Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output

Bogdan Babych
Centre for Translation Studies
University of Leeds
Department of Computer Science
University of Sheffield
bogdan@comp.leeds.ac.uk

Anthony Hartley
Centre for Translation Studies
University of Leeds
a.hartley@leeds.ac.uk

Abstract

We report the results of an experiment on automatic NE recognition from Machine Translations produced by five different MT systems. NE annotations are compared with the results obtained from two high-quality human translations. The experiment shows that for recognition of a large class of NEs (Person Names, Locations, Dates, etc.) MT output is almost as useful as a human translation. For other types of NEs (Organisation Names) Precision figures are close to the results for human annotation, although Recall is seriously distorted by the degraded quality of MT. The success rate of NE recognition doesn't strongly correlate with human or automatic MT evaluation scores, which suggests that the quality criteria needed for measuring MT usability for dissemination purposes are not pertinent for assimilation tasks such as Information Extraction.

1. Dissemination *vs* assimilation

Since the 1960's the 'Holy Grail' of Machine Translation technology has been Fully Automatic High Quality Translation (FAHQT), which aims at creating accurate and fluent texts in a target language suitable for dissemination (i.e. publication) purposes – a goal which has yet to be achieved.

However, there are successful attempts and suggestions to use 'crummy' MT output (Church and Hovy, 1993) for assimilation (i.e. comprehension) tasks: text classification, relevance rating, information extraction (White et al., 2000), for NLP tasks such as Cross-Language Information Retrieval (Gachot et al., 1998), and Multilingual Question Answering (a new task set up for CLEF 2003).

Multilingual Information Extraction is one such assimilation task and consequently an area where imperfect MT output is potentially useful. On the one hand MT can extend the reach of existing monolingual IE systems by translating a text before running IE; on the other hand, results of IE (identified Named Entities, template elements or scenario templates) can be translated into a foreign language after IE processing (Wilks, 1997: 7-8). The first scenario is more demanding for MT, because the performance of an automatic IE system may be influenced by MT quality.

There is an open question: Which aspects of MT quality are important for different IE tasks and may substantially influence the performance of IE?

MT quality is often benchmarked from the viewpoint of human users (White et al., 1994), focusing still on the goal of FAHQT for dissemination. As a result, automatic evaluation scores, such as BLEU (Papineni et al., 2002), are validated according to how well they correlate with human intuitive judgements of translation quality.

Using edit distances between MT output and a human reference translation to evaluate MT (Akiba et al., 2001) also makes an implicit assumption that MT should be suitable for dissemination purposes.

However, MT has created its own demand precisely in the area where otherwise there would be no translation at all. Where it is primarily used for assimilation purposes, the evaluation of NLP performance on MT output might give a better indication of its usefulness than dissemination criteria. Therefore there is a need for: (1) systematically benchmarking NLP technologies, such as IE (and its sub-tasks, e.g., NE recognition), on MT output; (2) developing and calibrating automatic MT evaluation scores for these *primary* uses of 'crummy' MT; (3) assessing quantitatively the extent to which certain human and automatic MT evaluation scores predict the performance of automatic systems on different NLP tasks.

2. Set-up of the experiment

We addressed some of the above issues by conducting a comparative evaluation of the performance of the ANNIE NE recognition module of Sheffield's GATE IE system (Gaizauskas et al., 1995; Cunningham et al., 1996, 2002). We used the DARPA-94 corpus of French-English MT and human translations (White et al., 1994). The MT systems were Candide, Globalink, Metal, and Systran (participants in DARPA), plus Reverso. Specifically, we focused on whether there is a significant divergence between NE recognition performance and the results of human and automatic evaluation of the MT systems. This indicates to what extent MT quality criteria may differ for human use and for the needs of NLP systems.

In the first stage NEs were annotated in translations of 100 news reports (each text is about 350 words), produced by each MT system.

NEs were also annotated in the two independent human translations of the same 100 texts: the Reference and the Expert translations.

