Translating a Language You Don't Know in the Chinese Room

Ulf Hermjakob, Jonathan May, Michael Pust, Kevin Knight
Information Sciences Institute & Department of Computer Science
University of Southern California
{ulf, jonmay, pust, knight}@isi.edu

Abstract

In a corruption of John Searle's famous AI thought experiment, the Chinese Room (Searle, 1980), we twist its original intent by enabling humans to translate text, e.g. from Uyghur to English, even if they don't have any prior knowledge of the source language. Our enabling tool, which we call the Chinese Room, is equipped with the same resources made available to a machine translation engine. We find that our superior language model and world knowledge allows us to create perfectly fluent and nearly adequate translations, with human expertise required only for the target language. The Chinese Room tool can be used to rapidly create small corpora of parallel data when bilingual translators are not readily available, in particular for low-resource languages.

1 Introduction

Domain adaptation for machine translation is a well-studied problem.[1] Most works assume a system-builder has an adequate amount of out-of-domain or 'general' domain parallel sentence training data and some smaller corpus of in-domain data that can be used, depending on the size of the in-domain corpus, for additional training, for parameter estimation, or, if the in-domain corpus is very small, simply for system evaluation. Very little, however, is said of the scenario where there is *no* in-domain parallel data available, and yet an in-domain system must be built.

In such scenarios one may try to mine parallel data from comparable corpora (Munteanu and Marcu, 2005), but in cases where even scant (but not zero) in-domain *monolingual* resources are available this is not a feasible strategy and the only way to obtain any reliably measure of quality is to solicit human translations. However, it may be difficult to recruit translators to prepare such data, if the language is underrepresented or politically sensitive.

Al-Onaizan et al. (2002) describe an experiment where individual humans translated 10 sentences from Tetun to English, without any prior knowledge of Tetun, based solely on an in-domain bitext of 1,102 sentences. Without any prior tools, translation was very tedious, inefficient, and impractical for the 10 sentences, taking about one sentence per hour. But the experiment successfully showed in principle the feasibility of human translation without prior knowledge of the source language.

We introduce a tool, the *Chinese Room*, to facilitate efficient human translation without prior knowledge of a source language. The name is inspired from Searle (1980) who envisioned a monolingual English-speaking human equipped with instructions for answering Chinese questions by manipulating symbols of a Chinese information corpus and the question text to form answers. While Searle used this idea to argue against 'strong' AI, we thought the setup, i.e. giving a human the tools an NLP model is given (in this case, a machine translation model), was a good one for rapidly generating useful translation data.

Apart from generating human translation data, an additional use of the Chinese Room is to support computational linguists in identifying the challenges of machine translation for a specific language pair and language resources. By placing humans in the role of the MT, we may better understand the nature and magnitude of out-of-vocabulary gaps, and whether they might be due to morphological complexity, compounding, assimi-

[1] See http://www.statmt.org/survey/Topic/DomainAdaptation for a survey of methodologies.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 62–67
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

lation, spelling variations, insufficient or out-of-domain parallel corpora or dictionaries, etc. We found that the Chinese Room can be a useful tool to help generate new ideas for machine translation research.

1.1 Features

Our Chinese Room tool has the following features:

1. Glosser accommodates a variety of NLP and source language resources
2. User can explore alternative translations
3. Grammar support (such as prefixes, suffixes, function words)
4. Optional romanization of source text
5. Robust to spelling variations
6. Optional confidence levels
7. Propagation of user translations
8. Dictionary search function (allowing regular expressions)
9. User accounts with login, password, worksets, separate workspaces
10. Web-based

2 System Description

2.1 Dictionary and T-table Lookup

The principal glossing resources are dictionaries and translation probability tables (t-tables) that are automatically computed from parallel corpora (Brown et al., 1993). The Chinese Room tool will present the top 10 t-table entries and all dictionary entries, including multi-word entries.

2.2 Out-of-Vocabulary Words

However, particularly for low-resource languages, words will frequently not be found that easily. Due to morphological inflection, affixes, compounding, assimilation, and typos, a source word might not occur in a dictionary or t-table.

