Assisting Translators in Indirect Lexical Transfer

Bogdan Babych, Anthony Hartley, Serge Sharoff
Centre for Translation Studies
University of Leeds, UK
{b.babych,a.hartley,s.sharoff}@leeds.ac.uk

Olga Mudraya
Department of Linguistics
Lancaster University, UK
o.mudraya@lancs.ac.uk

Abstract

We present the design and evaluation of a *translator's amenuensis* that uses comparable corpora to propose and rank non-literal solutions to the translation of expressions from the general lexicon. Using distributional similarity and bilingual dictionaries, the method outperforms established techniques for extracting translation equivalents from parallel corpora. The interface to the system is available at: http://corpus.leeds.ac.uk/assist/v05/

1 Introduction

This paper describes a system designed to assist humans in translating expressions that do not necessarily have a literal or compositional equivalent in the target language (TL). In the spirit of (Kay, 1997), it is intended as a *translator's amenuensis* "under the tight control of a human translator … to help increase his productivity and not to supplant him".

One area where human translators particularly appreciate assistance is in the translation of expressions from the general lexicon. Unlike equivalent technical terms, which generally share the same part-of-speech (POS) across languages and are in the ideal case univocal, the contextually appropriate equivalents of general language expressions are often indirect and open to variation. While the transfer module in RBMT may acceptably undergenerate through a many-to-one mapping between source and target expressions, human translators, even in non-literary fields, value legitimate variation. Thus the French expression *il faillit échouer* (lit.: he faltered to fail) may be variously rendered as *he almost/nearly/all but failed; he was on the verge/brink of failing/failure; failure loomed*. All of these translations are indirect in that they involve lexical shifts or POS transformations.

Finding such translations is a hard task that can benefit from automated assistance. 'Mining' such indirect equivalents is difficult, precisely because of the structural mismatch, but also because of the paucity of suitable aligned corpora. The approach adopted here includes the use of comparable corpora in source and target languages, which are relatively easy to create. The challenge is to generate a list of usable solutions and to rank them such that the best are at the top.

Thus the present system is unlike SMT (Och and Ney, 2003), where lexical selection is effected by a translation model based on aligned, parallel corpora, but the novel techniques it has developed are exploitable in the SMT paradigm. It also differs from now traditional uses of comparable corpora for detecting translation equivalents (Rapp, 1999) or extracting terminology (Grefenstette, 2002), which allows a one-to-one correspondence irrespective of the context. Our system addresses difficulties in expressions in the general lexicon, whose translation is context-dependent.

The structure of the paper is as follows. In Section 2 we present the method we use for mining translation equivalents. In Section 3 we present the results of an objective evaluation of the quality of suggestions produced by the system by comparing our output against a parallel corpus. Finally, in Section 4 we present a subjective evaluation focusing on the integration of the system into the workflow of human translators.

2 Methodology

The software acts as a decision support system for translators. It integrates different technologies for

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 136–143,
Prague, Czech Republic, June 2007.

extracting indirect translation equivalents from large comparable corpora. In the following subsections we give the user perspective on the system and describe the methodology underlying each of its sub-tasks.

2.1 User perspective

Unlike traditional dictionaries, the system is a *dynamic translation resource* in that it can successfully find translation equivalents for units which have not been stored in advance, even for idiosyncratic multiword expressions which almost certainly will not figure in a dictionary. While our system can rectify gaps and omissions in static lexicographical resources, its major advantage is that it is able to cope with an open set of translation problems, searching for translation equivalents in comparable corpora in runtime. This makes it more than just an extended dictionary.

Contextual descriptors

From the user perspective the system extracts indirect translation equivalents as sets of *contextual descriptors* – content words that are lexically central in a given sentence, phrase or construction. The choice of these descriptors may determine the general syntactic perspective of the sentence and the use of supporting lexical items. Many translation problems arise from the fact that the mapping between such descriptors is not straightforward.

The system is designed to find possible indirect mappings between sets of descriptors and to verify the acceptability of the mapping into the TL. For example, in the following Russian sentence, the bolded contextual descriptors require indirect translation into English.

