Original scientific paper

Evaluation of the Statistical Machine Translation

Marija Brkić
Department of Informatics, University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
mbrkic@uniri.hr

Tomislav Vičić
Freelance teacher of economics and translator
Zagreb, Croatia
ssimonsays@gmail.com

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Summary

Much thought has been given in an endeavour to formalize the translation process. As a result, various approaches to MT (machine translation) were taken. With the exception of statistical translation, all approaches require co-operation between language and computer science experts. Most of the models use various hybrid approaches. Statistical translation approach is completely language independent if we disregard the fact that it requires huge parallel corpus that needs to be split into sentences and words. This paper compares and discusses state-of-the-art statistical machine translation (SMT) models and evaluation methods. Results of statistically-based Google Translate tool for Croatian-English translations are presented and multilevel analysis is given. Three different types of texts are manually evaluated and results are analysed by the χ2-test.

Key words: SMT (statistical machine translation), online, Google Translate, MT, Croatian-English, manual evaluation, fluency, adequacy, χ2-test

Introduction

The translation process is conducted differently by different translators and results are, therefore, not uniform (MT Marathon, 2008). As Knight (2003) points

out, the non-existence of right answers in translation does not imply the non-existence of wrong answers.
MT was first conceived as a technology that significantly speeds up the translation process and offers human-like quality translations (Valderrábanos, 2003). Nowadays, it is seen as a tool of limited use. Current computational models of MT can address a number of non-literary translation tasks, like tasks for which a rough translation is adequate, tasks where a human post-editor is needed and tasks limited to sub-domains in which fully automatic high quality translation is achievable (FAHQT) (Jurafsky & Martin, 2009).
The basic idea behind the development of MT (compared to the human translation) is to find a way for busting up speed while reducing the cost of the translation process (i.e. removing human component as much as possible) (Awatef, 2005). Further development focuses itself on the precision and overall quality of the output.
So far, the MT has been directly associated with (and mostly restricted to) the translation of the written language. This is probably due to the fact that most of contemporary communication (legal, commercial, Internet and so forth) is in written form and still too often on paper.
There are various approaches to MT, such as word-for-word translation, syntactic transfer, interlingual approaches, controlled language, example-based translation, and SMT (MT Marathon, 2008). SMT has low development cost and it is portable across languages (Valderrábanos, 2003). The only requirement SMT imposes is a large parallel corpus.
The paper explores the development of MT. A particular attention is paid to the Google Translate system, which exemplifies SMT. The system is tested for Croatian-English language pair. MT systems need to be evaluated in order to be ranked. For that purpose, different evaluation methods are introduced and the results of a conducted manual evaluation method are given and discussed.

MT

Although most of the MT approaches integrate different methods (e.g. integration of statistical MT and syntactic transfer, or example-based MT with rule-based method), basic approaches in MT are, according to Hutchins[1] the following: "syntactic transfer", "example-based" and "statistical systems".

Syntactic transfer

Syntactic transfer approach applies linguistic rules to some extent, analyzing source text and creating translated text accordingly, involving some variety of intermediary linguistic representation, with morphological, syntactic and semantic analysis (Lavie, 2006). Since the 1980s, many new operational MT sys-

[1] Hutchins, J.: *Machine Translation: past, present, future* (Ellis Horwood, UK, 1986)

tems appeared and included this approach: the French multilingual system TITUS; the Chinese-English CULT system; the Spanish-English SPANAM; the Russian-English system Systran which was adopted by the US Air Force and the European Community; the System of Logos Corporation. (Awatef, 2005) In Europe, the Commission of the European Communities (CEC) supported a lot of work on the English-French version of the Systran. In Germany it was SUSY (Saarbrucker Ubersetzungssystem), the French-German System (ASCOF) and (SEMSYN) for the translation of Japanese scientific articles into German. A more ambitious and reputable system developed in this era is the EUROTRA project of the European Communities, which aimed at developing multilingual transfer system for translating among all the Community languages. In the 1980s, according to Hutchins (1992), Japan maintained the greatest commercial activity where most computer companies developed software for computer-aided translation mainly for the Japanese-English market. According to WTEC Hyper Librarian (1994), MT in Japan is viewed as an "important strategic technology that is expected to lay a key role in Japan's increasing participation in the world economy". The most sophisticated commercially available system was METAL, a German-English system, which originated from the research at the University of Texas at Austin and supported by Siemens, which obtained commercial rights for marketing it (Lehmann 2000: 162).