Low-resource languages often lack consistent spelling due to dialects, lack of spelling standards, or lack of education. For example, even a small Uyghur corpus included six different spellings for the Uyghur word for *kilometer*: kilometer, kilometir, kilomitir, kilometr, kilomitr, klometir.

It is therefore critical to be able to identify dictionary and t-table entries that approximately match a word or a part hereof. We address this challenge with a combination of multiple indexes and a weighted string similarity metric.

2.3 Multiple Indexes For String Matching

We currently use the following indexing heuristics: (1) stemming, (2) hashing, (3) drop-letter, and (4) long substring. Inspired by phonetic matching (Philips, 2000), our current hash function first removes duplicate letters and then removes vowels, except for any leading vowels that get mapped to a canonical *e*. For example, both *break* and *broke* are hashed to *brk*.

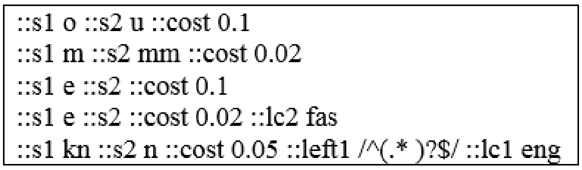
The drop-letter heuristic allows to find entries for words with typos due to letter deletion, addition, substitution and juxtaposition. For example, “crocodile” and “cocodrile” share the drop-letter sequence “cocodile”.

The long (7+ letters) substring heuristic finds dictionary entries that contain additional content.

2.4 Weighted String Distance Metric

Traditional edit distance metrics (Levenshtein, 1966) do not consider the particular characters being added, subtracted, or substituted, and will therefore typically assign a higher cost to (*gram, gramme*) than to (*gram, tram*). Such uniform edit distance costs are linguistically implausible.

The Chinese Room Editor therefore uses a modified metric that leverages a resource of edit distance costs. In particular, costs for vowels and duplicate letters are cheap.



```
::s1 o ::s2 u ::cost 0.1
::s1 m ::s2 mm ::cost 0.02
::s1 e ::s2 ::cost 0.1
::s1 e ::s2 ::cost 0.02 ::lc2 fas
::s1 kn ::s2 n ::cost 0.05 ::left1 /^(.* )?$/ ::lc1 eng
```

Table 1: String similarity rule examples.
::s1 = string 1; ::left1 = left context of string 1; ::lc1 = language code of string 1.

The first rule in Table 1 assigns a cost of 0.1 for o/u substitution, well below the default cost of 1. The second and third rule reduce the string distance cost of (*gram, gramme*) to 0.12. Cost entries for pairs of substrings can be restricted to specific left and right contexts or to specific languages. The last rule in Table 1 assigns a low cost to word-initial silent *k* in English. The manually created resource currently has 590 entries, including a core set of 252 language-independent cost entries that are widely applicable.

2.5 Romanization

For a similarity metric to be widely practical, the strings need to be in the same script. We therefore