Comparative evaluation of this NE annotation is different from standard evaluation procedure for NE recognition in two respects. The first difference is that in our experiment there is no gold standard NE annotation for any of the human translations or MT outputs. The second difference is that the annotated text is no longer constant.

2.1 Absence of a gold standard

Since all seven sets of texts are different, it would be too expensive to produce a gold standard annotation for each of them. However, all these texts have the same origin: all are translations of the same collection of French source texts, so it can be expected that there will be a great overlap between extracted NEs, namely for those typical cases when French NEs have a standard translation into English. While we expect that most types of NEs stay the same across different translations, we also have to account for possible variations.

Two main things can go wrong when NEs are extracted from MT output (which is generally regarded to be of lower quality than a human translation):

– NE recognition often relies on certain contextual conditions being met, so if a lexical or morpho-syntactic context is distorted in MT output, NEs will be not extracted, resulting in NE 'undergeneration'; likewise the distorted context may give rise to false NEs, leading to NE 'overgeneration'.
– If NEs are wrongly translated despite the context meeting the requirements of the NE recognition system, they are of no use in any other NLP tasks.

The goal of our comparative evaluation is to estimate to what extent the output of different MT systems and the alternative human translation are 'robust' against these two pitfalls, i.e., to what extent they may be useful for the IE purposes. This means that we are less interested in absolute performance figures for the NE recognition system, than in the comparison between its runs on the output of different MT systems.

Furthermore, the accuracy scores for leading NE recognition systems are relatively high. The default settings of ANNIE NE modules produce between 80-90% Precision & Recall on news texts originally written in English (Cunningham et al., 2002). We assume that for comparable texts – human translations of news reports into English – NE recognition performance is similar.

Therefore, for our purposes it is possible to use the NE annotation in one of the human translations as a reference, which will serve as a 'silver standard' for benchmarking NE recognition performance from 'low quality' MT texts. The baseline for such comparisons will be the NE annotation in the other human translation: it will indicate what difference in accuracy may be

expected if an alternative high-quality translation is used. This allows us to: (1) estimate the relative performance of the NE recognition system on texts with variable quality; (2) compare these relative figures with human and automatic MT evaluation scores; (3) answer the question whether usefulness of MT for IE should be characterised by criteria other than Adequacy and Fluency, or whether these correctly predict the potential performance of NE recognition.

2.2 Legitimate variation in translation

In our research the annotated text is no longer constant – it becomes changeable; on the contrary, it is the NE recognition system that is constant.

This requires a different interpretation of the figures for Precision, Recall and F-score: strictly speaking they only characterise *differences* rather than the degree of *perfection*. Annotation mismatches do not necessarily mean deterioration; they may be also due to the improved performance of NE recognition on the test file, or due to choosing a legitimate alternative translation.

For example, we expect that NEs normally have a standard translation and will not vary across different human translations; therefore the quality of MT systems depends on how well this standard is followed. The only exceptions to this rule should be less well known organisations which do not have an established translation. But surprisingly, some degree of legitimate variation was found in human translations for well-known institutions also:

ORI: *De son côté, le <u>département d'Etat américain</u>, dans un communiqué, a déclaré: 'Nous ne comprenons pas la décision' de Paris.*

HT-Expert: *For its part, the* <Organization> *<u>American Department of State</u>* </Organization> *said in a communique that 'We do not understand the decision' made by* <Location> *Paris* </Location>.

HT-Reference: *For its part, the* <Organization> *<u>American State Department</u>* </Organization> *stated in a press release: We do not understand the decision of* <Location> *Paris* </Location>.

MT-Systran: *On its side, the* <Organization> *<u>American State Department</u>* </Organization>, *in an official statement, declared: 'We do not include/understand the decision' of* <Location> *Paris* </Location>.

This indicates the need to identify classes of NEs which may undergo legitimate translation variation, similarly to other words or phrases in language, and to account for the legitimate translation variation in our experiment.

