Evaluating machine translation for assimilation via a gap-filling task

Ekaterina Ageeva
School of Linguistics
Higher School of Economics
Moscow, Russia
evageeva_2@edu.hse.ru

Mikel L. Forcada
Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
mlf@dlsi.ua.es

Francis M. Tyers
HSL-fakultetet
UiT Norgga árktalaš universitehta
9017 Romsa, Norway
francis.tyers@uit.no

Juan Antonio Pérez-Ortiz
Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
japerez@dlsi.ua.es

Abstract

This paper provides additional observations on the viability of a strategy independently proposed in 2012 and 2013 for evaluation of machine translation (MT) for assimilation purposes. The strategy involves human evaluators, who are asked to restore keywords (to fill gaps) in reference translations. The evaluation method is applied to two language pairs, Basque–Spanish and Tatar–Russian. To reduce the amount of time required to prepare tasks and analyse results, an open-source task management system is introduced. The evaluation results show that the gap-filling task may be suitable for measuring MT quality for assimilation purposes.

1 Introduction

As suggested by Church and Hovy (1993), modern machine translation (MT) systems may be divided into two broad categories according to their purpose: post-editing and assimilation systems. The output of the former is intended to be transformed into text comparable to human translation; the latter systems' goal is to enhance user's comprehension of text. Both kinds may be evaluated, either to control for quality in the development process or to compare the systems. Importantly, according to Church and Hovy (1993), the evaluation methods must closely consider the system's primary purpose.

Despite the fact that, as a result of widespread usage of online MT, assimilation (or gisting) is currently the most frequent application of MT (in 2012, daily output of Google Translate matched the yearly output of human translations[1]), few methodologies are established for assimilation evaluation of MT. The methods include post-editing and comparison by bilingual experts (Ginestí-Rosell et al., 2009), and multiple choice tests (Jones et al., 2007; Trosterud and Unhammer, 2012). These approaches are often costly and prone to subjectivity: see the discussion by O'Regan and Forcada (2013). As an alternative, the modification of *cloze* testing (Taylor, 1953) was introduced for assimilation evaluation, first by Trosterud and Unhammer (2012) as a supplementary technique, and then by O'Regan and Forcada (2013) as a stand-alone method. Prior to this, cloze tests have been used to evaluate raw MT quality (Van Slype, 1979; Somers and Wild, 2000). While these authors ask informants to fill gaps in MT output, Trosterud and Unhammer (2012) and O'Regan and Forcada (2013) ask informants to fill gaps in the reference (human) translation. A designated number of keywords is removed from the human-translated sentences. The evaluators are then asked to fill the gaps with suitable words with and without the help of MT output. The gap-filling task models how well users comprehend the key points of the text, as it is roughly equivalent with answering questions. Thus, the method does not directly evaluate the quality of machine-produced text, but rather its usefulness in understanding the meaning of the original text.

The gap-filling method has been successfully used to evaluate the Basque–English Apertium language pair. In this work we extend the evalua-

[1] http://googleblog.blogspot.co.uk/2012/04/breaking-down-language-barriersix-years.html

tion to two more language pairs: Basque–Spanish and Tatar–Russian. The former pair, while not producing output suitable for post-editing, is a good example of an assimilation MT system. In addition, Basque and Spanish are not mutually understandable, and therefore constitute a good pair for evaluation. For the latter pair, the evaluation served as a quality check in the period of active development during the Google Summer of Code 2014 programme. In addition to evaluating, we explore the previously unconsidered aspects of the experiment: the correlation between evaluators' scores, and the effects of the linguistic domain of texts and the percentage of gaps in a sentence. To facilitate the evaluation, we introduce an automated system which creates task sets from parallel corpora given a range of parameters (number of gaps in a sentence, hint type, gap filler, etc.), checks evaluators' answers, and calculates and reports generalized results. This system is integrated into the Appraise MT evaluation platform (Federmann, 2012); the code is open-source and is available on GitHub.[2]

We anticipate that the assessed MT systems will contribute to the users' understanding of text, that is, the users will show better results in gap-filling tasks when assisted with MT. We also expect to see different results depending on text domain and the relative number of gaps in a sentence.