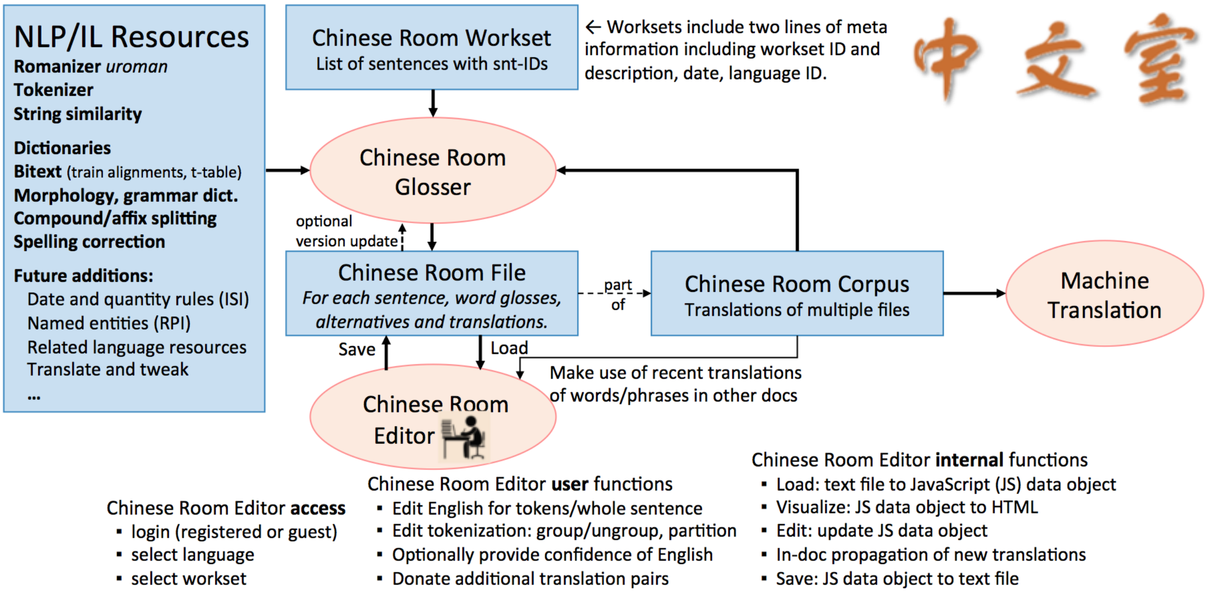
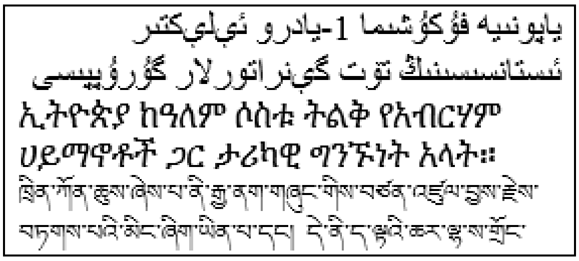

NLP/IL Resources
Romanizer *uroman*
Tokenizer
String similarity
Dictionaries
Bitext (train alignments, t-table)
Morphology, grammar dict.
Compound/affix splitting
Spelling correction
Future additions:
Date and quantity rules (ISI)
Named entities (RPI)
Related language resources
Translate and tweak
...
Chinese Room Workset
List of sentences with snt-IDs
← Worksets include two lines of meta information including workset ID and description, date, language ID.
中文室
Chinese Room Glosser
optional version update
Chinese Room File
For each sentence, word glosses, alternatives and translations.
part of
Chinese Room Corpus
Translations of multiple files
Machine Translation
Save
Load
Make use of recent translations of words/phrases in other docs
Chinese Room Editor
Chinese Room Editor **access**
login (registered or guest)
select language
select workset
Chinese Room Editor **user** functions
Edit English for tokens/whole sentence
Edit tokenization: group/ungroup, partition
Optionally provide confidence of English
Donate additional translation pairs
Chinese Room Editor **internal** functions
Load: text file to JavaScript (JS) data object
Visualize: JS data object to HTML
Edit: update JS data object
In-doc propagation of new translations
Save: JS data object to text file

Figure 1: Chinese Room process. Blue rectangles represent data, pink ovals programs and processes.

romanize before computing string similarity.

An additional motivation for romanization in the Chinese Room is based on the observation that foreign scripts present a massive cognitive barrier to humans who are not familiar with them. See Table 2 for examples.



ياپونىيە فۇكۇشىما 1-يادرو ئېلېكتىر
ئىستانسىسىنىڭ تۆت گېنراتورلار گۇرۇپپىسى
ኢትዮጵያ ከዓለም ሶስቱ ትልቅ የአብርሃም
ሀይማኖቶች ጋር ታሪካዊ ግንኙነት አላት።
ཁྲིན་གོན་རྒྱལ་ཞེས་པ་ནི་རྒྱ་ནག་གཞུང་གིས་བཙན་འཛུལ་བྱས་རྗེས་
བཏགས་པའི་མིང་ཞིག་ཡིན་པ་དང་། དེ་ནི་ད་ལྟའི་ཆར་ལྷ་ས་གྲོང་

Table 2: Texts in Uyghur, Amharic and Tibetan.