2.3 Evaluation parameters and procedure

We note that the results of NE annotation from human translations and from MT output are rather distant for some types of NEs. For the two human translations there is a norm for the number of differences in annotation, but MT output sometimes goes far beyond this norm. We suggest that the extent of deviation from these baseline norms characterises the usefulness of MT systems for IE. We compute parameters which are interpretable from this point of view and account for the problems of the standard accuracy measures.

– Counts of different types of annotated NEs

This parameter is very robust against legitimate translation variation. It shows in how many cases *any* NE has been identified in a particular context, i.e., whether MT output preserves the contextual conditions for identifying an NE. On the other hand, this parameter does not take into account cases where conditions for identification of new (either spurious or genuine) NEs are created in tested MT output. It characterises only the 'upper bound' of cases where conditions for identification of an NE have been met.

– Precision on the union *of NE annotations for two human translations*

This parameter is sensitive to legitimate translation variation: it rewards annotations that match at least one of the alternatives found in two independent human translations. If no match is found for a particular NE, the case is treated as 'over-generation'. For a given MT output, this parameter shows how successfully over-generation of NEs may be avoided. This parameter uses a similar approach to the BLEU method for MT evaluation, which computes precision on the union of n-gram units from several human translations.

– Recall on the <u>intersection</u> of NE annotations for two human translations

This parameter rewards annotations that match a set of NEs, which are constant across different human translations. The intuition is that, if a given NE is present in both human translations, it is very

likely to have some 'standard', obligatory translation, which it is necessary to preserve in MT. Such NE needs to be extracted exactly in the form used in both human translations. This parameter shows how successfully 'under-generation' for the set of most 'standard', uniformly translated NEs has been avoided.

In order to determine whether any human or automatic MT evaluation scores could predict the performance of NE recognition on MT output, we tested the performance figures for correlation with each of the evaluation scores, as follows.

The corpus was divided into 10 chunks each containing 10 texts. Human evaluation scores for Adequacy, Fluency and Informativeness are available for each machine-translated text in the DARPA corpus (not including Reverso, therefore). We also generated automatic BLEU scores for each text. Average scores for chunks of 10 texts in the corpus were computed. The resulting sets of scores contained 40 samples each (10 for each MT system).

For corresponding sets we computed Pearson's correlation coefficient *r*. We tested the statistical significance of this correlation with *t* distribution.

3. Results of NE recognition on MT output

The counts of extracted NEs are summarised in Table 1 and Figure 1. It can be seen that for Organisation Names there is a significant difference in the number of extracted NEs for texts produced by humans and for MT output: many more Organisation Names are extracted from human translations. The difference is much smaller for Titles, but the tendency is similar. However, this tendency reverses for Job Titles: MT output tends to give rise to a greater number of this type of NEs [1].

Results for other types of NEs for human translations and MT output come very close together. This gives an indication that distorted MT quality seriously affects the results of NE recognition for a specific type of Named Entities – Organisation Names, which are more context-dependent and less distinguishable from other types of words than other NE types. The latter may have some explicit mark-up or clearly defined boundaries, so they tend to be less affected by MT and may be extracted from MT output more successfully.

	Reference HT	**Expert HT**	**Candide**	**Globalink**	**Metal**	**Reverso**	**Systran**
paragraph	826	802	813	804	805	798	804
Organization	523	561	272	240	218	271	324
Title	254	215	138	80	101	159	150
JobTitle	213	248	303	341	299	312	321
{Job}Title	467	463	441	421	400	471	471
FirstPers.	515	528	518	530	519	504	537
Person	612	629	598	660	599	603	608
Date	577	572	562	541	556	597	554
Location	521	503	474	460	475	508	526
Money	101	108	117	80	81	99	100
Percent	72	71	72	71	71	72	72

Table 1. Number of extracted NEs

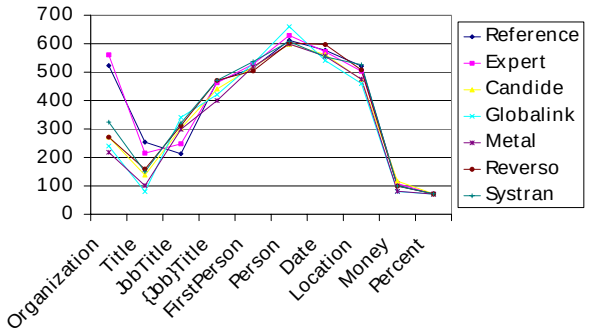

Figure 1. Number of extracted NEs

The Precision (P) and Recall (R) figures for each type of NE are summarised in the following tables and figures.