Example-based translation

Example-based MT was first suggested by Nagao Makoto in 1984[2]. He suggested the method which may be called *MT by example-guided inference* or MT by the analogy principle. One of the strong reasons for this approach has been that the detailed analysis of a source language sentence is of no use for the translation between languages that have completely different structure (for example, English and Japanese). In this approach, the translation unit is a block of words. This is accomplished by storing *varieties of example sentences in the dictionary* and deploying a mechanism for finding analogical example sentences.

The process of mechanical translation by analogy is time-consuming in its primary structure. Therefore, the process is divided into substages and the system

[2] "Problems inherent in current MT systems are shown to be inherently inconsistent. The present paper defines a model based on a series of human language processing and in particular the use of analogical thinking. Machine translation systems developed so far have a kind of inherent contradiction in themselves. The more detailed a system has become by the additional improvements, the clearer the limitation and the boundary will be for the translation ability. To break through this difficulty we have to think about the mechanism of human translation, and have to build a model based on the fundamental function of language processing in the human brain. The following is an attempt to do this based on the ability of analogy finding in human beings." *in ARTIFICIAL AND HUMAN INTELLIGENCE (A. Elithorn and R. Banerji, editors). Elsevier Science Publishers. B.V., NATO, 1984*

is fed with all the information available in the initial system construction. The learning comes in only during the augmentation stage of the system, which mainly refers to the increase of example sentences and the improvement of the thesaurus (Nagao, 1984). Examples of this approach are translation memories, which are often integrated with language-dependant approach.

Statistically-based translation

Further development in MT took place in the 1990s as computers became more powerful and storage capacities much larger and cheaper. The new development shifts from syntactic transfer to what has been called "statistical approaches" with provenance from the "corpus linguistics". Statistical translation systems do not depend on underlying grammatical rules any longer. Statistically-based MT systems rely on statistical models whose parameters are derived from bilingual corpus.

Put very simply, as Farah (2003) put it in an article for the *New York Times* (reprinted in the *International Herald Tribune*), traditional MT relied heavily on bilingual programmers entering the vast wealth of information, needed by the computer, in the lexicon and syntax. A team from IBM in the 1990s tried to make the computer learn the second language by feeding a computer with English text and its translation in a different language, and then analyzing it statistically. The example given by Farah (2003) is revealing:

"Compare two simple phrases in Arabic: "rajl kabir" and "rajl tawil. If a computer knows that the first phrase means "big man" and the second means "tall man," the machine can compare the two and deduce that *rajl* means "man," while *kabir* and *tawil* mean "big" and "tall," respectively." Phrases like these, called N-grams (with "N" representing the number of terms in a given phrase), are the basic building blocks of SMT.

Mackin (2003), in an article interestingly entitled "Romancing the Rosetta Stone," reports on work on translation using statistical approaches. Mackin quotes the computer scientist Franz Joseph Och boasting: "Give me enough parallel data, and you can have a translation system in hours." The new approach for translation uses huge volumes of "matched bilingual texts" which are the encoded equivalents. Och (Makin 2003) asserts that the new approach uses statistical models to find "the *most likely* translation for a given input." The new approach ignores explicit grammatical rules and traditional dictionary lists of the lexicon in order to have the computer itself match up patterns between original texts and translations. Och's work (Makin 2003) is an improvement of the earlier work on the statistical approach that started back in the late 1980s and early 1990s by Peter F. Brown and his colleagues at IBM's Watson Research Center.

Statistical approach to MT tackles the MT problem by finding the maximum likelihood solution (Watanabe & Sumita, 2002). According to Wang and Waibel (1997), SMT systems deal with the following problems:

- the modelling problem (in order to create language and translation models, with problems involving idioms, compounds, morphology and different word order),
- the learning problem (in order to estimate parameters from bilingual corpora), and
- the decoding problem (which essentially comes down to finding an efficient way of searching for a target language sentence).