We found that when we asked native English speakers to use the Chinese Room to translate text from languages such as Uyghur or Bengali to English, they strongly preferred working on a romanized version of the source language compared to its original form and indeed found using the native, unfamiliar script to be a nearly impossible task.

By default, we therefore romanize non-Latin-script text, using the universal romanizer *uroman*[2] (Hermjakob et al., 2018). The Chinese Room Editor includes the option to display the original text or both the original and romanized source text. The Uyghur text in Table 2 is romanized as

yaponie fukushima 1-yadro elektir
istansisining toet genratorlar guruppisi

which facilitates the recognition of cognates.

2.6 Grammar Resource Files

An additional optional resource is a set of grammar entries for affixes and function words that dictionaries and t-tables do not cover very well. Table 3 shows examples for five Hungarian affixes and two Tagalog function words.

```
::hun ak ::synt plural suffix ::eng -s, -es
::hun at ::synt accusative case suffix
::hun ból ::synt case suffix ::eng from, out of
::hun ből ::synt case suffix ::eng from, out of
::hun el ::synt verb prefix ::eng away; mis-
::tgl ang ::synt definite article ::eng the
::tgl mga ::synt particle ::function plural ::eng -s
```

Table 3: Grammar entries for Hungarian, Tagalog.

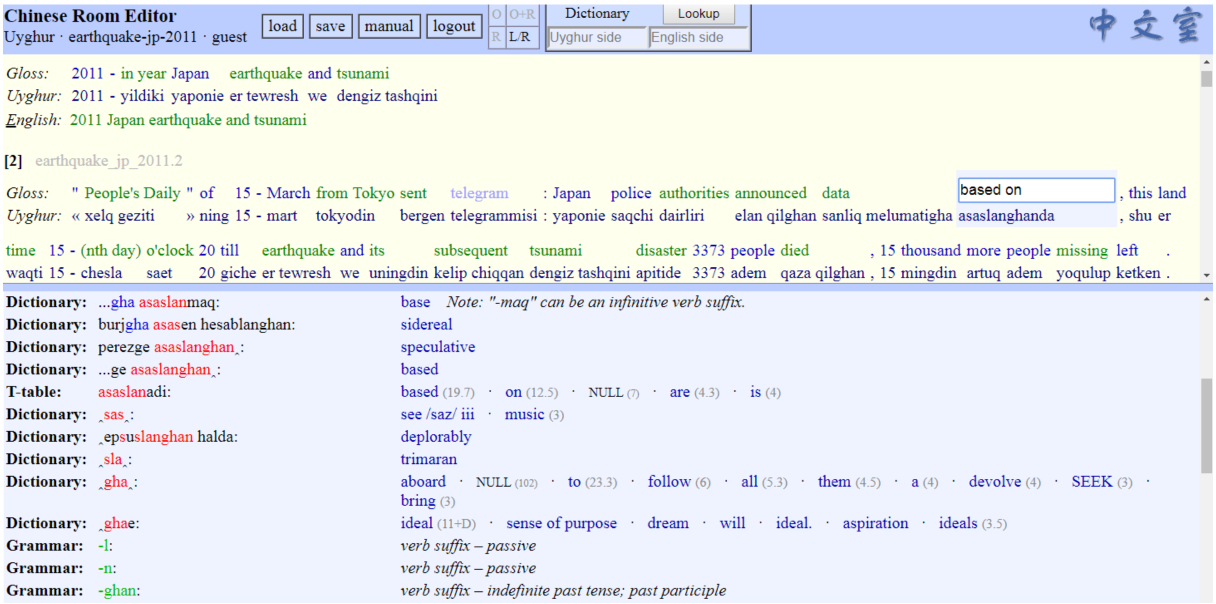
The grammar files have been built manually, typically drawing on external resources such as Wiktionary.[3] The size is language specific, ranging from a few dozen entries to several hundred entries for extremely suffix-rich Hungarian.