[1] These clashing tendencies for Titles and Job Titles have a simple technical explanation: cases where human translators capitalise the initial letter (e.g. 'Colonel') normally go into the Title category; where MT renders them in lower case (e.g. 'colonel'), they are often annotated as Job Titles. The category {Job}Title joins these two categories and shows no significant differences between human translations and MT output.

3.1 Organisation names

Figures for Organisation Names, taking as reference the NE annotations from the Reference and the Expert human translations and the union / intersection of these sets, are presented in Table 2 and Figure 2.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.5745	1	1
P.HT-ref	1	0.6172	1
P.candide	0.4924	0.5229	0.5763
P.globalink	0.4979	0.5319	0.5745
P.ms	0.5423	0.5672	0.607
P.reverso	0.5709	0.5709	0.6552
P.systran	0.5096	0.5223	0.5892
R.HT-exp.	0.6172	1	1
R.HT-ref	1	0.5745	1
R.candide	0.252	0.2491	0.3639
R.globalink	0.2285	0.2273	0.3386
R.ms	0.2129	0.2073	0.3196
R.reverso	0.291	0.2709	0.4019
R.systran	0.3125	0.2982	0.4399

Table 2. Precision, Recall – Organisations

In all cases scores for annotations of human translations are the highest, but the contrast between human translations and MT is the largest for Recall, and there is a very little difference for Precision. The improvement in Precision when the union of both translations is used as a reference is very moderate. The improvement in Recall when the intersection of the two translations is used is much higher.

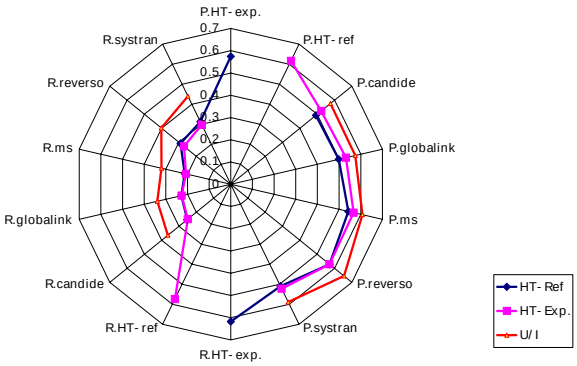

Figure 2. Precision, Recall – Organisations

This shows that in MT output over-generation of Organisation Names is very limited; the main problems are related to under-generation. Recall is the major aspect affected by the degraded quality of MT output; Precision results for NE recognition from MT output are almost unaffected.

The results demonstrate that Organisation Names constitute a broad and highly dynamic class of NE whose identification is very sensitive to MT quality, hence the low Recall figures. In this typical example, ANNIE fails to identify the string 'Egyptian Diplomacy' in the MT output as an Organisation Name, since this is not an expected way of expressing this concept in English.

ORI: … *le chef de la diplomatie égyptienne*
HT: *the* `<Title>`Chief`</Title>` *of the* `<Organization>`*Egyptian Diplomatic Corps* `</Organization>`
MT-Systran: *the* `<JobTitle>` *chief* `</JobTitle>` *of the Egyptian diplomacy*

Such occurrences are frequent, so generally far fewer Organisation Names are identified in MT output as compared to a human translation.

