Statistical modelling of MT output corpora for Information Extraction

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

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

Eric Atwell — School of Computing, University of Leeds, UK — eric@comp.leeds.ac.uk

Abstract

The output of state-of-the-art machine translation (MT) systems could be useful for certain NLP tasks, such as Information Extraction (IE). However, some unresolved problems in MT technology could seriously limit the usability of such systems. For example robust and accurate word sense disambiguation, which is essential for the performance of IE systems, is not yet achieved by commercial MT applications. In this paper we try to develop an evaluation measure for MT systems that could predict their possible usability for some IE tasks, such as scenario template filling, or automatic acquisition of templates from texts. We focus on statistically significant words for a text in a corpus, which are used now for some IE tasks such as automatic template creation (Collier, 1998). Their general importance for IE was also substantiated by our material, where they often include name entities and other important candidates for filling IE templates. We suggest MT evaluation metrics which are based on comparing the distribution of statistically significant words in corpora of MT output and in human reference translation corpora. We show that there are substantial differences in such distributions between human translations and MT output, which could seriously distort IE performance. We compare different MT systems with respect to the proposed evaluation measures and look into their relation to other MT evaluation metrics. We also show that the statistical model suggested could highlight specific problems in MT output that are related to conveying factual information. Dealing with such problems systematically could considerably improve the performance of MT systems and their usability for IE tasks.

1. Introduction

State-of-the-art commercial Machine Translation (MT) systems do not yet achieve fully automatic high quality MT, but their output can still be used as input to some NLP tasks, such as Information Extraction (IE). IE systems, such as GATE (Cunningham et al., 1996), are mainly used for "scenario template filling": processing texts in a specific subject domain (such as management succession events, satellite launches, or football match reports) and filling a predefined template for each text with strings taken from it. On the one hand, IE systems usually do local analysis of the input text and it is reasonable to assume that they tolerate low scores for MT fluency (besides it is the most difficult aspect to achieve in MT output). But in certain cases mistranslation could inhibit IE performance. In this paper we try to develop MT evaluation metrics that capture this aspect of MT quality, and relate them to other evaluation measures, such as MT adequacy scores.

On the other hand, some aspects of IE technology impose a specific set of requirements on MT output. These requirements are important for the general performance of IE systems. For example, named entities (strings of proper names) have to be accurately identified by MT systems: an IE system for Russian will not be able to correctly fill the template if a person name like "Bill Fisher" had been translated from English into Russian as "*выставить счет рыбаку*" ('to send a bill to a fisher'). Moreover, IE requires adequate translation of specific words which are significant for template filling tasks. These words are usually not highly frequent and have a very precise meaning. Therefore it is difficult to substitute such words with synonymous words. For example, the French phrase (1) was translated into English by one of our MT systems:

(1)	French original:	*un montant <u>global</u> de 30 milliards de francs*
	Human translation:	*a <u>total</u> amount of 30 billion francs*
	Machine translation:	*a <u>global</u> 30 billion franc amount*

The correct meaning of the word '*<u>global</u>*' could be guessed by a human post-editor, but the phrase could be misinterpreted by a template-filling module of an IE system, e.g, as an 'amount related to company's global operations', etc. Similarly in the translation of the French sentence (2):

(2)	French original:	*La reprise, de l'<u>ordre</u> de 8%, n'a pas été suffisante pour compenser la chute européenne.*
	Human translation:	*The recovery, <u>about</u> 8%, was not enough to offset the European decline.*
	Machine translation:	*The resumption, of the <u>order</u> of 8 %, was not sufficient to compensate for the European fall.*

The word '*<u>order</u>*' could be misinterpreted by a template-filling IE module as related to ordering of products, but not to uncertainty of information.

Developers of commercial MT systems often do not have sufficient resources to properly disambiguate such words, partly because they rarely occur in corpora that are used for the development and testing of MT systems, and partly because it is difficult to distinguish these problems from other types of issues in MT development. Therefore, it would be useful to have a reliable statistical criterion to highlight MT problems that are related to mismatches in factual information between human translation and MT output. This could be essential for improving the performance of IE systems that run on MT output.

Another important problem for present-day IE research is automatic acquisition of templates, which is aimed to making IE technology more adaptive (Wilks and Catizone, 1999). There have been suggestions to use lexical statistical models of a corpus and a text for IE to automatically acquire templates: statistically significant words (i.e., words in a text that have considerably higher frequencies than expected from their frequencies in a reference corpus) could be found in the text; templates could be built around sentences where these words are used (Collier, 1998).