2.7 Process

Figure 1 provides an overview of the Chinese Room process. Given a set of NLP resources and a workset of source language sentences, the Chinese Room Glosser builds a Chinese Room File, which can be edited in the Chinese Room Editor. The resulting Chinese Room Corpus can be used for machine translation and other NLP applications.

[2] bit.ly/uroman

[3] https://en.wiktionary.org, e.g. https://en.wiktionary.org/wiki/Appendix:Hungarian_suffixes



Gloss: 2011 - in year Japan earthquake and tsunami
Uyghur: 2011 - yildiki yaponie er tewresh we dengiz tashqini
English: 2011 Japan earthquake and tsunami

[2] earthquake_jp_2011.2

Gloss: " People's Daily " of 15 - March from Tokyo sent telegram : Japan police authorities announced data based on , this land
Uyghur: « xelq geziti » ning 15 - mart tokyodin bergen telegrammisi : yaponie saqchi dairliri elan qilghan sanliq melumatigha asaslanghanda , shu er

time 15 - (nth day) o'clock 20 till earthquake and its subsequent tsunami disaster 3373 people died , 15 thousand more people missing left .
waqti 15 - chesla saet 20 giche er tewresh we uningdin kelip chiqqan dengiz tashqini apitide 3373 adem qaza qilghan , 15 mingdin artuq adem yoqulup ketken .

Dictionary:	...gha asaslanmaq:	base *Note: "-maq" can be an infinitive verb suffix.*
Dictionary:	burjgha asasen hesablanghan:	sidereal
Dictionary:	perezge asaslanghan‸:	speculative
Dictionary:	...ge asaslanghan‸:	based
T-table:	asaslanadi:	based (19.7) · on (12.5) · NULL (7) · are (4.3) · is (4)
Dictionary:	‸sas‸:	see /saz/ iii · music (3)
Dictionary:	‸epsuslanghan halda:	deplorably
Dictionary:	‸sla‸:	trimaran
Dictionary:	‸gha‸:	aboard · NULL (102) · to (23.3) · follow (6) · all (5.3) · them (4.5) · a (4) · devolve (4) · SEEK (3) · bring (3)
Dictionary:	‸ghae:	ideal (11+D) · sense of purpose · dream · will · ideal. · aspiration · ideals (3.5)
Grammar:	-l:	*verb suffix – passive*
Grammar:	-n:	*verb suffix – passive*
Grammar:	-ghan:	*verb suffix – indefinite past tense; past participle*

Figure 2: Screenshot of Chinese Room Editor with Uyghur example. Demo site: bit.ly/chinese-room

2.8 Chinese Room Example

Figure 2 shows an example from a Uyghur article about an earthquake. For the (romanized) Uyghur word *asaslanghanda*, the tool shows several relevant entries that guide the translator to the correct gloss *based (on)*. Note the information regarding the suffixes *-maq, -ghan,* and *-da*.

2.9 Gloss Propagation

Words and expressions often occur multiple times in a document. When a translator edits a gloss, the edited gloss is propagated to other yet unedited glosses of the same word(s) in a document. The propagated glosses can be overwritten, but that is rarely necessary.

Additionally, the edited glosses are collected as an additional translation resource, which can be compiled and propagated to other documents. This allows the sharing of newly discovered translations between translators.

At times, some sentences will be difficult to fully translate, particularly if there are multiple unknown words. The meaning of some of those words will become apparent in other sentences with a stronger context, which in turn will help comprehension of the original difficult sentence.

The discovery of morphological bridge forms is one such case. In (romanized) Uyghur, for example, a translator might struggle with the meaning of *panahliniwetiptu*, but later in the text find a related word *panahlinish*, which in turn is similar enough to the dictionary entry *panalinish* = *shelter* to be found by the tool. With additional grammar guidance for the suffixes *-wet, -ip, -tu,* and *-sh*, the originally hard word can now be glossed and the sentence translated.