3.2 Person names

In general, the accuracy for Person Names (Table 3 and Figure 3) is much higher than for Organisation Names. However, the figures are more dependant on a particular MT system and do not characterise any general tendency for MT output as compared to human translations. The results of leading MT systems for this type of NE are practically undistinguishable from the results obtained from human translations, in terms of both Precision and Recall.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.785	1	1
P.HT-ref	1	0.8056	1
P.candide	0.7525	0.7425	0.8161
P.globalink	0.4932	0.5099	0.5478
P.ms	0.6868	0.6834	0.7437
P.reverso	0.6083	0.6333	0.6783
P.systran	0.7169	0.7318	0.7897
R.HT-exp.	0.8056	1	1
R.HT-ref	1	0.785	1
R.candide	0.7353	0.707	0.8235
R.globalink	0.531	0.535	0.6085
R.ms	0.6699	0.6497	0.7586
R.reverso	0.5964	0.6051	0.6856
R.systran	0.7075	0.7038	0.8073

Table 3. Precision, Recall – Person

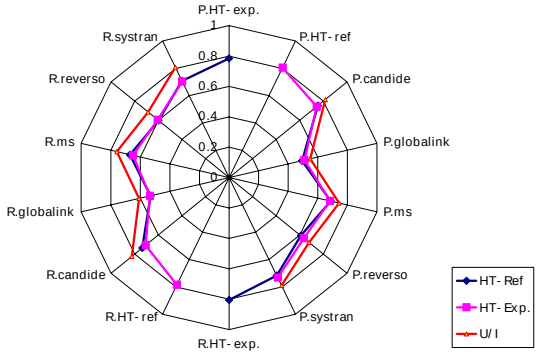

Figure 3. Precision, Recall – Person

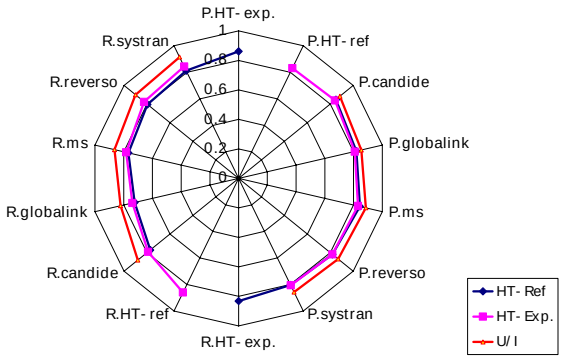

Figure 4. Precision, Recall – Location

But there is no correlation between human evaluation scores for the performance of an MT system and the accuracy figures for recognition of Person Names. This indicates that recognition of these NEs is not directly influenced by other translation problems, such as ambiguity between Person Names and common nouns (e.g. *Bill Fisher*) These cases are relatively rare and easily identifiable; current MT technology has successfully solved this problem.

3.3 Location names

The data for Location Names (Table 4 and Figure 4) shows a more even performance across different MT systems as compared to other types. This may be due to the fact that this type of NE is less ambiguous than the other types of NE. Once again, NE recognition from MT output is practically undistinguishable from NE recognition from a human translation, in terms of both Precision and Recall.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.8608	1	1
P.HT-ref	1	0.8311	1
P.candide	0.8481	0.8397	0.884
P.globalink	0.8196	0.8087	0.8543
P.ms	0.8439	0.8312	0.8861
P.reverso	0.8185	0.8185	0.8679
P.systran	0.8042	0.8061	0.8574
R.HT-exp.	0.8311	1	1
R.HT-ref	1	0.8608	1
R.candide	0.7716	0.7913	0.8799
R.globalink	0.7236	0.7396	0.8222
R.ms	0.7678	0.7833	0.8637
R.reverso	0.7965	0.825	0.9007
R.systran	0.8119	0.8429	0.9145

Table 4. Precision, Recall – Location

3.4 Overgeneration

The results for other types of NE lie within the range of scores described above. Each type of NE behaves differently across MT systems and human reference translations. Nevertheless, for all these types of NEs, the best MT systems give results comparable with the accuracy of NE annotation from human translations.

