Translating from under-resourced languages: Comparing direct transfer against pivot translation

Bogdan Babych, Anthony Hartley, Serge Sharoff

Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT, UK
{b.babych,a.hartley,s.sharoff}@leeds.ac.uk}

Abstract

In this paper we compare two methods for translating into English from languages for which few MT resources have been developed (e.g. Ukrainian). The first method involves direct transfer using an MT system that is available for this language pair. The second method involves translation via a cognate language, which has more translation resources and one or more advanced translation systems (e.g. Russian for Slavonic languages). The comparison shows that it is possible to achieve better translation quality via the pivot language, leveraging on advanced dictionaries and grammars available for it and on lexical and syntactic similarities between the source and pivot languages. The results suggest that MT development efforts can be efficiently reused for families of closely related languages, and investing in MT for closely related languages can be more productive than developing systems from scratch for new translation directions. We also suggest a method for comparing the performance of a direct and pivot translation routes via automated evaluation of segments with varying translation difficulty.

1. Introduction

The number of translation resources existing for some languages is far greater than for others. There are commercial systems for translation into English from well-resourced languages, such as French or Russian, that can achieve acceptable quality for many practical applications of machine translation. At the same time there are many more languages for which good quality translation resources are not available. For some of those languages MT systems have occasionally been developed, but their lexical and syntactic coverage is very far from what has been achieved for better-resourced languages.

This bottleneck can be opened by using Statistical Machine Translation (Och and Ney, 2003), (Marcu and Wong, 2002), which can be trained on parallel corpora for any language pair. However, development of a good quality SMT system requires the use of large collections of parallel texts aligned at the sentence level, amounting to at least several million words. At the same time, parallel corpora of this size tend to be very rare, especially for under-resourced languages. Even for well-resourced languages such resources also tend to be specialised, e.g. Europarl (Koehn, 2005), which covers the language of debates in the European Parliament, so their performance degrades significantly when the system is applied to a slightly different domain, e.g. news (Babych et al., 2007).

In this paper we investigate the performance of translation from an under-resourced language into English via a closely-related, or cognate, pivot language with well-developed translation resources. Typically any language can be used as the pivot if it covers the bridge for a language pair that is not available in a given MT system. For instance, if no system translating from French to Japanese is available, English can serve as the pivot for translation from French into English and then from English into Japanese. Sometimes ‘pivot’ is understood as an interlingua, an artificial language implemented with the intention of making MT systems portable between languages (Hutchins, 1995).

The use of a natural pivot language is frequently discouraged because of the argument concerning the loss in translation quality in the process of double translation. This argument is confirmed by anecdotal experience, but to our knowledge there has been no published evaluation of the actual drop in quality. The method proposed in this paper is novel in two respects. First, our pivot is closely related to the source language. Second, we use a parallel corpus to evaluate and compare the output quality of a direct MT process with that of a pivot MT process.

MT between closely related languages has been very successful, achieving near-publishable quality (which needs very little or no post-editing) for a number of historically and structurally-related languages, such as Czech and Slovak (Hajic et al., 2000b), Catalan and Spanish (Alonso, 2005), Ukrainian and Russian (Gryaznukhina, 2004). Such engines explore similarities between the related languages (Dvorak et al., 2006) and typically rely on shallow processing techniques and knowledge-light linguistic resources (Armentano-Oller et al., 2005). High quality makes such MT systems useful in the pivot-based MT framework, which we take here to mean that the text is translated in several stages via one or more intermediate natural languages, or pivots. Overall translation quality crucially depends on the quality of the weakest link in the pipeline, which is usually the stage between more distant languages. From an engineering perspective, therefore, it is beneficial to use the best available MT system for that stage, even if there is no access to its source code.

The only existing reference to an approach involving pivot-based translation via related languages is the work of Hajič and his colleagues on Česílko, an MT system for translation of software manuals from English into and between Czech and Slovak (Hajic et al., 2000b). However, their system is designed for high quality translation to Slovak and is supplemented with a translation memory system. Hajic et al. (2000b) dismiss the quality of automatic pivot translation but do not give any figures to support their position.

In Section 2 of this paper we present the design of our experiment for translating via a pivot language and the methodology for evaluating its quality. In Sections 3 and 4 we discuss the results and implications for pivot-based MT via closely related language. Then in Section 5 we outline the prospects for the development from scratch of pivot MT systems using comparable corpora.

