Semi-Automatic Acquisition of Domain-Specific Translation Lexicons

Philip Resnik
Dept. of Linguistics and UMIACS
University of Maryland
College Park, MD 20742 USA
resnik@umiacs.umd.edu

I. Dan Melamed
Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104 USA
melamed@unagi.cis.upenn.edu

Abstract

We investigate the utility of an algorithm for translation lexicon acquisition (SABLE), used previously on a very large corpus to acquire general translation lexicons, when that algorithm is applied to a much smaller corpus to produce candidates for domain-specific translation lexicons.

1 Introduction

Reliable translation lexicons are useful in many applications, such as cross-language text retrieval. Although general purpose machine readable bilingual dictionaries are sometimes available, and although some methods for acquiring translation lexicons automatically from large corpora have been proposed, less attention has been paid to the problem of acquiring bilingual terminology specific to a domain, especially given domain-specific parallel corpora of only limited size.

In this paper, we investigate the utility of an algorithm for translation lexicon acquisition (Melamed, 1997), used previously on a very large corpus to acquire general translation lexicons, when that algorithm is applied to a much smaller corpus to produce candidates for domain-specific translation lexicons. The goal is to produce material suitable for post-processing in a lexicon acquisition process like the following:

1. Run the automatic lexicon acquisition algorithm on a domain-specific parallel corpus.
2. Automatically filter out "general usage" entries that already appear in a machine readable dictionary (MRD) or other general usage lexical resources.
3. Manually filter out incorrect or irrelevant entries from the remaining list.

Our aim, therefore, is to achieve sufficient recall and precision to make this process — in particular the time and manual effort required in Step 3 — a viable alternative to manual creation of translation lexicons without automated assistance.

The literature on cross-lingual text retrieval (CLTR) includes work that is closely related to this research, in that recent approaches emphasize the use of dictionary- and corpus-based techniques for translating queries from a source language into the language of the document collection (Oard, 1997). Davis and Dunning (1995), for example, generate target-language queries using a corpus-based technique that is similar in several respects to the work described here. However, the approach does not attempt to distinguish domain-specific from general usage term pairs, and it involves no manual intervention. The work reported here, focusing on semi-automating the process of acquiring translation lexicons specific to a domain, can be viewed as providing bilingual *dictionary* entries for CLTR methods like that used by Davis in later work (Davis, 1996), in which dictionary-based generation of an ambiguous target language query is followed by corpus-based disambiguation of that query.

Turning to the literature on bilingual terminology identification *per se*, although monolingual terminology extraction is a problem that has been previously explored, often with respect to identifying relevant multi-word terms (e.g. (Daille, 1996; Smadja, 1993)), less prior work exists for bilingual acquisition of domain-specific translations. *Termight* (Dagan and Church, 1994) is one method for analyzing parallel corpora to discover translations in technical terminology; Dagan and Church report accuracy of 40% given an English/German technical manual, and observe that even this relatively low accuracy permits the successful application of the system in a translation bureau, when used in conjunction with an appropriate user interface.

The *Champollion* system (Smadja, McKeown, and Hatzivassiloglou, 1996) moves toward higher accuracy (around 73%) and considerably greater flexibility in the handling of multi-word translations, though the algorithm has been applied primarily to very large corpora such as the Hansards (3-9 million words; Smadja et al. observe that the method has difficulty handling low-frequency cases), and no

attempt is made to distinguish corpus-dependent translations from general ones.

Daille et al. (1994) report on a study in which a small (200,000 word) corpus was used as the basis for extracting bilingual terminology, using a combination of syntactic patterns for identifying simple two-word terms monolingually, and a statistical measure for selecting related terms across languages. Using a manually constructed reference list, they report 70% precision.

The SABLE system (Melamed, 1996b) makes no attempt to handle collocations, but for single-word to single-word translations it offers a very accurate method for acquiring high quality translation lexicons from very large parallel corpora: Melamed reports 90+% precision at 90+% recall, when evaluated on sets of Hansards data of 6-7 million words. Previous work with SABLE does not attempt to address the question of domain-specific vs. general translations.