SMT systems produce a general model of the translation process. Specific rules are acquired automatically from bilingual and monolingual text corpora. Although all of these systems share the same underlying principle, they differ in the structures and sources of their translation models.

In a *word-based approach*, words are treated like tokens, independently from other words. This poor handling of morphology is one of the major drawbacks of this approach. A word-based system may recognize one form of a word, but not the other form of the same word (MT Marathon, 2008). This is particularly apparent with morphologically rich languages, as shall be seen in our study. IBM models 1-5 fall into this category (Koehn, Och, & Marcu, 2003).

Nowadays, many systems implement *phrase-based models*. What differentiates them from word-based models is a lexicon, which is not single-word-based, but phrase-based (Och & Ney, 2004). In addition, phrase length should not exceed three words (Koehn, et al., 2003). Phrase-based models translate small word sequences at a time and do not use explicit syntactic or morphological information (MT Marathon, 2008). Moreover, as Koehn, et al. (2003) report, imposing syntactic restrictions on phrases does not lead to better system performance. The number of useful phrases grows with the size of the training corpora. Log-linear models are variations to a standard model. However, since phrase-based models cannot model grammaticality and long-distance dependencies, they are not suitable for large-scale restructuring of sentences. Furthermore, they cannot generalize.

Syntax-based models can be classified according to the underlying syntactic formalism. Representatives of this approach, tree-based models, have proved to have performance comparable to phrase-based models (MT Marathon, 2008).

Bearing in mind that SMT is language-independent and that existent language resources are sparse, moderate results should be expected. According to Sepesy Maucec and Kacic (2007), a hybrid approach, which combines SMT with rule-based MT, would presumably give much better results.

Evaluation

MT evaluation is not a straightforward task. Different translators translate the very same sentence differently. Evaluation methods can be manual or automatic. Nevertheless, both categories are extremely subjective (Jurafsky & Martin, 2009).

The correlation between two metrics is usually computed using the Pearson correlation coefficient in (1), whereas sample means and variances are expressed in (2) and (3), respectively (MT Marathon, 2008).

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}. \quad (1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (2)$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2. \quad (3)$$

Manual evaluation

Unfortunately, bilingual evaluators, which are best suited for manual evaluation task, are not always available. If that is the case, monolingual target language speaking evaluators are given reference translations and employed for the task. In this case study, two criteria are taken into consideration:

- *fluency,* which refers to grammaticality and word choices (MT Marathon, 2008); according to Jurafsky and Martin (2009), there are three aspects of *fluency – clarity*, *naturalness* and *style*, and
- *adequacy*, which, on the other hand, questions whether any part of a message is lost, added, or distorted; Jurafsky and Martin (2009) group *adequacy* and *informativeness* into another dimension – *fidelity*.

In manual evaluation task, evaluators are asked to score output on a 1-5 scale according to both criteria. It is advisable that evaluators read the output prior to reading the reference translation, because human mind tends to fill in the missing information if reference translation is read first or evaluators are acquainted with the domain. Judgements of *fluency* and *adequacy* are usually related, which either points to the difficulty in distinguishing the two criteria or just to the fact that ungrammatical sentences and wrong word choices carry less meaning (MT Marathon, 2008).

Besides the described procedure for measuring *fluency* and *adequacy*, *fluency* can also be measured through the time needed for reading the translation (Jurafsky & Martin, 2009) or through cloze test (Taylor, 1953, 1957 in Jurafsky & Marin, 2009).

Furthermore, described dimensions can be measured through the edit cost of post-editing the MT output into a satisfying translation. This can be done on word-level, time-level or keystrokes-level (Jurafsky & Martin, 2009).

Hajič, Homola, and Kuboň (2003) present a way of exploiting TM (translation

memory) tools for MT manual evaluation. A TM is created by aligning source text and corresponding MT output. The source text is then translated by a human translator, and with the aid of the newly-built TM. Finally, MT system is used to determine the percentage of similarity between the MT output and the human translation of the same sentence (reference translation), which is stored in the TM.
Evaluation procedure is of crucial importance in comparing different translation models. Manual evaluation methods are too expensive and time-consuming (Papineni et al., 2001). Hence, automatic evaluation methods are needed.