2. Method

To test the impact of the pivot framework on MT quality, we first established a baseline for pivot MT: from Russian into English via French and German (relatively distant languages). We then performed pivot MT via closely related language: from Ukrainian into English via Russian. We compared the results with direct MT from Russian and Ukrainian into English.

We used a parallel corpus from a Ukrainian political newspaper *Mirror Weekly* (http://www.mirror-weekly.com), which is published on-line in Ukrainian, Russian and English. All texts selected for our corpus appeared between January and March 2007, but describe a broad range of topics: domestic politics, international relations, financial policy, science, information technology, etc. The majority of articles are originally written in Ukrainian, some originate in Russian and two texts – in English (these are an interview and an article by a British diplomat). Table 1 presents the characteristics of the corpus.

Language	**Texts**	**Paras**	**Sentences**	**Words**
Ukrainian	35	1449	4675	64575
Russian	35	1449	4528	65181
English	35	1449	3513	68445

Table 1: Parameters of MT evaluation corpus

The size of our corpus is almost twice that of the DARPA 94 MT evaluation corpus of 36k words (White et al., 1994), which has been widely used for such tasks and has been shown to be sufficient for automated MT evaluation methods (e.g., BLEU) to ensure high correlation with human evaluation scores for translation adequacy and fluency (Babych et al., 2004). The corpus was aligned on the paragraph level and MT-translated into English using commercial MT systems available for Ukrainian, Russian and English. Table 2 gives the characterisitics of the MT systems used for the experiment.

MT	**Version / Dev**	**Source L**	**Target L**
Pragma	2.0 (2002) *Trident Soft*	Ukrainian	English Russian
		Russian	English
Plaj-Ruta	5.0 (2003) *ProLingLtd.*	Ukrainian	Russian
ProMT XP	3.0 (2002) *ProMT*	Russian	English
			German
			French
Systran	5.0 (2004) *Systran S.A.*	Russian	English
		German	
		French	

Table 2: MT systems

The quality of MT was measured using the standard BLEU metric (Papineni et al., 2002), as well as the less commonly used WNM (Weighted N-gram Model), which on large corpus has been shown to produce a better correlation with human *adequacy* judgments (Babych and Hartley, 2004). BLEU and WNM are in some sense complementary, measuring different quality parameters: WNM assigns salience scores (similar to *tf.idf*) to N-grams, which rewards matches of those content words that are most important for the general text structure. So its correlation with *adequacy* can be expected to be higher. BLEU, however, is a better predictor for *fluency*, since it does not disregard matching sequences of function words. BLEU was computed with one reference and N-gram size 5 (BLEUr1n5).

The automated scores were computed for direct translation from Russian and Ukrainian into English, then for each pivot pipeline and for intermediate stages of the pivot translation. We also computed these scores for each text and for each paragraph in the corpus, ranked the segments by the difference in direct translation and for pivot scores, and manually checked some extreme examples with the biggest difference. No formal human evaluation was carried out; however, since the corpus is large enough and homogeneous in terms of text genres and MT architectures (all systems are Rule-Based) we can interpret the differences in automated evaluation scores as likey differences in MT quality of the evaluated systems and pivot pipelines.

3. Results

In our experiment we compared the results for the pivot translation and the (best available) direct translation for two types of pivot. The first type is translation via a traditional *distant* language pivot from a *well-resourced* language, here Russian (for which a wide range of linguistic resources is available in the public domain, as are several commercial MT systems which translate between Russian and various other European languages). The second type is translation via a *closely related* pivot from a relatively under-resourced language, here Ukrainian (for which there are fewer freely available resources, MT systems or MT translation directions).

The purpose of the experiment was to establish whether pivot introduces in all cases a loss of quality large enough to justify the development of a direct system, or whether any loss of quality can be within acceptable bounds, allowing the developers to effectively reuse existing MT systems for supported translation directions within the pivot framework, and to concentrate on the supposedly easier tasks of developing MT between closely related languages.

Determining the size of the MT evaluation corpus

Automated evaluation scores such as BLEU can be used for comparison of different MT systems only on a sufficiently large corpus: on smaller texts there is little correlation with human judgments about translation quality. In addition systems compared should have been developed with the same type of MT architecture – rule-based, statistical, etc.). Otherwise, the scores can be useful for monitoring the development of the same MT system over time, but not for comparing one system with another. In our first experiment we tried to establish whether the size of our MT evaluation corpus is sufficient for comparing MT systems, to ensure a high correlation between automated scores and human judgments.