The paper is organised as follows: in section 2 we describe the gap-filling method for assimilation evaluation: the task layout, the choice of words, and how the tasks are generated. Section 3 introduces the experimental material, the evaluators, the distribution of tasks and the evaluation procedure. In section 4 we describe and discuss the experiment results. Finally, section 5 draws some conclusions. This paper is concerned primarily with assimilation evaluation; for a deeper discussion on evaluation see e.g. (Koehn, 2010, ch. 8).

2 Methodology

This section discusses the reasoning behind the gap-filling method and task structure. The gap-filling method of evaluating machine translation for assimilation purposes is based on the following hypothesis: a reader's understanding of a given text correlates with the number of words they are able to correctly restore in the text. Therefore, the base of an assimilation task is a (reference) sentence, where some of the words are blacked out, or removed. The sentence is produced by a human (as opposed to machine-translated), and it is in the language known to evaluators, which is also the *target language* of the machine translation system. The additional elements of the task are what we call hints, or extra sentences that help the participant to understand the main sentence. There are two types of hints: first, the *source*, which is semantically equivalent to the *reference*, also human-produced, but in the source language of the pair. The second type is the *machine-translated* hint, which comes from the machine translation of the *source* sentence. Table 1 shows a sample task, and Figure 1 shows the task in the online evaluation environment.

In the course of the experiment, following O'Regan and Forcada (2013), we offer these hint combinations:

Reference sentence only: The participants are asked to fill the gaps without being given any context. This task serves as a baseline score and as an indicator of gaps that can be completed using common knowledge or language intuition (e.g. idioms and strong collocations). For example, in an English phrase 'Jack ordered <...> and chips', one of the natural answers would be 'fish'. Such an answer, however, may be unrelated to the meaning of the source text, and may be given on the basis of collocation only.

Reference sentence and source sentence: By setup, the participants have no command of the source language, however, it may help them to fill in proper nouns or loan words.

Reference sentence and MT hint: In addition to the reference sentence, the participants see the source sentence translated via the MT system, in this case Apertium (Forcada et al., 2011). This type of task is used for measuring the contribution of machine translation to understanding the gist of the text.

Reference sentence and both hints: This task is added to check whether MT and source provide complementary hints.

In order to prepare the evaluation questions, we determine and remove keywords from the reference sentences. We consider two parameters: the list of allowed parts of speech (PoS), and the number of gaps relative to sentence length ("gap den-

