Automated error analysis for multiword expressions: Using BLEU-type scores for automatic discovery of potential translation errors

Bogdan Babych
University of Leeds

Anthony Hartley
University of Leeds

We describe the results of a research project aimed at automatic detection of MT errors using state-of-the-art MT evaluation metrics, such as BLEU. Currently, these automated metrics give only a general indication of translation quality at the corpus level and cannot be used directly for identifying gaps in the coverage of MT systems. Our methodology uses automatic detection of frequent multiword expressions (MWEs) in sentence-aligned parallel corpora and computes an automated evaluation score for concordances generated for such MWEs which indicates whether a particular expression is systematically mistranslated in the corpus. The method can be applied both to source and target MWEs to indicate, respectively, whether MT can successfully deal with source expressions, or whether certain frequent target expressions can be successfully generated. The results can be useful for systematically checking the coverage of MT systems in order to speed up the development cycle of rule-based MT. This approach can also enhance current techniques for finding translation equivalents by distributional similarity and for automatically identifying features of MT-tractable language.

1. Introduction

Automated MT evaluation methods such as BLEU, NIST and Meteor have been shown to be useful for monitoring progress in MT development, for parameter optimisation of statistical systems and, in some controlled circumstances, for comparing the performance of different MT systems. All such MT evaluation experiments rely on a corpus of human translations which are used as a reference for the MT output. Automated evaluation scores correlate with human scores and correctly establish the ranking of systems only if this corpus is relatively large, i.e. more than 6,000-7,000 words (Estrella et al., 2007; Babych et al., 2007b). Smaller samples of data are too noisy for reliably predicting a system's performance, since individual lexical mismatches between MT output and human reference are not informative on their own: they may be attributable either to errors of

translation or to choices of different legitimate translation variants. While human judgements are meaningful at any granularity for which they are generated (the levels of syntactic constituent, sentence, paragraph, text and corpus as a whole), automated scores are generally not meaningful at any level below that of the corpus. As a result, automated evaluation scores are currently uninformative for error analysis tasks—specifically, for discovering typical translation errors and prioritising them for the purposes of MT development—since they give only a very general, 'birds-eye' view of MT performance.

Moreover, MT developers are often less interested in such non-specific performance figures than in a more detailed analysis and ranking of typical problems for their MT system whose resolution will improve the system's performance generally. As a result, developers of industry-standard (especially rule-based) systems consider these core automated evaluation metrics to be of little help in the MT development cycle (Thurmair, 2007), noting that they are not designed to provide direction to R&D (Miller & Vanni, 2005). Although human evaluation scores can be much more useful in this respect, they are expensive to obtain and are not available for significantly large corpora. Thus it is not feasible to rely on them for determining the range, frequency and seriousness of errors and, especially, for monitoring the progress of an MT system over time.

From this perspective, the challenge for automatic MT evaluation research is to develop methodology suitable for differentiated and fine-grained error analysis along the lexical, grammatical and stylistic dimensions. Our paper reports on a project for automatically discovering and ranking errors in translating multiword expressions (MWEs). We use the term in the sense of 'phraseological units' proposed by Vinogradov, as discussed in Cowie (1998). MWEs are defined as *repeated (continuous or discontinuous) combinations of words which are* ***re-constructed (****rather than* ***constructed)*** *in speech and are part of the 'mental lexicon'*. This definition includes both 'compositional MWEs' (e.g. *washing machine*) and 'non-compositional MWEs', or idioms in the broad sense (e.g. *meet the demand*, etc.).

While at this stage our methodology targets only the lexical dimension, we argue that it is a useful step towards more informative MT evaluation for developers and users of state-of-the-art MT systems.

2. Methodology

Our method is based on automatic evaluation of the translation of concordances for frequent MWEs extracted from aligned corpora. The methodology follows the following five stages.

Firstly, we generate automatically frequency-ranked lists of MWEs, using the approach described by Babych et al. (2007a), which relies on a

combination of part-of-speech and frequency filters. The idea behind this approach is to collect all possible multiword candidates found inside a sliding window of a certain length (usually up to five words) and to compute the frequency of every candidate. Larger windows can also be used, but these result in smaller sets of MWEs passing the frequency threshold, since N-gram frequency quickly drops with longer N-grams.

Candidates which are above a specified frequency threshold and conform to certain part-of-speech patterns are typically found to be meaningful MWEs. Part-of-speech patterns can be specified either as a list of permitted configurations, or as a set of restrictions on them.

We modified this approach in order not to depend on morphological annotation, thereby making it knowledge-light and language-independent. The idea came from an observation that part-of-speech filters typically prevent the appearance of function words at one or both edges of MWEs. For example, the sequences *visual processing to, visual processing in, visual processing and* are filtered out, leaving only *visual processing* as a candidate MWE, which is selected if it passes a certain frequency threshold in a corpus. But rather than using a filter that relies on prior knowledge of parts-of-speech, we filter instead by log IDF scores, which distinguish content words from function words:

$$logIDF_i = \log\left(\frac{N}{df_i}\right)$$

where N is the number of texts in the corpus and df_i is the number of texts in which $word_i$ is found. The value of *logIDF* which best distinguishes content words from function words must be established experimentally for each corpus. It depends on the size of the texts and the total number of documents in the collection. For a corpus which contains 100 texts, each of about 350 words, the threshold $logIDF > 1$ yields a relatively good distinction between content and function words.

Function words can be included inside candidate continuous MWEs (a productive pattern in Romance languages especially, e.g. *Fr: discrimination fondée* ***sur la*** *race – 'racial discrimination'*), but normally do not appear at the edges.

Thus, in our experiment we used a simple frequency filter and a statistical differentiation between content and function words for extracting MWEs. Other researchers have used different word association measures: mutual information, Dice's coefficient, t-score, chi-square and log likelihood (Baldwin, 2006). However, according to Evert and Krenn (2001), simple frequency can be as good as a wide range of such association measures for this task.

For continuous MWEs, a lower frequency filter can also yield good results, e.g. $Freq(MWE) > 1$ (Sharoff et al., 2006). However, since our methodology uses MWEs to generate concordances for which BLEU scores will be computed, a higher threshold was chosen in order to enhance the reliability of the automated scores by using a larger concordance sample.