For this task we used the DARPA 94 corpus, for which human evaluation scores are available. Chart 1 summarises the correlation with human *adequacy* judgments and standard deviation of scores on data samples of different sizes taken from the corpus: for chunks of 1, 5, 10, 20 33, 50 and 100 text (each text about 360 words). Chart 2 presents the same results for correlation with human *fluency* judgments.

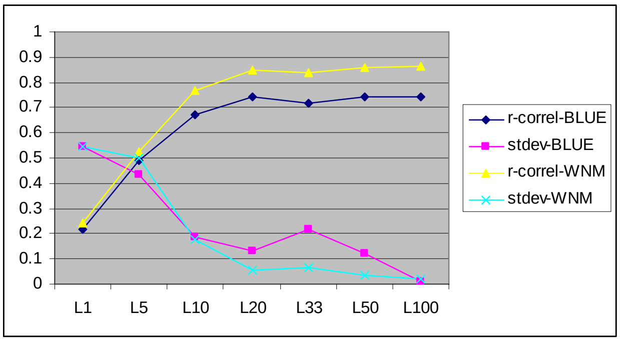

Chart 1: R correlation and stdev of r – adequacy

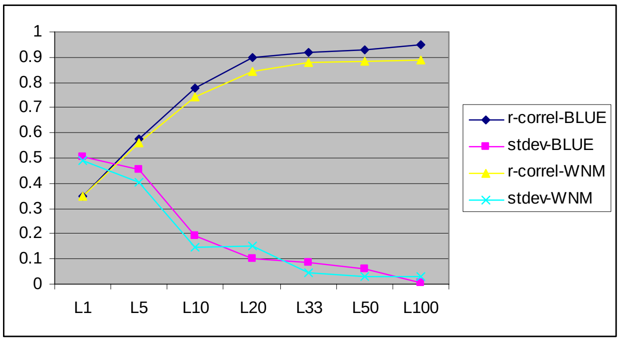

Chart 2: R correlation and stdev of r – fluency

It can be seen from the charts that the correlation and standard deviation achieve high marks and that the lines start to flatten on a corpus of about 20 texts (7200 words), which is thus the minimum size for automated evaluation. Correlation with human judgments on larger corpora get even better, but the improvement is not as fast. We can safely expect that the evaluation experiment on the corpus of our size – about 65k words (which is also in the same subject domain as the DARPA 94 corpus, used for calibration) gives a good prediction of human intuitions about translation quality.

System comparison on direct translation task

Our starting point is the comparison between different MT engines which translate between Ukrainian and Russian and from these languages into English. Some of these systems are used in the pivot pipelines, and others give a general indication of MT quality achieved for specific translation directions – between distant *vs* between closely related languages. Table 3 summarises these results. (No WNM scores are reported for translation into Russian since no reliable salience scores for Russian lexicon were available at the time of writing).

Firstly, it can be seen from the table that both automated evaluation scores rank Ukrainian–English translation lower than any Russian–English translation, which suggests that availability of development resources and the amount of the development effort. This is possibly guided by commercial considerations, like the size of potential market for the system and competition with other systems. It can be decisive for the quality of MT: for Ukrainian fewer resources are available, there is little competition and there is smaller market than for Russian.

System	**Direction**	**BLEUr1n5**	**WNM5**
MT between closely related languages			
Plaj-Ruta	ua>ru	**0.5783**	–
Pragma	ua>ru	**0.6193**	–
MT between distant languages			
Pragma	ua>en	0.0387	0.1827
Pragma	ru>en	0.0429	0.1945
ProMT	ru>en	**0.0574**	**0.2053**
Systran	ru>en	0.0511	0.1935

Table 3: MT evaluation scores for direct translation

Secondly, according to both automated scores the best direct translation quality for English–Russian direction is achieved by ProMT (which is not surprising for a mainstream translation direction developed by a well-resourced Russian team working for many years).

Thirdly, BLEU scores for closely related translation (ua>ru) are much higher than for distant translation (ua>en and ru>en). Even though BLEU scores for translation into different languages (English vs. Russian) are not directly comparable – the difference in scores does not necessarily correspond to a difference in human judgment about translation quality (Babych et al., 2005) – there is still no doubt that for MT between closely related languages the number of N-gram matches between MT output and human reference is much higher, especially for longer N-grams.