Of course, some additional (spurious or genuine) NEs may appear in MT output despite this tendency. The rarity of such an event is because distorted MT output makes it much harder to create new conditions for identifying an NE than to reconstruct necessary conditions similar to those which often are created in an alternative human translation.

Nevertheless, in a few cases genuine new NEs are found in MT output. In these cases, MT is more favourable for IE tasks: it is more literal (therefore less misleading) than human translation, e.g.:

ORI : *"Il a été fait chevalier dans l'ordre national du Mérite en mai 1991"*

HT: *"He was made a Chevalier in the National Order of Merit in May, 1991."*

MT-Systran: *"It was made* `<JobTitle>` *knight*`</JobTitle>` *in the national order of the Merit in May 1991"*.

MT-Candide: *"He was knighted in the national command at Merite in May, 1991"*.

The human translator used the borrowed French word *Chevalier*, which 'distracted' ANNIE. Although this translation might be more adequate from the human point of view, it is less useful for the NE recognition system. Interestingly, the more idiomatic translation of this sentence produced by the statistical MT system Candide removed this

NE from the sentence. The literal output of rule-based MT systems, such as Systran, proves most favourable for identifying the NE.

For MT output, Recall is highest for 'constant' NEs which have a standard translation (and are identified in both human translations). The figures for Recall correlate highly with the counts of extracted Organisation Names (Table 1): Pearson's correlation coefficient is 0.9561 (there is no correlation for Precision). This confirms the suggestion that over-generation does not affect counts of NEs.

4. Correlation with MT evaluation scores

The second problem addressed in our experiment is determining whether human or automatic MT evaluation scores correlate with accuracy of NE recognition.

r(38)=	**ADE**	**FLU**	**INF**
ref=Exp ; Precision	-0.0047	-0.0232	0.1100
ref=Exp ; Recall	**0.2558** *p>0.05*	0.0671	-0.0128
ref=Ref ; Precision	0.1887	-0.0994	0.0011
ref=Ref ; Recall	**0.3997** *p<0.01*	0.0084	-0.0633
ref=u(ER); Precision	0.1804	-0.1284	-0.0071
ref=i(ER); Recall	**0.3465** *p<0.05*	-0.1070	-0.0458

Table 5. Pearson's *r* coefficient: (Organisations *vs* Adequacy, Fluency, Informativeness).

Only weak or moderate correlation was found in a number of cases, which suggests that human judgements or automatic MT evaluation scores do not show how useful MT output may be for assimilation purposes, e.g., for NLP tasks such as NE recognition.

For Organisation Names the highest correlation figures were between Recall and human scores for Adequacy – in cases when we used the Reference human translation or the intersection between the two human annotations (Table 5).

This weak correlation is also statistically significant (at the levels $p<0.01$ and $p<0.05$) for the Reference human translation and the intersection between the two references.

Other cases of moderately strong positive correlation were also found, although it is difficult to give linguistically meaningful interpretation to these correlations. We suggest that they may be due to indirect links between the overall quality of an MT system and the attention that particular groups of developers pay to specific NE problems.

The highest positive correlation for human evaluation scores was found between Recall on Date NEs and human scores for Fluency for the case when the Expert human translation was used as a reference:

$r(38)=0.4847, p<0.001; (t=3.8392)$.

With the other human reference translation the correlation is much weaker:

$r(38)=0.3559$; $p<0.05$; ($t=2.6388$)

The highest positive correlation for BLEU scores is close to these figures. Again, it is difficult to give any meaningful linguistic interpretation to this correlation. The correlation of BLEU scores is more consistent across different references, e.g. the correlation between BLEU and Recall figures for Title NEs with Expert and Reference human translations and the intersection of these annotations:

ref=Exp: $r(48)= 0.4844; p<0.001; (t=3.8364)$

ref=Ref: $r(48)= 0.4025; p<0.01; (t=3.0467)$

ref=i(ER): $r(48)= 0.3251; p<0.05; (t=2.3819)$

5. Conclusions and future work

We conclude that neither human evaluation scores nor automatic BLEU scores reliably predict the performance of NE recognition for any of the NE types. Still, in a few cases the Pearson's *r* coefficient is significantly different from zero, which indicates a positive link between better performance in some aspects of NE recognition (e.g., boosting recall for Organisation Names) and better quality (e.g., higher Adequacy scores) of MT from the point of view of human evaluators. Nevertheless, for many other aspects of NE recognition there is no such link, so the usability of MT output cannot be judged just by intuitive human criteria, which usually assume dissemination purposes for MT output.