Other methods of identifying MWEs may use linguistic annotation (e.g. part-of-speech tags) and apply different settings to the selection parameters, which will yield different types of expressions: discontinuous MWEs, expressions underspecified for certain lexical or morphological features, certain types of linguistic constructions, such as light verb constructions (*make decision, take into account, put pressure*) or phrasal verbs (*look after, come along*). The choice of setting is determined by the aspect of MT performance the research is intended to address.

In the second stage, and for the most frequent MWEs in a sentence-aligned parallel corpus, we automatically generate concordances that contain the MWEs themselves and several words in their local context. Thus the concordances can be viewed as sub-corpora selected by a specific MWE, intended to characterise the successfulness of their translation by MT. Moreover, concordances can be generated either for the source language (SL) or for the target language (TL): SL concordances are generated from original source texts while TL concordances are generated from human reference translations. Both are used for evaluating the quality of MT output sentences aligned with them.

In the third stage, the SL concordances are translated by the MT systems which are under evaluation. Interestingly, TL concordances can be used even if there is no access to the MT engine itself, that is, if only its MT-generated corpus is available. This is the case for some old systems which are no longer maintained, and for some in-house systems for which the developers choose not to give the evaluators direct access to their engine. In practical MT-evaluation scenarios, the users of MT systems often have no access to the working MT engine, and can use only an MT-translated corpus. Such scenarios typically occur when the evaluations of a system that is no longer maintained are intended to serve as an 'historic baseline', or when the MT system to be evaluated does not offer remote access and cannot be installed on the evaluators' local machines.

The reason that such use of the TL corpus is nonetheless possible is that the dynamic data (MWE concordances) is generated from human reference translations and not from texts translated by MT. Thus it is possible to use a 'frozen' corpus of previously translated MT output. This property of TL concordances proved useful for normalising the proposed methodology using human scores associated with the DARPA-94 MT evaluation corpus (White et al., 1994), even though the MT engines which translated the source texts are no longer available.

In the fourth stage, we compute BLEU scores (Papineni et al., 2002) based on both types of concordance. The scores for the translations of each SL concordance indicate how well a particular SL expression and its immediate context are translated, while the scores for the MT-generated versions of each TL concordance show whether a particular TL expression can be successfully generated by MT.

There are two important technical issues with using BLEU as a metric for this type of concordance-based MT evaluation. In the case of SL

concordances, since word alignment may be too noisy we take the whole sentences (or even paragraphs) aligned with the concordance segments as the reference. As a result, the reference texts may be much longer than the tested concordances. This, however, is not a problem for BLEU, which is an asymmetric, precision-based metric and which therefore characterises the ability of MT to avoid generation of redundant N-grams. With the brevity penalty switched off, BLEU is only interested in whether a test file contains any spurious items which are not found in the reference. Therefore, the reference text can be arbitrarily large.

In the case of TL concordances, the MT output may be longer: it contains complete sentences rather than the immediate context of specific MWEs. In this case, we either use a recall-oriented metric – e.g. WNM (Babych and Hartley 2004) – or, if we prefer to use a precision-oriented metric, we swap the test and the reference files such that the MT output becomes a reference.

In the final stage, we generate the evaluation results in the form of tables, where particular MWEs are ranked by BLEU or other automated scores. MT developers can use the resulting tables similarly to how they use traditional risk-analysis tables: they can focus on highly-probable (i.e. most frequent) lexical errors with the greatest impact on quality (i.e. lowest BLEU for the concordance).

3. Experiments

We extracted MWEs from two aligned parallel corpora – a section of about 700k words from the Europarl corpus (Koehn, 2005) and the French/English section of the DARPA-94 corpus (35k words). The DARPA-94 data contains two human translations of the SL texts, named 'reference' and 'expert'. Despite the DARPA-94 corpus being much smaller, it is useful for normalising the proposed evaluation method because it offers two independent professional translations of the same text and human scores for *adequacy, fluency* and *informativeness*.

Our first group of experiments characterises the performance of the state-of-the-art rule-based system Systran 6.0 in translating between English and French/German/Spanish. The second group of experiments focuses on translations between English and French produced by several MT systems, (both rule-based and statistical) and on a meta-evaluation of the proposed methodology.

3.1. Extracting MWEs

From both corpora we extracted continuous MWEs with a high *logIDF* threshold, which produced lists of terminological or near-terminological expressions and proper names.

The selected part of the Europarl corpus was divided into 20 sections, each containing up to 1,500 segments, corresponding to approximately half a day of a plenary session of the Parliament. The sessions took place on different days between February 2000 and July 2001, and so the extracted terms and named entities reflect the topics discussed over this period. Since the sections were relatively large (up to 50k words), and the number of sections (treated as 'documents' in the collection) was small, we set a threshold of *logIDF>0.4* and applied a frequency filter of *Freq>4*. For the DARPA-94 corpus, which includes 100 relatively short news stories, we set *logIDF>1.0* and kept the same frequency threshold of *Freq>4*. Since MWEs were extracted only from original SL texts and from human reference translations, not from MT output, the set of MWEs was the same for all evaluated MT systems. Table 1 shows the number of MWEs extracted for each translation direction under these settings.

Table 1: Size of corpora and number of extracted MWEs

Corpora		French		*German*		*Spanish*	
		en>fr	fr>en	en>de	de>en	en>es	es>en
Europarl section	Words (tokens)	675k	706k	670k	625k	661k	683k
	MWEs (types)	279	333	249	283	287	273
DARPA94 (en transl.: 'expert'/ 'reference'	Words (tokens)	39k	39k	-	-	-	-
	MWEs (types)	58/68	54	-	-	-	-

For the Europarl, 154 English MWEs were found in all three sets aligned with French, German and Spanish (with different frequencies), and 106 other English MWEs occurred in two of the three sets. We used these common MWEs to investigate the quality of the translation of MWEs out of English into different target languages.

The majority of discovered MWEs consist of two words, but some have up to five. Frequencies of MWEs are in the range of 42-5 for the DARPA corpus and in the range of 86-5 for the Europarl corpus, and have the usual Zipfian distribution with a steeper hyperbolic curve typical for MWEs. Figure 1 illustrates the frequency distributions of MWEs, here in the DARPA-94 corpus.

Our selection settings (relatively high *logIDF* and frequency thresholds) in the experiments described here yielded primarily named entities and terminology or near-terminological expressions, and these provide the material for illustrating our error-analysis methodology. However, as we noted earlier, the range of evaluated constructions can potentially be much wider.

