'Irrefragable answers'
using comparable corpora to retrieve translation equivalents

Serge Sharoff, Bogdan Babych and Anthony Hartley
Centre for Translation Studies, University of Leeds

Abstract. In this paper we present a tool that uses comparable corpora to find appropriate translation equivalents for expressions that are considered by translators as difficult. For a phrase in the source language the tool identifies a range of possible expressions used in similar contexts in target language corpora and presents them to the translator as a list of suggestions. In the paper we discuss the method and present results of human evaluation of the performance of the tool, which highlight its usefulness when dictionary solutions are lacking.

Keywords: Large comparable corpora, translation equivalents, multiword expressions, distributional similarity

1. Introduction

There is no doubt that both professional and trainee translators need access to authentic data provided by corpora. With respect to polysemous lexical items, bilingual dictionaries list several translation equivalents for a headword, but words taken in their contexts can be translated in many more ways than indicated in dictionaries. For instance, the Oxford Russian Dictionary (ORD) lacks a translation for the Russian expression исчерпывающий ответ ('exhaustive answer'), while the Multitran Russian-English dictionary suggests that it can be translated as *irrefragable answer*. Yet this expression is extremely rare in English; on the Internet it occurs mostly in pages produced by Russian speakers.

On the other hand, translations for polysemous words are too numerous to be listed for all possible contexts. For example, the entry for *strong* in ORD already has 57 subentries and yet it fails to mention many word combinations frequent in the British National Corpus (BNC), such as *strong {feeling, field, opposition, sense, voice}*. *Strong voice* is also not listed in the Oxford French, German or Spanish Dictionaries.

There has been surprisingly little research on computational methods for finding translation equivalents of words from the general lexicon. Practically all previous studies have concerned detection of terminological equivalence, e.g. (Dagan and Church, 1997; Grefenstette, 2002; Daille and Morin, 2005). At the same time, translators often experience more difficulty in dealing with expressions from the gen-

eral lexicon because of their polysemy, which is reflected differently in the target language, thus causing the dependency of their translation on the corresponding context. Such variation is often not captured by dictionaries.

Because of their importance, words from the general lexicon are studied by translation researchers, and comparable corpora are increasingly used in translation practice and training (Zanettin, 1998). However, such studies are mostly confined to lexicographic exercises, which compare the contexts and functions of potential translation equivalents once they are known, for instance, *absolutely* vs. *assolutamente* in Italian (Partington, 1998). Such studies do not provide a computational model for *finding* appropriate translation equivalents for expressions that are not listed or are inadequate in dictionaries.

Parallel corpora, consisting of original texts and their exact translations, provide a useful supplement to decontextualised translation equivalents listed in dictionaries. However, parallel corpora are not representative. Many of them are in the range of a few million words, which is simply too small to account for variations in translation of moderately frequent words. Those that are a bit larger, such as the Europarl corpus, are restricted in their domain. For instance, all of the 14 instances of *strong voice* in the English section of Europarl are used in the sense of 'the opinion of a political institution'. At the same time the BNC contains 46 instances of *strong voice* covering several different meanings.

In this paper we propose a computational method for using comparable corpora to find translation equivalents for source language expressions that are considered as difficult by trainee or professional translators. The model is based on detecting frequent multi-word expressions (MWEs) in the source and target languages and finding a mapping between them in comparable monolingual corpora, which are designed in a similar way in the two languages.

The described methodology is implemented in ASSIST, a tool that helps translators to find solutions for difficult translation problems. The tool presents the results as lists of translation suggestions (usually 50 to 100 items) ordered alphabetically or by their frequency in target language corpora. Translators can skim through these lists and identify an example which is most appropriate in a given context.

In the following sections we outline our approach, evaluate the output of the prototype of ASSIST and discuss future work.

2. Finding translations in comparable corpora

The proposed model finds potential translation equivalents in three steps, which include

1. expansion of words in the original expression using related words;
2. translation of the resultant set using existing bilingual dictionaries and its further expansion of the set using related words in the target language;
3. filtering of the set according to expressions frequent in the target language corpus.