Automatic evaluation
All automatic evaluation metrics use one or more reference translations. These reference translations are used for comparison with MT output or candidate translations (MT Marathon, 2008). Automatic method is considered to be better if it has higher degree of correlation with human judgements. There are a number of automatic methods, such as Bilingual Evaluation Understudy (BLEU), NIST, TER, Precision and Recall, and METEOR. Although they differ in the way they measure similarity, they all rank better the candidate translation which is closer to human translation (Jurafsky & Martin, 2009).

Experimental study

Google Translate Service

The tool Google Translate is chosen in this case study for two basic reasons: it is statistically-based and only of a kind that offers Croatian as one of the languages in the translation pairs. Furthermore, Google developed its own statistical software for translation. According to Och (now head of Google MT department), a solid base for the development of a usable SMT system for a new language pair from scratch, would consist in having a bilingual text corpus (or parallel collection) of more than a million words and two monolingual corpora of each more than a billion words. Statistical models built from this data would then used for translating between those languages.
Google acquired the initial amount of linguistic data from United Nations' documents, which are available in six official UN languages (Arabic, Chinese, English, French, Russian and Spanish). To quote Google: "Our system takes a different approach: we feed the computer billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model. We've achieved very good results in research evaluations."[3]
This service now (2009) offers following languages for bidirectional translation

[3] http://www.google.com/intl/en/help/faq_translation.html#statmt

(alphabetically): Arabic, Bulgarian, Catalan, Chinese (Simplified), Chinese (Traditional), *Croatian*, Czech, Danish, Dutch, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian and Vietnamese.

Croatian-English Language Pair

The translation process is hindered by the fact that languages involved often differ culturally, stylistically, syntactically, and lexically. These differences are called translation divergences and they can be, according to (Jurafsky & Martin, 2009):

- systematic,
- idiosyncratic, and
- lexical.

While *systematic* differences can be modelled in a general way, idiosyncratic and lexical differences must be dealt one by one (Jurafsky & Martin, 2009). Croatian language is essentially very different from English.

Croatian, in comparison to English, has relatively free word order, does not have articles and uses fewer pronouns. Languages which omit pronouns are called pro-drop, referentially sparse or cold languages because they require the hearer to do more inferential work to recover antecedents. Translating from pro-drop languages into non-pro-drop languages is exhaustive because each zero has to be identified and anaphor recovered.

Idiosyncratic differences also have to be tackled for the translation process to succeed. For example, 'existential *there*' is the name of an English idiosyncratic construction used to introduce a new scene (Jurafsky & Martin, 2009). Croatian does not have a similar construction.

Finally, there are *lexical divergences* which further complicate the translation process. Besides difficulties in disambiguating homonymous and polysemous expressions, divergences can also be grammatical. For example, part-of-speech (POS) tags between source words and corresponding target words do not have to overlap. Another divergence is that Croatian marks gender, number and case on adjectives, while English does not. Nevertheless, one of the languages may have a lexical gap (the meaning of a word or phrase cannot be conveyed in another language because there is no corresponding word or phrase) (Jurafsky & Martin, 2009).

Examples and translations

Croatian-English translation is done on three different types of texts:

- text on corpus linguistics, annotation and research methods,
- text on small, medium and large enterprises and Government's plan for reform, and

- text on purchasing washing machine.

The Croatian texts and the reference texts, i. e. English translations are taken from the Internet and used without any modifications.

Comparison and analysis

The task in these examples was to compare human and MTs from Croatian to English, using Google Translate service. Source texts on Croatian and reference translations on English, taken from the Internet[4], had no restrictions for use and have not been modified in any way.

The comparison and analysis of translations has been done on lexical, morphological, syntactic and semantic level. The usage of punctuation marks has also been analysed.

On the *lexical level* (i.e. wrong translation / misuse of words), the lack of translation indicates that the system does not "recognize" single words, even repeatedly used, although these words are internationalisms (e.g. *leksičko, inherentno, kontekstno*). These untranslated units are called zerotones according to Sepesy Maucec and Kacic in (2007).