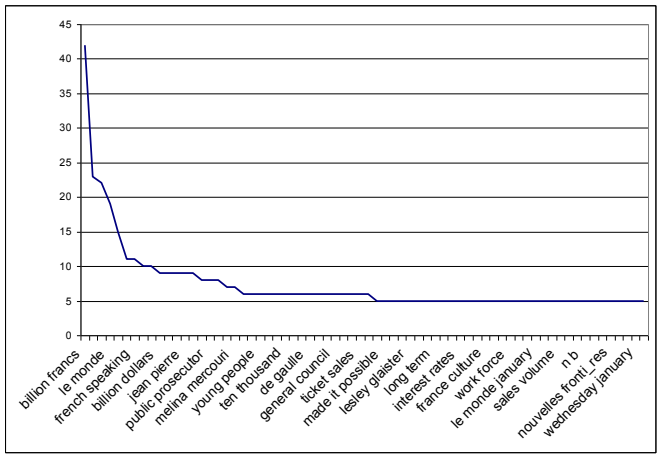

Figure 1: Frequency distribution of MWEs in DARPA-94

3.2. Generating aligned concordances, MT output and BLEU scores

For each extracted MWE, we generated aligned concordances. The concordances included the MWE itself and up to four words to the left and to the right. Each of these lines was aligned with a full segment (typically a sentence in the Europarl corpus, or a paragraph in the DARPA-94 corpus).

In our methodology, the use of both Source and Target Language concordances is designed to characterise two different aspects of MT quality. SL concordances identify problems mainly on the *analysis* side and highlight SL MWEs that are not translated properly. TL concordances identify problems on the *generation* side, listing TL MWEs that should be used in translation, but are not produced by the MT systems.

3.2.1. Source Language concordances

We use SL concordances to check the quality of MT for the immediate contexts of source MWEs. The concordances generated on the SL side are translated by MT and then aligned with the corresponding segments in the human reference translation (by their segment IDs). Table 2 illustrates these original concordances, their translation generated by the Systran 6.0 MT system ('Syst'), and the aligned human reference translations of the corresponding segments.

The rationale for our approach is that BLEU penalises disfluencies in MT output like *Minister* ***for the*** *European* ***businesses*** *(20-2), Minister* ***for the interior matters*** *(28-4), minister* ***for the*** *social affairs (35-2)*. Since these contexts are selected systematically and in a controlled way, if an SL expression is systematically mistranslated, this has a measurable effect on the BLEU score for the concordance. Despite the evaluated concordance being much smaller than the texts normally evaluated by BLEU, the scores prove meaningful in that they allow MT evaluators to prioritise errors in the

contexts of individual MWEs using the 'risk-analysis' framework, which we propose in Section 4

Table 2: Fr>En: SL concordance: French MWE *ministre des affaires*

seg	Aligned concordance	
12-2	ori.:	t il feint d'être **ministre des affaires** culturelles auprès du général
	Syst.:	it pretends to be a Minister for the cultural affairs near the general
	hum.:	[...] Malraux pretended to be minister of cultural affairs under General de Gaulle [...]
3-3	ori.:	et un représentant du **ministre des affaires** étrangères de même que
	Syst.:	and a representative of the Foreign Minister just as
	hum.:	[...] and a representative of the Ministry of Foreign Affairs," as well as with General Rahimi [...]
20-2	ori.:	théodore pangalos **ministre des affaires** européennes du gouvernement papandréou
	Syst.:	Theodore pangalos Minister for the European businesses of the government papandréou
	hum.:	[...] Theodore Pangalos, Minister of European Affairs in the Papandreou government [...]
28-4	ori.:	mathot (également du ps) **ministre des affaires** intérieures du même gouvernement
	Syst.:	mathot also of the PS Minister for the interior matters of the same government
	hum.:	[...] Guy Mathot (also a SP member), the minister of internal affairs of the same regional government.
35-2	ori.:	de vote. simone veil **ministre des affaires** sociales, de la santé
	Syst.:	of vote. Simone Veil, Minister for the social affairs, of health
	hum.:	[...] the right to vote. Simone Veil, Minister of Social Affairs, Health, and Cities [...]

As noted previously, in the case of SL concordances, the human reference segments are longer than their corresponding concordance segments and MT-generated translations (the table shows only part of these segments, which are one paragraph long and typically contain several sentences). To account for this, we switch off the brevity penalty when computing BLEU scores for each concordance. With these settings, the scores become meaningful thanks to the asymmetric nature of BLEU, which calculates only the *precision* of the N-gram matches, such that the scores are affected only by spurious items in MT output but not by missing items.

For the example in Table 2, the raw BLEU precision score (without brevity penalty) is 0.2563; the brevity penalty value is 0.0010 (an unusually low value for the standard text-level evaluation) and the final BLEUr1n4 score (the score with a single reference and N-gram size = 4, which takes into account the brevity penalty) is 0.0003. In our experiments we use the raw BLEU *PrecScore* alone as the only meaningful score under these settings.

Since the SL concordance lines are short and do not form complete sentences, the MT output may not be exactly the same for a particular concordance line as for the whole sentence from which it was extracted. However, MT systems usually take into account only local context of words and expressions, and normally the output is close to the sentence-level MT. It is not possible to use full sentences instead of concordances on the source side, because there will also be full sentences on the target side, and therefore BLEU scores will be influenced by errors in other parts of the sentence and will not characterise the quality of translation of particular individual MWEs.

3.2.2. Target Language concordances

We use TL concordances to check whether particular TL MWEs and their immediate contexts are accurately generated by the evaluated MT systems. The concordances generated on the TL side are aligned with the segments in the MT output produced by different MT systems. Table 3 illustrates the aligned concordance for the target English MWE *once again*, aligned with MT output from Systran 6.0 RBMT system ('Syst.') and the Google on-line SMT system ('gSMT'). The French original is given for explanatory purposes only and plays no part in the evaluation.

