# Comparative human and automatic evaluation of glass-box and black-box approaches to interactive translation prediction

**Daniel Torregrosa**, Juan Antonio Pérez-Ortiz, Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, Spain

EAMT2017

# Outline

# Abstract


- Interactive translation prediction (ITP) offers suggestions as the translation is being written by the translator
- We compare black-box and glass-box ITP for the first time
- Translators can potentially save 20–50% keystrokes using ITP
- All software used for the comparison is free/open-source

# Outline

# Translation technologies

- Professional translators often use translation technologies such as
  - dictionaries
  - bilingual concordancers (Context Reverso, Linguee)
  - translation memories
  - machine translation
  - post-editing
  - **interactive translation prediction**

  to achieve better, faster translations

# Computer assisted translation

## Interactive translation prediction

- Interactive translation prediction (ITP) focuses on offering suggestions as the translator types the translation
- The approaches in the literature use a **glass-box** approach, where the inner workings of a SMT system are queried to provide ITP
- We have proposed a **black-box** approach[1] that can use any bilingual resource to provide ITP

[1] Torregrosa Rivero, Daniel, Mikel L. Forcada, and Juan Antonio Pérez-Ortiz. "An Open-Source Web-Based Tool for Resource-Agnostic Interactive Translation Prediction." (2014).

# Glass-box ITP

- Glass-box ITP typically uses a modified or tailor-made SMT system that is also able to provide additional information, such as word alignments, alternative translations and translation probabilities
- Recently, neural MT has been used to provide ITP
  - Unlike with SMT, access to the internals is not needed
  - The decoding process is modified so it can accept a prefix

# Glass-box ITP

## Example

### Source sentence

er geht ja nicht nach hause

### Target translation

he does not go home

- In this example, we will use the decoder of a modified statistical machine translation system
- The translator types the prefix of the translation, and gets the best path as a suggestion

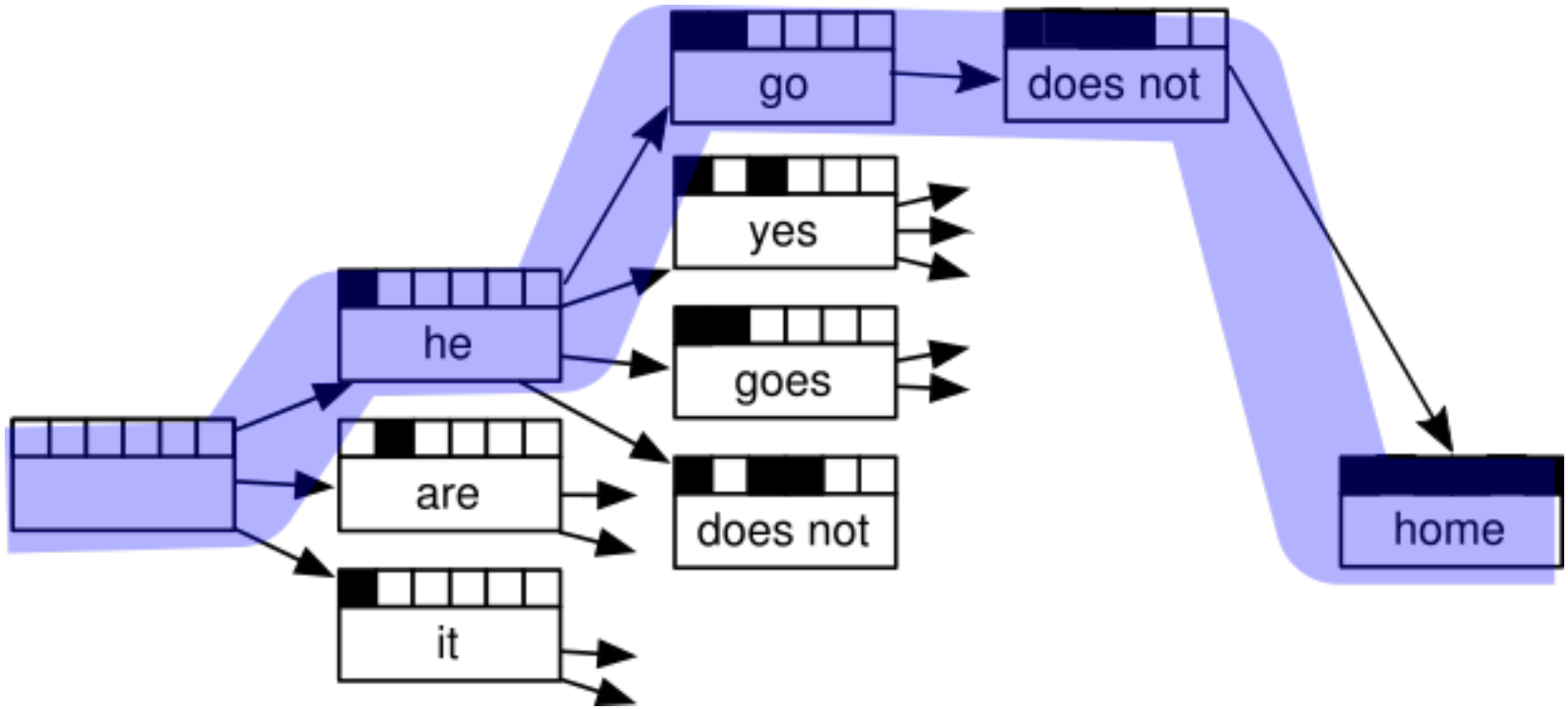
Based on *Statistical Machine Translation*(2009) by Philipp Koehn