However, it is not clear whether this method would be effective if applied to a corpus of MT output texts. On the one hand, the output of traditional knowledge-based MT systems produces significantly different statistical models from the models built on "natural" English texts (either original texts or human translations of texts, done by native speakers). It has been shown that N-gram precision of MT output text (in relation to a human reference translation) is significantly lower then the N-gram precision of some other human translation (in relation to the same reference) (Papineni et al., 2001). This is due to the fact that translation equivalence in MT output texts is triggered primarily by source-language structures, not by balancing the adequacy of the target text on the pragmatic level with its fluency, which depends on statistical laws in target language – as is the case for professional human translation. Structures that are treated by knowledge-based MT systems as translation equivalents could have a different distribution in "natural" source and target corpora. As a result, many words that are not statistically significant in "natural" English texts become significant in MT output, and vice versa. Subsequently, different sentences may be selected as candidates for a template pattern based on MT output and one based on human translation.

On the other hand, even if corresponding sentences are selected, the value of template patterns could be diminished by errors in word sense disambiguation, made by MT systems, e.g.:

(3)	French original:	*la <u>reddition</u> des armées allemandes*
	Human translation:	*the <u>surrender</u> of the German armed forces*
	Machine translation:	*the <u>rendering</u> of the German armies*

Words '*<u>surrender</u>*' and '*<u>rendering</u>*' could induce different IE templates, even if corresponding sentences in MT output have been correctly identified as statistically significant. Therefore the requirement of proper word sense disambiguation of statistically significant words is central to usability of MT output corpora for IE tasks.

High quality word sense disambiguation for large vocabulary systems is a complex task, which requires interaction of different knowledge sources and where "best results are to be obtained from optimisation of a combination of types of lexical knowledge" (Stevenson and Wilks, 2001). However, it is also important to find out to what extent the output of different state-of-the-art MT systems is now usable for IE tasks.

In this paper we report the results of an experiment for establishing an evaluation measure for MT systems which contrasts the distribution of statistically significant words in MT output and in human translation and gives an indication of how usable the output of particular MT systems could be for IE tasks. The remainder of this paper is organised as follows: in Section 2 we describe the set-up of our experiment, establish the evaluation measure for MT output and discuss linguistic intuitions behind this measure. In Section 3 we present the results of evaluation of the output of 5 MT systems and a human "expert" translation on the data of the DARPA94 MT evaluation exercise, and compare these results with other measures of MT evaluation, available for this corpus. In section 4 we discuss conclusions and future work.

2. Experiment set-up and evaluation metrics

We developed and compared statistical models for a corpus which has been developed for the DARPA94 MT evaluation exercise (White et al., 1994). This corpus contains 100 human reference translations of newspaper articles, alternative human "expert" translations, and the output of 5 French-English MT systems for each of these texts. The length of each original French text is 300–420 words, with an average length of 370 words. For 4 of these systems scores of "fluency", "adequacy" and "informativeness" are also available.

We suggest the following method of measuring MT quality for IE tasks.

1. In the first stage we develop a statistical model for the corpus of MT output and for a parallel corpus of human translations. These models highlight statistically significant words for each text in the corpus and give a certain score of statistical significance for each highlighted word.
2. In the second stage we compare statistical models for MT output and for human translation corpora. In particular,
   - 2.a - we establish which words in the MT output are "over-generated" – are marked as statistically significant, even though they are absent or not marked as significant in human translation – and what is the overall score of "statistical significance" for such words;
   - 2.b - we establish which words in MT output are "under-generated" – are absent or not marked as statistically significant, even though they are significant in human translation of the same text – and what is the overall score of "statistical significance" of these words;
   - 2.c- we establish which words are marked as significant both in MT and human translation, but which have different scores of statistical significance. Then we calculate the overall difference in the score for each pair of texts in the corpora;
   - 2.d - we compute 3 measures that characterise differences in statistical models for MT and human translation of each text: a measure of "avoiding over-generation" (which is linked to the standard "precision" measure); a measure of "avoiding under-generation" (which is linked to the "recall" measure); and finally – a combined score based on these two measures (calculated similarly to the F-measure).
   - 2.e - we compute the average scores for each MT system.

Besides general scores of translation quality, this method allows us to automatically generate lists of statistically significant words which have a problematic translation in MT output. Such lists could be directly useful for MT development and tuning MT systems for a particular subject domain. Further we present formulae used to compute the scores and we illustrate this process with examples from our corpus.