The rationale for evaluating TL concordances is that MT should be able to generate idiomatic TL expressions used by human translators, even if they come from a variety of different contexts in the source language. As can be seen from Table 3, for Systran and for Google SMT, the English MWE *once again* is only generated if it comes from the French source *une fois encore*, but not from the expressions *à nouveau, de nouveau,* nor from the lexical sources of this meaning like *redevenir ('to become once again'), revenir ('to come back').*

The table also shows that while Systran usually preserves a trace of all SL lexical items, the SMT system sometimes drops 'awkward' expressions which do not fit the target fluency model (segments 22-7, 73-5). In the standard BLEU evaluation scenario, only the brevity penalty accounts for these omissions, and at the text level they can pass practically unnoticed. However, our approach of using TL concordances reveals and penalises such omissions. To account for the fact that, in the case of TL concordances, the MT output is longer than the human reference, we again compute BLEU without the brevity penalty. In addition, we submit the TL concordances (i.e. the human reference translations) as *test* files and the

aligned MT output segments as *reference* files. This may seem counter-intuitive (usually the MT output is the *test),* but it is done because BLEU, as a precision-based metric, basically counts how many N-grams from the *test* file are not in the *reference* and penalises these omissions. Since in our experiments we want to know whether TL expressions like *once again* have been omitted or mistranslated by MT, these TL expressions need to be in the *test* file when they are processed by the BLEU script.

Table 3: Fr>En: TL concordance: English MWE *once again*

seg	Aligned concordance	
22-7	hum.:	united states hopes to **once again** dominate the communications satellite
	Syst.:	Thanks to this experimental apparatus of 363 million dollars, the United States hopes to *again* dominate the market of the communications satellites [...]
	gSMT:	With this experimental device of 363 million dollars, the U.S. hopes to dominate the market for communications satellites [...]
	fr.ori.:	Grâce à cet appareil expérimental de 363 millions de dollars, les Etats-Unis espèrent dominer *à nouveau* le marché des satellites de communication. [...]
73-5	hum.:	hostile posture and become **once again** that affable champion the
	Syst.:	[...] Johann Koss could get rid of its quarrelsome airs. And to become *again* this gracious champion, [...]
	gSMT:	[...] Johann Koss could get rid of its air war. And become the champion affable, [...]
	fr.ori.:	[...] Johann Koss pouvait se débarrasser de ses airs belliqueux. Et redevenir ce champion affable, [...]
81-3	hum.:	also by declining prices. once again gains were realized by
	Syst.:	[...] but also by a fall of the prices. The profits *once again* came from the branch health [...]
	gSMT:	[...] but also by lower prices. Gains *once again* came from the health branch [...]
	fr.ori.:	[...] mais aussi par une baisse des prix. Les gains *une fois encore* sont venus de la branche santé [...]

Table 3: Fr>En: TL concordance: English MWE *once again* (continued)

81-4	hum.:	its available self-financing. **once again** stable in 1992, it now
	Syst.:	[...] the group which chairs Jean-Rene Fourtou improved its self-financing available. *Returned* with balance in 1992, it is from now on surplus of 2,15 billion. [...]
	gSMT:	[...] the group chaired by Jean-Rene Fourtou has improved its cash available. *Income* balance in 1992, it is now surplus of 2.15 billion. [...]
	fr.ori.:	[...] le groupe que préside Jean-René Fourtou a amélioré son autofinancement disponible. Revenu à l'équilibre en 1992, il est désormais excédentaire de 2,15 milliards. [...]
83-2	hum.:	the rail line, connecting **once again** with the casa dei
	Syst.:	He was always among those which, twenty-two years later, on the same way, inaugurated at the summer 1993 the rebirth of the way, rejoining *again* put it dei Puy-de-Dôme,
	gSMT:	He was always those who, twenty-two years later on the same route, inaugurated in the summer of 1993 the revival of the road, rallying *again* casa dei du Puy de Dome[...]
	fr.ori.:	Il était toujours de ceux qui, vingt-deux ans plus tard, sur le même trajet, inaugurèrent à l'été 1993 la renaissance de la voie, ralliant *de nouveau* la casa dei du Puy-de-Dôme [...]

Furthermore, TL concordances use MT output generated from complete sentences and texts (not just from very short concordance lines), so the result is not influenced by missing or inadequate contexts (a potential problem for the evaluation of SL concordances that we acknowledged above).

Finally, in the case of SL concordances mistranslated MWEs can usually be corrected by extending dictionary coverage, e.g. adding entries for *ministre des affaires culturelles, ministre des affaires étrangères, ministre des affaires européennes, ministre des affaires intérieures*. In contrast, the evaluation of TL concordances usually reveals more subtle translation problems, which may not be easy to rectify directly, e.g. that of generating the phrase *once again* from implicit semantic components of the verbs *redevenir* and *revenir*, while restricting this to appropriate contexts only.

4. Evaluation results

In this section we describe the results of SL and TL concordance-based evaluation for different MT systems before presenting the results of normalising the automated scores using human evaluation scores. The

distribution of BLEU scores for the 260 MWEs identified in the Europarl corpus is shown Figure 2. BLEU scores are shown on the vertical axis, and ranks of MWEs on the horizontal axis. In this distribution there are fewer MWEs (39%) with scores below the average value of BLEU=0.11.

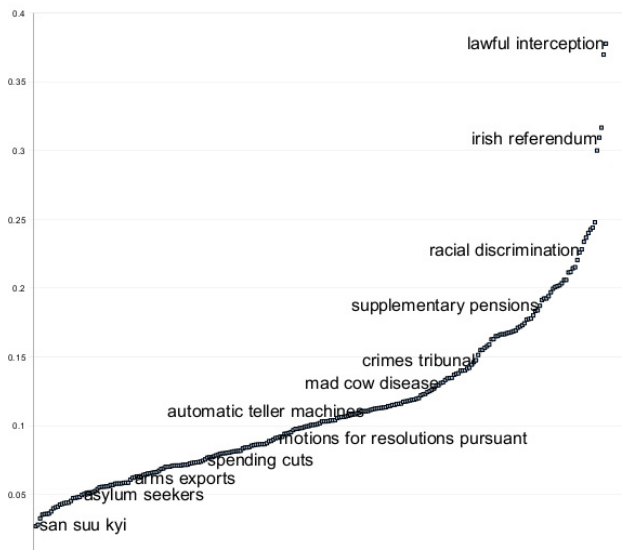

Figure 2. Distribution of BLEU scores for MWEs in the Europarl corpus

4.1. Evaluation of SL concordances of MWEs

We summarise the evaluation results for the contexts of all the identified MWEs within the framework of risk analysis. Traditionally, this framework is used to prioritise 'risks' for a particular project using two dimensions: the likelihood that some unfortunate event will occur, and the magnitude of its impact on the project. Thus the most likely events with the highest detrimental impact can be addressed first. In our approach we interpret frequencies of particular MWEs as the likelihood of events, and BLEU scores for their concordances as the magnitude of their impact. Our framework prioritises MWEs for MT developers, who can in the first instance deal with the most frequent MWEs with the lowest BLEU scores. For presentation purposes we plot *log(Frequency)* against *exp(BLEU)*, which scatters the evaluated MWEs more evenly across the risk analysis chart.