Interestingly, the distribution of BLEU scores for N-grams of different length is different for MT between closely related languages and MT for distant languages. Chart 3 illustrates these distributions for N-grams N=1 to N=5. The most surprising fact is not the even greater N-gram precision for closely related translation, but the different rates of decline in precision for longer N-grams: the decline is close to linear for ua>ru translation and exponential for ru>en (so the logarithm of the ru>en scores will show linear decline).

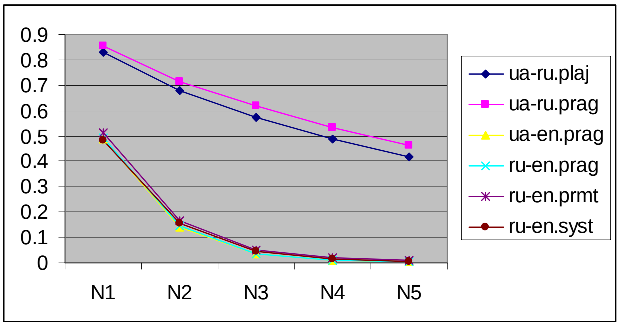

Chart 3: BLEU scores distribution for N=1 to N=5

Table 4 quantifies this intuition in terms of correlation between the size of N-grams and corresponding BLEU scores. It presents correlation figures for the scores in Chart 3. (When the correlation is close to –1 the relation is linear).

It can be seen from the table that with larger N-grams the decline in BLEU scores for closely related **ua>ru** MT is almost linear, but for distant **ru>en** MT the decline is exponential (the decline of logarithms of the scores is linear). The linguistic interpretation of this fact is that MT between closely related languages is radically different from MT between distant languages: here it takes advantage of structural similarity between Ukrainian and Russian and often successfully follows source structural patterns without change. It can also carry much of structural and lexical ambiguity into the target without the need for disambiguation, hence longer N-gram sequences are subject to much smaller variation in the output text relative to the reference translation. Because of this, closely related MT can rely on shallow processing techniques rather than keeping track of the entire sentence structure.

***r* corr with**	BLEU	log(BLEU)
ua-ru.plaj	**-0.9894**	-0.9990
ua-ru.prag	**-0.9906**	-0.9989
ua-en.prag	-0.8479	**-0.9985**
ru-en.prag	-0.8512	**-0.9979**
ru-en.prmt	-0.8611	**-0.9969**
ru-en.syst	-0.8640	**-0.9987**

Table 4: *r* correlation between N-gram size and BLEU

It can be seen from the table that with larger N-grams the decline in BLEU scores for closely related **ua>ru** MT is almost linear, but for distant **ru>en** MT the decline is exponential (the decline of logarithms of the scores is linear). Linguistic interpretation of this fact is that MT between closely related languages is radically different from MT between distant languages: it takes advantage of structural similarity between Ukrainian and Russian and often successfully follows source structural patterns without change, and can also carry much of structural and lexical ambiguities into the target without the need for disambiguation, so longer N-gram sequences are subject to much smaller variation in target and reference. Because of this, closely related MT can rely on shallow processing techniques rather than keep track of the entire sentences structure.

This suggests that pivot MT via well-resourced closely related language has the potential for achieving high translation quality, provided the quality of the closely related stage is sufficiently good.

Direct MT and pivot MT

This section presents evaluation scores which compare direct MT with different pivot MT pipelines for translation from Ukrainian and Russian into English. The baseline in the experiment is *traditional* pivot MT – from a well-resourced language (Russian) into distant languages (French and German), and then into another distant language (English) using the best MT systems available. This baseline experiment confirms our expectations that there is a substantial loss of quality in such a pivot pipeline (since errors on each stage naturally tend to accumulate rather than recover each other). As a result, the scores for the target text are consistently lower than for the direct translation using the best available direct MT system from Russian into English.

Our experiment tested whether the same unacceptable decline in quality happens for MT from a relatively under-resourced language (Ukrainian) via a closely related pivot (Russian), especially when direct MT quality from Ukrainian into English is not as good as from Russian, the better resourced pivot language. Our question is whether the quality of direct translation is always superior and the quality decline in pivot pipelines is consequently unavoidable. Alternatively, can we make this decline negligible and achieve very close evaluation scores for under-resourced Ukrainian and well-resourced Russian languages, and beat the scores for the direct MT route?