# Statistical Machine Translation

## Decoder

### Source sentence

er geht ja nicht nach hause



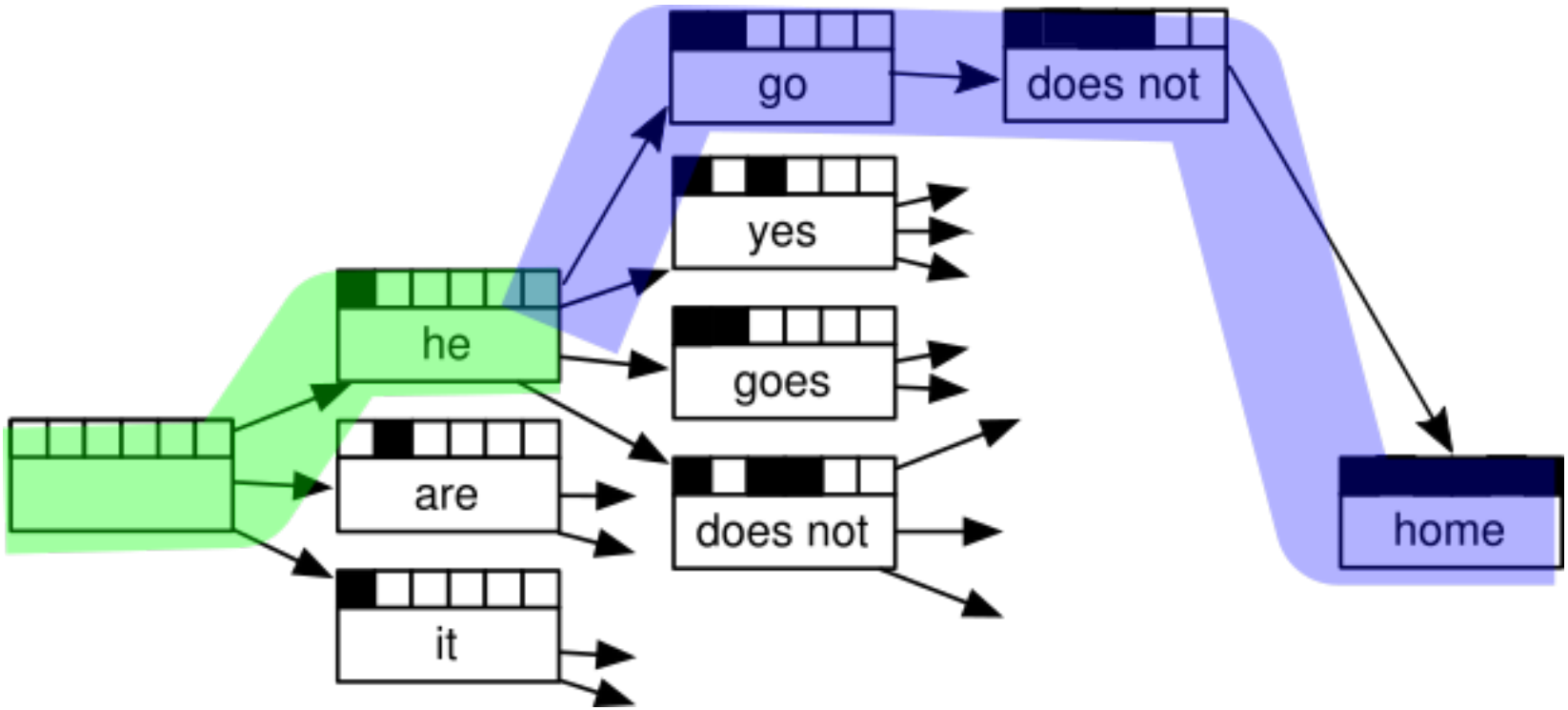

Based on *Statistical Machine Translation* by Philipp Koehn

# Glass-box ITP

## Example I

### Typed prefix

he



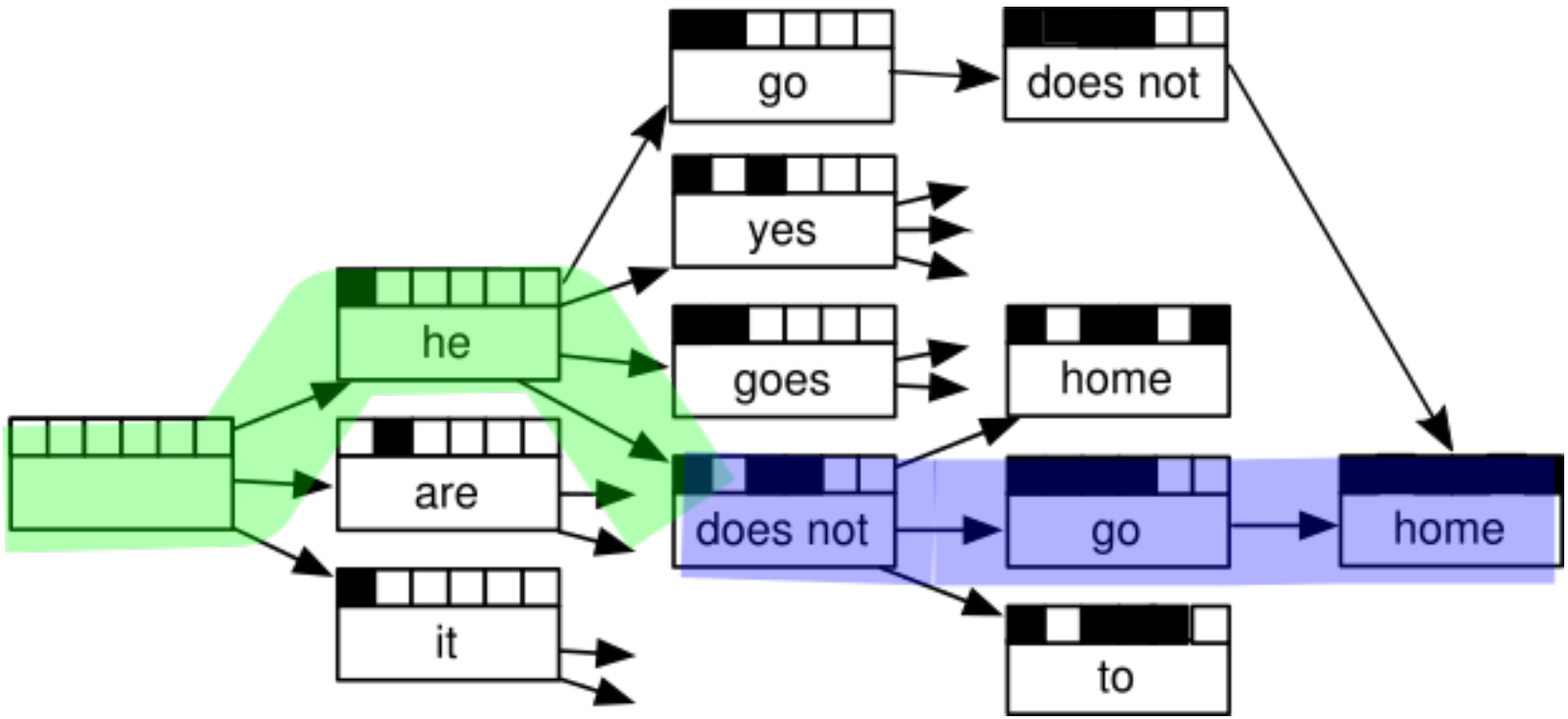