3 Chinese Room Editor User Interface[4]

The Chinese Room URL is bit.ly/chinese-room. Temporary visitors are invited to login as *guest*.

3.1 Loading a Workset

To get started, click the *load* button, wait a moment, select a source language (e.g. Uyghur), and a workset (e.g. earthquake-jp-2011-wo-cr-corpus).

3.2 Exploring and Editing Glosses

The initial gloss will often provide a good first idea of what the sentence is about. To explore alternatives to the glosses provided, hover the mouse over a gloss that you want to explore. A blue info box will appear in the lower part of the screen, providing you with a range of entries related to the word. To edit a gloss, or just to fix an info box for subsequent scrolling, click on a gloss, and the gloss becomes editable. To move on to another gloss, click that gloss. To exit gloss editing mode, press *Enter* (while the cursor is in a gloss box). Alternatively, you can click on a translation in the info box to select that translation as a gloss. Double-clicking on a source word will copy it to the gloss; this is

[4] For more details, please consult the Chinese Room Editor manual at bit.ly/chinese-room-manual.

useful for words that don't need translations, such as names.

3.3 Editing Sentence Translations

To edit the translation of the full sentence, click on the current translation (in green), initially *empty*. Type text, or adopt a gloss by clicking on it. Press *Enter* to exit sentence editing.

3.4 Grouping, Ungrouping, Confidence

In the *Special ops* section, click on *group* to combine words to a multi-word expression, or *ungroup* to undo. You may optionally assign a confidence level to glosses and sentence translations, which allows you flag uncertainty, for later review by you or somebody else, or to inform a subsequent user (such as a learning algorithm). For more info, hover over a special-op name.

4 Experiments

We have built Chinese Rooms for Bengali, Hungarian, Oromo, Somali, Swahili, Tagalog, Tigrinya, and Uyghur.

For Bengali, two of the authors of this paper translated an article of 10 Bengali sentences to English, without any prior knowledge of Bengali, using the Chinese Room. To evaluate the results, we asked a native speaker from Bangladesh, a graduate student living in the US who is not a professional translator, to first translate the same 10 sentences independently and then to evaluate our translations. According to the native speaker our translations were better; we only missed one Bengali word in translation, and were actually aware of it, but were unable to decode it with the resources at hand.

We used the Chinese Room to create small corpora of parallel data in a time-constrained MT system-building scenario. In this scenario we were required to translate documents from Uyghur to English describing earthquakes and disaster relief efforts. However, we had no parallel data dealing with this topic, and our use of an unrelated test set (see Figure 3) to estimate overall task performance was not reliable. We thus wanted to construct an in-domain Uyghur-English parallel corpus.

In the scenario we were given a small number of one-hour sessions with a *native informant* (NI), a Uyghur native who spoke English and was not a linguistics or computer science expert. We initially asked the NI use the time to translate documents, one sentence at a time. This was accomplished at a rate of 360 words per hour, but required another 30-60 minutes of post-editing to ensure fluency. We next tried typing for the NI (and ensured fluency); this yielded 320 words/hr but did not require post-editing. Finally we used the Chinese Room to translate and asked the NI to point out any errors. This hour yielded 480 words. Machine translation quality on the resulting in-domain set tracked much better with performance on the evaluation set. Later on we built a second in-domain set but did not have any further access to the NI. Using this set of approximate translation to tune parameters yielded a 0.3 BLEU increase in system performance.

We have trained more than 20 people to use the Chinese Room with very good results for the training test case, Somali. We are similarly confident in our translations for Hungarian and Uyghur. Tagalog and Swahili are recent builds, and translations look very promising.

However, we found the dictionary and bitext resources for Tigrinya (to a lesser degree) and Oromo (to a larger degree) to be too small to confidently translate most sentences. We were able to translate some sentences completely, and many others partially, but had to rely on the support of non-professional native speakers to complete the translations. The Chinese Room nevertheless proved to be very useful in this very-low-resource scenario. We could already build glosses for many words and provide a partial translations, so that the native speaker could finish a sentence faster than starting from scratch. The Chinese Room also helped the native speaker to more easily find the English words he/she was looking for, and allowed us to make sure that the translation covered all essential parts of the original text.