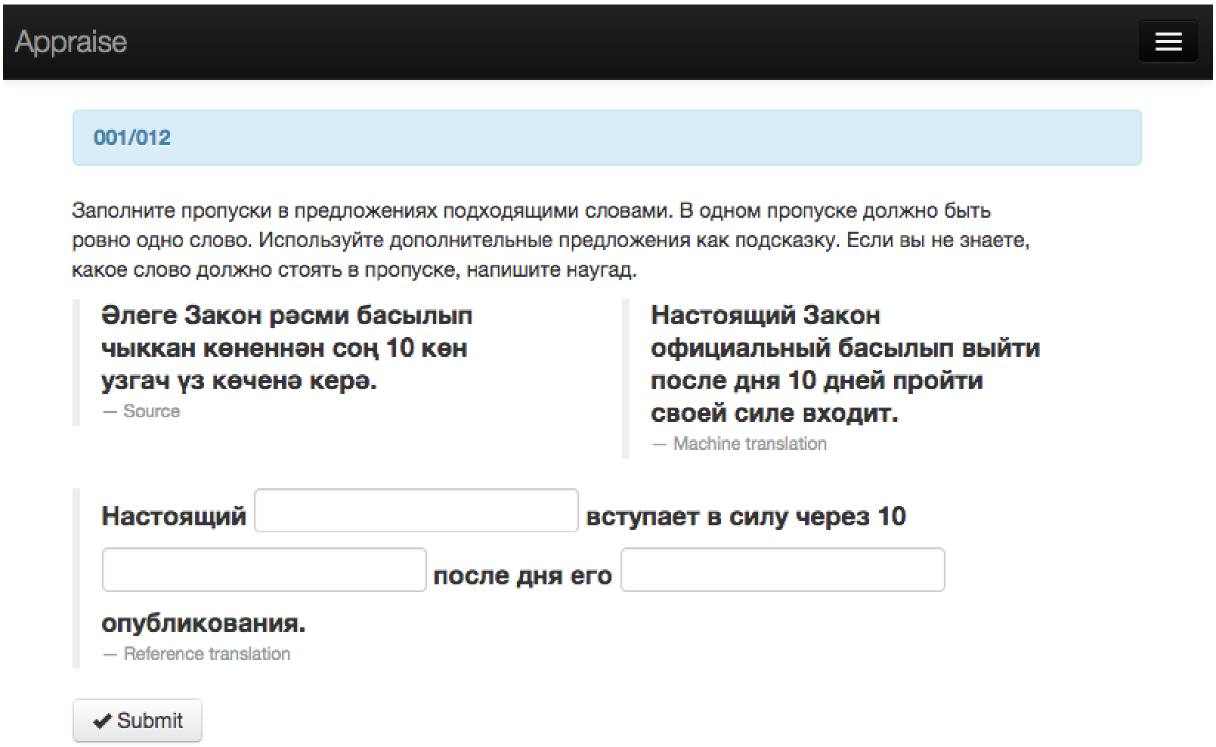
[2] https://github.com/Sereni/Appraise

Ref	Ayudas económicas para el tratamiento de toxicomanías en comunidades terapéuticas no concertadas.
Task	Ayudas económicas para el { } de toxicomanías en comunidades terapéuticas no concertadas.
Src	Komunitate terapeutiko itundu gabeetan toxikomaniak tratatzeko diru-laguntzak ematea.
MT	Comunidad terapéutico pactar gabeetan toxikomaniak las-ayudas de dinero para tratar dar.

Table 1: An example group of sentences showing the gapped sentence and hint types. Reference, MT and task sentences are in Spanish, the source sentence is in Basque.

Ref	Примерно полчаса; вам нужно выйти через 7 остановок, потом пройти ещё около 100 метров.
10%	Примерно полчаса; вам нужно выйти через 7 { }, потом пройти ещё около 100 метров.
20%	{ } полчаса; вам нужно { } через 7 остановок, { } пройти ещё около 100 метров.
30%	Примерно полчаса; вам нужно { } через 7 { }, потом пройти { } около 100 { }.

Table 2: Example of different gap percentage settings for a Russian reference sentence.



Заполните пропуски в предложениях подходящими словами. В одном пропуске должно быть ровно одно слово. Используйте дополнительные предложения как подсказку. Если вы не знаете, какое слово должно стоять в пропуске, напишите наугад.

Әлеге Закон рәсми басылып чыккан көненнән соң 10 көн узгач үз көченә керә.

— Source

Настоящий Закон официальный басылып выйти после дня 10 дней пройти своей силе входит.

— Machine translation

Настоящий ______ вступает в силу через 10 ______ после дня его ______ опубликования.

— Reference translation

Figure 1: An example set of sentences in the online environment. The task is Russian legal text with 30% gaps.

sity"). For the evaluations described in this paper we use gap densities of 10, 20 and 30 percent (Table 2), and the following parts of speech: noun (including proper nouns), adjective, adverb and lexical verb (as opposed to auxiliary verb).

For each sentence, the list of candidate keywords is prepared. It is composed of all the words that fall into the allowed PoS list. The number of gaps in the sentence is calculated based on sentence length and specified gap density. All reference sentences are longer than 10 words. Finally, the required number of keywords is selected from the candidate list in such a manner that the gaps are distributed evenly throughout the sentence. We start at a random word in a sentence and check whether it is a keyword candidate. If yes, we remove it, and move *n* words forward, going back to the beginning of sentence if necessary. The step length *n* is the sentence length divided by the desired number of gaps. If the word is not a keyword, or has already been removed, we look at the next word instead. The process is repeated until the designated number of words has been removed, or until there are no more words in the keyword list.