1. The score of statistical significance is computed for each word (with absolute frequency ≥ 2 in the particular text) for each text in the corpus, as follows:

$$S_{word[text]} = \ln \frac{\left(P_{word[text]} - P_{word[rest-corp]}\right) \times N_{word[txts-not-found]}}{P_{word[all-corp]}}$$

where:

$S_{word[text]}$ is the score of statistical significance for a particular word in a particular text

$P_{word[text]}$ is the relative frequency of the word in the text;

$P_{word[rest-corp]}$ is the relative frequency of the same word in the rest of the corpus, without this text;

$N_{word[txt-not-found]}$ is the proportion of texts in the corpus, where this word is not found (number of texts, where it is not found divided by number of texts in the corpus)

$P_{word[all-corp]}$ is the relative frequency of the word in the whole corpus, including this particular text
"relative frequency" is (number of tokens of this word-type) / (total number of tokens).

The first factor ($P_{word[text]} - P_{word[rest-corp]}$) in this formula is the difference of relative frequencies in a particular text and in the rest of the corpus. Its value is very high for proper names, which tend to re-occur in one text, but have a very low (often 0) frequency in the rest of the corpus. The higher the difference, the more significant is the word for this text.

The second factor $N_{word[txt-not-found]}$ describes how evenly the word is distributed across the corpus: if it is concentrated in a small number of texts, the value is high and the word has more chances of becoming statistically significant for this particular text.

The third factor ($1 / P_{word[all-corp]}$) boosts statistical significance of low-frequent words. The intuition behind it is that if a word occurs in a particular text more then 2 times (and we consider only words with absolute frequency in the text ≥ 2), it becomes more significant if its general relative frequency in the corpus is low.

We use the natural logarithm of the computed score to scale down the range of its values.

Here we give an example of words ranked according to coefficient of statistical significance in Text 1 of the DARPA94 corpus:

Word	$S_{[word]}$ [text1]	N_{word} [txt-not-found]	$(P_{word[text]} - P_{word[rest-corp]})$ * 100%	$P_{word[all-corp]}$ * 100%
urba-gracco	4.620857	0.99	1.098901	0.010710
pezet	4.620857	0.99	0.824176	0.008032
sanmarco	4.620857	0.99	0.549451	0.005355
laignel	4.620857	0.99	0.549451	0.005355
hearing	4.620857	0.99	0.549451	0.005355
facet	4.620857	0.99	0.549451	0.005355
emmanuelli	4.620857	0.99	0.549451	0.005355
presiding	4.200307	0.98	0.546747	0.008032
marseille	4.190050	0.97	1.093494	0.016065
deputies	3.907667	0.98	0.544043	0.010710
lyon	3.897411	0.97	0.544043	0.010710
directors	3.897411	0.97	0.544043	0.010710
confrontation	3.897411	0.97	0.544043	0.010710
appeals	3.729578	0.96	0.813361	0.018742
forgeries	3.679541	0.98	0.541339	0.013387
sp	3.592717	0.96	0.810657	0.021420
henri	3.481956	0.97	0.538635	0.016065
questioned	3.301939	0.95	0.535932	0.018742
confronted	3.301939	0.95	0.535932	0.018742
research	3.019206	0.93	0.530524	0.024097
affair	3.019206	0.93	0.530524	0.024097
former	2.714896	0.82	1.578053	0.085678
director	2.647501	0.83	1.047529	0.061581
socialist	2.641580	0.94	0.519709	0.034807
brought	2.575622	0.88	0.519709	0.034807
criminal	2.529820	0.91	0.517005	0.037484
department	2.444534	0.90	0.514301	0.040162
judge	2.418210	0.94	0.511597	0.042839
companies	2.396704	0.92	0.511597	0.042839
officials	2.340823	0.87	0.511597	0.042839
wednesday	2.263339	0.86	0.508894	0.045517
political	2.261380	0.84	0.764692	0.066936
case	2.206641	0.83	0.761988	0.069614
court	2.110550	0.85	0.753877	0.077646
together	1.970650	0.81	0.498078	0.056226
part	1.736603	0.78	0.487263	0.066936
three	0.837934	0.68	0.427780	0.125840
were	0.800100	0.59	0.656540	0.174034
also	0.658376	0.60	0.422372	0.131195
these	0.525725	0.66	0.398038	0.155292
but	-0.478429	0.47	0.314220	0.238293
an	-0.766620	0.30	0.671701	0.433747
from	-1.536841	0.18	0.601402	0.503360
by	-2.715982	0.10	0.548968	0.830009
which	-3.039982	0.14	0.210413	0.615813
it	-3.216353	0.23	0.081693	0.468553
with	-3.230189	0.11	0.218525	0.607781
for	-3.839087	0.03	0.691207	0.963881
and	–	0.0	2.259603	2.158023
of	–	0.0	2.210549	4.404402
a	–	0.0	0.183472	2.016118

Expert translation, text 1:

In the ***Marseille Facet*** of the ***Urba-Gracco Affair***, Messrs. ***Emmanuelli***, ***Laignel***, ***Pezet***, and ***Sanmarco Confronted*** by the ***Former Officials*** of the ***SP Research Department***

On ***Wednesday***, February 9, the ***presiding judge*** of the ***Court*** of ***Criminal Appeals*** of ***Lyon***, ***Henri*** Blondet, charged with investigating the ***Marseille facet*** of the ***Urba-Gracco affair***, proceeded with an extensive ***confrontation*** among several ***Socialist deputies*** and ***former directors*** of ***Urba-Gracco***. Ten persons, including ***Henri Emmanuelli*** and Andre ***Laignel***, ***former*** treasurers of the ***SP***, Michel ***Pezet***, and Philippe ***Sanmarco***, ***former deputies*** (***SP***) from the Bouches-du-Rhône, took ***part*** in a ***hearing*** which lasted more than seven hours.

Besides these ***political*** personalities, three ***former*** Urba ***directors***, Gérard Monate, chairman and managing ***director*** of Urbatechnic, Joseph Delcroix (editor of the "journals" detailing the internal operation of this exceptional ***research department***), and Bruno Desjoberts, ***director*** of the ***Marseille*** regional delegation, participated in this confrontational ***hearing***, which also ***brought together*** Bernard Pigamo, ***former*** campaign ***director*** for Mr. ***Pezet*** and ***director*** for "supporting associations" and a company head. All were ***questioned*** as ***part*** of a ***case*** bearing on acts of bribery, influence peddling, ***forgeries*** and the use of ***forgeries***, and complicity in, or concealment of, these major crimes.

Questions and answers turned mainly on the relationship and the operating methods implemented between ***Urba-Gracco*** and the ***Socialist*** Party. It was an opportunity for the examining magistrate to go further toward illuminating an organized financing system, since local decision makers and national ***political officials***, but also beneficiaries and intermediaries for sums paid by many ***companies*** were ***confronted*** with each other. The thirty-eight heads of ***companies questioned*** in the ***case*** had already been heard, but three of them were ***brought together Wednesday*** following the "***political***" ***confrontation***.

The ***presiding judge*** of the ***Court*** of ***Criminal Appeals*** is to render a closing opinion, thus establishing a twenty-day deadline for requests from the various parties, followed by a "may it be communicated" order for settlement of the ***case*** by the ***Lyon*** public prosecutor's office. Considering the thickness of the file, which results from a long procedural battle in the ***Court*** of ***Appeals*** and the Council of State, initiated by an ecologist deputy from ***Marseille***, a trial is not foreseen before 1995.

Table 1: expert translation of Text 1 and word list

$S_{word[text]}$ is computed for all words with a positive difference $P_{word[text]} - P_{word[rest-corp]}$. However, many function words also receive this score simply due to the fact that their frequency in a particular text happened to be somewhat higher than their general frequency in the rest of the corpus. So, for comparing statistical models of different MT systems, we established a threshold – $S_{word[text]} > 1$. This threshold separates content words and function words rather accurately, and words just above the threshold ("part" and "together" in the above example) are general "low-content" open-class words. The words with $S_{[word][text1]} > 1$ are highlighted in the text.

2. In the second stage, the lists of statistically significant words for corresponding texts together with their $S_{word[text]}$ scores are compared across different MT systems. Comparison is done in the following way:

For all words which are present in lists of statistically significant words both in the human reference translation and in the MT output, we compute the sum of changes of their $S_{word[text]}$ scores:

$$S_{text.diff} = \sum \left| \left(S_{word[text.reference]} - S_{word[text.MT]} \right) \right|$$

The score $S_{text.diff}$ is added to the scores of all "over-generated" words (words that do not appear in the list of statistically significant words for human reference translation, but are present in such list for MT output). The resulting score becomes the general "over-generation" score for this particular text:

$$S_{over-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.over-generated[text]}$$

The opposite "under-generation" score for each text in the corpus is computed by adding $S_{text.dif}$ and all $S_{word[text]}$ scores of "under-generated" words – words present in the human reference translation, but absent from the MT output.

$$S_{under-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.undergenerated[text]}$$

It is more convenient to use inverted scores, which increases as the MT system improves. These scores, $S_{o.text}$ and $S_{u.text}$, could be interpreted as scores for ability to avoid "over-generation" and "under-generation" of statistically significant words. The combined (o&u) score is computed similarly to the F-measure, where Precision and Recall are equally important:

$$S_{o.text} = \frac{1}{S_{over-generation.text}}; \quad S_{u.text} = \frac{1}{S_{under-generation.text}}; \quad S_{o\&u.text} = \frac{2 S_{o.text} S_{u.text}}{S_{o.text} + S_{u.text}}$$