5 Related Work

Callison-Burch (2005); Albrecht et al. (2009); Koehn (2010) and Trados[5] have built computer-aided translation systems for high-resource languages, with an emphasis on post-editing.

Hu et al. (2011) describe a *monolingual translation* protocol that combines MT with not only monolingual target language speakers, but, unlike the Chinese Room, also monolingual source language speakers.

[5] https://www.sdltrados.com

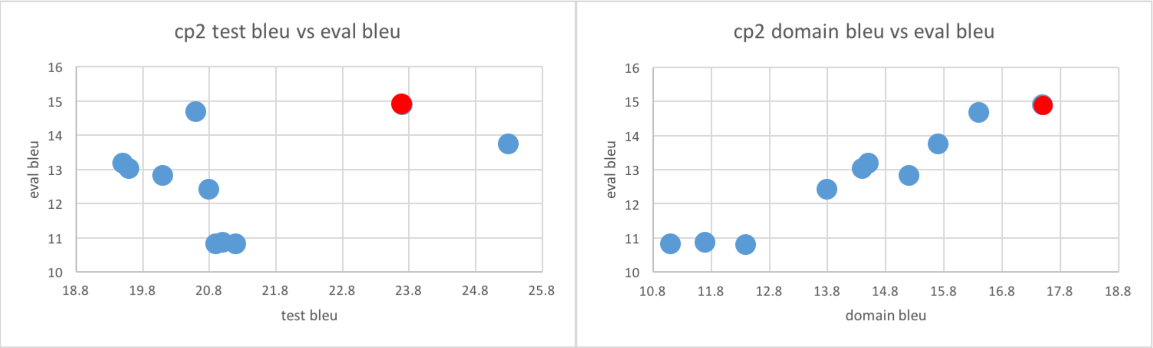


Figure 3: MT performance on an out-of-domain corpus (‘test’) does not predict performance on the evaluation (’eval’) set but performance on our ‘domain’ data set which comprises NI translations and Chinese Room post-edits, is predictive.

6 Future Work

We have observed that by using the Chinese Room, human translators start to learn some of the vocabulary and grammar of the source language. It might therefore be worthwhile to explore how the Chinese Room tool, with a few modifications, could be used in foreign language learning.

7 Conclusion

We have established the feasibility of a practical system that enables human translation from an unfamiliar language, supporting even low-resource languages. We found that we were able to create perfectly fluent and nearly adequate translations, far exceeding the quality of a state-of-the-art machine translation system (Cheung et al., 2017) using the same resources as the Chinese Room, by exploiting the human translators’ target language model and their world knowledge, both of which are still far superior to those of a computer.

Acknowledgments

This work was sponsored by DARPA (HR0011-15-C-0115).

References

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada. 2002. Translation with scarce bilingual resources. *Machine translation*, 17(1):1–17.

Joshua S Albrecht, Rebecca Hwa, and G Elisabeta Marai. 2009. Correcting automatic translations through collaborations between MT and monolingual target-language users. In *EACL*, pages 60–68.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chris Callison-Burch. 2005. Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Leon Cheung, Ulf Hermjakob, Jonathan May, Nima Pourdamghani, Michael Pust, Kevin Knight, et al. 2017. Elisa system description for LoReHLT 2017.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool *uroman*. In *Proceedings of the 56th Annual Meeting of Association for Computational Linguistics, Demo Track*.

Chang Hu, Benjamin Bederson, Philip Resnik, and Yakov Kronrod. 2011. Monotrans2: A new human computation system to support monolingual translation. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 1133–1136. ACM.

Philipp Koehn. 2010. Enabling monolingual translators: Post-editing vs. options. In *NAACL Human Language Technologies*, pages 537–545.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Lawrence Philips. 2000. The double Metaphone search algorithm. *C/C++ Users J.*, 18(6):38–43.

John Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–442.