In this study we use several comparable corpora for English and Russian, including large reference corpora (the BNC and the Russian Reference Corpus) and corpora of major British and Russian newspapers. All corpora used in the study are quite large, i.e. the size of each corpus is in the range of 100-200 million words (MW), so that they provide enough evidence to detect such collocations as *strong voice* and *clear defiance*.

Although the current study is restricted to the English-Russian pair, the methodology does not rely on any particular language. It can be extended to other languages for which large comparable corpora, POS-tagging and lemmatisation tools, and bilingual dictionaries are available. For example, we conducted a small study for translation between English and German using the Oxford German Dictionary and a 200 MW German corpus derived from the Internet (Sharoff, 2006).

2.1. Query expansion

The problem with using comparable corpora to find translation equivalents is that there is no obvious bridge between the two languages. Unlike aligned parallel corpora, comparable corpora provide a model for each individual language, while dictionaries, which can serve as a bridge, are inadequate for the task in question, because the problem we want to address involves precisely translation equivalents that are not listed there.

Therefore, a specific query needs first to be generalised in order to then retrieve a suitable candidate from a set of candidates. One way to generalise the query is by using *similarity classes*, i.e. groups of words with lexically similar behaviour. In his work on distributional similarity (Lin, 1998) designed a parser to identify grammatical relationships between words. However, broad-coverage parsers suitable for processing BNC-like corpora are not available for many languages (including

Russian). Another, resource-light approach treats the context as a bag of words (BoW) and detects the similarity of contexts on the basis of collocations in a window of a certain size, typically 3-4 words. Using a parser can increase precision in identification of contexts in the case of long-distance dependencies (e.g. *to cook Alice a whole meal*). However, we can find a reasonable set of relevant terms returned using the BoW approach, cf. the results of human evaluation for English and German by (Rapp, 2004).

For each source word s_0 we produce a list of similar words: $\Theta(s_0) = s_1, \ldots, s_N$ (in our tool we use $N = 20$ as the cutoff). We can also produce a more reliable similarity class $S(s_0)$ using the assumption that the similarity classes of similar words must have common members:

$w \in S(s_0) \iff w \in \Theta(s_0) \wedge w \in \bigcup \Theta(s_i)$

i.e. w is also in the similarity class of at least one of other words, so that occasional irrelevant words in $\Theta(s_0)$ are removed, as they do not produce similarity classes consistent with other words.

This yields for *experience* the following similarity class: *knowledge, opportunity, life, encounter, skill, feeling, reality, sensation, dream, vision, learning, perception, learn* (ordered according to the cosine distance score as implemented by Rapp). Even if there is no requirement in the BoW approach that words in the similarity class are of the same part of speech, it happens quite frequently that most words have the same part of speech because of the similarity of contexts.

2.2. Query translation and further expansion

In the next step we produce a translation class by translating all words from the similarity class into the target language using a bilingual dictionary ($T(w)$ for the translation of w). Then for expanding the translation set in the target language we have two options: a full translation class (TF) and a reduced one (TR).

$TF = S(T(S(s_0)))$ consists of similarity classes produced for all translations. However, this causes a combinatorial explosion. If a similarity class contains N words (the average figure is 16) and a dictionary lists on average M equivalents for a source word (the average figure is 11), this procedure outputs on average $M \times N^2$ words in the full translation class. For instance, the complete translation class for *experience* contains 998 words. What is worse, some words from the full translation class do not refer to the domain implied in the original expression because of the ambiguity of the translation operation. For instance, the word *dream* belongs to the similarity class of *experience*. Since it can be translated into Russian as сказка ('fairy-tale'), the latter Russian word will be expanded in the full translation class with words

referring to legends and stories. In the later stages of the project, word sense disambiguation in corpora could improve precision of translation classes. However at the present stage we attempt to trade the recall of the tool for greater precision by translating words in the source similarity class, and generating the similarity classes of translations only for the source word:

$TR(s_0) = S(T(s_0)) \cup T(S(s_0))$.

The reduced translation class of *experience* contains 128 words.

This step crucially relies on a wide-coverage machine readable dictionary. The bilingual dictionary resources we use are derived from the source file for the Oxford Russian Dictionary, provided by OUP.