Thus our results confirm the idea that the criteria most widely used for assessing MT quality fail to reflect the needs of subsequent NLP processing in general and of NE recognition in particular; these criteria tend to underestimate the usefulness of MT

for automated assimilation tasks, for which MT in most aspects may be as useful as a human translation.

There remain open questions why just for one particular type of NEs – Organisation Names – Recall of NE recognition of the rule-based ANNIE system is substantially distorted by the degraded MT output, and why the Recall weakly correlates with human scores for MT Adequacy. On the one hand, the existence of this link may suggest that the Recall figures give some indirect indication of the translation Adequacy, so technological improvements in recognition of Organisation Names in MT systems will boost translation quality (in eyes of human evaluators) and, therefore, will enhance the usability of MT for dissemination purposes as well. For related work see (Babych and Hartley, 2003).

On the other hand it may be the case that for the other type of NE recognition systems, namely the ones based on statistical Machine Learning, the MT output would appear much more useful (e.g., the distortion of Recall for Organisation Names would be significantly smaller), since the recognition could then adapt to any regular patterns produced by the MT system, even if they differed from the natural form. Whether or not any NE error patterns in MT output could be learnt by statistical IE systems is an interesting problem, which has important implications for MT and IE technology.

Future work will include research on usability of MT for other IE tasks, such as scenario template filling, co-reference resolution, automatic summarisation, etc. Also the suitability of the MT output for a range of learning IE systems will be investigated, and typologically different language pairs will be involved in the evaluation. Further direction of research will include moving from the "black-box" to the "glass-box" evaluation and examining the ways in which particular MT systems treat different kinds of NE.

This research will lead to theoretical generalisations about the nature of dynamic quality criteria for translation, which correctly predict the usability of human translations and MT for different purposes.

References

Akiba Y., K. Imamura and E. Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. *Procs. MT Summit VIII* 15–20.

Babych, B. and A. Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools. Improving MT throught other language technology tools. Recourses and tools for building MT.* Budapest, Hungary. p. 1–8.

Church, K.W. and E.H. Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*. 8. 239-258.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for robust NLP Tools and Applications. *Procs. ACL'02.* Philadelphia, July 2002.

Cunningham, H., Y. Wilks and R. Gaizauskas. 1996. GATE – a General Architecture for Text Engineering. *Procs. COLING-96.* Copenhagen, August 1996.

Gachot D., E. Lange and J. Yang. 1998. The Systran NLP browser: an application of Machine Translation technology in Cross-Language Information Retrieval. In: Grefenstette, G. (ed.), 1998. Cross-Language Information Retrieval. Kluwer. 105-118.

Gaizauskas, R., T. Wakao, K. Humphreys, H. Cunningham, Y. Wilks. 1995. University of Sheffield: Description of the LaSIE system as used for MUC-6. *Procs. MUC-6*. 207-220.

Papineni, K, S. Roukos, T. Ward, W-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM research report RC22176 (W0109-022) September 17, 2001.

White, J., T. O'Connell and F. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Procs. 1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD, October 1994. 193-205.

White, J., J. Doyon and S. Talbott. 2000. Determining the tolerance of text-handling tasks for MT output. *Procs. LREC-2000.* Athens, May-June 2000.29-32.

Wilks, Y. 1997. Information Extraction as a core language technology. In: Pazienza, M.T. (ed.) 1997. Information Extraction. A multidisciplinary approach to an emerging technology. Springer. 1-9.