The usage of "not appropriate" words in the translation (i.e. synonyms or words that do not warp the meaning) does not significantly affect intelligibility (e.g. *rewriting* instead of *processing*, *certain* instead of *determined by* or *stipulated by*), since the rest of the translation provides an understandable message. There is also an issue with personal pronouns (*he* instead of *it)* or expressions (*great body* instead of *large corpora*). On the other hand, if something is not translated and cannot be "deducted" from the similarity in expression (for example, in a language the user does not speak at all), it can make the message undecodable, although a partial translation is available.

As for *syntax*, the word order in Croatian is relatively free, and in English is basically determined with the rule SVO. This order (along with formal structure) is also common in "bureaucratic languages". Therefore, it should not come as a surprise that the best results were achieved in texts 2 and 3, since Google Translate obtained its basic language corpora from the official documents of UN and EU. On the other hand, most mistakes are found in text 1, written in more of a "scientific language" (longer and somehow more complicated sentences).

Morphological analysis shows results similar to those of syntactic analysis. The most notable mistake is the frequent misuse of singular/plural. Some mistakes are due to the "odd" use of expressions (e.g. "environment friendly" should be used in general, but MT translated the original almost literally ("not burden the environment"). Another issue is the usage of cases in Croatian, which explains the lack of translation for "already translated" words.

[4] Respectively: (1.a) http://ling.unizd.hr/znanost/projekti/index.hr.html, (2.a) http://www.mingorp.hr/default.aspx?id=8 and (3.a) http://products.gorenje.si

On the *semantic* level (i.e. preservation of original message), Google MT shows some "effort", although in some cases the user has a lot of inferential work. It is also obvious that statistical MT lacks in taking context into account, which could significantly affect original message. The usage of *punctuation* marks is mainly taken over from the original text.

Manual evaluation

We employed manual evaluation method in order to obtain results which could later be used in evaluating automatic methods and determining their correlation with human judgements.

Six evaluators were kindly asked to score 21 machine-translated sentences according to a scale given in Table 1, and with regard to corresponding reference translations.

Table 1: *Fluency* and *adequacy* scale

	Fluency	Adequacy
1	incomprehensible	none
2	disfluent English	little meaning
3	non-native English	much meaning
4	good English	most meaning
5	flawless English	all meaning

Source: MT Marathon, 2008

The results are as follows. The average *fluency* judgement per judge ranges from 2.14 to 3.57, while the average *adequacy* judgement per judge ranges from 2.71 to 3.67. The average of a set of judgements is calculated according to the formula in (2). The averages are 2.98 for *fluency* and 3.36 for *adequacy*. The standard deviation of experimental data is calculated using the formula in (4), where n stands for the number of different values and n_i for the total frequency of each value. The standard deviation per question ranges from 0.52 to 1.03 for *fluency* and from 0.41 to 1.05 for *adequacy*, while standard deviation per judge according to the *fluency* criterion ranges from 0.60 to 1.06, and according to the *adequacy* criterion from 0.60 to 1.32.

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}\,. \tag{4}$$

We used the χ^2-test to determine whether there is a difference in the distribution of grade 3 among different evaluators for *fluency* and *adequacy* separately. We had to pool categories in the above mentioned way. Otherwise, one or more of the expected frequencies would fall below five, which would invalidate the chi-square test results. The χ^2 formula is given in (5), where O stands for observed

frequencies and *E* for expected frequencies (6). When χ^2 is used as a test of association, the expected frequencies are calculated directly from the observed frequencies by assuming independence between the categories. We applied the test to the data in tables 2 and 3.

$$\chi^2 = \sum \frac{(O-E)^2}{E}. \tag{5}$$

$$E = \frac{rowTotal \times columnTotal}{overallTotal}. \tag{6}$$

Table 2: Score frequencies according to *fluency* criteria

Fluency	Eval1	Eval2	Eval3	Eval4	Eval5	Eval6	Total
score 3	5	10	6	9	3	10	43
other scores	16	11	15	12	18	11	83
Total	21	21	21	21	21	21	126

Table 3: Score frequencies according to *adequacy* criteria

Fluency	Eval1	Eval2	Eval3	Eval4	Eval5	Eval6	Total
score 3	5	10	7	11	5	10	48
other scores	16	11	14	10	16	11	78
Total	21	21	21	21	21	21	126

The number of the degrees of freedom is 5 ((*rowTotal* - 1) x (*columnTotal* - 1)). The table value for the χ^2 with 5 degrees of freedom at the 5 per cent significance level is 11.070. We obtained χ^2 values for *fluency* and *adequacy*, 9.073 and 7.269 respectively. Since these values are smaller than the appropriate table value, we can conclude that the evaluators do not significantly differ in assigning score 3, neither for *fluency*, nor for *adequacy*. The same counts at the 1 per cent significance level because the appropriate table value is 9.236.