Table 5 summarises corpus-level BLEU and WNM evaluation scores (BLEU scores are shaded) for the direct and pivot routes for the baseline and test scenarios.

Firstly, the data in the table confirm that there is a substantial decline in MT evaluation scores for our baseline – pivot MT from Russian via distant languages. Distant pivot scores are consistently lower than scores for the best available direct MT system (ProMT), most noticeably for BLEU, by 32%-40% .In fact, pivot BLEU scores are lower than the scores for *any* direct MT (c.f. Table 3), which suggests that the structural level is the worst affected during distant pivot translation.

However, a completely different picture can be observed in our scenario testing pivot MT via a closely related language: the scores for the pivot pipelines are consistently better than for the direct Ukrainian–English Pragma MT: the decline in quality in pivot translation is small enough to remain ahead of the direct system. Only in one of the pipelines are WNM scores slightly lower for pivot translation, for the three other routes they are higher. BLEU scores are always higher (by 18% to 37%).

Baseline pivot (between distant languages)			
Stage 1: **ru>fr/de** Stage 2: **fr/de>en**	ProMT ru>de (% diff. with direct)	ProMT ru>fr (% diff. with direct)	Best direct translation **ru>en (ProMT)**
Systran De/fr > ru	**0.0345** (–40%) **0.1961** (–4.5%)	**0.0392** (–32%) **0.1980** (–3.5%)	0.0574 **0.2053**
Pivot from via a closely related language			
Stage 1: **ua>ru** Stage 2: **ru>en**	Plaj-Ruta ua>ru	Pragma ua>ru	Direct translation **ua>en (Pragma)**
ProMT Ru>en	**0.0498** (+29%) **0.2040** (+12%)	**0.0532** (+37%) **0.2024** (+11%)	0.0387 0.1827
Systran Ru>en	**0.0458** (+18%) **0.1881** (+3%)	**0.0472** (+22%) **0.1785** (–2%)	

Table 5: Automated evaluation scores for pivot pipelines

The following example illustrates better translation via the Russian pivot as compared to direct translation from Ukrainian:

(1) **Source:** Для розв'язання кризи сторони конфлікту звертаються до глави держави.
(For solving the crisis the conflicting parties seek the mediation of the head of the state.)

(2) **Direct:** *For permission of crisis of side of conflict address country's leader.*

(3) **Pivot:** *Для решения кризиса стороны конфликта обращаются к главе государства.*

(4) **From pivot:** *For solving the crisis the sides of conflict are turned to the Head of The State.*

Even though translation errors on each stage of pivot translation accumulate, their effects are substantially smaller during the stages between closely related languages. Overall translation quality crucially depends on the quality of the weakest link in the pipeline, which is usually the stage between more distant languages. Since many lexical ambiguities are shared between related source and pivot languages, an MT system translating into a related pivot does not have to make a decision about solving them. For instance, *розв'язання* in example (1) is ambiguous between two readings: `permission' and `solving'. Similarly *сторони* can be either genitive singular or nominative/accusative plural. The Ukrainian-English module of Pragma made both translations wrong (2), while its Ukrainian-Russian module left the ambiguities intact (3). Later they have been successfully resolved by a more highly developed MT system (SYSTRAN in this case). At the same time, translation via a pivot can lead to less fluent translations. For instance, the original expression *звертаються* in (1) has been rendered in direct translation by *to address*, which is a better solution than *are turned to* (4).

This source example also shows that the task of MT translation between closely related languages is relatively simple, as many sentences can be mapped to each other keeping the same word order and POS tags:

(5)	*Для*	*розв'язання*	*кризи*	*сторони*
	Для	решения	кризиса	стороны
	For	solution$_{gen}$	crisis$_{gen}$	sides$_{nom}$
	конфлікту	*звертаються*	*до*	*глави*
	конфликта	обращаются	к	главе
	conflict$_{gen}$	ask$_{3p,pl,refl}$	to	head$_{dat}$
	держави			
	государства			
	state$_{gen}$			

An even more surprising finding is that some closely related pivot pipelines from Ukrainian get higher scores than some of the direct systems translating from Russian. This can be inferred from Table 6. This table compares the scores for the ua>ru>en pivot pipelines and for the corresponding direct ru>en systems, which are used in those pipelines on the final stages.