Based on *Statistical Machine Translation*(2009) by Philipp Koehn

# Glass-box ITP

## Example II

### Typed prefix

he d



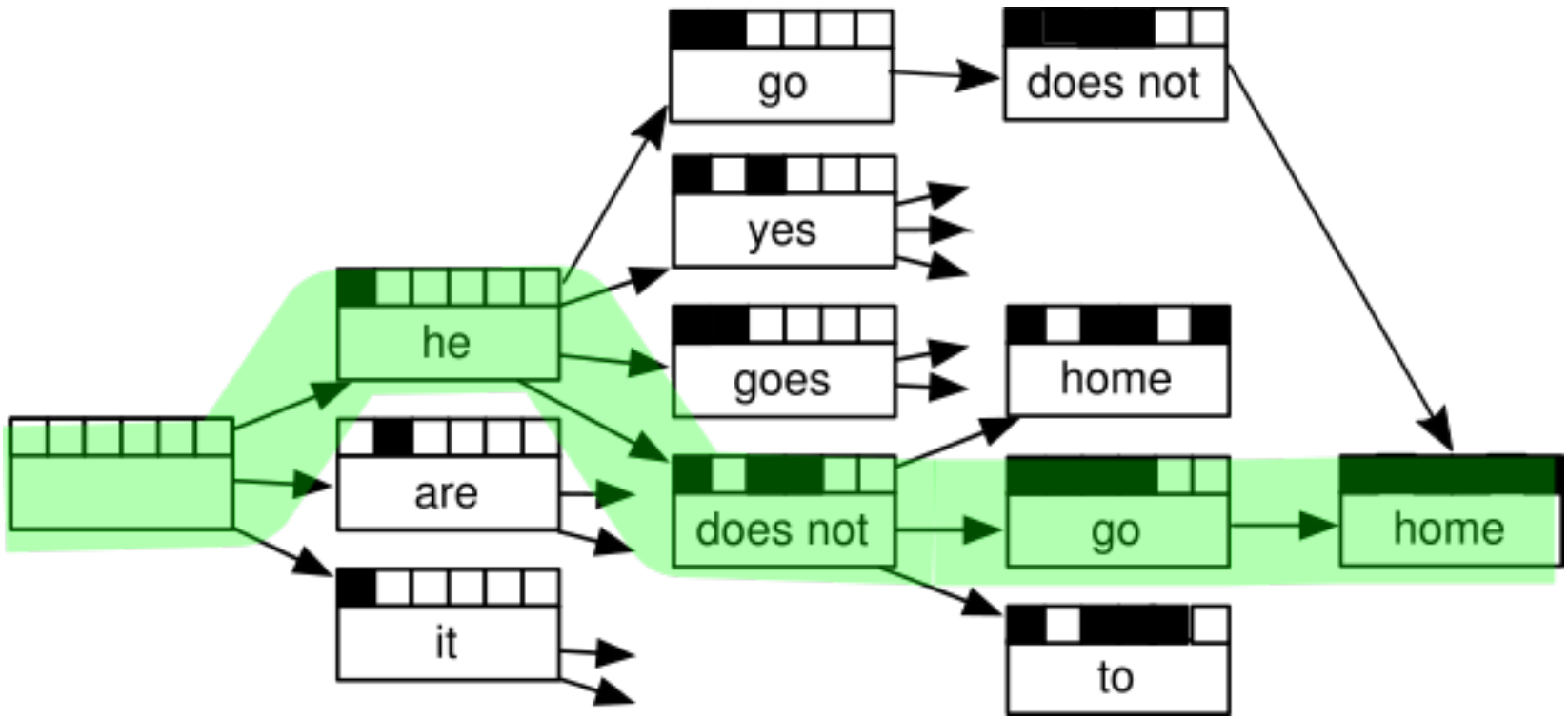

Based on *Statistical Machine Translation*(2009) by Philipp Koehn

# Glass-box ITP

## Example III

### Typed prefix

he does not go home




Based on *Statistical Machine Translation*(2009) by Philipp Koehn

# Black-box ITP

- To generate translation suggestions, black-box ITP can use any kind of bilingual resource that provides one or more translations for a sentence
- This lets us to seamlessly integrate any kind of resource without needing to know how they work