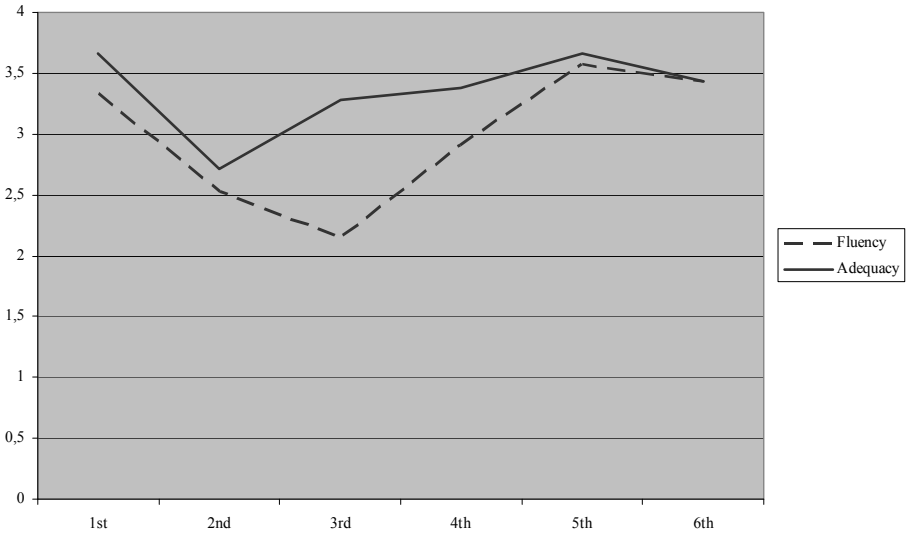
Table 4: Association of two criteria with regard to different evaluators

	Eval1	Eval2	Eval3	Eval4	Eval5	Eval6
χ^2	7.118	0.067	4.571	0.077	13	0.048

We performed the same test to see whether there is a significant difference in assigning scores for the two criteria per each evaluator applying the same pooling strategy. The results are shown in table 4. Since the table value with one degree of freedom at the 5 per cent significance level is 3.841, we may conclude that there is a significant difference in assigning *fluency* and *adequacy* scores for the first, third and fifth evaluator, while there is almost no difference for the remaining evaluators.

In general terms, *adequacy* scored slightly better than *fluency*, as evident in chart 1. Histograms of *adequacy* judgements show that different human evaluators use the scale 1-5 differently. Histograms of *fluency* judgements point to the same phenomenon.

Chart 1: *Fluency* and *adequacy* average judgements



Results of the language independent statistically-based MT service could be improved by the integration of the language-dependant module, which already exists for a number of languages. Human intervention in the post-editing step could certainly improve the output, although even the raw output, taken *cum grano salis,* could be useful and even usable for the basic information transfer and personal use.

Conclusion

In this case study the statistically-based MT service has been evaluated on the Croatian-English language pair. The results of the χ^2 test show that different evaluators do not significantly differ in assigning score 3, neither for *fluency*, nor for *adequacy*. Furthermore, the same test points that half of the evaluators find *fluency* and *adequacy* criteria to be closely related as far as grade 3 is concerned, while the other half of them can better distinguish between these criteria, and, therefore, rates them differently. In order to perform the χ^2 test, the pooling strategy had to be applied. This highlighted the need for the greater number of evaluators, and, accordingly, higher frequencies.

Since user expectations are of considerable importance (including education, intelligence, culture), it should be pointed out that one should be aware of MT

limitations and possibilities, even though SMT service could be improved by the integration of language-dependant module or by introducing the post-editing step.