This paper applies the SABLE system to a much smaller (approximately 400,000 word) corpus in a technical domain, and assesses its potential contribution to the semi-automatic acquisition process outlined above, very much in the spirit of Dagan and Church (1994) and Daille et al. (1994), but beginning with a higher accuracy starting point and focusing on mono-word terms. In the remainder of the paper we briefly outline translation lexicon acquisition in the SABLE system, describe its application to a corpus of technical documentation, and provide a quantitative assessment of its performance.

2 SABLE

SABLE (Scalable Architecture for Bilingual LExicography) is a turn-key system for producing clean broad-coverage translation lexicons from raw, unaligned parallel texts (bitexts). Its design is modular and minimizes the need for language-specific components, with no dependence on genre or word order similarity, nor sentence boundaries or other "anchors" in the input.

SABLE was designed with the following features in mind:

- *Independence from linguistic resources*: SABLE does not rely on any language-specific resources other than tokenizers and a heuristic for identifying word pairs that are mutual translations, though users can easily reconfigure the system to take advantage of such resources as language-specific stemmers, part-of-speech taggers, and stop lists when they are available.
- *Black box functionality*: Automatic acquisition of translation lexicons requires only that the user provide the input bitexts and identify the two languages involved.
- *Robustness*: The system performs well even in the face of omissions or inversions in translations.
- *Scalability*: SABLE has been used successfully on input bitexts larger than 130MB.
- *Portability*: SABLE was initially implemented for French/English, then ported to Spanish/English and to Korean/English. The porting process has been standardized and documented (Melamed, 1996c).

The following is a brief description of SABLE's main components. A more detailed description of the entire system is available in (Melamed, 1997).

2.1 Mapping Bitext Correspondence

After both halves of the input bitext(s) have been tokenized, SABLE invokes the *Smooth Injective Map Recognizer (SIMR)* algorithm (Melamed, 1996a) and related components to produce a bitext map. A bitext map is an injective partial function between the character positions in the two halves of the bitext. Each point of correspondence (x, y) in the bitext map indicates that the word centered around character position x in the first half of the bitext is a translation of the word centered around character position y in the second half. SIMR produces bitext maps a few points at a time, by interleaving a point generation phase and a point selection phase.

SIMR is equipped with several "plug-in" matching heuristic modules which are based on cognates (Davis et al., 1995; Simard et al., 1992; Melamed, 1995) and/or "seed" translation lexicons (Chen, 1993). Correspondence points are generated using a subset of these matching heuristics; the particular subset depends on the language pair and the available resources. The matching heuristics all work at the word level, which is a happy medium between larger text units like sentences and smaller text units like character n-grams. Algorithms that map bitext correspondence at the phrase or sentences level are limited in their applicability to bitexts that have easily recognizable phrase or sentence boundaries, and Church (1993) reports that such bitexts are far more rare than one might expect. Moreover, even when these larger text units can be found, their size imposes an upper bound on the resolution of the bitext map. On the other end of the spectrum, character-based bitext mapping algorithms (Church, 1993; Davis et al., 1995) are limited to language pairs where cognates are common; in addition, they may easily be misled by superficial differences in formatting and page layout and must sacrifice precision to be computationally tractable.

SIMR filters candidate points of correspondence using a geometric pattern recognition algorithm. The recognized patterns may contain non-monotonic sequences of points of correspondence, to account for