Keyword removal could be one of the most time-consuming steps in task preparation. It normally requires human effort, because we would like to determine the words that contribute the most to understanding the text as opposed to removing random words. In our automatic setup, the above procedure is performed by a script integrated into the task generation pipeline. Parts of speech are determined with Apertium's morphological analysers. To control for homonymy, we only allow the word into the candidate list if all of its possible part of speech attributions are on the PoS list. For example, if we only allow nouns on the word list, and the word "fly" receives two possible part of speech attributions from the tagger, noun and verb, it is not considered for the candidate list.

Having prepared the sentence sets, we assemble them into XML formatted for the Appraise platform.

3 Experimental set-up

In this section we will discuss the evaluators, the evaluation procedure, and the tasks in more detail.

For each experiment we called for native speakers of target language of the language pair (i.e. Spanish and Russian) who had no command of source language of the pair (Basque and Tatar, respectively). The knowledge was self-reported, and the participants were not asked about any other languages they may know. Eleven evaluators participated in the Basque–Spanish experiment, and 28 in Tatar-Russian (although not everyone completed the task in full, see discussion). The majority of Russian participants were aged 20–25, with university degrees or in the process of obtaining them. Although we have not asked the participants about their knowledge of languages other than Tatar and Russian, it is reasonable to assume that most Russian participants knew English to some extent. The Spanish participants were university staff with background in computer science.

By design, our gap-filling tasks require a human translation (reference) of source sentences. Calling for a human translator, however, would significantly increase the resources needed for evaluation. We therefore use parallel text sources, which provide the same sentence in two languages simultaneously:

1. For Basque–Spanish, from the corpus of legal texts "Memorias de traducción del Servicio Oficial de Traductores del IVAP";[3]

2. For Tatar-Russian, from the following sources on three different topics:

   (a) Casual conversations, from a textbook[4] of spoken Tatar;

   (b) Legal texts, from the Constitution and laws[5] of Tatarstan;

   (c) News, from the President of Tatarstan website[6].

Each set features 36 pairs of sentences. For the Basque–Spanish experiment the pairs were drawn randomly from the corpora; for Tatar–Russian, compiled by hand by the developer of the language pair in Apertium. The Basque–Spanish experiment featured 94, 181 and 272 gaps in the 10, 20 and 30 % tasks, respectively. For Tatar–Russian these numbers are 272, 396 and 724, due to longer sentences used in task creation.

[3] http://tinyurl.com/ivaptm2

[4] Литвинов И.Л. Я начинаю говорить по-татарски. Казань: Татарское кн. изд-во, 1994. — 320 с. ISBN 5–298–00463–6 (стр. 219, 220, 232, 233, 234)

[5] http://tatarstan.ru

[6] http://president.tatarstan.ru/

3.1 Procedure

The evaluations took place online, in a system called Appraise (Federmann, 2012), which is designed specifically for various MT evaluation tasks. We adapted the code of Appraise to accommodate for the gap-filling tasks. The tasks were uploaded into the system and manually distributed between the participants by the following rules:

1. Each participant evaluates every sentence (understood as a succession of words), a total of 36;
2. these sentences are divided into 4 groups of 9, one for each evaluation mode (see section 2);
3. in total, all sentences of the set are evaluated with 10, 20 and 30% of words removed;
4. each participant may encounter a given sentence in only one of the percentage variations;
5. each sentence-mode-percentage combination is evaluated by more than one participant.

The participants are given the instructions in their native language; these instructions are repeated above each task in the evaluation system. For the participants' convenience, the body of questions is split into smaller groups which allow multiple evaluation sessions. The instructions are the following: read all the available hints and fill each gap with one suitable word, guessing if unsure. Participants' answers are recorded and marked correct or incorrect automatically. In addition, the time taken to fill the gaps in one sentence is recorded.

