terms 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. The longer keyword phrases provide richer contextual support for the translations. Our evaluations examine the validity of this context assumption and demonstrate that the technique yields acceptable translations.

[1] Even after a term is assumed to be translated, there will be keyword phrases containing that term which contain other high translation value terms not yet translated. In some cases, the sum of the translation value of the untranslated terms will be high enough to warrant addition of the keyword phrase to the prioritized list, despite the already translated term.

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

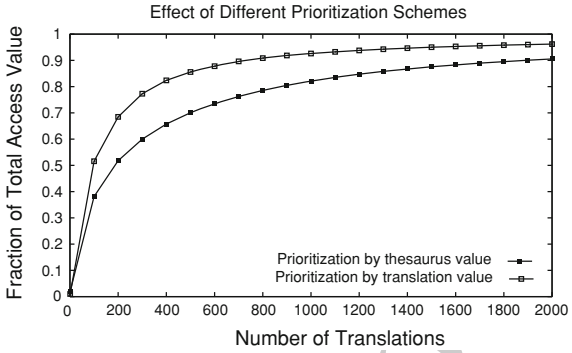

Fig. 3 Efficiency of different prioritization schemes

Following the process described above, the most important elements of the thesaurus will be translated first, and the most important vocabulary terms will quickly become available for automatic translation of those keyword phrases with high thesaurus value that do not make it onto the prioritized list for manual translation. To evaluate our prioritization scheme we need to quantify the accessibility of the collection (via the translated thesaurus) at different levels of human translation. We measure *access value* as the sum of the thesaurus value of translated keywords—whether by manual or automatic means. Access value represents the utility of the thesaurus after machine translation in providing multilingual access to the contents of the archive. Figure 3 plots the rate of gain in access value after eliciting translations. It can be seen that prioritizing elicited translations based on translation value yields a more efficient process than prioritization based on thesaurus value.

3.2 Caveats, alternatives, and possible improvements

We introduced three measures in this work: *Thesaurus value* is a measure of the contribution each keyword phrase makes to the overall value of the thesaurus. *Translation value* is a measure of the contribution each translated vocabulary term makes to the overall translation of the thesaurus. *Access value* is a measure of the collection access facilitated by a translated thesaurus. These measures come from a careful analysis of the problem of prioritized partial translation. Nevertheless, there are some operational assumptions which deserve further discussion.

In our definition of thesaurus value, we did not attempt to quantify the relative importance of browsing the concept hierarchy vs. accessing collection content. In some settings it may be more important to facilitate content access over concept browsing, or vice versa. It would be possible to add a weighting constant to Eq. 1, giving emphasis to either one or the other.

Descriptors in a thesaurus have two functions: indexers use them to index the collection, and patrons use them to retrieve collection contents. We chose not to model the frequency of their use to retrieve content. Patrons' use of English keyword phrases in queries could be used to estimate the expected use of keyword phrases in

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

another language. However, this would reflect only those interests the collection has served in the past. With a large and growing collection of oral histories it is impossible to know what interests it will serve in the future. Smaller collections and narrower domains may profit from an analysis of the patrons' queries.

We mention earlier that each video segment was assumed to be of equal importance. In some settings this may not be the case. For example, most patrons will have a preference for video testimonies in their native language. We found that the frequency distribution of keyword assignments to Czech content is similar to that of English, but there may be a biased distribution for other languages. For these it might be advantageous to give higher importance to segments in that language. Purpose of use is also a factor: for example, some of the collection content is cleared for broad use, some only for limited use. If the purpose of thesaurus translation were to provide access only for broad use, limited use segments should be discounted or excluded when calculating thesaurus value.

Our measure for the translation value of vocabulary terms is based on the thesaurus value of the keyword phrases, and so it inherits assumptions of that measure. Translation value has the further operational assumption that terms are equally informative about their language. This is obviously false—some terms will carry a great deal of information about the coding scheme of their language (i.e. morphology, syntax, etc.), while others will not. From this view, translations of complex terms may be more valuable. Quantifying that value, however, requires a means of identifying complex terms and of weighing the value of different language features.

Access value also inherits the assumptions of translation value and comes with certain caveats. It is an approximation based solely on collection content and is used here only to compare different prioritization schemes. It does not include an assessment of usability and is not intended to measure the translation output in an absolute sense. We report on quality of translations later in this paper.