2.3. Filtering equivalence classes

In the final step we check all possible combinations of words from the translation classes for their frequency in target language corpora.

The number of elements in the set of theoretically possible combinations is usually very large: $\prod T_i$, where T_i is the number of words in the translation class of each word of the original MWE. This number is much larger than the set of word combinations which is found in the target language corpora. For instance, *daunting experience* has 202,594 combinations for the full translation class of *daunting experience* and 6,144 for the reduced one. However, in the target language corpora we can find only 2,256 collocations with frequency > 2 for the full translation class and 92 for the reduced one.

Each theoretically possible combination is generated and looked up in a database of MWEs (which is much faster than querying corpora for frequencies of potential collocations). The MWE database was precompiled from corpora using a method of filtering, similar to part-of-speech filtering suggested in (Justeson and Katz, 1995): in corpora each N-gram of length 2, 3 and 4 tokens was checked against a set of filters. However, instead of pre-defined patterns for entire expressions our filtering method uses sets of *negative* constraints, which are usually applied to the edges of expressions. This change boosts recall of retrieved MWEs and allows us to use the same set of patterns for MWEs of different length. The filter uses constraints for both lexical and part-of-speech features, which makes configuration specifications more flexible.

The whole procedure is relatively language-independent. The original development has been done for the English-Russian translation pair and has been later extended to the English-German pair. For instance, given an expression like исчерпывающий ответ ('exhaustive answer') the system produces a range of expressions for исчерпывающий, e.g. *reliable, accessible, effective, truthful, unlimited*, another range for ответ, e.g.

argument, response, reply, however, a filter through the list of English MWEs leaves expressions like *comprehensive answer, comprehensive response, definite answer, truthful answer*, etc. Similarly for German an expression like *schlecht wegkommen* (lit. 'come out badly', for instance, in the context of elections) produces *poor performance, make bad, poor service, appalling record, inadequate performance*, etc. The use of distributionally similar words for translation between English and French has been studied in (Ploux and Ji, 2003).

3. Evaluation

There are several attributes of our system which can be evaluated, and many of them are crucial for its efficient use in the workflow of professional translators, including: usability, quality of final solutions, trade-off between adequacy and fluency across usable examples, precision and recall of potentially relevant suggestions, as well as real-text evaluation, i.e. "What is the coverage of difficult translation problems typically found in a text that can be successfully tackled?"

In this paper we focus on evaluating the quality of potentially relevant translation solutions, which is the central point for developing and calibrating our methodology. The evaluation experiment discussed below was specifically designed to assess the usefulness of translation suggestions generated by our tool – in cases where translators have doubts about the usefulness of dictionary solutions. In this paper we do not evaluate other equally important aspects of the system's functionality, which will be the matter of future research.

3.1. Set-up of the experiment

For each translation direction we collected ten examples of possibly recalcitrant translation problems – words or phrases whose translation is not straightforward in a given context. Some of these examples were sent to us by translators in response to our request for difficult cases. For each example, which we included in the evaluation kit, the word or phrase either does not have a translation in ORD (which is a kind of a baseline standard reference for Russian translators), or its translation has significantly lower frequency in a target language corpus in comparison to the frequency of the source expression. If an MWE is not listed in available dictionaries, we took translations for individual words from ORD. In order to remove a possible bias towards a specific dictionary, we also checked translations in Multitran, an on-line translation dictionary, which was often quoted as one of the best resources for translation from and into Russian.

Table I. Scores to translation equivalents

Translation	t1	t2	t3	t4	t5	σ
clear plan	5	5	3	4	4	0.84
clear policy	5	5	3	4	4	0.84
clear programme	5	5	3	4	4	0.84
clear strategy	5	5	5	5	5	0.00
concrete plan	1	5	3	3	5	1.67
Best Dict	5	5	3	4	4	0.84
Best Syst	5	5	5	5	5	0.00

For each translation problem five solutions were presented to translators for evaluation. One or two of these solutions were taken from a dictionary (usually from Multitran, and if available and different, from ORD). The other suggestions were manually selected from lists of possible solutions returned by ASSIST. Again, the criteria for selection were intuitive: we included those suggestions which made best sense in the given context. Dictionary suggestions and the output of ASSIST were indistinguishable in the questionnaires to the evaluators. The segments were presented in sentence context and translators had an option of providing their own solutions and comments.