This variety of the gap-filling task requires open answers, and it is therefore possible that the participants may provide words that fit the gaps well, but do not match the original answer. To account for these cases, we process all the answers to detect possible synonyms (a method suggested by O'Regan and Forcada (2013)). An answer is considered a candidate synonym if it is given by two or more evaluators, and it does not match the answer word. We record each candidate synonym along with the answer key and the context sentence. For example, the word *asumir* is the original answer in the Spanish sentence *Aprender a jugar y divertirse en el agua sin asumir riesgos* ('Learning to play and have fun in the water without taking risks'). However, two or more evaluators gave a different answer, *correr* (*correr riesgos*, 'running risks'). Based on this data, an expert, who is native speaker of the target language and who has not participated in the evaluations, decides whether the candidate synonym is an acceptable replacement to the answer key in the given context. We then check participants' results against the compiled synonym list and increase scores where appropriate. On average, the scores improve by three percentage points in all evaluation modes. Candidate synonyms are extracted automatically from the evaluators' responses, and each individual score is automatically updated according to the synonym list.

The synonym lists for Basque–Spanish and Tatar–Russian contain 52 and 25 words, respectively. The time taken to compile each list depends on the number of candidate synonyms, and in our case was approximately 30 minutes.

4 Results and discussion

The results are presented in this section. Table 3 shows the proportion and standard deviation of correct answers depending on evaluation mode and gap density. The evaluators' correct answer percentage is averaged over the number of evaluators. In addition to the percentage of correct answers we kept a record of the time taken to fill the gaps in one sentence. To reduce the noise from participants who were distracted during evaluation, when calculating times we remove all the results over 6 minutes (the statistical mode is approximately two minutes). The typical time taken to complete one question varies from under one minute for tasks without hints and few gaps, to approximately two minutes for tasks with more hints and gaps.

We expect the scores obtained in different task modes inside one gap density to decrease when going from tasks with MT and source hint to tasks with MT hint only, to tasks with source hint only, and finally, to tasks with no hint. We also expect that with the increase in gap density, the time taken to fill the gaps should also increase, and the percentage of correct answers should decrease.

The latter trend holds: the average time taken to fill the gaps increases and the average percentage of correct answers decreases as the relative number of gaps goes up. The larger number of gaps in the sentence makes it more difficult to predict the answer based on the context, and also leaves more room for mistakes. Exploring different percentage-mode combinations, we may note that the 10% no-hint tasks take the least time to complete. We would have expected longer completion time, since the participant must come up

Density	Basque–Spanish				Tatar–Russian			
	MT & Src	**MT**	**Src**	**No hint**	**MT & Src**	**MT**	**Src**	**No hint**
10%	62 ± 32	58 ± 28	40 ± 39	49 ± 40	57 ± 42	64 ± 41	54 ± 43	46 ± 41
20%	65 ± 30	70 ± 27	31 ± 28	31 ± 30	65 ± 31	60 ± 33	46 ± 31	39 ± 32
30%	48 ± 26	40 ± 24	26 ± 20	18 ± 18	59 ± 28	56 ± 26	40 ± 28	35 ± 30

Table 3: Average number of gaps successfully filled (%), using a synonym list, for each language pair in all four task modes.

with their own answer unassisted. However, in the no-hint task the participant is required to read only one (reference) sentence, as opposed to two or three (reference and hints) in other tasks. Also, the number of gaps in 10%-gap tasks is low, as it never exceeds three. We found that, as opposed to trying to devise the best word for no-hint gaps, the participants often resorted to filing these gaps with random words, which takes little time.

We will now discuss the percentage of correct answers based on task type. In general, tasks with MT hints score higher than tasks without MT hints. This aligns well with our expectations and suggests than machine translation helps to understand the provided text. In addition, tasks with source hints are completed better than tasks without hints, and the same relation holds between MT+source and MT-only types of tasks. In view of the relatively large standard deviations, the significance of the hints' contribution was tested using a linear regression model. The data points (y) were represented as an individual evaluator's average score (the number of correct answers divided by the total number of answers) in each of the percentage-hint combinations. Two separate models were created: one for no-hint ($x = 0$) vs MT-hint ($x = 1$) tasks, and another for no-hint ($x = 0$) vs source-hint ($x = 1$) tasks. Given the null hypothesis that the slope b of the regression line $y = a + bx$ equals zero, the contribution of MT hint is found to be significant on the $p < 0.001$ level, while the contribution of the source hint is significant only with $p < 0.162$.