Each of these measures could be expanded in different ways to suit different purposes. Changes to thesaurus value or to translation value will result in changes to the prioritized list of phrases to be translated. Choice of different utility functions for measuring access value will give different views of the success of prioritization. Researchers wishing to expand these measures should be careful to justify any added complexity with clear purpose.

3.3 Human translation, alignment, and decomposition

Following the prioritization scheme described above, we obtained professional Czech translations for the top 3,000 keyword phrases. We tokenized these translations and presented them to another bilingual Czech speaker for alignment. This second informant linked equivalent Czech and English words using a GUI. Multiple links conveyed the relationship between a single word in one language and a phrase in the other. Details of the alignment step can be found in Murray et al. 2006a. Human translation of the keyword phrases took approximately 70 h, and the alignments took 55 h. The overall cost of human input (translation and alignment) was less than 1,000€.

Journal : 10579 | Dispatch : 20-9-2008 | Pages : 14
Article No. : 9074 | LE | TYPESET
MS Code : CHUM641R2 | CP | DISK

From the human output, we constructed a probabilistic English-Czech phrase dictionary based on the distribution of the alignments. For example, in the top 3,000 keyword phrases "stills" appeared 29 times. It was aligned with "statické snímky" 28 times and only once with "statické záběry", giving us a translation probability of 28/29 = 0.966 for "statické snímky".

3.4 Automatic translation

To demonstrate the effectiveness of our approach, we show that a probabilistic dictionary, generated using the process we just described, facilitates high quality automatic translation. Our translation system implements a greedy algorithm with a simple back-off strategy. It first scans the English input to find the longest matching substring in our dictionary, and replaces the substring with the most likely Czech translation. For example, given the phrase "monasteries and convents (stills)", the system first looks for the entire phrase in the dictionary, but finds no translation. Then, the system backs off to "monasteries and convents" and finds the translation "kláštery". Next, the system tries to find a match for "stills" in the same manner.

If the system fails to find a match in our lexical resources, it backs off to a dictionary induced from the PCEDT (Čmejrek et al. 2004). If no match is found in either dictionary for the full token, the process is repeated with the stem. Failing a match on the stem, terms are simply passed through untranslated. A minimal set of heuristic rules is then applied to reorder the Czech tokens, but the output is primarily word-by-word lookup translation.

As this work focuses primarily on processes for human translation, we were not concerned about the simplicity of our system (compared to state-of-the-art statistical MT technology). In our case, we interpret measures of translation accuracy as quality measures for lexical resources. The simplicity of our system ensures that improvements to lexical coverage are not conflated with other factors. More sophisticated systems will no doubt also benefit from these resources.

4 Evaluation

We evaluated our translation process in two different ways. First, we compared automatic translations with human reference translations using BLEU (Papineni et al. 2002) and TER (Snover et al. 2005), two commonly-used metrics for automatic evaluation of MT output. Second, we presented automatic translations to Czech speakers and gathered subjective judgments of fluency and accuracy.

For evaluation, we selected 418 keyword phrases using a stratified sampling technique so that items with a broad range of thesaurus values would be represented. However, we ensured that there was no overlap between these keyword phrases and the 3,000 manually-translated keyword phrases used to build our lexicon.

For the automatic evaluation, we obtained two separate sets of reference translations. First, prior to automatic translation, we gathered at least two independent human translations for each keyword phrase. We refer to this as the "independent

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

reference" set. Second, we asked our informants to correct automatic translations into fluent Czech, preserving as much of the original machine output as possible. For these, we automatically translated the test set using a probabilistic dictionary that was generated using the first 2,500 prioritized translations. The machine output was then corrected by native Czech speakers, who adjusted word order, word choice, morphology, etc. We refer to this as the "human corrected" set. These translations often differed from the independent references, since there are multiple ways to translate the same phrase.