> *Дети посещают* ***плохо отремонтированные*** *школы, в которых* ***недостает*** *самого* ***необходимого***
> *(Children attend* ***badly repaired*** *schools, in which [it] is* ***missing*** *the most* ***necessary****)*

Combining direct translation equivalents of these words (e.g., translations found in the Oxford Russian Dictionary – ORD) may produce a non-natural English sentence, like the literal translation given above. In such cases human translators usually apply structural and lexical transformations, for instance changing the descriptors' POS and/or replacing them with near-synonyms which fit together in the context of a TL sentence (Munday, 2001: 57-58). Thus, a structural transformation of *плохо отремонтированные* (badly repaired) may give *in poor repair* while a lexical transformation of *недостает самого необходимого* ([it] is missing the most necessary) gives *lacking basic essentials*.

Our system models such transformations of the descriptors and checks the consistency of the resulting sets in the TL.

Using the system

Human translators submit queries in the form of one or more SL descriptors which in their opinion may require indirect translation. When the translators use the system for translating into their native language, the returned descriptors are usually sufficient for them to produce a correct TL construction or phrase around them (even though the descriptors do not always form a naturally sounding expression). When the translators work into a non-native language, they often find it useful to generate concordances for the returned descriptors to verify their usage within TL constructions.

For example, for the sentence above translators may submit two queries: *плохо отремонтированные* (badly repaired) and *недостает необходимого* (missing necessary). For the first query the system returns a list of descriptor pairs (with information on their frequency in the English corpus) ranked by distributional proximity to the original query, which we explain in Section 2.2. At the top of the list come:

```
bad repair = 30         (11.005)
bad maintenance = 16    (5.301)
bad restoration = 2     (5.079)
poor repair = 60        (5.026)…
```

Underlined hyperlinks lead translators to actual contexts in the English corpus, e.g., *poor repair* generates a concordance containing a desirable TL construction which is a structural transformation of the SL query:

in such a	**poor**	state of **repair**
bridge in as	**poor**	a state of **repair** as the highways
building in	**poor**	**repair**.
dwellings are in	**poor**	**repair**;

Similarly, the result of the second query may give the translators an idea about possible lexical transformation:

```
missing need = 14       (5.035)
important missing = 8   (2.930)
missing vital = 8       (2.322)
lack necessary = 204    (1.982)…
essential lack = 86     (0.908)…
```

The concordance for the last pair of descriptors contains the phrase *they lack the three essentials*, which illustrates the transformation. The resulting translation may be the following:

Children attend schools that are in ***poor repair*** *and* ***lacking*** *basic* ***essentials***

Thus our system supports translators in making decisions about indirect translation equivalents in a number of ways: it suggests possible structural and lexical transformations for contextual descriptors; it verifies which translation variants co-occur in the TL corpus; and it illustrates the use of the transformed TL lexical descriptors in actual contexts.

2.2 Generating translation equivalents

We have generalised the method used in our previous study (Sharoff et al., 2006) for extracting equivalents for continuous multiword expressions (MWEs). Essentially, the method expands the search space for each word and its dictionary translations with entries from automatically computed thesauri, and then checks which combinations are possible in target corpora. These potential translation equivalents are then ranked by their similarity to the original query and presented to the user. The range of retrievable equivalents is now extended from a relatively limited range of two-word constructions which mirror POS categories in SL and TL to a much wider set of co-occurring lexical content items, which may appear in a different order, at some distance from each other, and belong to different POS categories.

The method works best for expressions from the general lexicon, which do not have established equivalents, but not yet for terminology. It relies on a high-quality bilingual dictionary (en-ru ~30k, ru-en ~50K words, combining ORD and the core part of Multitran) and large comparable corpora (~200M En, ~70M Ru) of news texts.