Two records in the data do not align with our expectations: the no-hint 10% sentences in Basque–Spanish, which scored significantly higher than the source-hint in the same category, and MT+source 10% sentences in Tatar–Russian, which we would have expected to score higher than the corresponding MT-only task. In the first case, this is largely due to the use of synonyms list. Before taking synonyms into account, the scores were 32 and 35 percent for source and no-hint tasks, respectively. This still shows a small difference in favor of no-hint tasks. However, the latter percentage increases significantly after we extend the answer list with synonyms. Such an increase suggests that, in this case, the content words were restored by semantic context rather than through strong collocation. The second pattern, low scores in Tatar–Russian 10% MT+source, does not stem from the task content. Instead, it is the result of the fixed order of tasks: the participants have always been given MT+source 10% sentences first, followed by other task types. The participants have not received any training tasks before the main evaluations. Therefore, it is possible that the accommodation period is responsible for lower-than-expected scores in this mode of evaluation.

It remains questionable whether we can compare results for different gap densities. The 10%, 20%, and 30% sets contained the same sentences. However, in each case different words were removed. It appears that some content words are easier to fill than the others. This may explain why in Basque–Spanish the 20% MT tasks are completed with better accuracy than 10% tasks.

It is worth noting that many participants reported feeling frustrated in the course of evaluations, especially while working on the no-hint tasks. The latter required suggesting the words with very little context, which led some of the participants to giving random words for answers, or leaving the space blank. 6 out of 49 participants quit the experiment before completing it. Considering the importance of receiving the full set of evaluations, we must address the issue of participant motivation in the upcoming experiments. It may be beneficial to offer monetary compensation for the evaluators' efforts (in our case, they were volunteers).

4.1 Annotator agreement

After obtaining the results we calculated Krippendorff's alpha (Krippendorff, 1970) measure to represent annotator agreement, shown in Table 4.

We selected this measure because of its compatibility with more than two annotators per task

Density	Basque–Spanish				Tatar–Russian			
	MT & Src	**MT**	**Src**	**No hint**	**MT & Src**	**MT**	**Src**	**No hint**
10%	0.496	0.517	0.400	0.124	0.598	0.459	0.711	0.517
20%	0.714	0.700	0.358	0.275	0.740	0.667	0.473	0.261
30%	0.559	0.430	0.406	0.300	0.534	0.581	0.411	0.412

Table 4: Krippendorff Alpha measure of annotator agreement, for each language pair in all four task modes.

and missing data (not all the gaps were evaluated). To calculate Krippendorff's alpha we used an algorithm implementation by Thomas Grill,[7] dividing the answers in each gap into two categories: correct and incorrect. The previously obtained synonym lists were taken into account, i.e. if the two answers are different but both correct, they fall into one category. The measure was calculated separately for each hint and percentage combination.

The interpretation of Krippendorff's alpha varies depending on the application. One of the general guidelines suggested by Landis and Koch (1977) for kappa-like measures (which includes Krippendorff's Alpha) is as follows: $k < 0$ indicates "poor" agreement, 0 to 0.2 "slight", 0.21 to 0.4 "fair", 0.41 to 0.6 "moderate", 0.61 to 0.8 "substantial", and 0.81 to 1 "near perfect".

In general, the level of annotator agreement is relatively high. As the MT and MT+source hints are introduced, the agreement increases (measures closer to 1): the annotators are more consistently correct or incorrect in each given sentence. The agreement measure for the same sentences without hints is closer to zero, which attests to the reliability of our methodology. We note the outlier score in Tatar–Russian 10% source tasks, which has the most contribution from the news texts. This set of sentences contains many loan words, which have similar form in Tatar and Russian (e.g. president, minister, championship), and are understood by Russian speakers. The gaps with loan words have mostly been filled correctly, while there was some disagreement in other gaps.