Pivot MT ua>ru>en	**BLEU r1n5**	**Diff w. stage 2**	**Stage2 ru>en**	**Stage2 BLEU**
plaj-prmt	0.0498	(–13%)	prmt	0.0574
prag-prmt	**0.0532**	(–7%)		
plaj-syst	0.0458	(–10%)	syst	**0.0511**
prag-syst	0.0472	(–8%)		
not used in pivot			prag	0.0429
direct: **ua>en**			prag	0.0387

Table 6: Scores for ua>ru>en pivot MT and ru>en stage

The differences between the pivot MT scores and the scores for stage 2 (i.e., for MT from human translation from Russian) can be interpreted as loss of quality in the pivot pipeline on the first stage. The smallest loss amounts to –7% of the corresponding direct MT BLEU score, but even the biggest loss (–13%) does not make the translation unusable, since the score is still higher than the direct ua>en and even ru>en MT by the Pragma system.

In addition, note that the pivot translation from Ukrainian that uses ua>ru Pragma and ru>en ProMT outperforms direct translation from Russian done by Systran (by +4.1% of Systran's BLEU score – the highlighted figures in Table 6), which shows that pivot MT from under-resourced languages can achieve industrial-level translation quality if the pivot translation is via a well-resourced closely related language.

Interestingly, the Pragma system, which shows the worst results for distant translation, is the best candidate for the closely related stage of the pivot MT, yielding best-performing pivot pipelines (in conjunction with good ru>en MT systems).

Text-level evaluation

There is one further dimension for comparing MT quality in the direct and pivot routes between distant languages. It is based on BLEU scores for different segments in the corpus (in our case – for different texts). It can be noted that texts with higher scores for the direct route also receive higher scores for the pivot route and vice versa (so BLEU scores for the same texts translated via different routes correlate highly with each other).

This can be explained by the fact that some texts are objectively more difficult for MT than others, and thus the score is much more dependent on this objective difficulty of a text than on particular system or the route taken for translation.

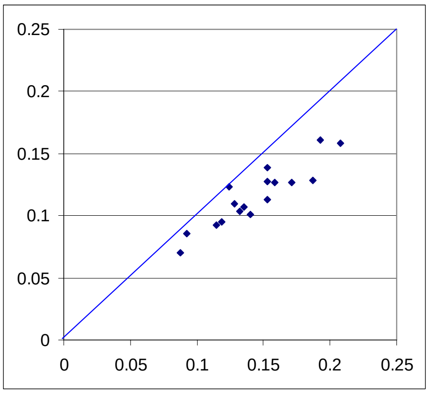

Chart 4: Baseline pivot (distant) ru>de>en.prmt-syst (Y) *vs* direct ru>en.syst (X)

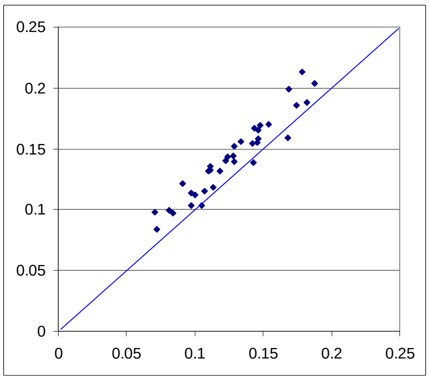

Chart 5: Pivot (closely related) ua>ru>en.plaj-prmt (Y) *vs* direct ua>en.prag (X)

A segment-level perspective on the BLEU scores for the corpus is shown on Chart 4 and Chart 5, where on the X axis we show the range of scores for individual texts translated via the direct route, and on the Y axis the range of scores translated via the pivot route. The diagonal on the charts is a guideline for comparing the scores thus represented at the segment level: if there are more points above the diagonal, Y is higher than X and the pivot route is better, and vice versa.

Note that the points are positioned along a straight line, so X and Y correlate highly with each other (illustrating our *text difficulty* concept). Table 7 shows correlation figures and also figures for the regression parameters of the line – the slope and intercept for the best-fit line through these points.