One direction for future research is developing an experimental meta-evaluation procedure for the proposed MT evaluation method, which will enable us to determine different scaling and weighting factors for the risk analysis framework.

Figure 3 shows a risk analysis plot for SL concordances of English MWEs from the DARPA-94 corpus translated by Systran 6.0 into French.

To simplify the presentation we show only selected MWEs; *exp(BLEU)* is on the vertical axis and *log(Frequency)* on the horizontal axis. Items in the bottom right quadrant are the most 'risky', since they have the highest frequency and the lowest BLEU score.

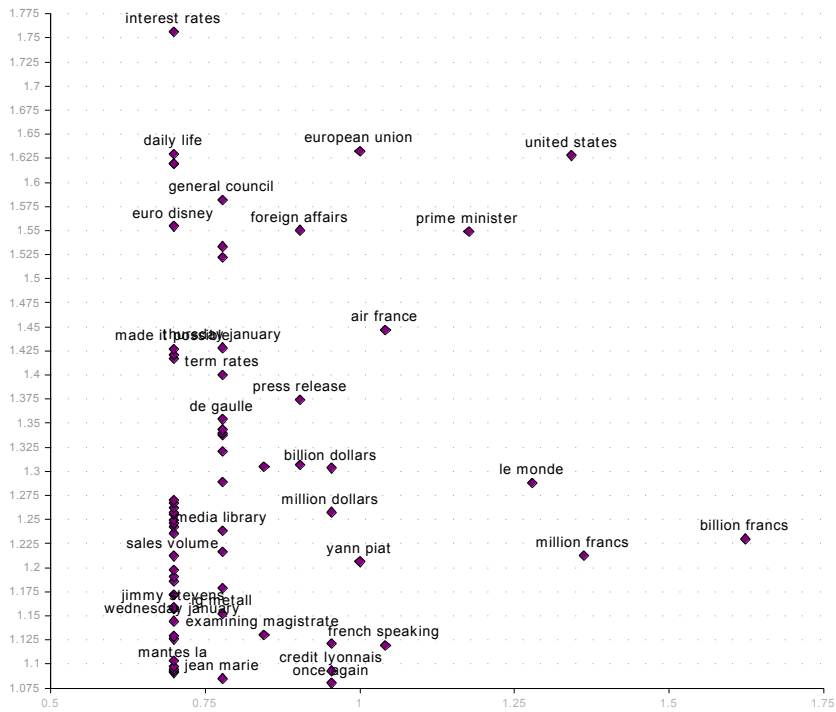

Figure 3. DARPA-94 MWE risk analysis chart: *x=log(Freq), y=exp(BLEU)*

Priority lists of MWEs can be generated by combining the two plotted parameters in different ways, e.g. *log(Freq)/exp(BLEU)=Priority* (possibly with different weights for Frequency and inverse BLEU scores). Table 4 shows the top of one such priority list.

Error analysis of these items identifies the following problems (we focus solely on the linguistically most interesting examples):

- MWE *billion francs* in the context of numerals is often translated as *milliard de francs*, while the reference contains *milliard**s** de francs*.
- MWE *french speaking* is consistently translated as *de langue française* by Systran, instead of *francophone(s)* or *francophonie*.
- MWE *once again* is always translated as *de nouveau* by Systran, while in the reference translation it is variously rendered as: *les Etats-Unis espèrent dominer **à nouveau** le marché des satellites de communication; ... ils s'étaient glissés **sous le nouveau** record du monde; Les gains **une fois encore** sont venus de la branche santé...;*

***Revenu** à l'équilibre en 1992, il est désormais excédentaire de 2,15 milliards.* This expression displays greater variation in the ways it is translated in different contexts.

Table 4: Priority list of English MWEs in the DARPA-94 corpus

MWE	*FRQ*	*log(frq)*	*exp(BLEU)*	*BLEU*	*Priority*
billion francs	42	1.62	1.23	0.21	1.32
million francs	23	1.36	1.21	0.19	1.12
le monde	19	1.28	1.29	0.25	0.99
french speaking	11	1.04	1.12	0.11	0.93
once again	9	0.95	1.08	0.08	0.88
...					

Figure 4 and Table 5 show a risk analysis chart and the top of a priority list for English MWEs from the Europarl corpus, translated by Systran 6.0 into French, German and Spanish (using the average of BLEU across all three target languages).

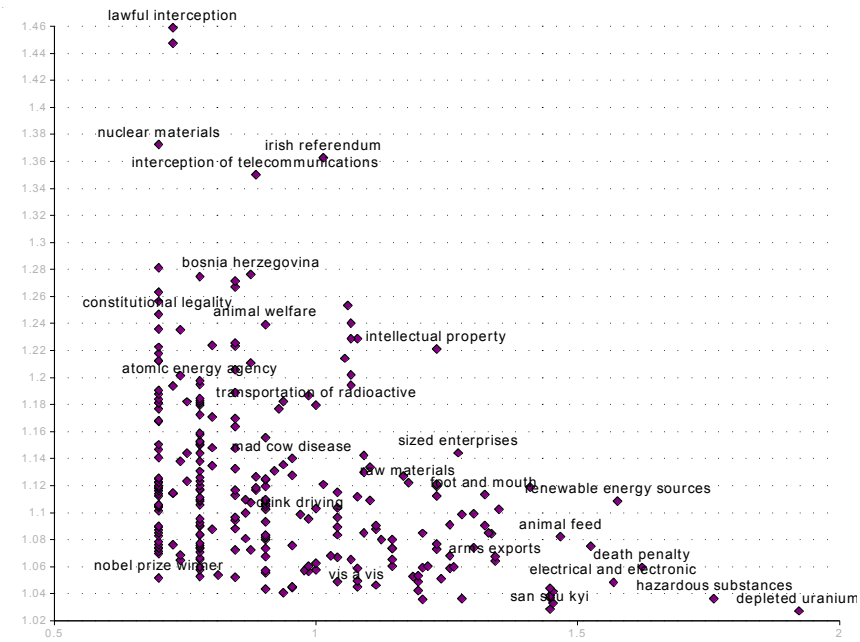

Figure 4. Europarl MWE risk analysis chart: *x=log(Freq), y=exp(BLEU)*

Table 5: Priority list of English MWEs in the Europarl corpus

MWEs	***frqAVE***	***log(FRQ)***	***exp(BLEU)***	***Priority***
depleted uranium	83.67	1.92	1.03	1.87
hazardous substances	57.5	1.76	1.04	1.7
death penalty	42	1.62	1.06	1.53
electrical and electronic	37	1.57	1.05	1.5
renewable energy sources	37.67	1.58	1.11	1.42
...				