We then asked professional translators affiliated to *ITI* (Institute of Translation and Interpreting) to rate these five potential equivalents using a five-point scale. We received responses from eleven translators. Some translators did not score all solutions, but there were at least seven independent judgements for each of the 100 translation variants. An example of the combined answer sheet for all responses to the Russian example четкая программа (lit. 'precise programme') is given in Table I (t1, t2,... denote translators; the dictionary translation is *clear programme*).

3.2. Interpretation of the results

The results were surprising in so far as for the majority of problems translators preferred very different translation solutions and did not agree in their scores for the same solutions. For instance, *concrete plan* in Table I received the score 1 from translator t1 and 5 from t2.

In general, the translators very often picked up on different opportunities presented by the suggestions from the lists, and most suggestions were equally legitimate ways of conveying the intended content, cf. the study of legitimate translation variation with respect to the BLEU score

in (Babych and Hartley, 2004). In this respect it may be unfair to compute average scores for each potential solution, since for most interesting cases the scores do not fit into the normal distribution model. So averaging scores would mask the potential usability of really inventive solutions.

In this case it is more reasonable to evaluate two *sets* of solutions – the one generated by ASSIST and the other found in dictionaries – but not each solution individually. In order to do that for each translation problem the best scores given by each translator in each of these two sets were selected. This way of generalising data characterises the general quality of suggestion sets, and exactly meets the needs of translators, who collectively get ideas from the presented sets rather than from individual examples. This also allows us to measure inter-evaluator agreement on the *dictionary* set and the *ASSIST* set, for instance, via computing the standard deviation σ of absolute scores across evaluators (Table I).

The range of scores given by individual translators vary for some translation problems more than for the other problems, which is shown by different standard deviation figures. Disagreement indicates that translators are not sure about the quality of dictionary solutions in certain contexts. This appeared to be a very informative measure for dictionary solutions.

In particular, there is a surprising relation between the inter-annotator disagreement on dictionary solutions and the quality of the *ASSIST* solutions: higher disagreement on the quality of dictionary translations correlates with the better quality of *ASSIST* solutions (i.e. the difference between the average dictionary and *ASSIST* scores), which peaks at a certain point, but then becomes unstable (see Figure 1 vs. 2).

Linguistic interpretation of this phenomenon could be that the *ASSIST* technology works best for a certain class of translation problems – the problems, whose translation is not straightforward, but on the other hand isn't too much idiosyncratic or controversial. Our system successfully finds non-literal translation equivalents, which are within a certain optimal range of distributional similarity from the original, but it cannot yet apply radical translation transformations and shifts which go beyond distributional similarity model.

It is possible to predict usefulness of the *ASSIST* solutions for translators on the basis of how much translators disagree on dictionary solutions for particular problems by defining the range of the standard deviation scores, where *ASSIST* scores normally are higher than dictionary scores. It can be seen from the figures that the optimal range is approximately between $\sigma = 0.75$ and $\sigma = 1.5$ – the difference of