The number of statistically significant words could be different in each text, so in order to make the scores compatible across texts we compute the average over-generation and under-generation scores per each statistically significant word in a given text. For the o_{text} score we divide $S_{o.text}$ by the number of statistically significant words in the MT text, for the u_{text} score we divide $S_{u.text}$ by the number of statistically significant words in the human (reference) translation:

$$o_{text} = \frac{S_{o.text}}{n_{statSignWordsInMT}}; \quad u_{text} = \frac{S_{u.text}}{n_{statSignWordsInHT}}; \quad u \,\&\, o_{text} = \frac{2 o_{text} u_{text}}{o_{text} + u_{text}}$$

The general performance of an MT system for IE tasks could be characterised by the average o-score, u-score and u&o-score for all texts in the corpus.

The use of contrasting statistical models for human translation and MT output is illustrated by the following example in Table 2:

MT Reverso;	**"Expert"human translation**
Overgenerated words: motor, 4,565274; obligation, 4,565274; tires, 4,565274; debts, 3,841254; global, 3,404379; 12th, 3,255370; actions, 3,234316; franc, 2,839973; order, 2,829043; first, 1,042027	Undergenerated words: tire, 4,564768; automobile, 4,143929; fiscal, 4,143929; bonds, 3,840742; stock, 3,612322; reduce, 3,601959; debt, 3,403861; six, 2,839444; 12; 2,817465; amount, 2,716706; per, 2,657005; rates, 2,448991; itself, 2,128073; total, 2,068308; months, 1,956732; beginning, 1,745085; any, 1,297940; can, 1,294282
To reduce the cost of its debt Michelin throws a bond issue for 3,5 billion francs	To Reduce The Cost of Its Debt, Michelin Is Launching a Bond Issue for 3.5 Billion Francs
Michelin decided to proceed, from Wednesday, January ***12th***, to a bond issue convertible into 3,5 billion ***franc actions***. The ***first*** world manufacturer of tyres so intends to relieve his short-term ***debts***, while bringing him capital necessary for his recovery in the middle of a crisis of the European ***motor*** market. This broadcast will be opened to the public on January ***12th*** at the 255-***franc*** price the ***obligation*** and will concern 9 445 700 titles. His annual interest rate will be 2,5 % and its rate of return actuariel raw product of 5,03 % in case of non-conversion. Of a duration of six years, eleven months and a day, he will be quoted in the Paris Stock Exchange.	Michelin has decided to begin issuing, ***beginning*** Wednesday, January ***12***, an issue ***bonds*** convertible into ***stock*** in the ***amount*** of 3.5 billion francs. In this way, the world's leading ***tire*** manufacturer wants to ***reduce*** its short-term ***debt*** while bringing in the capital needed to recover from the full-blown European ***automobile*** market crisis. This issue will be open to the public on January ***12*** at the price of 255 francs ***per*** bond, and will involve 9,445,700 ***bonds***. Its annual interest rate will be 2.5% and its gross actuarial yield rate will be 5.03% in the event of non-conversion. The issue will have a maturity period of six years, eleven months and one day and will be quoted on the Paris Stock Exchange.
According to Michelin, the conversion, at the rate of an action for an ***obligation***, can be made at any time from February 2nd, 1994. The loan will be altogether paid off itself on January 1st, 2001 at the 307-***franc*** price. A priority period of signature will be reserved for the shareholders, inclusive from 12 till 21 January, at the rate of an ***obligation*** for fifteen ***actions***.	According to Michelin, the conversion, at a rate of one share ***per*** bond can be made at any time ***beginning*** February 2, 1994. The loan itself will be repaid in full as of January 1, 2001 at the price of 307 francs. A subscription-priority period will be reserved for shareholders from January 12 through January 21, at the rate of one bond for fifteen shares.
This operation is going to allow Michelin not to weigh down too much its interest charges in this period of high interest rates, from which particularly suffered the clermontoise firm. A strong part of its debts, a ***global*** 30 billion ***franc*** amount,	This operation will enable Michelin to avoid burdening ***itself*** with finance costs during this period of high interest rates, which have hit the Clermont firm particularly hard. A large proportion of debt, in the ***total*** amount of 30 billion francs,