	***r* corr.**	**slope (DS)**	**intercept**
Baseline pivot: ru>fr/de>en : **correl with ru-en.syst**			
ru-de-en.prmt-syst	0.9123	**0.6683**	0.0196
ru-fr-en.prmt-syst	0.8851	**0.7606**	0.0206
Closely related pivot: ua>ru>en: **corr. with ua-en.prag**			
ua-ru-en.plaj-prmt	0.9567	**0.9820**	0.0175
ua-ru-en.plaj-syst	0.9459	**0.9360**	-0.0003
ua-ru-en.prag-prmt	0.9478	**0.9144**	-0.0085
ua-ru-en.prag-syst	0.9485	**0.9253**	-0.0031
Stage 1 of pivot: ua> ru: **corr. with ua-en.prag**			
ua-ru.plaj	0.2718	0.2614	-0.0513
ua-ru.prag	0.2637	0.2727	-0.0685

Table 7. Texts: correlation and regression parameters

We suggest that this high correlation can be used to compute another quality parameter which compares MT systems by their ability to handle a range of texts of varying difficulty. The intuition is that systems can be compared by the rate of increase in BLEU scores for easier texts and (equivalently) the rate of decline in the scores for harder texts.

The parameter which measures these rates of BLEU score changes is the slope of the line (the *slope* column in Table 7): if it is 1, then the decline of the scores for difficult texts (and the increase of the scores for easier texts) for both systems are approximately the same. If it is less then 1, then scores for pivot MT fall quicker on harder texts (and rise slower on easier texts) than the corresponding scores for the direct route, which could be interpreted to mean that noise from the additional pivot stage makes a considerable contribution to the decline in quality.

We will refer to the slope of the fitted line as the *Difficulty Slope (DS)*, which can be viewed as another quality parameter for MT, which can be used for comparing MT systems in terms of the relation between their BLEU scores for segments of varying difficulty. Note that this parameter is useful only when the scores for the same segments correlate: if there is no such agreement, the DS parameter is not meaningful. This is the case for the ua>ru and the ua>en translations (see the last two lines in Table 7). These scores cannot be compared by the DS, since here systems disagree on what is difficult and what is easy for translation.

In this way DS scores allow us to assess the impact of an additional pivot stage in terms of relative differences in BLEU scores rather than their absolute values, which is plausibly a more natural measure for this parameter. The smaller the number, the bigger is the quality degradation introduced at the pivot stage as compared to the direct translation route. (There is a theoretical possibility of having DS scores greater than 1, but this would mean that the pivot removes errors rather than introduces them.)

Note that according to the DS parameter the ua-ru-en.plaj-prmt pivot pipeline is the best (DS=0.9820), even though it received the second best BLEU score (0.0498) after the ua-ru-en.prag-prmt pipeline (BLEU=0.0532, DS=0.9144) This pipeline copes much better with variations in the difficulty of segments, almost as well as the direct ua>en translation.

The results of the evaluation experiment indicate that it may be more rewarding to invest development effort in good MT from under-resourced or less commercially viable languages into closely related languages, for which state-of-the-art commercial MT systems are available, and to use them in pivot pipelines, rather than to develop direct MT for under-resourced languages from scratch.

4. Discussion

The full potential of the pivot architecture lies not just in enabling new translation directions for MT, but also in paving the way to higher MT quality for under-resourced languages, than can be realistically achievable through the development of direct MT for these languages.

Usually academic or industrial development teams who work on new translation directions cannot spend several years developing high-quality in-house MT for distant languages in order to achieve a performance equal to state-of-the-art MT systems, such as Systran. On the other hand, established teams are often busy with improving MT for more commercially viable directions, and do not work on new directions especially in the case of non-commercial or under-resourced languages. Therefore, all developers will benefit from a clear methodology for testing the performance of commercial MT systems used in the pivot framework on large corpora. As a result, this methodology would enable the reuse of MT development effort for families of related languages, and a concentration on the easier and much more rewarding tasks of developing MT within closely related groups.

Our results suggest that translation between closely related languages pipelined together with advanced commercial MT systems for distant languages will yield better results and require less development effort than direct MT systems developed from scratch. Such a pivot MT architecture will enable potentially higher translation quality for a greater number of language pairs, including some under-resourced languages, for which no commercial MT is currently available.

5. Conclusions and future research

The results of using a closely related pivot language for the purposes of information assimilation from under-resourced languages are promising. Even though translation errors at each stage of pivot translation accumulate, their effects are substantially smaller at the stages between closely related languages. What is more, translation via the pivot can in principle utilise any more advanced bilingual dictionaries and grammars available for translation out of the pivot. However, the experiment reported here used two existing systems for translating from Ukrainian into Russian. Where few language resources are available, this frequently means that MT systems from such languages to a pivot are also not available. Parallel and comparable corpora, bilingual lexicons, part-of-speech taggers and lemmatisers might be limited or not available either. The next task in this research is to estimate the resources needed to develop an MT system for translation into the related pivot.