To assess the effectiveness of our translation process, we compared uncorrected automatic translations to the two different sets of reference translations. These results are shown in Fig. 4, with BLEU on the top and (1-TER) on the bottom. The *x* axis shows the number of aligned human translations used to construct the lexicon. The zero

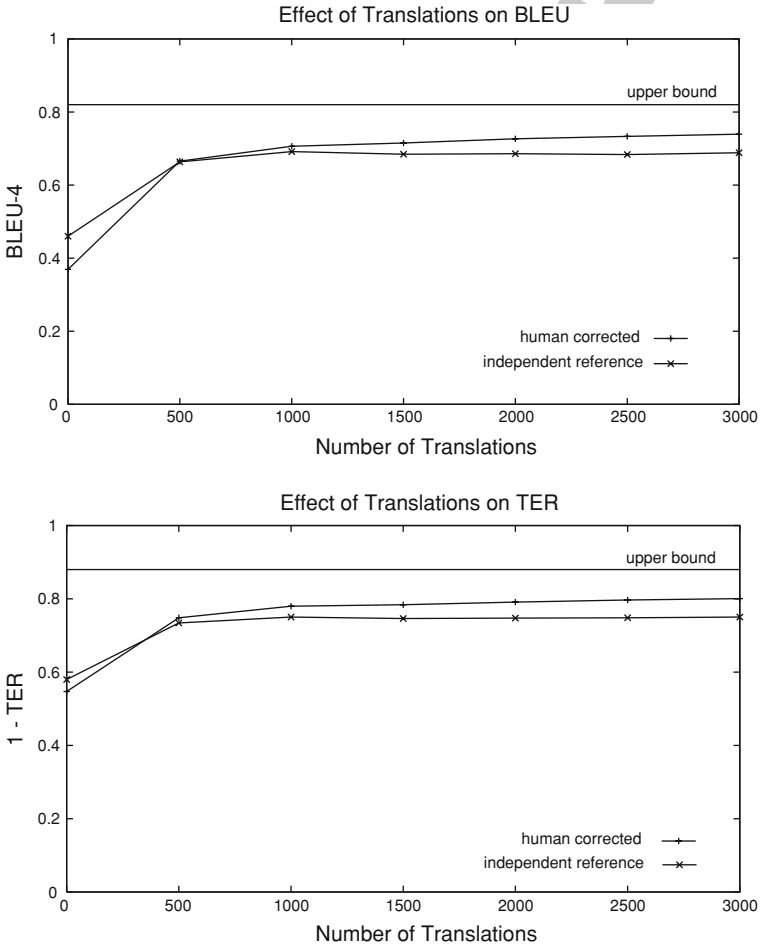

Fig. 4 Automatic evaluation of machine translations with BLEU (top) and TER (bottom) using two reference sets: "independent reference" and "human corrected"

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

condition represents our baseline: translations generated using only the dictionary available in the PCEDT. We take the performance of the human corrected translations with respect to the independent references as the upper bound, shown in both graphs. There is a big jump in both BLEU and (1-TER) scores after the first 500 translations are added to our probabilistic dictionary. Gains thereafter are smaller, but noticeable. In both cases, it appears that performance approaches the upper bound.

To determine the impact of external resources, we removed the PCEDT 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, we still achieved a BLEU score of 0.65 against the independent references. This suggests that even for languages where no resources are available, our process is capable of coping with vocabulary coverage issues.

In our subjective evaluation, we presented a random sample of automatic translations and corrected translations (i.e., the "human corrected" set described above) to seven native Czech speakers. They were asked to rate the fluency and accuracy of the phrases on a 5-point Likert scale (1 = good, 5 = bad). Results are shown in Fig. 5. In all cases, the mode is 1 (i.e., "good"). According to our judges, 59% of the uncorrected automatic translations were rated 2 or better for fluency; 66% were rated 2 or better for accuracy. Disfluencies were primarily caused by errors in morphology and word order; for more details, see (Murray et al. 2006b). We note that lexical accuracy is more important than grammatical fluency for providing information access.

5 Related work

The notion of human-assisted machine translation is not new, and human input has been used to great effect in the past. The Pangloss project (Frederking et al. 1994) developed an MT system where human assistance was solicited during the translation process. Other approaches to human-in-the-loop translation have involved more sophisticated symbolic representations (Olsen et al. 1998; Sabarís et al. 2001). These detail-oriented approaches tend to be knowledge-intensive and difficult to economize. Our study takes a cost-oriented approach.

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. Others have leveraged parallel corpora or bilingual dictionaries for lexical acquisition (Echizen-ya et al. 2006; Kaji and Aizono 1996; Rapp 1999; Tanaka and Iwasaki, 1996). However, our work deals with the fundamentally different task of translating a large thesaurus, where one can leverage the structural properties of the resource.