For each of the SL query terms q the system generates its *dictionary translation* $Tr(q)$ and its *similarity class* $S(q)$ – a set of words with a similar distribution in a monolingual corpus. Similarity is measured as the cosine between collocation vectors, whose dimensionality is reduced by SVD using the implementation by Rapp (2004). The descriptor and each word in the similarity class are then translated into the TL using ORD or the Multitran dictionary, resulting in $\{Tr(q) \cup Tr(S(q))\}$. On the TL side we also generate similarity classes, but only for dictionary translations of query terms $Tr(q)$ (not for $Tr(S(q))$, which can make output too noisy). We refer to the resulting set of TL words as a *translation class T*.

$$T = \{Tr(q) \cup Tr(S(q)) \cup S(Tr(q))\}$$

Translation classes approximate lexical and structural transformations which can potentially be applied to each of the query terms. Automatically computed similarity classes do not require resources like WordNet, and they are much more suitable for modelling translation transformations, since they often contain a wider range of words of different POS which share the same context, e.g., the similarity class of the word *lack* contains words such as *absence, insufficient, inadequate, lost, shortage, failure, paucity, poor, weakness, inability, need.* This clearly goes beyond the range of traditional thesauri.

For multiword queries, the system performs a consistency check on possible combinations of words from different translation classes. In particular, it computes the Cartesian product for pairs of translation classes T_1 and T_2 to generate the set P of word pairs, where each word (w_1 and w_2) comes from a different translation class:

$$P = T_1 \times T_2 = \{(w_1, w_2) \mid w_1 \in T_1 \text{ and } w_2 \in T_2\}$$

Then the system checks whether each word pair from the set P exists in the database D of discontinuous content word bi-grams which actually co-occur in the TL corpus:

$$P' = P \cap D$$

The database contains the set of all bi-grams that occur in the corpus with a frequency ≥ 4 within a window of 5 words (over 9M bigrams for each language). The bi-grams in D and in P are sorted alphabetically, so their order in the query is not important.

Larger N-grams ($N > 2$) in queries are split into combinations of bi-grams, which we found to be an optimal solution to the problem of the scarcity of higher order N-grams in the corpus. Thus, for the query *gain significant importance* the system generates $P'_{1(\texttt{significant importance})}$, $P'_{2(\texttt{gain importance})}$, $P'_{3(\texttt{gain significant})}$ and computes P' as:

$$P' = \{(w_1,w_2,w_3) \mid (w_1,w_2) \in P'_1 \;\&\; (w_1, w_3) \in P'_2 \;\&\; (w_2,w_3) \in P'_3\},$$

which allows the system to find an indirect equivalent *получить весомое значение* (lit.: receive weighty meaning).

Even though P' on average contains about 2% - 4% of the theoretically possible number of bigrams present in P, the returned number of potential translation equivalents may still be large and contain much noise. Typically there are several hundred elements in P', of which only a few are really useful for translation. To make the system usable in practice, i.e., to get useful solutions to appear close to the top (preferably on the first screen of the output), we developed methods of ranking and filtering the returned TL contextual descriptor pairs, which we present in the following sections.

2.3 Hypothesis ranking

The system ranks the returned list of contextual descriptors by their distributional proximity to the original query, i.e. it uses scores $cos(v_q, v_w)$ generated for words in similarity classes – the cosine of the angle between the collocation vector for a word and the collocation vector for the query or dictionary translation of the query. Thus, words whose equivalents show similar usage in a comparable corpus receive the highest scores. These scores are computed for each individual word in the output, so there are several ways to combine them to weight words in translation classes and word combinations in the returned list of descriptors.

We established experimentally that the best way to combine similarity scores is to multiply weights $W(T)$ computed for each word within its translation class T. The weight $W(P'_{(w1,w2)})$ for each pair of contextual descriptors $(w_1, w_2) \in P'$ is computed as:

$$W(P'_{(w1,w2)}) = W(T_{(w1)}) \times W(T_{(w2)});$$

Computing $W(T_{(w)})$, however, is not straightforward either, since some words in similarity classes of different translation equivalents for the query term may be the same, or different words from the similarity class of the original query may have the same translation. Therefore, a word w within a translation class may have come by several routes simultaneously, and may have done that several times. For each word w in T there is a possibility that it arrived in T either because it is in $Tr(q)$ or occurs n times in $Tr(S(q))$ or k times in $S(Tr(q))$.