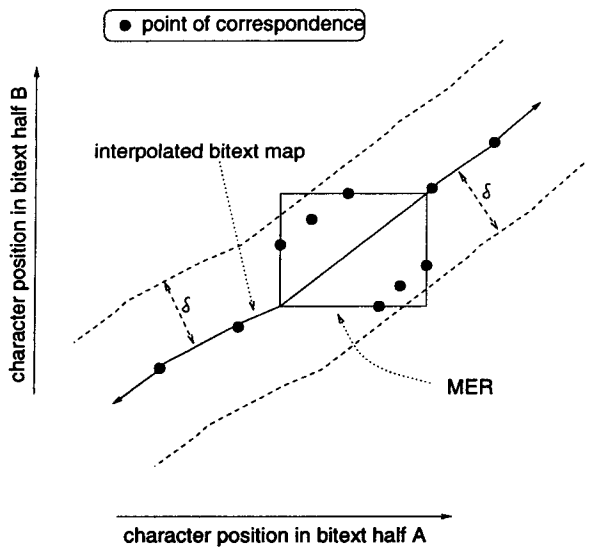

Figure 1: *Word token pairs whose co-ordinates lie between the dashed boundaries count as co-occurrences.*

word order differences between languages. The filtering algorithm can be efficiently interleaved with the point generation algorithm so that SIMR runs in linear time and space with respect to the size of the input bitext.

2.2 Translation Lexicon Extraction

Since bitext maps can represent crossing correspondences, they are more general than "alignments" (Melamed, 1996a). For the same reason, bitext maps allow a more general definition of token co-occurrence. Early efforts at extracting translation lexicons from bitexts deemed two tokens to co-occur if they occurred in aligned sentence pairs (Gale and Church, 1991). SABLE counts two tokens as co-occurring if their point of correspondence lies within a short distance δ of the interpolated bitext map in the bitext space, as illustrated in Figure 1. To ensure that interpolation is well-defined, minimal sets of non-monotonic points of correspondence are replaced by the lower left and upper right corners of their minimum enclosing rectangles (MERs).

SABLE uses token co-occurrence statistics to induce an initial translation lexicon, using the method described in (Melamed, 1995). The *iterative filtering* module then alternates between estimating the most likely translations among word tokens in the bitext and estimating the most likely translations between word types. This re-estimation paradigm was pioneered by Brown et al. (1993). However, their models were not designed for human inspection, and though some have tried, it is not clear how to extract translation lexicons from their models (Wu and Xia, 1995). In contrast, SABLE automatically constructs an explicit translation lexicon, the lexicon consisting

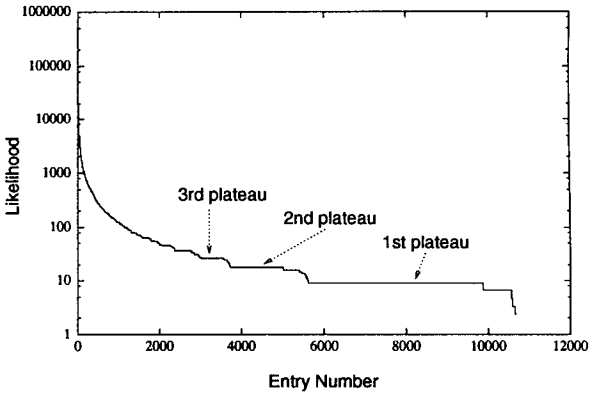

Figure 2: *Translation lexicon entries proposed by SABLE exhibit plateaus of likelihood.*

of word type pairs that are not filtered out during the re-estimation cycle. Neither of the translation lexicon construction modules pay any attention to word order, so they work equally well for language pairs with different word order.

2.3 Thresholding

Translation lexicon recall can be automatically computed with respect to the input bitext (Melamed, 1996b), so SABLE users have the option of specifying the recall they desire in the output. As always, there is a tradeoff between recall and precision; by default, SABLE will choose a likelihood threshold that is known to produce reasonably high precision.

3 Evaluation in a Technical Domain

3.1 Materials Evaluated

The SABLE system was run on a corpus comprising parallel versions of Sun Microsystems documentation ("Answerbooks") in French (219,158 words) and English (191,162 words). As Melamed (1996b) observes, SABLE's output groups naturally according to "plateaus" of likelihood (see Figure 2). The translation lexicon obtained by running SABLE on the Answerbooks contained 6663 French-English content-word entries on the 2nd plateau or higher, including 5464 on the 3rd plateau or higher. Table 1 shows a sample of 20 entries selected at random from the Answerbook corpus output on the 3rd plateau and higher. Exact matches, such as *cpio/cpio* or *clock/clock*, comprised roughly 18% of the system's output.