Many recent approaches to dictionary and thesaurus translation are geared toward providing domain-specific thesauri for specialists in a particular field, e.g., medical terminology (Déjean et al. 2005) or agricultural terminology (Chun and Wenlin 2002). Researchers on these projects are faced with the choice of either finding human domain experts to manage manual translation or applying automatic

Journal : 10579 Dispatch : 20-9-2008 Pages : 14
Article No. : 9074 LE TYPESET
MS Code : CHUM641R2 CP DISK

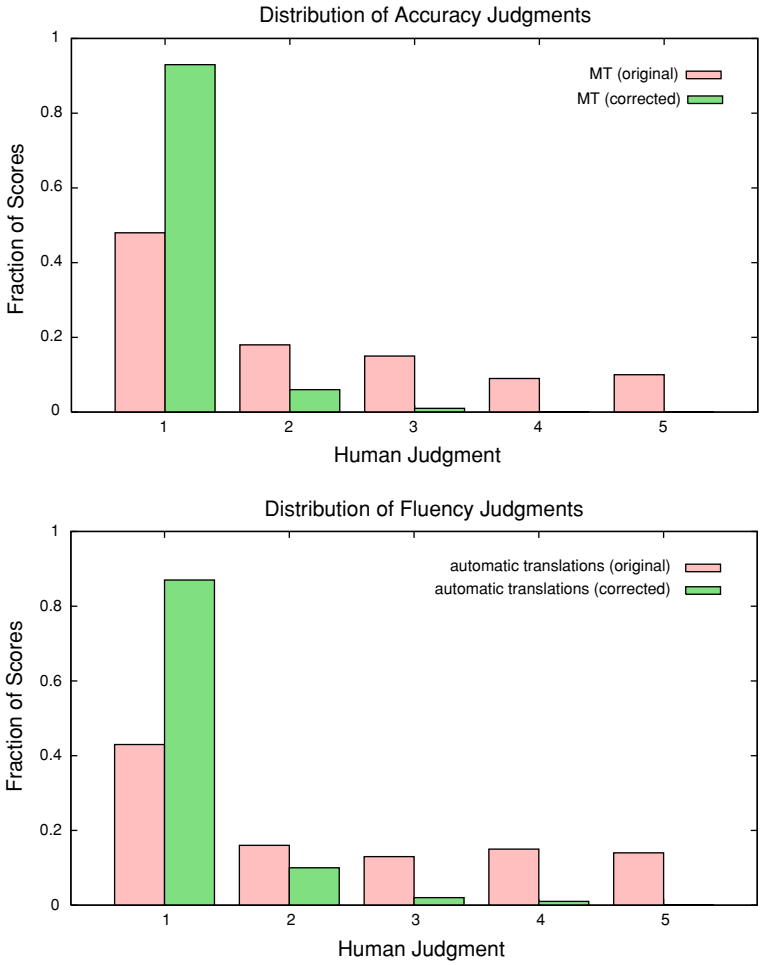

Fig. 5 Results of the subjective evaluation: fluency (left) and accuracy (right)

acquisition techniques, where data sparsity poses a problem for low-frequency terms. we balanced the need for human domain knowledge against the cost of human input.

Our process enriches the field of study with a hybrid alternative to full human translation or human assisted automated translation. We leverage the structure of the thesaurus and the recombinant nature of keyword phrases to prioritize the human input in advance. Then we set a threshold on the cost of translation, effectively switching from human translation to automated translation when the threshold is reached. In this way we tie into other research on both sides of the threshold, and introduce a cost function for managing the tradeoff.

Future work in this vein could explore individual influences of factors such as language complexity, domain specificity, or concept portability. The current work does not address variability of concepts across different cultures. The keyword

Journal : 10579 | Dispatch : 20-9-2008 | Pages : 14
Article No. : 9074 | LE | TYPESET
MS Code : CHUM641R2 | CP | DISK

phrases we translated are best seen as English labels on distinct concepts. Our translations produced Czech labels for these concepts. However, a thesaurus represents one particular view of the world. A cohesive concept that has clear boundaries in one language (and its culture) may be far less cohesive in another, and the labels for that concept may require more finesse in translation. Any interlingua inherently faces problems of mapping concepts from one culture to concepts in another. In translating thesauri it may be possible to leverage the structural arrangement of concepts to improve translations of other material. It may also be possible to leverage prioritized human input to learn structural mappings between dissimilar concept structures, e.g. competing ontologies. These questions warrant further investigation.