We found that the number of occurrences n and k of each word w in each subset gives valuable information for ranking translation candidates. In our experiments we computed the weight $W(T)$ as the sum of similarity scores which w receives in each of the subsets. We also discovered that ranking improves if for each query term we compute in addition a larger (and potentially noisy) space of candidates that includes TL similarity classes of translations of the SL similarity class $S(Tr(S(q)))$. These candidates do not appear in the system output, but they play an important role in ranking the displayed candidates. The improvement may be due to the fact that this space is much larger, and may better support relevant candidates since there is a greater chance that appropriate indirect equivalents are found several times within SL and TL similarity classes. The best ranking results were achieved when the original $W(T)$ scores were multiplied by 2 and added to the scores for the newly introduced similarity space $S(Tr(S(q)))$:

$$\begin{aligned}W(T_{(w)}) = {} & 2\times(1 \text{ if } w \in Tr(q))+ \\ & 2\times\sum(cos(v_q, v_{Tr(w)}) \mid \{w \mid w \in Tr(S(q)) \}) + \\ & 2\times\sum(cos(v_{Tr(q)}, v_w) \mid \{w \mid w \in S(Tr(q)) \}) + \\ & \sum(cos(v_q, v_{Tr(w)}) \times cos(v_{Tr(q)}, v_w) \mid \\ & \qquad \{w \mid w \in S(Tr(S(q))) \})\end{aligned}$$

For example, the system gives the following ranking for the indirect translation equivalents of the Russian phrase *весомое значение* (lit.: weighty meaning) – figures in brackets represent $W(P')$ scores for each pair of TL descriptors:

```
1. significant importance = 7 (3.610)
2. significant value = 128    (3.211)
3. measurable value = 6       (2.657)…
8. dramatic importance = 2    (2.028)
9. important significant = 70 (2.014)
10. convincing importance = 6 (1.843)
```

The Russian similarity class for *весомый* (weighty, ponderous) contains: *убедительный* (convincing) (0.469), *значимый* (significant) (0.461), *ощутимый* (notable) (0.452) *драматичный* (dramatic) (0.371). The equivalent of *significant* is not at the top of the similarity class of the Russian query, but it appears at the top of the final ranking of pairs in P', because this hypothesis is supported by elements of the set formed by $S(Tr(S(q)))$; it appears in similarity classes for *notable* (0.353) and *dramatic* (0.315), which contributed these values to the $W(T)$ score of *significant*:

W(T(`significant`)) =
2 × (Tr(значимый)=`significant` (0.461))
+ (Tr(ощутимый)=`notable` (0.452)
× S(`notable`)=`significant` (0.353))
+ (Tr(драматичный)=`dramatic` (0.371)
× S(`dramatic`)= `significant` (0.315))

The word *dramatic* itself is not usable as a translation equivalent in this case, but its similarity class contains the support for relevant candidates, so it can be viewed as useful noise. On the other hand, the word *convincing* does not receive such support from the hypothesis space, even though its Russian equivalent is ranked higher in the SL similarity class.

2.4 Semantic filtering

Ranking of translation candidates can be further improved when translators use an option to filter the returned list by certain lexical criteria, e.g., to display only those examples that contain a certain lexical item, or to require one of the items to be a dictionary translation of the query term. However, lexical filtering is often too restrictive: in many cases translators need to see a number of related words from the same semantic field or subject domain, without knowing the lexical items in advance. In this section we present the semantic filter, which is based on Russian and English semantic taggers which use the same semantic field taxonomy for both languages.

The semantic filter displays only those items which have specified semantic field tags or tag combinations; it can be applied to one or both words in each translation hypothesis in *P'*. The default setting for the semantic filter is the requirement for both words in the resulting TL candidates to contain any of the semantic field tags from a SL query term.

In the next section we present evaluation results for this default setting (which is applied when the user clicks the *Semantic Filter* button), but human translators have further options – to filter by tags of individual words, to use semantic classes from SL or TL terms, etc.

For example, applying the default semantic filter for the output of the query *плохо отремонтированные* (badly repaired) removes the highlighted items from the list:

```
 1. bad repair = 30        (11.005)
[2. good repair = 154      (8.884) ]
 3. bad rebuild = 6        (5.920)
[4. bad maintenance = 16   (5.301) ]
 5. bad restoration = 2    (5.079)
 6. poor repair = 60       (5.026)
[7. good rebuild = 38      (4.779) ]
 8. bad construction = 14 (4.779)
```

Items 2 and 7 are generated by the system because *good*, *well* and *bad* are in the same similarity cluster for many words (they often share the same collocations). The semantic filter removes examples with *good* and *well* on the grounds that they do not have any of the tags which come from the word *плохо* (badly): in particular, instead of tag `A5-` (Evaluation: Negative) they have tag `A5+` (Evaluation: Positive). Item 4 is removed on the grounds that the words *отремонтированный* (repaired) and *maintenance* do not have any tags in common – they appear ontologically too far apart from the point of view of the semantic tagger.