was it indeed with loans with floating interest rate.	was in fact borrowed at floating interest ***rates***.
Especially since Michelin can hardly count on the European ***motor*** market to raise its accounts. His losses amounted to 3,45 billion francs in the ***first*** half of the year and should border the 4 billion francs for the fiscal year 1993, according to certain analysts. This result succeeds three negative exercises (11 million from francs to 1992, 1 billion in 1991 and 5,3 billion francs in 1990), in spite of two recovery packages ending in more than 30 000 abolitions of employments on a ***global*** strength of the ***order*** of 125 000 persons.	Especially since Michelin can no hardly count ***any*** longer on the European ***automobile*** market to rehabilitate its books. Its losses rose to 3.45 billion francs for the ***first six months*** and should approach 4 billion francs for fiscal year 1993, according to some analysts. This result follows three negative ***fiscal*** years (11 million francs in 1992, 1 billion in 1991, and 5.3 billion in 1990), despite two recovery plans ending with the elimination of 30,000 jobs cut out of a ***total*** work force of approximately 125,000 persons.
In 1993, both the market of the ***tires*** of ***first*** horsemanship (for the new cars) and that of the ***tires*** of replacement collapsed in Europe. In the United States, where Michelin is very present thanks to the acquisition in April, 1990 of Uniroyal-Goodrich, the resumption, of the ***order*** of 8 %, was not sufficient to compensate for the European fall.	In 1993, both the new car ***tire*** and the ***tire*** replacement markets collapsed in Europe. In the United States, where Michelin has a strong presence because of its acquisition of Uniroyal-Goodrich in April 1990, the recovery, about 8%, was not enough to offset the European decline.

o_{text} = 0.612915 u_{text}=0.585990; $u\&o_{text}$ = 0.599452

Table 2: Overgenerated and undergenerated statistically significant words in texts

The words highlighted in Table 2 are different for MT output and for human translation. In many cases these differences signal important problems in lexical well-formedness of the MT output which are related to word sense disambiguation or to necessary lexical transformations in the target text, e.g.:

(4)	French original:	*marché automobile européen*
	Human translation:	"*European automobile market*"
	Machine translation:	"*European motor market*"
(5)	French original:	*une obligation pour quinze actions*
	Human translation:	"*one bond for fifteen shares*"
	Machine translation:	"*an obligation for fifteen actions*"
(6)	French original:	*Ce résultat succède a trois exercices négatifs*
	Human translation:	"*This result follows three negative fiscal years* "
	Machine translation:	"*This result succeeds three negative exercises*"
(7)	French original:	*sur un effectif global*
	Human translation:	"*out of a total work force*"
	Machine translation:	"*on a global strength* "
(8)	French original:	*le marché des pneus de première monte (pour les voitures neuves) que celui des pneus de remplacement*
	Human translation:	"*the new car tire and the tire replacement markets* "
	Machine translation:	"*the market of the tires of first horsemanship (for the new cars) and that of the tires of replacement*"

(Only statistically significant words are underlined). Differences in the statistical models of aligned MT output and human translation allow us to spot most serious factual mistakes automatically, and so improve an aspect of MT that is crucial for the performance of IE systems.

Note however, that the proposed scores could go beyond the range [0, 1], which makes them different from precision/ recall scores.

3. Results of MT evaluation based on statistical modelling

MT evaluation was performed using both human translations as a reference. But to have a complete picture, we also compared MT systems with each other, making each of them a reference system in turn. The results of comparing average scores for each of the MT systems and for "reference" and "expert" human translations are presented in Table 3.

It can be seen from the table that scores for human "expert" translation are the best in relation to the other human translation – the "reference" translation. Scores for MT systems are substantially lower, which reflects the fact that they produce many more cases of lexical "under-generation" and "over-generation" of statistically significant words.

Note, that the proposed metrics measure only one aspect of MT, which we consider important for IE purposes, in particular – semantic appropriateness in translations of statistically significant words. We do not measure any other aspects, e.g, syntactic well-formedness. However, a correlation could be noted between our metrics and some other MT evaluation measures. The best match has been found between our o-score (the score for avoiding lexical over-generation) and the adequacy scores in DARPA94 MT evaluation (Table 4). This close match could be interpreted as a fact that translation adequacy always involves avoiding over-generation: it requires that there were no "incorrect" or "misleading" meanings in translation.

There is also match between ranking of MT systems according to our "u&o" combined score, and the ranking produced by DARPA94 fluency measures, even though the extent of divergences in scores is different (Table 5). Because of this, it is not yet clear if the combined u&o metric could be interpreted as "lexical fluency".