References

Bhagat, Rahul; Ravichandran, Deepak. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. // *Proceedings of Association for Computational Linguistics (ACL).* OH, Columbus, 2008, 674-682

Brants, Thorsten; Popat, Ashok; Xu, Peng; Och, Franz, Dean, Jeffrey. Large Language Models in Machine Translation. // *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*. Czech Republic, Prague, 2007

Direct mailing with Mr. Josh Estelle, Google Translate, Google Inc.

Dovedan, Zdravko; Seljan, Sanja; Vučković, Kristina. Machine Translation as Help in the Communication Process. // *Informatologia*, vol. 4 (2002), 35, 283-291

Google Translate. http://translate.google.com/# (12.08.2009)

Hajič, Jan; Homola, Petr; Kuboň, Vladislav. A simple multilingual machine translation system. // *Proceedings of the MT Summit IX*. New Orleans, Louisiana, 2003

Hutchins, John; Somers, Harold. An Introduction to Machine Translation. UK, London : Academic Press, 1992

Jayaraman, Shyamsundar; Lavie, Alon. Multi-Engine Machine Translation Guided by Explicit Word Matching. // *10th Conference of the European Association for Machine Translation (EAMT)*, Hungary, Budapest, Hungary, 2005, 143-152

Jurafsky, Daniel; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. New Jersey : Pearson education, 2009

Knight, Kevin. Teaching Statistical Machine Translation. // *Proceedings of the MT Summit IX Workshop on Teaching Translation Technologies and Tools.* New Orleans, 2003, 17-19

Koehn, Philip; Och, Franz J.; Marcu, Daniel. Statistical Phrase-Based Translation. // *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1.* New York, Morristown : Association for Computational Linguistics, 2003, 48-54

Lin, Dekang, Wu, Xiaoyun. Phrase Clustering for Discriminative Learning. // *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Singapore, 2009, 1030-1038

Macherey, Wolfgang; Och, Franz J. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. // *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Czech Republic, Prague, 2007, 986-995.

Mohri, Mehryar. Statistical Natural Language Processing. In: M. Lothaire (Ed.), *Applied Combinatorics on Words*. Cambridge: Cambridge University Press, 2005

Nagao Magao, A framework of a mechanical translation between Japanese and English by analogy principle. // *Proceedings of the international NATO symposium on Artificial and human intelligence*. New York : Elsevier North-Holland, Inc., 1984, 173-180

Och, Franz J.; Ney, Hermann. The Alignment Template Approach to Statistical Machine Translation. // *Computational Linguistics.* 30 (2004), 4; 417-449

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), 2001

Pasca, Marius; Dienes, Peter. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. // *Proceedings of the 2nd International Joint Conference on Natural Language* Processing (IJCNLP-2005). Korea, Jeju Island, 2005, 119-130

Second Machine Translation Marathon, Lecture Notes, Germany : Berlin, 2008

Sepesy Maucec, Mirjam; Kacic, Zdravko. Statistical Machine Translation from Slovenian to English. // *Journal of Computing and Information Technology*. 15 (2007), 1; 47-59

Valderrábanos, Antonio S.; Esteban, José; Iraola, Luis. TransType2 - A New Paradigm for Translation Automation. // *Proceedings of the MT Summit IX*. New Orleans, 2003, 498-501

Wang, Ye-Yi; Waibel, Alex. Decoding Algorithm in Statistical Machine Translation. // *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* New York, Morristown : Association for Computational Linguistics, 1997, 366-372

Watanabe, Taro; Sumita, Eiichiro. Bidirectional Decoding for Statistical Machine Translation. // *Proceedings of the 19th international conference on Computational linguistics – Volume 1*. Taipei, Taiwan, 2002, 1-7

Zollmann, Andreas; Venugopal, Ashish; Och, Franz; Ponte, Jay. A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. // *Proceedings of 22nd International Conference on Computational Linguistics Coling*, Manchester 2008, 1145–1152

Zughoul, Muhammad R.; Abu-Alshaar, Awatef M. English/Arabic/English Machine Translation: A Historical Perspective. // *Journal des traducteurs / Meta: Translators' Journal*. 50 (2005), 3; 1022-1041