The core of the system's multilingual semantic tagging is a knowledge base in which single words and MWEs are mapped to their potential semantic field categories. Often a lexical item is mapped to multiple semantic categories, reflecting its potential multiple senses. In such cases, the tags are arranged by the order of likelihood of meanings, with the most prominent first.

3 Objective evaluation

In the *objective evaluation* we tested the performance of our system on a selection of indirect translation problems, extracted from a parallel corpus consisting mostly of articles from English and Russian newspapers (118,497 words in the R-E direction, 589,055 words in the E-R direction). It has been aligned on the sentence level by JAPA (Langlais et al., 1998), and further on the word level by GIZA++ (Och and Ney, 2003).

3.1 Comparative performance

The intuition behind the objective evaluation experiment is that the capacity of our tool to find indirect translation equivalents in comparable corpora can be compared with the results of automatic alignment of parallel texts used in translation models in SMT: one of the major advantages of the SMT paradigm is its ability to reuse indirect equivalents found in parallel corpora (equivalents that may never come up in hand-crafted dictionaries). Thus, automatically generated GIZA++ dictionaries with word alignment contain many examples of indirect translation equivalents.

We use these dictionaries to simulate the generator of translation classes *T*, which we recombine to construct their Cartesian product *P*, similarly to the procedure we use to generate the output of our system. However, the two approaches generate indirect translation equivalence hypotheses on the basis of radically different material: the GIZA dictionary uses evidence from parallel corpora of existing human translations, while our system recombines translation candidates on the basis of their distributional similarity in monolingual comparable corpora. Therefore we took GIZA as a baseline.

Translation problems for the objective evaluation experiment were manually extracted from two parallel corpora: a section of about 10,000 words of a corpus of English and Russian newspapers, which we also used to train GIZA, and a section of the same length from a corpus of interviews published on the Euronews.net website.

We selected expressions which represented cases of lexical transformations (as illustrated in Section 0), containing at least two content words both in the SL and TL. These expressions were converted into pairs of contextual descriptors – e.g., *recent success, reflect success* – and submitted to the system and to the GIZA dictionary. We compared the ability of our system and of GIZA to find indirect translation equivalents which matched the equivalents used by human translators. The output from both systems was checked to see whether it contained the contextual descriptors used by human translators. We submitted 388 pairs of descriptors extracted from the newspaper translation corpus and 174 pairs extracted from the Euronews interview corpus. Half of these pairs were Russian, and the other half English.

We computed recall figures for 2-word combinations of contextual descriptors and single descriptors within those combinations. We also show the recall of translation variants provided by the ORD on this data set. For example, for the query *недостает необходимого* ([it] is missing necessary [things]) human translators give the solution *lacking essentials*; the lemmatised descriptors are *lack* and *essential*. ORD returns direct translation equivalents *missing* and *necessary*. The GIZA dictionary in addition contains several translation equivalents for the second term (with alignment probabilities) including: *necessary* ~0.332*, need* ~0.226*, essential* ~0.023. Our system returns both descriptors used in human translation as a pair – *lack essential* (ranked 41 without filtering and 22 with the default semantic filter). Thus, for a 2-word combination of the descriptors only the output of our system matched the human solution, which we counted as one hit for the system and no hits for ORD or GIZA. For 1-word descriptors we counted 2 hits for our system (both words in the human solution are matched), and 1 hit for GIZA – it matches the word *essential* ~0.023 (which also illustrates its ability to find indirect translation equivalents).

	2w descriptors		1w descriptors	
	news	**interv**	**news**	**interv**
ORD	6.7%	4.6%	32.9%	29.3%
GIZA++	13.9%	3.4%	35.6%	29.0%
Our system	**21.9%**	**19.5%**	**55.8%**	**49.4%**

Table 1 Conservative estimate of recall

It can be seen from **Table 1** that for the newspaper corpus on which it was trained, GIZA covers a wider set of indirect translation variants than ORD. But our recall is even better both for 2-word and 1-word descriptors.