These data identify the following problems with MWE translation:

- MWE *depleted uranium* is translated into German by Systran as *verbrauchtes Uran*, while the human reference translation uses *abgereichertes Uran* or in some contexts integrates the meaning into nominal compounds: *die Affäre um die Urangeschosse; uranhaltiger Munition*. This MWE is translated by Systran into French as *uranium épuisé,* while human translators always use *uranium appauvri.* The Spanish translation produced by Systran is always *uranio agotado,* while human translators use *uranio empobrecido.*
- MWE *death penalty* is translated by Systran into French as *pénalité de mort*, while human translators always use *peine de mort*.

4.2. Evaluation of TL concordances MWEs

To evaluate the TL concordances, we used four MT systems and the human 'expert' translation from the DARPA-94 MT evaluation corpus. For all five, we computed BLEU scores for each of our 68 concordances, using the (single) 'reference' translation and N-gram size up to 4. Table 6 presents the scores for some interesting MWEs for each MT system and for the 'expert' translation. The MWEs are sorted by the BLEU score for Systran. The headings in the table show the names of evaluated MT systems in DARPA-94 corpus: Human Expert translation, Candide SMT system, and Globalink, Metal, Reverso, Systran RBMT systems.

For MT output, low scores for the concordance of an MWE mean that it is not generated properly by the particular MT system. So we suggest that the highlighted MWEs are problematic for Systran and require the developers' attention. The threshold is set at the system's average BLEU score of 2.7, which also coincides with a jump in the series of values.

Table 6: BLEU scores for MWEs

	Hum (exp)	cand	glbl	ms	rev	syst
credit lyonnais	0.33	0.16	0.16	0.10	0.12	0.10
work force	0.37	0.35	0.10	0.10	0.12	0.11
ticket sales	0.26	0.24	0.09	0.11	0.2	0.11
once again	0.12	0.09	0.09	0.15	0.09	0.11
french speaking	0.48	0.11	0.15	0.23	0.26	0.12
sales volume	0.18	0.13	0.10	0.11	0.11	0.12
public prosecutor	0.21	0.17	0.16	0.12	0.30	0.18
take place	0.32	0.17	0.14	0.15	0.34	0.18
term rates	0.37	0.25	0.12	0.2	0.35	0.19
press release	0.23	0.22	0.19	0.15	0.17	0.19
daily life	0.39	0.17	0.23	0.17	0.45	0.20
so-called	0.38	0.20	0.15	0.19	0.16	0.21
young people	0.32	0.10	0.10	0.18	0.16	0.28
managing director	0.42	0.22	0.19	0.42	0.21	0.31
minister of foreign affairs	0.63	0.59	0.29	0.54	0.18	0.33
examining magistrate	0.36	0.13	0.14	0.29	0.25	0.34
media library	0.50	0.17	0.11	0.16	0.32	0.34
other hand	0.37	0.16	0.66	0.46	0.63	0.39
prime minister	0.54	0.33	0.44	0.24	0.44	0.39
interest rates	0.70	0.39	0.20	0.44	0.52	0.41
made it possible	0.23	0.21	0.10	0.11	0.18	0.41
european union	0.44	0.33	0.45	0.5	0.46	0.45
general council	0.43	0.21	0.49	0.45	0.48	0.48
united states	0.56	0.28	0.41	0.35	0.53	0.62

…

	Hum (exp)	cand	glbl	ms	rev	syst
Average	0.38	0.22	0.22	0.25	0.29	0.27

Note that average scores can characterise the general performance of any translation system, e.g. scores for human translation are higher than for MT output. Remember, however, that these scores are computed very differently than standard BLEU scores. The correlation of the average with human judgements is lower than the figures reported for BLEU, which are in the region of 0.98 (Babych & Hartley, 2004). Nevertheless, these

concordance-based scores show a high positive correlation with adequacy, and a slightly lower correlation with fluency, despite the corpus size being much smaller. Table 7 shows these correlation figures.

Table 7: Correlation of average for all MWEs

	r correl
Adequacy	0.883
Fluency	0.620
Informativeness	0.380

We checked contexts for some expressions in Table 6 in order to determine whether lower BLEU scores are due to sporadic mismatches (since the size of the evaluation sub-corpus in this case is much smaller than for a standard BLEU evaluation), or whether lower scores indeed correspond to translation problems for these particular MWEs. In the majority of cases, lower BLEU scores correspond to consistently less fluent translations or mistranslations. Tables 8 and 9 illustrate such cases by comparing concordances for the human reference translation and MT output.

As can be seen from the tables, the MWEs were consistently translated less adequately than in the case of human translation. However, for MWEs with higher BLEU scores this was not the case: their translation was still adequate. Table 10 illustrates this fact for the MWE *minister of foreign affairs*, which is above the threshold of BLEU – 0.27.

Table 8: MWE *work force*

*Fr:… Depuis le début du siècle, ses **effectifs** sont passés de 15000 à 2500 emplois…*	
Human Ref	Systran
its **work force** has fallen from	its **manpower** passed from
believes that reducing the **work force** would	estimates that to touch **manpower** would
continues to reduce its **work force** in Europe	continues the reduction of its **manpower** in Europe
reducing its **work force** from	bringing back its **manpower** in

These results are surprising, given that BLEU is generally used only at 'higher' levels of evaluation: it offers high correlation with human judgments only at the level of an entire corpus, but not for individual texts or sentences. Yet it appears from our experiments that these scores present an additional 'island of stability' at the level of individual lexicogrammatic constructions. Concordance-based evaluation appears to provide an approach to these constructions that is sufficiently focused for BLEU scores

to become meaningful also at the micro-level. A possible explanation for this is that the sub-corpus used for evaluating MWEs is collected in a very controlled way, which limits the noise factor.