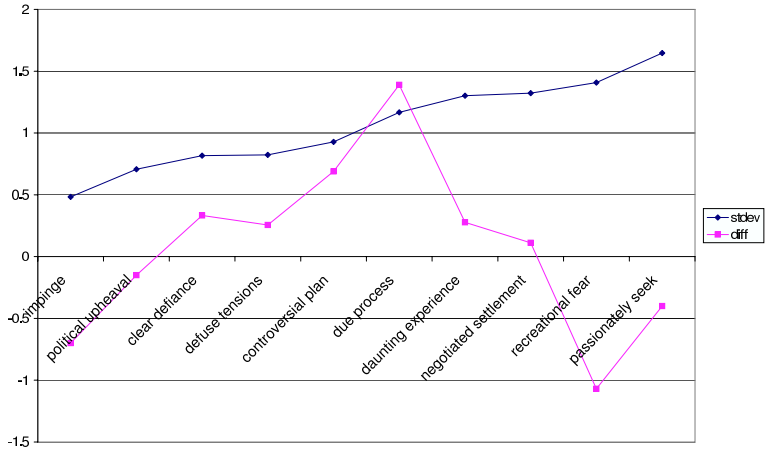

Figure 1. Agreement scores: dictionary

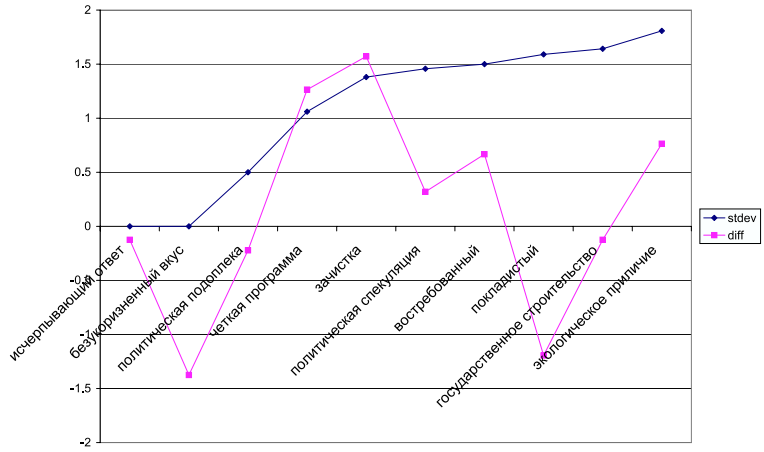

Figure 2. Agreement scores: ASSIST

ASSIST and dictionary scores then is usually positive, and outside this range it is more often negative.

The two groups of translation problems and the respective evaluation scores are shown in Table II. The first group is outside the optimal range of standard deviation figures for dictionary solutions, so here *ASSIST* average scores are about 9% lower than dictionary scores. The second group shows problems within the range of $\sigma = 0.75$ and $\sigma = 1.5$,

Table II. Groups of translation equivalents

Problems outside range	dict	σ	system	σ
impinge	4.7	0.483	4	1
political upheaval	4.5	0.707	4.35	1.334
исчерпывающий ответ	5	0	4.875	0.354
покладистый	3.444	1.590	2.25	1.165
экологическое приличие	3.125	1.808	3.889	1.269
Average	4.107		3.714	*-9.5%*
clear defiance	4	0.816	4.333	1.323
defuse tensions	4.3	0.823	4.556	0.527
controversial plan	4.111	0.928	4.8	0.632
due process	3.111	1.167	4.5	0.707
daunting experience	3.222	1.302	3.5	1.269
recreational fear	2.625	1.408	1.556	1.130
четкая программа	3.625	1.061	4.889	0.333
зачистка	2.286	1.380	3.857	1.464
востребованный	3.333	1.5	4	1
Average	3.431		3.959	*+15.4%*

and here *ASSIST* solutions scored about 15% higher than dictionary solutions.

Having said this, solutions from our system are really not in competition with dictionary solutions: they provide less literal translations, which often emerge in later stages of the translation task, when translators correct and improve an initial draft, where they have usually put more literal equivalents. It is a known fact in translation studies that non-literal solutions are harder to see and translators often find them only upon longer reflection. Yet another fact is that non-literal translations often require re-writing other segments of the sentence, which may not be obvious at first glance.

4. Conclusions and future work

The results of evaluation show that the tool is successful in finding translation equivalents for a range of examples. What is more, in cases where the problem is genuinely difficult, for each query ASSIST consistently provides a solution that scores around 4 – "minor adaptations

needed". At the same time the precision of the tool is low, it suggests 50-100 examples with only 2-4 useful for the current context. However, recall of the output is more relevant than precision, because translators typically need just one solution for their problem, and often have to look through reasonably large lists of dictionary translations and examples to find something suitable for a problematic expression. Even if no immediately suitable translation can be found in the list of suggestions, it frequently contains a hint for solving the problem in the absence of adequate dictionary information.