However, note that GIZA's ability to retrieve from the newspaper corpus certain indirect translation equivalents may be due to the fact that it has previously seen them frequently enough to generate a correct alignment and the corresponding dictionary entry.

The Euronews interview corpus was not used for training GIZA. It represents spoken language and is expected to contain more 'radical' transformations. The small decline in ORD figures here can be attributed to the fact that there is a difference in genre between written and spoken texts and consequently between transformation types in them. However, the performance of GIZA drops radically on unseen text and becomes approximately the same as the ORD.

This shows that indirect translation equivalents in the parallel corpus used for training GIZA are too sparse to be learnt one by one and successfully applied to unseen data, since solutions which fit one context do not necessarily suit others.

The performance of our system stays at about the same level for this new type of text; the decline in its performance is comparable to the decline in ORD figures, and can again be explained by the differences in genre.

3.2 Evaluation of hypothesis ranking

As we mentioned, correct ranking of translation candidates improves the usability of the system. Again, the objective evaluation experiment gives only a conservative estimate of ranking, because there may be many more useful indirect solutions further up the list in the output of the system which are legitimate variants of the solutions found in the parallel corpus. Therefore, evaluation figures should be interpreted in a comparative rather then an absolute sense.

We use ranking by frequency as a baseline for comparing the ranking described in Section 2.3 – by distributional similarity between a candidate and the original query.

Table 2 shows the average rank of human solutions found in parallel corpora and the recall of these solutions for the top 300 examples. Since there are no substantial differences between the figures for the newspaper texts and for the interviews, we report the results jointly for 556 translation problems in both selections (lower rank figures are better).

	Recall	**Average rank**
2-word descriptors		
frequency (baseline)	16.7%	*rank=93.7*
distributional similarity	19.5%	*rank=**44.4***
sim. + semantic filter	14.4%	*rank=**26.7***
1-word descriptors		
frequency (baseline)	48.2%	*rank=42.7*
distributional similarity	52.8%	*rank=**21.6***
sim. + semantic filter	44.1%	*rank=**11.3***

Table 2 Ranking: frequency, similarity and filter

It can be seen from the table that ranking by similarity yields almost a twofold improvement for the average rank figures compared to the baseline. There is also a small improvement in recall, since there are more relevant examples that appear within the top 300 entries.

The semantic filter once again gives an almost twofold improvement in ranking, since it removes many noisy items. The average is now within the top 30 items, which means that there is a high chance that a translation solution will be displayed on the first screen. The price for improved ranking is decline in recall, since it may remove some relevant lexical transformations if they appear to be ontologically too far apart. But the decline is smaller: about 26.2% for 2-word descriptors and 16.5% for 1-word descriptors. The semantic filter is an optional tool, which can be used to great effect on noisy output: its improvement of ranking outweighs the decline in recall.

Note that the distribution of ranks is not normal, so in **Figure 1** we present frequency polygons for rank groups of 30 (which is the number of items that fit on a single screen, i.e., the number of items in the first group (r030) shows solutions that will be displayed on the first screen). The majority of solutions ranked by similarity appear high in the list (in fact, on the first two or three screens).

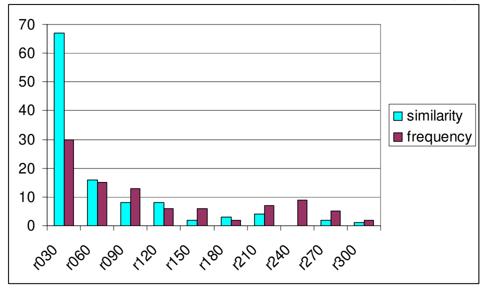

Figure 1 Frequency polygons for ranks

4 Subjective evaluation

The objective evaluation reported above uses a single reference translation and is correspondingly conservative in estimating the coverage of the system. However, many expressions studied have more than one fluent translation. For instance, *in poor repair* is not the only equivalent for the Russian expression *плохо отремонтированные*. It is also possible to translate it as *unsatisfactory condition, bad state of repair, badly in need of repair*, and so on. The objective evaluation shows that the system has been able to find the suggestion used by a particular translator for the problem studied. It does not tell us whether the system has found some other translations suitable for the context. Such legitimate translation variation implies that the performance of a system should be studied on the basis of multiple reference translations, though typically just two reference translations are used (Papineni, et al, 2001). This might be enough for the purposes of a fully automatic MT tool, but in the context of a translator's amanuensis which deals with expressions difficult for human translators, it is reasonable to work with a larger range of acceptable target expressions.