Table 9: MWE *ticket sales*

*Fr:… Soit 53 % des **entrées** avec 40 % des écrans…* *La famille-fantôme fait mieux que la famille saint-bernard avec, respectivement, 75 000 (près de 160 000 en quinze jours) et 67 000 **entrées** (200 000 en trois semaines).*	
Human Ref	Systran
this would be 53% of **ticket sales** with 40% of the screens	That is to say 53% of the **entries** with 40% of the screens
and 67 000 **ticket sales** (200 000 in three weeks	and 67 000 **entries** (200 000 in three weeks
with another 43,000 **ticket sales** during its fifth week	with 43 more 000 **entries** in fifth week

Table 10: MWE *minister of foreign affairs*

*Fr:… Les négociations actuelles, patronnées par les Etats-Unis, sont menées par le **ministre** croate **des affaires étrangères**, Mate Granic, et le premier ministre bosniaque, Haris Silajdzic*	
Human Ref	Systran
in paris the **minister of foreign** affairs stated friday	In Paris, the **Foreign Minister** declared, Friday
the israeli **minister of foreign** affairs Shimon Peres thought	The Israeli **Foreign Minister** Shimon Peres estimated
led by the **croat minister of foreign affairs** Mate Granic	carried out by the Croatian **Minister for the Foreign Affairs**, Mate Granic
the nigerian **minister of foreign affairs** babangana kingibe left	**The minister** Nigerian **of the Foreign Affairs**, Babangana Kingibe, flied away

To conclude, we can define our risk analysis measure for MWE expressions as a (possibly weighted) combination of MT evaluation score for an MWE concordance and its frequency.

4.3. Normalisation for translation variation

As noted earlier, in the case of MT output, low BLEU scores for the concordance of an MWE mean that the MWE is not generated properly.

However, we included in our evaluation set the second human translation provided by DARPA-94 (the 'expert' translation) and for this human translation the meaning of lower BLEU scores is very different. If we suppose that professional human translators cannot frequently be wrong, then lower scores for a given MWE mean that there are other legitimate ways to express the intended meaning. Therefore, generating that specific MWE is not essential for the content. Such expressions typically belong to the general lexicon and can be freely re-phrased in the same context.

On the other hand, if a given MWE has a high BLEU score, then it was consistently inserted into the text by both human translators. Thus, it is more stable and possibly even obligatory for such contexts. Such expressions are usually terms or other stable constructions which require specific and invariable translation equivalents.

Table 11 presents MWEs sorted by the BLEU scores for the 'expert' human translation. The table shows that general language expressions with greater contextual variability are at the top, while more stable terminological units are at the bottom. (Highlighting of problematic expressions for Systran is preserved from Table 6.)

This finding suggests that MT systems should be rewarded for having higher BLEU scores for more stable constructions but allowed greater freedom to deviate from less stable equivalents. Accordingly, we should take into account not only absolute values of BLEU for a given construction, but also how different the score for an MT system is from the corresponding BLEU score for a human translation. In the general case, BLEU for MT and for human translations are independent, but the measure of MT quality is precisely how close they are: in other words, whether we can reliably predict the difference between the MT and human scores given the raw MT score.

Figure 5 illustrates this point. The horizontal axis shows values for human translation, and the vertical axis shows values for Systran.

Table 11: MWEs sorted by 'expert' human BLEU

	Hum (exp)	cand	glbl	ms	rev	syst
once again	0.12	0.09	0.09	0.15	0.09	0.11
sales volume	0.18	0.13	0.1	0.11	0.11	0.12
public prosecutor	0.21	0.17	0.16	0.12	0.30	0.18
press release	0.23	0.22	0.19	0.15	0.17	0.19
made it possible	0.23	0.21	0.10	0.11	0.18	0.41
ticket sales	0.26	0.24	0.09	0.11	0.20	0.11
take place	0.32	0.17	0.14	0.15	0.34	0.18
young people	0.32	0.1	0.10	0.18	0.16	0.28
credit lyonnais	0.33	0.16	0.16	0.10	0.12	0.10
examining magistrate	0.36	0.13	0.14	0.29	0.25	0.34
work force	0.37	0.35	0.1	0.10	0.12	0.11
term rates	0.37	0.25	0.12	0.20	0.35	0.19
other hand	0.37	0.16	0.66	0.46	0.63	0.39
so-called	0.38	0.20	0.15	0.19	0.16	0.21
daily life	0.39	0.17	0.23	0.17	0.45	0.20
managing director	0.42	0.22	0.19	0.42	0.21	0.31
general council	0.43	0.21	0.49	0.45	0.48	0.48
european union	0.44	0.33	0.45	0.50	0.46	0.45
french speaking	0.48	0.11	0.15	0.23	0.26	0.12
media library	0.50	0.17	0.11	0.16	0.32	0.34
prime minister	0.54	0.33	0.44	0.24	0.44	0.39
united states	0.56	0.28	0.41	0.35	0.53	0.62
minister of foreign affairs	0.63	0.59	0.29	0.54	0.18	0.33
interest rates	0.70	0.39	0.20	0.44	0.52	0.41

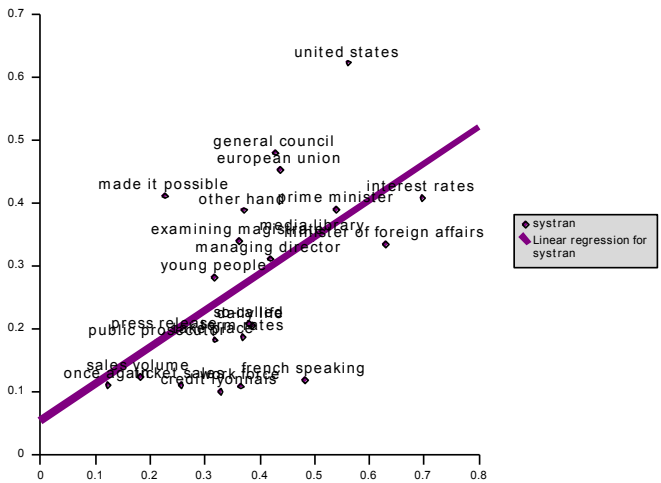

Figure 5: BLEU for human translation vs Systran MT

MWEs in this chart are located along two dimensions: MWEs closer to the right are more stable (more terminological), while those closer to the left belong to the general lexicon and can be more frequently rephrased. On the other hand, MWEs at the top are less problematic for Systran MT, while those at the bottom are more difficult. In an ideal case, the points of the chart should be close to the diagonal line. Deviations from this line mean either that an MT output matches the human translation of a variable term (e.g. MWE *made it possible* in the top-left corner of Figure 5), or that it does not cover specific stable terms (e.g. MWE *French-speaking* in the bottom-right corner of Figure 5 – there is a gap in Systran's dictionary: *...of the Flemish francophonie...* instead of *... of the Flemish French-speaking community...*).