The current implementation of the model is restricted in several respects. First, the majority of target language constructions mirror the syntactic structure of the source language example. Even if the procedure for producing similarity classes does not impose restrictions on POS properties, nevertheless words in the similarity class tend to follow the POS of the original word, because of the similarity of their contexts of use. Furthermore, dictionaries also tend to translate words using the same POS. This means that the existing method finds mostly NPs for NPs, verb-object pairs for verb-object pairs, etc, even if the most natural translation uses a different syntactic structure, e.g. *I like doing X* instead of *I do X gladly* (when translating from German *ich mache X gerne*).

These issues can be addressed by introducing a model of the semantic context of situation, e.g. 'changes in business practice' as in the example above, or 'unpleasant situation' as in the case of *daunting experience*. This will allow less restrictive identification of possible translation equivalents, as well as reduction of suggestions irrelevant for the context of the current example. Currently we are working on an option to identify semantic contexts by means of 'semantic signatures' obtained from a broad-coverage semantic parser, such as USAS (Rayson et al., 2004). The semantic tagset used by USAS is a language-independent multi-tier structure with 21 major discourse fields, subdivided into 232 sub-categories (such as `I1.1-` = Money: lack; `A5.1-` = Evaluation: bad), which can be used to detect the semantic context.

Another possibility of representing semantics in similarity classes is to utilise the notion of lexical functions (Mel'čuk, 1996). For instance, examples like *strong feeling, opposition, sense* indicate a high degree of a quality. Such cases have been generalised by Mel'čuk under the notion of a lexical function, e.g. **Magn**(*feeling*)=*strong*. If knowledge of this sort is encoded in the lexicon (or inferred automatically from corpora), the output can be filtered to include only words that can be used in this function.

Acknowledgements

This research is supported by EPSRC grant EP/C005902. We are grateful to the anonymous reviewers for their insightful comments and links to relevant research.

References

Babych, B. and A. Hartley: 2004, 'Extending the BLEU MT Evaluation Method with Frequency Weightings'. In: *Proceedings of the 42^d Annual Meeting of the Association for Computational Linguistics*. Barcelona.

Dagan, I. and K. Church: 1997, 'Termight: Coordinating Humans and Machines in Bilingual Terminology Acquisition'. *Machine Translation* **12**(1/2), 89–107.

Daille, B. and E. Morin: 2005, 'French-English terminology extraction from comparable corpora'. In: *Proceedings IJCNLP 2005: Second International Joint Conference*, Vol. 3651 of *Lecture Notes in Computer Sciences (LNCS)*. pp. 707–719.

Grefenstette, G.: 2002, 'Multilingual corpus-based extraction and the Very Large Lexicon'. In: L. Borin (ed.): *Language and Computers, Parallel corpora, parallel worlds*. Rodopi, pp. 137–149.

Justeson, J. S. and S. M. Katz: 1995, 'Techninal terminology: some linguistic properties and an algorithm for identification in text'. *Natural Language Engineering* **1**(1), 9–27.

Lin, D.: 1998, 'Automatic Retrieval and Clustering of Similar Words'. In: *Proc. Joint COLING-ACL-98*. Montreal, pp. 768–774.

Mel'čuk, I. A.: 1996, 'Lexical Functions: a tool for the description of lexical relations in a lexicon'. In: L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins, pp. 37–102.

Partington, A.: 1998, *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Ploux, S. and H. Ji: 2003, 'A model for matching semantic maps between languages (French/English, English/French)'. *Computational Linguistics* **29**(2), 155–178.

Rapp, R.: 2004, 'A Freely Available Automatically Generated Thesaurus of Related Words'. In: *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*. Lisbon, pp. 395–398.

Rayson, P., D. Archer, S. Piao, and T. McEnery: 2004, 'The UCREL semantic analysis system'. In: *Proc. Beyond Named Entity Recognition Workshop in association with LREC 2004*. Lisbon, pp. 7–12.

Sharoff, S.: 2006, 'Creating general-purpose corpora using automated search engine queries'. In: M. Baroni and S. Bernardini (eds.): *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit. http://wackybook.sslmit.unibo.it.

Zanettin, F.: 1998, 'Bilingual comparable corpora and the training of translators'. *Meta* **XLIII**(4).