Existing research with Czech and Slovak (Hajic et al., 2000a) shows that simple transfer systems operating at the word level can produce reasonable results for closely related languages. The induction of transfer rules for closely related languages can be achieved using comparable corpora: bootstrapping from a small initial bilingual lexicon or the set of orthographic cognates, the system can identify words of the two languages that occur in contexts with a large number of words that are known mutual translations from the seed lexicon. As shown in (Rapp, 1999) this automatic procedure can produce a reliable bilingual lexicon without resorting to parallel corpora. This procedure relies on the availability of sufficiently large comparable corpora (of the size of 20-100 million words). The feasibility of semi-automatic acquisition of such corpora has already been demonstrated (Sharoff, 2006).

References

Alonso, Juan Alberto. 2005. Machine translation for Catalan <->Spanish: the real case for productive MT. Paper presented at 10th EAMT conference "Practical applications of machine translation", Budapest.

Armentano-Oller, Carme, Corbi-Bellot, Antonio M., Forcada, Mikel L., Ginesti-Rosell, Mireia, Bonev, Boyan, Ortiz-Rojas, Sergio, Perez-Ortiz, Juan Antonio, Ramirez-Sanchez, Gema, and Sanchez-Martinez, Felipe. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. Paper presented at OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X, Phuket, Thailand.

Babych, Bogdan, Elliott, Debbie, and Hartley, Anthony. 2004. Extending MT evaluation tools with translation complexity metrics. Paper presented at COLING 2004. Proceedings of the 20th International Conference on Computational Linguistics, University of Geneva, Switzerland.

Babych, Bogdan, and Hartley, Anthony. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. Paper presented at ACL-04: the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Babych, Bogdan, Hartley, Anthony, and Elliott, Debbie. 2005. Estimating the predictive power of n-gram MT evaluation metrics across language and text types. Paper presented at MT Summit X, Phuket, Thailand.

Babych, Bogdan, Sharoff, Serge, Hartley, Anthony, and Mudraya, Olga. 2007. Assisting Translators in Indirect Lexical Transfer. Paper presented at ACL 2007, Prague, Czech Republic. (In press).

Dvorak, Bostan, Homola, Petr, and Kubon, Vladislav. 2006. Exploiting similarity in the MT into a minority language. Paper presented at LREC-2006. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages, Genoa, Italy.

Gryaznukhina, Tetyana. 2004. The Structure and Functions of the Automation Multilingual Translation Dictionary. Paper presented at Dialog-2004. Computational linguistics and intelligent technologies, Verkhnevolzhskiy, Russia.

Hajic, Jan, Hric, Jan, and Kubon, Vladislav. 2000a. Machine Translation of Very Close Languages. Paper presented at ANLP, Seattle, Washington.

Hajic, Jan, Kubon, Vladislav, and Hric, Jano. 2000b. CESILKO - an MT system for closely related languages. Paper presented at ACL2000, Tutorial Abstracts and Demonstration Notes.

Hutchins, W.John. 1995. Machine Translation: A Brief History. In Concise history of the language sciences: from the Sumerians to the cognitivists, eds. E.F.K.Koerner and R.E.Asher, 431-445. Oxford: Pergamon Press, 1995.

Koehn, Philipp. 2005. Europarl: a parallel corpus for statistical machine translation. Paper presented at MT Summit X, Phuket, Thailand.

Marcu, Daniel, and Wong, William. 2002. A phrase-based, joint probability model for statistical machine translation. Paper presented at ACL-02 conference on Empirical methods in natural language processing.

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

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, WeiJing. 2002. Bleu: a method for automatic evaluation of machine translation. Paper presented at ACL-02: the 40th Annual Meeting of the Association for Computational Linguistics.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. Paper presented at 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland.

Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. In WaCky! Working papers on the Web as Corpus, eds. Marco Baroni and Silvia Bernardini. Bologna: Gedit.

White, J., O'Connell, T., and O'Mara, F. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. Paper presented at 1st Conference of the Association for Machine Translation in the Americas., Columbia, MD.