In order to eliminate likely general usage entries from the initial translation lexicon, we automatically filtered out all entries that appeared in a French-English machine-readable dictionary (MRD) (Cousin, Allain, and Love, 1991). 4071 entries remained on or above the 2nd likelihood plateau, including 3135 on the 3rd likelihood plateau or higher.

French	English
constantes	constants
multi-fenêtrage	windows
risque	may
extensions	extensions
exemple	such
relâché	released
rw-r	r
reçus	received
préalable	first
cpio	cpio
sont	will
defaults	defaults
fn	fn
alphabétique	alphabetically
activée	activates
machine	workstation
mettre	turns
connectés	connected
bernard	spanky
superutilisateur	root

Table 1: *Random sample of SABLE output on software manuals.*

In previous experiments on the Hansard corpus of Canadian parliamentary proceedings, SABLE had uncovered valid general usage entries that were not present in the Collins MRD (e.g. *pointillés/dotted*). Since entries obtained from the Hansard corpus are unlikely to include relevant technical terms, we decided to test the efficacy of a second filtering step, deleting all entries that had also been obtained by running SABLE on the Hansards. On the 2nd plateau or higher, 3030 entries passed both the Collins and the Hansard filters; 2224 remained on or above the 3rd plateau.

Thus in total, we evaluated four lexicons derived from all combinations of two independent variables: cutoff (after the 2nd plateau vs. after the 3rd plateau) and Hansards filter (with filter vs. without). Evaluations were performed on a random sample of 100 entries from each lexicon variation, interleaving the four samples to obscure any possible regularities. Thus from the evaluator's perspective the task appeared to involve a single sample of 400 translation lexicon entries.

3.2 Evaluation Procedure

Our assessment of the system was designed to reasonably approximate the post-processing that would be done in order to use this system for acquisition of translation lexicons in a real-world setting, which would necessarily involve subjective judgments. We hired six fluent speakers of both French and English at the University of Maryland; they were briefed on the general nature of the task, and given a data sheet containing the 400 candidate entries (pairs containing one French word and one English word) and a "multiple choice" style format for the annotations, along with the following instructions.

1. If the pair clearly cannot be of help in constructing a glossary, circle **"Invalid"** and go on to the next pair.
2. If the pair can be of help in constructing a glossary, choose *one* of the following:[1]
   - **V**: The two words are of the "plain vanilla" type you might find in a bilingual dictionary.
   - **P**: The pair is a case where a word changes its part of speech during translation. For example, "to have protection" in English is often translated as "être protégé" in Canadian parliamentary proceedings, so for that domain the pair protection/protégé would be marked P.
   - **I**: The pair is a case where a direct translation is incomplete because the computer program only looked at single words. For example, if French "immédiatement" were paired with English "right", you could select I because the pair is almost certainly the computer's best but incomplete attempt to be pairing "immédiatement" with "right away".
3. Then choose *one* or *both* of the following:
   - **Specific**. Leaving aside the relationship between the two words (your choice of P, V, or I), the word pair would be of use in constructing a *technical* glossary.
   - **General**. Leaving aside the relationship between the two words (your choice of P, V, or I), the word pair would be of use in constructing a *general usage* glossary.

   Notice that a word pair could make sense in both. For example, "corbeille/wastebasket" makes sense in the computer domain (in many popular graphical interfaces there is a wastebasket icon that is used for deleting files), but also in more general usage. So in this case you could in fact decide to choose both "Specific" and "General". If you can't choose either "Specific" *or* "General", chances are that you should reconsider whether or not to mark this word pair "Invalid".

[1] Since part-of-speech tagging was used in the version of SABLE that produced the candidates in this experiment, entries presented to the annotator also included a minimal form of part-of-speech information, e.g. distinguishing nouns from verbs. The annotator was informed that these annotations were the computer's best attempt to identify the part-of-speech for the words; it was suggested that they could be used as a hint as to why that word pair had been proposed, if so desired, and otherwise ignored.