We suggest that we can measure certain aspects of MT quality by the degree of agreement between BLEU scores for human translation and for MT. Such agreement can be captured by the correlation coefficient *r*. We compute it between two arrays of scores: the array of raw BLEU figures for an MT system, and the array of differences between these scores and BLEU for the human translation (for corresponding MWEs):

$$N = r\frac{BLEU_{MT}^{MWE}}{BLEU_{MT}^{MWE} - BLEU_{HumanTr}^{MWE}}$$

We found that there is a high correlation between human judgments for informativeness and the N (normalised variation) score. Table 12 illustrates the correlation between N and each of the human evaluation parameters available for the DARPA corpus.

Table 12: Correlation: N-score vs human scores

	cand	glbl	ms	rev	syst	**r corr with N**
N-score	0.13	0.38	0.25	0.45	0.38	
[ade]	0.68	0.71	0.71		0.79	0.72
[flu]	0.45	0.38	0.38		0.50	-0.02
[inf]	0.64	0.75	0.66		0.76	**0.97**

The table suggests that for better, more informative MT systems there is better agreement between BLEU scores for MT and the difference {MT vs human}: if BLEU is low, then the difference should also be low, which means that the human score is low as well. Thus MT is allowed to have low scores only for re-phrasable, highly variable expressions from the general lexicon.

To summarise, the proposed N-score is the measure of how well MT translates stable (e.g. terminological or idiomatic) expressions, which are repeatable and highly recognisable by human users of MT, especially for particular subject domains, genres or types of texts. Normalisation for legitimate translation variation for N-scores comes at a cost, as it is essential to have more than one human translation for MT evaluation.

5. Applications

The proposed approach can be useful in two main ways, without the need for human scores. Firstly, it can discover MWEs on the SL side or on the TL side which are, respectively, poorly translated by one or several MT systems, or not properly generated. Along these lines, our method is useful for MT developers in their efforts to discover the most typical lexical errors and improve the quality of their systems. It is equally useful for MT users who wish to extend their dictionaries before launching production in a new subject domain.

Secondly, our approach can also highlight MWEs which are usually translated correctly by MT systems. This information can be useful in the specification of features of MT-tractability (Bernth et al., 2001) using large-scale corpus data, and based on the performance of a particular state-of-the-art MT system.

Finally, we have shown that the N-score, which is a correlation coefficient between standard and normalised BLEU scores for individual MWEs, is a good predictor of human judgements about informativeness at the corpus level. Previously, no automated metrics could approximate this particular quality parameter.

6. Future work

Future work will involve determining an optimal size of immediate context for the concordances, selecting the most revealing automatic metrics, the (meta-)evaluation of the approach using, for example, corpus-level human scores, and determining those classes of MT error which most influence human evaluation scores.

Acknowledgements

The work is supported by the Leverhulme Trust.

Bibliography

Babych, B. & Hartley, A. (2004). Extending the BLEU MT evaluation method with frequency weightings. In ACL 2004: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (pp. 621-628); Barcelona, Spain, July 21-26, 2004.

Babych, B., Sharoff, S., Hartley, A., & Mudraya, O. (2007a). Assisting Translators in Indirect Lexical Transfer. In ACL 2007: Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (pp. 136-143); Prague, Czech Republic, June 23-30 2007.

Babych, B., Hartley, A., & Sharoff, S. (2007b). Translating from under-resourced languages: comparing direct transfer against pivot translation. In Proceedings of Machine Translation Summit XI (pp. 412-418); Copenhagen, Denmark, September 10-14, 2007.

Baldwin, T. (2006, July). *Compositionality and multiword expressions: Six of one, half a dozen of the other?* Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, Australia.

Bernth, A., & Gdaniec, C. (2001). MTranslatability. *Machine Translation, 16*, 175-218.

Cowie, A.P. (1998). Introduction. In A.P. Cowie, (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 1-20). Oxford: Oxford University Press.

Estrella, P., Hamon, O., & Popescu-Belis, A. (2007). How much data is needed for reliable MT evaluation? Using bootstrapping to study human and automatic metrics. In Proceedings of Machine Translation Summit XI (pp. 167-174); Copenhagen, Denmark September 10-14, 2007.

Evert, S., & Brigitte K. (2001). Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter (ACL-EACL 2001) (pp. 188–195); Toulouse, France, July 7, 2001.

Koehn P. (2005) Europarl: A parallel corpus for statistical machine translation. In Proceedings of Machine Translation Summit X (pp. 79-86); Phuket, Thailand, September 13-15, 2005.

Miller, K.J. & Vanni, M. (2005). Inter-rater agreement measures, and the refinement of metrics in the PLATO MT evaluation paradigm. In Proceedings of Machine Translation Summit X (pp. 125-132); Phuket, Thailand, September 13-15, 2005.

Papineni, K., Roukos, S., Ward, T., & Zhu, W-J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318); Philadelphia, PA, July 6-12, 2002.

Sharoff, S., Babych, B., & Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In Proceedings of COLING/ACL 2006 Conference (pp. 739-746);, Sydney, Australia, July 17-21, 2006.

Thurmair, G. (2007, September). *Automatic evaluation in MT system production*. Invited talk given at Machine Translation Summit XI workshop: Automatic Procedures in MT Evaluation, Copenhagen, Denmark.

White, J.S., O'Connell, T., & O'Mara, F. (1994). The ARPA MT evaluation methodologies: evolution lessons and future approaches. In Technology partnerships for crossing the language barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas (pp. 193-205); Columbia, MD, USA, October 5-8, 1994.