	HT ref	**HT expert**	**MT systran**	**MT reverso**	**MT candide**	**MT ms**	**MT globalink**
HT ref		u=0.951 *o=0.957* **uo=0.954**	u=0.786 *o=0.763* **uo=0.774**	u=0.727 *o=0.714* **uo=0.721**	u=0.800 *o=0.629* **uo=0.714**	u=0.715 *o=0.699* **uo=0.707**	u=0.675 *o=0.651* **uo=0.663**
HT expert	u=0.957 o=0.951 **uo=0.954**		u=0.776 o=0.752 **uo=0.764**	u=0.719 o=0.707 **uo=0.713**	u=0.811 o=0.634 **uo=0.723**	u=0.693 o=0.677 **uo=0.685**	u=0.677 o=0.651 **uo=0.664**
MT systran	u=0.763 o=0.786 **uo=0.774**	u=0.752 o=0.776 **uo=0.764**		u=0.931 o=0.940 **uo=0.936**	u=0.824 o=0.659 **uo=0.742**	u=0.852 o=0.865 **uo=0.859**	u=0.902 o=0.879 **uo=0.891**
MT reverso	u=0.714 o=0.727 **uo=0.721**	u=0.707 o=0.719 **uo=0.713**	u=0.940 o=0.931 **uo=0.936**		u=0.764 o=0.619 **uo=0.692**	u=0.833 o=0.837 **uo=0.835**	u=0.835 o=0.809 **uo=0.822**
MT candide	u=0.629 o=0.800 **uo=0.714**	u=0.634 o=0.811 **uo=0.723**	u=0.659 o=0.824 **uo=0.742**	u=0.619 o=0.764 **uo=0.692**		u=0.621 o=0.761 **uo=0.691**	u=0.608 o=0.732 **uo=0.670**
MT ms	u=0.699 o=0.715 **uo=0.707**	u=0.677 o=0.693 **uo=0.685**	u=0.865 o=0.852 **uo=0.859**	u=0.837 o=0.833 **uo=0.835**	u=0.761 o=0.621 **uo=0.691**		u=0.784 o=0.764 **uo=0.774**
MT globalink	u=0.651 o=0.675 **uo=0.663**	u=0.651 o=0.677 **uo=0.664**	U=0.879 o=0.902 **uo=0.891**	u=0.809 o=0.835 **uo=0.822**	u=0.732 o=0.608 **uo=0.670**	u=0.764 o=0.784 **uo=0.774**	
DARPA scores		I=0.795 *A=0.920* **F=0.850**	I=0.758 *A=0.789* **F=0.508**	I=NA A=NA F=NA	I=0.638 *A=0.677* **F=0.454**	I=0.663 *A=0.718* **F=0.382**	I=0.747 *A=0.710* **F=0.381**

Table 3: MT evaluation scores for statistically significant words

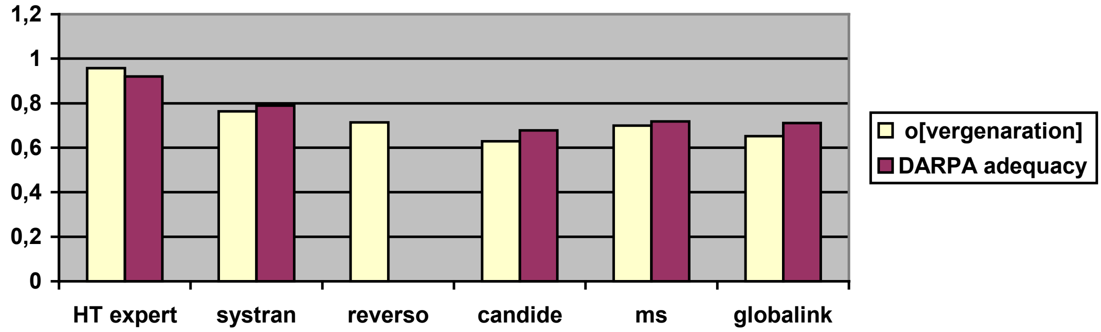

Table 4: o-scores and DARPA 94 adequacy scores

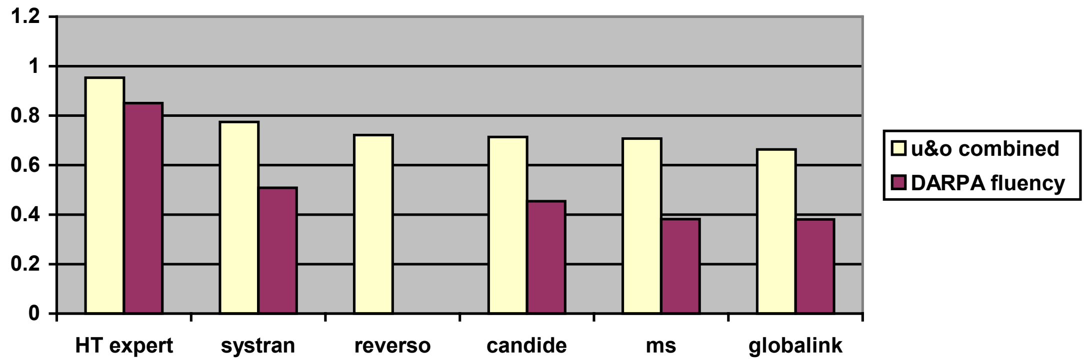

Table 5: combined u&o-scores and DARPA 94 Fluency scores

No correlation has been found between our u-score and any of the DARPA 94 human evaluation scores, and between the DARPA 94 "informativeness" score and any of our evaluation scores.