4. If you're completely at a loss to decide whether or not the word pair is valid, just put a slash through the number of the example (the number at the beginning of the line) and go on to the next pair.

Annotators also had the option of working electronically rather than on hardcopy.

The assessment questionnaire was designed to elicit information primarily of two kinds. First, we were concerned with the overall accuracy of the method; that is, its ability to produce reasonable candidate entries whether they be general or domain specific. The "Invalid" category captures the system's mistakes on this dimension. We also explicitly annotated candidates that might be useful in constructing a translation lexicon, but possibly require further elaboration. The *V* category captures cases that require minimal or no additional effort, and the *P* category covers cases where some additional work might need to be done to accommodate the part-of-speech divergence, depending on the application. The *I* category captures cases where the correspondence that has been identified may not apply directly at the single-world level, but nonetheless does capture potentially useful information. Daille et al. (1994) also note the existence of "incomplete" cases in their results, but collapse them together with "wrong" pairings.

Second, we were concerned with domain specificity. Ultimately we intend to measure this in an objective, quantitative way by comparing term usage across corpora; however, for this study we relied on human judgments.

3.3 Use of Context

Melamed (1996b) suggests that evaluation of translation lexicons requires that judges have access to bilingual concordances showing the contexts in which proposed word pairs appear; however, out-of-context judgments would be easier to obtain in both experimental and real-world settings. In a preliminary evaluation, we had three annotators (one professional French/English translator and two graduate students at the University of Pennsylvania) perform a version of the annotation task just described: they annotated a set of entries containing the output of an earlier version of the SABLE system (one that used aligned sub-sentence fragments to define term co-occurrence; cf. Section 2.2). No bilingual concordances were made available to them.

Analysis of the system's performance in this pilot study, however, as well as annotator comments in a post-study questionnaire, confirmed that context is quite important. In order to quantify its importance, we asked one of the pilot annotators to repeat the evaluation on the same items, this time giving her access to context in the form of the bilingual concordances for each term pair. These concordances contained up to the first ten instances of that pair as used in context. For example, given the pair *déplacez/drag*, one instance in that pair's bilingual concordance would be:

> Maintenez SELECT enfoncé et **déplacez** le dossier vers l' espace de travail .
>
> Press SELECT and **drag** the folder onto the workspace background .

The instructions for the in-context evaluation specify that the annotator should look at the context for every word pair, pointing out that "word pairs may be used in unexpected ways in technical text and words you would not normally expect to be related sometimes turn out to be related in a technical context."

Although we have data from only one annotator, Table 2 shows the clear differences between the two results.[2] In light of the results of the pilot study, therefore, our six annotators were given access to bilingual concordances for the entries they were judging and instructed in their use as just described.

4 Results

4.1 Group Annotations

A "group annotation" was obtained for each candidate translation lexicon entry based on agreement of at least three of the six annotators. "Tie scores" or the absence of a 3-of-6 plurality were treated as the absence of an annotation. For example, if an entry was annotated as "Invalid" by two annotators, marked as category V and Specific by two annotators, and marked as category P, Specific, and General by the other two annotators, then the group annotation would contain an "unclassified valid type" (since four annotators chose a valid type, but there was no agreement by at least three on the specific subclasification) and a "Specific" annotation (agreed on by four annotators). All summary statistics are reported in terms of the group annotation.

4.2 Precision

SABLE's precision on the Answerbooks bitext is summarized in Figure 3.[3] Each of the percentages being derived from a random sample of 100 observations, we can compute confidence intervals under a normality assumption; if we assume that the observations are independent, then 95% confidence intervals are narrower than one twentieth of a percentage point for all the statistics computed.

The results show that up to 89% of the translation lexicon entries produced by SABLE on or above the

[2] Again, this sample of data was produced by an older and less accurate version of SABLE, and therefore the percentages should only be analyzed relative to each other, not as absolute measures of performance.

[3] The exact numbers gladly provided on request.