With this in mind we evaluated the performance of the tool with a panel of 12 professional translators. Problematic expressions were highlighted and the translators were asked to find suitable suggestions produced by the tool for these expressions and rank their usability on a scale from 1 to 5 (not acceptable to fully idiomatic, so 1 means that no usable translation was found at all).

Sentences themselves were selected from problems discussed on professional translation forums proz.com and forum.lingvo.ru. Given the range of corpora used in the system (reference and newspaper corpora), the examples were filtered to address expressions used in newspapers.

The goal of the subjective evaluation experiment was to establish the usefulness of the system for translators beyond the conservative estimate given by the objective evaluation. The intuition behind the experiment is that if there are several admissible translations for the SL contextual descriptors, and system output matches any of these solutions, then the system has generated something useful. Therefore, we computed recall on sets of human solutions rather than on individual solutions. We matched 210 different human solutions to 36 translation problems. To compute more realistic recall figures, we counted cases when the system output matches any of the human solutions in the set. **Table 3** compares the conservative estimate of the objective evaluation and the more realistic estimate on a single data set.

	2w default	**2w with sem filt**
Conservative	*32.4%; r=53.68*	*21.9%; r=34.67*
Realistic	***75.0%; r=7.48***	***61.1%; r=3.95***

Table 3 Recall and rank for 2-word descriptors

Since the data set is different, the figures for the conservative estimate are higher than those for the objective evaluation data set. However, the table shows the there is a gap between the conservative estimate and the realistic coverage of the translation problems by the system, and that real coverage of indirect translation equivalents is potentially much higher.

Table 4 shows averages (and standard deviation σ) of the usability scores divided in four groups: (1) solutions that are found both by our system and the ORD; (2) solutions found only by our system; (3) solutions found only by ORD (4) solutions found by neither:

	system (+)	**system (–)**
ORD (+)	4.03 (0.42)	3.62 (0.89)
ORD (–)	4.25 (0.79)	3.15 (1.15)

Table 4 Human scores and σ for system output

It can be seen from the table that human users find the system most useful for those problems where the solution does not match any of the direct dictionary equivalents, but is generated by the system.

5 Conclusions

We have presented a method of finding indirect translation equivalents in comparable corpora, and integrated it into a system which assists translators in indirect lexical transfer. The method outperforms established methods of extracting indirect translation equivalents from parallel corpora.

We can interpret these results as an indication that our method, rather than learning individual indirect transformations, models the entire family of transformations entailed by indirect lexical transfer. In other words it learns a translation strategy which is based on the distributional similarity of words in a monolingual corpus, and applies this strategy to novel, previously unseen examples.

The coverage of the tool and additional filtering techniques make it useful for professional translators in automating the search for non-trivial, indirect translation equivalents, especially equivalents for multiword expressions.

References

Gregory Grefenstette. 2002. Multilingual corpus-based extraction and the very large lexicon. In: Lars Borin, editor, Language and Computers, *Parallel corpora, parallel worlds*, pages 137-149. Rodopi.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3-23.

Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and practical issues in evaluating alignment techniques. In *Proc. Joint COLING-ACL-98*, pages 711-717.

Jeremy Munday. 2001. *Introducing translation studies. Theories and Applications*. Routledge, New York.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation, *RC22176 W0109-022*: IBM.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Procs. the 37th ACL,* pages 395-398.

Reinhard Rapp. 2004. A freely available automatically generated thesaurus of related words. In *Procs. LREC 2004*, pages 395-398, Lisbon.

Serge Sharoff, Bogdan Babych and Anthony Hartley 2006. Using Comparable Corpora to Solve Problems Difficult for Human Translators. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 739-746.