Several systems have a better "u&o" combined scores in relation to "reference" translation than in relation to "expert" translation. This might be due to the fact that the quality of the human "reference" translation is lower than that of the "expert" translation, so "reference" contains more cases of literal translation that better match MT output.

The exception to this rule is "Candide", which has a better u&o combined score for the "expert" translation. It also for some reason has a very high u-score, and considerably lower o-score. Such exceptionality of "Candide" can be explained by the fact that this system implements the IBM statistical approach to MT (Berger et al., 1994), and (as it might be expected) produces a substantially different output, partially determined by the statistical structure of the target language. Our analysis allows us to see that the IBM statistical approach does not really improve the score for "avoiding over-generation", which has been found to closely match the DARPA "Adequacy" score. Instead, it considerably improves the score for "avoiding under-generation", which does not directly correspond to any of the DARPA evaluation scores (it influences the combined u&o score, which has been found to match (to some extent) the DARPA "Fluency" score, but more work needs to be done to determine if it really correspond to any important aspect in the quality of MT).

This observation provides additional evidence for the suggestion made in (Wilks, 1994) that there are fundamental limits for improving pure statistically-based systems: "Candide" showed lowest scores for "avoiding over-generation of statistically significant words" among all tested MT systems. Over-generation and possibly other "precision-based" measures seem to be the weakest point for statistical MT. At the same time the measure of translation adequacy (which is found to be related to our "over-generation" scores) is considered to be the most important aspect of the translation quality in general.

4. Conclusion

We have investigated a word-significance measure S which compares word frequency within the current text against frequency across the rest of the corpus; by setting a suitable threshold, S>1, we can eliminate high-frequency function words, leaving significant content words which characterise the text. A comparison of words flagged by this S metric in MT output and human translation highlights factual mistakes. Statistical modelling of MT output corpora has shown substantial differences in distribution of significant words with respect to human translation, which imply that the usability of MT systems for IE technology is still substantially limited. However, the suggested evaluation methodology also allows us to highlight the problems of MT which might be important for the IE task, if MT output is to be used for template filling or acquiring templates automatically. It might also help developers of the state-of-the-art MT systems to identify specific problems relevant for preserving factual information in MT. We proposed measures of lexical match for statistically significant words, and found that these correlate to DARPA MT evaluation measure of "adequacy". This should allow prediction of the degree to which particular MT systems might be usable for IE tasks.

Future work will look at the problem of investigating stochastic models for the output of example-based MT systems, and comparing them with models for traditional knowledge-based applications and statistical MT. This could provide insights to establishing the formal properties of intuitive judgements about translation equivalence, adequacy and fluency both for human translation and for MT, and to investigating possible limits on improving MT quality with certain methodologies.

Other prospective directions of research would be investigating the actual performance of different modules of IE system (such as named entity recognition, template element filling and scenario template filling, summary generation) which use MT output of different quality. We will try to establish if this performance actually correlates with MT evaluation measures proposed in this paper and with other metrics proposed previously.

References

Berger A, Brown P, Cocke J, Pietra S, Pietra V, Gillett J, Lafferty J, Mercer R, Printz H, Ures L 1994 The Candide system for Machine Translation. *Proceedings of the ARPA workshop on Human Language Technology*. San Mateo, Morgan Kaufmann. pp. 152-157.

Collier R 1998 Automatic template creation for information extraction. PhD thesis. UK.

Cunningham H, WilksY, Gaizauskas R 1996 GATE -- a General Architecture for Text Engineering. *Proceedings of the 16th Conference on Computational Linguistics (COLING-96),* Copenhagen, Aug, 1996.

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

Stevenson M, Wilks Y 2001 The integration of knowledge sources in word sence disambiguation. *Computational Linguistics* 27(3):321-349.

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

Wilks Y 1994 Developments in MT research in the US. *Aslib Proceedings,* vol.46, no.4, April 1994. pp.111-116.

Wilks Y, Catizone R 1999 Can we make information extraction more adaptive? In M. Pazienza (ed.) *Proceedings of the SCIE99 Workshop,* Springer-Verlag, Berlin. Rome.