	% V	% P	% I	All Valid Entries	Domain-Specific Only	General Usage Only	Both
Out-of-Context	39.5	9.25	5.5	57.75	29.75	23.5	1
In-Context	46.75	5	13	69.5	38	23.25	3.5

Table 2: *Effect of in-context vs. out-of-context evaluation. All numbers are in %. n = 400.*

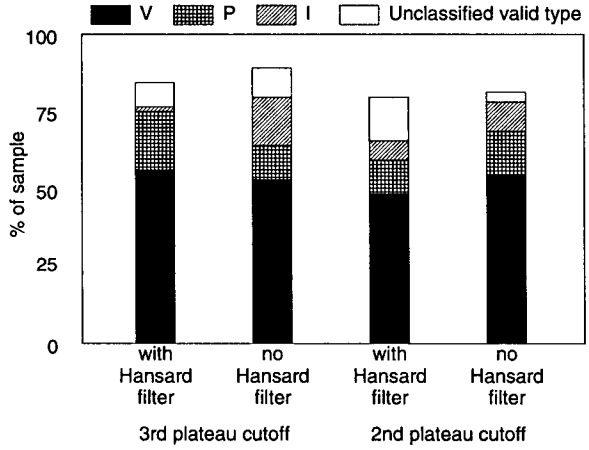

Figure 3: *Summary of filtered translation lexicon validity statistics.*

3rd likelihood plateau "can be of help in constructing a glossary." Up to 56% can be considered useful essentially as-is (the *V* category alone). Including all entries on the 2nd plateau or higher provides better coverage, but reduces the fraction of useful entries to 81%. The fraction of entries that are useful as-is remains roughly the same, at 55%. At both recall levels, the extra Hansards-based filter had a detrimental effect on precision.

Note that these figures are based on translation lexicons from which many valid general usage entries have been filtered out (see Section 3). We can compute SABLE's precision on unfiltered translation lexicons for this corpus by assuming that entries appearing in the Collins MRD are all correct.[4] However, these are not the real figures of interest here, because we are mainly concerned in this study with the acquisition of domain-specific translation lexicons.

4.3 Recall

Following Melamed (1996b), we adopt the following approach to measuring recall: the upper bound is defined by the number of different words in the bitext. Thus, perfect recall implies at least one entry containing each word in the corpus. This is a much more conservative metric than that used by Dailles et al. (1994), who report recall with respect to a relatively small, manually constructed reference set. Although we do not expect to achieve perfect recall on this criterion after general usage entries have been filtered out, the number is useful insofar as it provides a sense of how recall for this corpus correlates with precision. We have no reason to expect this correlation to change across domain-specific and general lexicon entries. For the unfiltered translation lexicons, recall on the 3rd likelihood plateau and above was 30.4%. When all entries on and above the 2nd plateau were considered, recall improved to 37.0%.

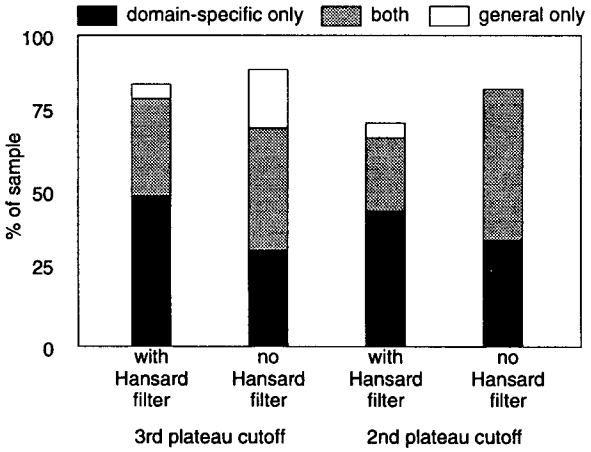

Figure 4: *Summary of filtered translation lexicon domain-specificity statistics.*

Hansards Filter?	Plateau Cutoff	% Domain Specific	% General Usage	% Both
Yes	3rd	82	37	35
No	3rd	71	53	35
Yes	2nd	66	27	22
No	2nd	81	47	47

Table 3: *Domain-specificity of filtered translation lexicon entries.*

4.4 Domain Specificity

Figure 4 demonstrates the effectiveness of the MRD- and corpus-based filters, with details in Table 3. If we assume that translation pairs in the Collins MRD are not specific to our chosen domain, then domain-specific translation lexicon entries constituted only

[4] Result: 88.4% precision at 37.0% recall or 93.7% precision at 30.4% recall.