6 Conclusion

The task of thesauri translation can be recast as the problem of implementing a cost-effective process for acquiring domain-specific lexical resources. We developed a process for eliciting human translations. From 3,000 manually translated keyword phrases, we induced a probabilistic dictionary. Using this resource, we achieved acceptable automatic translation of the complete 56,000-concept thesaurus. As a rough calculation, the overall cost of human input was less than $1,000€$. Had we paid for human translation of the entire thesaurus it would have cost close to $20,000€$. Naturally, this is a biased comparison since manual translation of the entire thesaurus would have yielded a product much higher in quality. Nevertheless, we are able to implement a solution that *approximates* a gold standard, at a small fraction of the cost.

The value of our work lies in the process model we developed for cost-effective acquisition of lexical resources. We have shown that careful prioritization of human translations can efficiently yield reusable lexicons for automatic translation. The development of a utility function that accurately models both the direct and indirect value of a particular concept is the key to a cost-effective prioritization. Our process model aims to address the most critical deficiencies in vocabulary coverage first, such that the value obtained from each additional human translation becomes successively smaller. Under such a framework, choosing the number of human translations to elicit becomes a function of the financial and human resources available for the task.

Although this work focuses on thesaurus translation, the process we developed can be extended to other types of structured texts as well. For example, ontologies and knowledge bases have poly-hierarchic structures similar to the typonomic relations in the Shoah thesaurus. Our objective function was based on access to multimedia, but similar objective functions could be developed for different types of structural nodes to guide the translation process. We believe that our process is ideal for languages with scarce resources. Resources tend to be scarce for exactly the same languages and cultures which stand to gain the most from translated structural knowledge representations. The end result of this work will be a step toward a rich multilingual dictionary of Holocaust terms. Similar resources could be developed

Journal : 10579 | Dispatch : 20-9-2008 | Pages : 14
Article No. : 9074 | LE | TYPESET
MS Code : CHUM641R2 | CP | DISK

for legal terms, medical terms, etc. These in turn could serve to educate and empower the peoples of many nations.

Acknowledgements Our thanks to Doug Oard for helpful discussions; to our Czech informants; 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, #LC536 and #MSM0021620838, and the Grant Agency of the Czech Republic #GA405/06/0589. The first author would like to thank Esther and Kiri for their kind support.

References

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

Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., & Kuboň, V. (2004). Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In *Proceedings of LREC 2004*.

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. (2006). Automatic extraction of bilingual word pairs using inductive chain learning in various languages. *Information Processing and Management, 42*(5), 1294–1315.

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Hovy, E., Knight, K., Beale, S., Domashnev, C., Attardo, D., Grannes, D., & Brown, R. (1994). The Pangloss Mark III machine translation system. In *Proceedings of the 1st AMTA Conference*.

Gustman, S., Soergel, D., Oard, D. W., Byrne, W. J., Picheny, M., Ramabhadran, B., & Greenberg, D. (2002). Supporting access to large digital oral history archives. In *Proceedings of JCDL 2002* (pp. 18–27).

Kaji, H., & Aizono, T. (1996). Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of COLING 1996* (pp. 23–28).

Murray, G. C., Dorr, B., Lin, J., Hajič, J., & Pecina, P. (2006a). Leveraging recurrent phrase structure in large-scale ontology translation. In *Proceedings of EAMT 2006*.

Murray, G. C., Dorr, B., Lin, J., Hajič, J., & Pecina, P. (2006b). Leveraging reusability: Cost-effective lexical acquisition for large-scale ontology translation'. In *Proceedings of COLING/ACL 2006* (pp. 945–952).

Olsen, M., Dorr, B., & Thomas, S. (1998). Enhancing automatic acquisition of thematic structure in a large-scale lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002* (pp. 311–318).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German Corpora. In *Proceedings of ACL 1999* (pp. 519–526).

Sabarís, M., Alonso, J., Dafonte, C., & Arcay, B. (2001). Multilingual authoring through an artificial language. In *Proceedings of MT Summit VIII*.

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 SIGIR 2003* (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 Report LAMP-TR-126/CS-TR-4755/UMIACS-TR-2005-58, University of Maryland, College Park.

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

USC. (2006). USC Shoah Foundation Institute for Visual History and Education.