4.2 Results for different domains

For the Tatar–Russian language pair the participants were offered texts from three different domains (in equal proportions): casual conversations, legal texts and news. The results by domains are displayed in table 5. The MT system used in the evaluation has been targeted to translate texts from all three of the domains. Taking into consideration the above discussion of 10% MT+Source tasks, we observe similar results across the three categories. Note that the source sentences paired with MT improve participants' performance in casual texts, compared to MT-only task mode. This may be due to the fact that many words are borrowed from Russian into Tatar, and are in fact understood by Russian speakers.

5 Conclusions

We have conducted assimilation evaluation of two Apertium translation directions: Basque–Spanish and Tatar–Russian. The results suggest that this evaluation method reflects the contribution of MT to users' understanding of text. The version of the toolkit used in this experiment may be downloaded from our repository.[8]

The experiments may easily be repeated for any language pair (provided a parallel corpus) and any machine translation system. Based on our experience, we would like to suggest the following amendments to the procedure:

1. As reported by O'Regan and Forcada (2013), unless the evaluation is targeted at a specific text domain, it may be beneficial to include a stylistic variety of texts in the initial corpus. Neighboring sentences on the same topic may assist the users in gap-filling tasks;
2. If possible, increase the number of evaluators, or reduce the number of questions per participant. In the above experiments each participant filled from 110 to 187 gaps, divided into small groups. Reducing the amount of work may increase task completion rate;
3. To account for the adaptation period, provide training tasks before the main evaluations take place.

As a consideration for future work, it may be beneficial to compare the results of evaluation by

[7] http://grrrr.org/data/dev/krippendorff_alpha/

[8] https://github.com/Sereni/Appraise/tree/1e9d735faee64d1b97fb343ab111ace6a64509d7

		Evaluation mode			
Domain	**Gap percentage**	**MT & Src**	**MT**	**Src**	**No hint**
Casual	10%	64 ± 45	64 ± 43	62 ± 46	53 ± 44
	20%	73 ± 32	63 ± 36	41 ± 28	38 ± 31
	30%	70 ± 31	60 ± 24	39 ± 27	38 ± 33
Legal	10%	53 ± 40	68 ± 35	39 ± 38	33 ± 35
	20%	61 ± 25	66 ± 24	50 ± 34	48 ± 34
	30%	50 ± 26	48 ± 29	40 ± 27	34 ± 29
News	10%	53 ± 38	60 ± 44	57 ± 42	49 ± 39
	20%	59 ± 34	49 ± 35	47 ± 32	29 ± 29
	30%	58 ± 22	61 ± 22	41 ± 30	35 ± 27

Table 5: Tatar–Russian Average number of gaps successfully filled (%), using a synonym list, for three different domains, in all four task modes.

gap-filling method with the traditional evaluation metrics, as well as with human evaluation.

Acknowledgments: This work has been partly funded by the Spanish Ministerio de Economía y Competitividad through project TIN2012-32615. Ekaterina Ageeva's work has been supported by Google Summer of Code 2014 through the Apertium project. We thank the volunteers who participated in the evaluations.

References

Church, K. W. and Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

Federmann, C. (2012). Appraise: an open-source toolkit for manual evaluation of MT output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Ginestí-Rosell, M., Ramırez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., and Forcada, M. L. (2009). Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, 43:187–195.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M. (2007). ILR-based MT comprehension test with multi-level questions. In *HLT 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

O'Regan, J. and Forcada, M. L. (2013). Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on `apertium.org`. *Procesamiento del Lenguaje Natural*, 51:15–22.

Somers, H. and Wild, E. (2000). Evaluating machine translation: the Cloze procedure revisited. In *Translating and the Computer 22: Proceedings of the Twenty-second International Conference on Translating and the Computer*.

Taylor, W. L. (1953). "Cloze procedure": a new tool for measuring readability. *Journalism quarterly*.

Trosterud, T. and Unhammer, K. B. (2012). Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*.

Van Slype, G. (1979). Critical study of methods for evaluating the quality of machine translation. *Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR*, 19142.