κ	A1	A2	A3	A4	A5	A6
κ_1	0.70	0.44	0.59	0.82	0.90	0.82
κ_2	0.62	0.67	0.72	0.74	0.55	0.73
κ_3	0.28	0.19	0.50	0.00	0.00	0.56
κ_4	0.67	0.69	0.68	0.74	0.61	0.81

Table 4: *Inter-annotator agreement.*

49% of SABLE's unfiltered output on or above the 2nd plateau and 41% on or above the 3rd plateau. The MRD filter increased this ratio to 81% and 71%, respectively. As noted in Section 4.2, the second filter, based on the Hansard bitext, reduced the overall accuracy of the translation lexicons. Its effects on the proportion of domain-specific entries was mixed: an 11% increase for the entries more likely to be correct, but a 15% decrease overall. The corpus-based filter is certainly useful in the absence of an MRD. However, our results suggest that combining filters does not always help, and more research is needed to investigate optimal filter combination strategies.

4.5 Consistency of Annotations

In order to assess the consistency of annotation, we follow Carletta (1996) in using Cohen's κ, a chance-corrected measure of inter-rater agreement. The κ statistic was developed to distinguish among levels of agreement such as "almost perfect, substantial, moderate, fair, slight, poor" (Agresti, 1992), and Carletta suggests that as a rule of thumb in the behavioral sciences, values of κ greater than .8 indicate good replicability, with values between .67 and .8 allowing tentative conclusions to be drawn. For each such comparison, four values of κ were computed:

- κ_1: agreement on the evaluation of whether or not a pair should be immediately rejected or retained;
- κ_2: agreement, for the retained pairs, on the type V, P, or I assigned to the pair;
- κ_3: agreement, for the retained pairs, on whether to classify the pair as being useful for constructing a domain-specific glossary;
- κ_4: agreement, for the retained pairs, on whether to classify the pair as being useful for constructing a general usage glossary.

In each case, the computation of the agreement statistic took into account those cases, if any, where the annotator could not arrive at a decision for this case and opted simply to throw it out. Resulting values for inter-rater reliability are shown in Table 4; the six annotators are identified as A1, A2, ... A6, and each value of κ reflects the comparison between that annotator and the group annotation.

With the exception of κ_3, these values of κ indicate that the reliability of the judgments is generally reasonable, albeit not entirely beyond debate. The outlandish values for κ_3, despite high rates of absolute agreement on that dimension of annotation, are explained by the fact that the κ statistic is known to be highly problematic as a measure of inter-rater reliability when one of the categories that can be chosen is overwhelmingly likely (Grove et al., 1981; Spitznagel and Helzer, 1985). Intuitively this is not surprising: we designed the experiment to yield a predominance of domain-specific terms, by means of the MRD and Hansards filters. Our having succeeded, there is a very high probability that the "Specific" annotation will be selected by any two annotators, because it appears so very frequently; as a result the actual agreement rate for that annotation doesn't actually look all that different from what one would get by chance, and so the κ values are low. The values of κ_3 for annotators 4 and 5 emphasize quite clearly that κ is measuring not the level of absolute agreement, but the distinguishability of that level of agreement from chance.

5 Conclusion

In this paper, we have investigated the application of SABLE, a turn-key translation lexicon construction system for non-technical users, to the problem of identifying domain-specific word translations given domain-specific corpora of limited size. Evaluated on a very small (400,000 word) corpus, the system shows real promise as a method of processing small domain-specific corpora in order to propose candidate single-word translations: once likely general usage terms are automatically filtered out, the system obtains precision up to 89% at levels of recall very conservatively estimated in the range of 30–40% on domain-specific terms.

Of the proposed entries not immediately suitable for inclusion in a translation lexicon, many represent part-of-speech divergences (of the *protect/protégé* variety) and a smaller number incomplete entries (of the *immédiatement/right* variety) that would nonetheless be helpful if used as the basis for a bilingual concordance search — for example, a search for French segments containing *immédiatement* in the vicinity of English segments containing *right* would most likely yield up the obvious correspondence between *immédiatement* and *right away*. Going beyond single-word correspondences, however, is a priority for future work.

6 Acknowledgments

The authors wish to acknowledge the support of Sun Microsystems Laboratories, particularly the assistance of Gary Adams, Cookie Callahan, and Bob Kuhns, as well as useful input from Bonnie Dorr, Ralph Grishman, Marti Hearst, Doug Oard, and three anonymous reviewers. Melamed also acknowledges grants ARPA N00014-90-J-1863 and ARPA N6600194C 6043.

References

Alan Agresti. 1992. Modeling patterns of agreement and disagreement. *Statistical methods in medical research*, 1:201–218.

P. F. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics 19*:2.

Jean Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

S. Chen. 1993. "Aligning Sentences in Bilingual Corpora Using Lexical Information". *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

K. W. Church. 1993. "Char_align: A Program for Aligning Parallel Texts at the Character Level". *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

P. H. Cousin, L. Sinclair, J. F. Allain, and C. E. Love. 1991. *The Collins Paperback French Dictionary.* Harper Collins Publishers, Glasgow.

Ido Dagan and Ken W. Church. 1994. TERMIGHT: Identifying and translating technical terminology. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (13–15 October 1994, Stuttgart).* Association for Computational Linguistics, October.

I. Dagan, K. Church, and W. Gale. 1993. "Robust Word Alignment for Machine Aided Translation". *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, available from the ACL.

Béatrice Daille. 1994. *Combined approach for terminology extraction: lexical statistics and linguistic filtering.* Ph.D. thesis, University Paris 7.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* MIT Press.

Mark Davis. 1996. "New experiments in cross-language text retrieval at NMSU's Computing Research Lab". Fifth Text Retrieval Conference (TREC-5). NIST.

Mark Davis and Ted Dunning. 1995. "A TREC evaluation of query translation methods for multilingual text retrieval". Fourth Text Retrieval Conference (TREC-4). NIST.

Mark Davis, Ted Dunning, and William Ogden. 1995. Text alignment in the real world: improving alignments of noisy translation using common lexical features, string matching strategies, and n-gram comparisons. In *EACL-95.*

W. Gale and K. W. Church. 1991. "Identifying Word Correspondences in Parallel Texts". *Proceedings of the DARPA SNL Workshop*, 1991.

W. Grove, N. Andreasen, P. McDonald-Scott, M. Keller, and R. Shapiro. 1981. Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry*, 38, April.

I. Dan Melamed, 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Massachusetts.

I. Dan Melamed. 1996a. A geometric approach to mapping bitext correspondence. In *Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pennsylvania.

I. Dan Melamed. 1996b. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, Montreal, Canada.

I. Dan Melamed. 1996c. Porting SIMR to new language pairs. IRCS Technical Report 96-26. University of Pennsylvania.

I. Dan Melamed. 1997. A scalable architecture for bilingual lexicography. Dept. of Computer and Information Science Technical Report MS-CIS-97-01. University of Pennsylvania.

Douglas W. Oard. 1997. "Cross-Language Text Retrieval Research in the USA". Third DELOS Workshop. European Research Consortium for Informatics and Mathematics. March.

M. Simard, G. F. Foster and P. Isabelle. 1992. "Using Cognates to Align Sentences in Bilingual Corpora". In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1), March.

E. Spitznagel and J. Helzer. "A proposed solution to the base rate problem in the kappa statistic". *Archives of General Psychiatry*, 42. July, 1985.

D. Wu and X. Xia. 1994. "Learning an English-Chinese Lexicon from a Parallel Corpus". *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, MD.