# Black-box ITP

## Generating suggestions

<table>
<tr><th>Source sentence</th><th>This</th><th>studio</th><th>is</th><th>spacious</th></tr>
<tr><td>Subsegments of length 1</td><td>este</td><td>estudio</td><td>es</td><td>espacioso</td></tr>
<tr><td rowspan="3">Subsegments of length 2</td><td colspan="2">este estudio</td><td colspan="2"></td></tr>
<tr><td></td><td colspan="2">estudio es</td><td></td></tr>
<tr><td colspan="3"></td><td>es amplio</td></tr>
<tr><td rowspan="2">Subsegments of length 3</td><td colspan="3">este estudio está</td><td></td></tr>
<tr><td></td><td colspan="3">estudio es amplio</td></tr>
</table>

# Black-box ITP

## Offering suggestions

| Typed prefix | Este | e |
|---|---|---|
| Proposals | | estudio es |
| | | estudio es amplio |
| | | estudio |
| | | este |
| | | es amplio |
| | | espacioso |

- We need to rank and select which suggestions to show.

# Black-box ITP

## Example

| | |
|---|---|
| Source sentence | this studio is spacious |
| Target sentence | este estudio es amplio |
| Prefix | e |
| Suggestions | este estudio está<br>este<br>estudio es amplio<br>estudio |

# Black-box ITP

## Example

| | |
|---|---|
| Source sentence | this studio is spacious |
| Target sentence | este estudio es amplio |
| Prefix | e |
| Suggestions | este estudio está<br>**este**<br>estudio es amplio<br>estudio |

# Black-box ITP

## Example

| | |
|---|---|
| Source sentence | this studio is spacious |
| Target sentence | este estudio es amplio |
| Prefix | e |
| Suggestions | este estudio está<br>**este**<br>estudio es amplio<br>estudio |
| Prefix | e*ste* e |
| Suggestions | estudio es amplio<br>estudio<br>es amplio<br>es |

# Black-box ITP

## Example

| | |
|---|---|
| Source sentence | this studio is spacious |
| Target sentence | este estudio es amplio |
| Prefix | e |
| Suggestions | este estudio está<br>**este**<br>estudio es amplio<br>estudio |
| Prefix | *este* e |
| Suggestions | **estudio es amplio**<br>estudio<br>es amplio<br>es |
| | |

# Black-box ITP

## Example

| | |
|---|---|
| Source sentence | this studio is spacious |
| Target sentence | este estudio es amplio |
| Prefix | e |
| Suggestions | este estudio está<br>**este**<br>estudio es amplio<br>estudio |
| Prefix | *este* e |
| Suggestions | **estudio es amplio**<br>estudio<br>es amplio<br>es |
| Prefix | *este estudio es amplio* |

# Outline

# Software

## Glass-box ITP

- As a glass-box implementation, we use Thot
- Suggests one translation completion that automatically updates as the user types the prefix
- Can also be used as an SMT system
- Trained using 1 000 000 sentences from the United Nations Parallel Corpus v1.0[2]
  - Motivated by lack of resources
  - A bilingual domain adaptation technique[2] has been used to minimize the impact of reducing the size of the corpus
  - Excerpts of this corpus will be used for testing

[2] `http://conferences.unite.un.org/UNCorpus`

[2] Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. "Domain adaptation via pseudo in-domain data selection." (EMNLP 2011)

# Software

## Black-box ITP

- As a black-box implementation, we use Forecat
- Using Thot SMT as the only bilingual resource
- We use a multilayer perceptron[2] for ranking the black-box model suggestions.
- Some features
  - Source and target position and lengths of the suggestion
  - Alignment model
  - Position with respect the last used suggestion: before, after, overlapping...

[2] With $\approx 10^4$ parameters.

# Automatic evaluation

## Method

- We simulate the behaviour of a professional translator
  - who has a planned, immutable translation in mind
  - who writes monotonically
  - who makes no mistakes
  - who reads all the proposed suggestions, evaluates them all, and uses the longest suggestion or suggestion prefix that fits (if any)

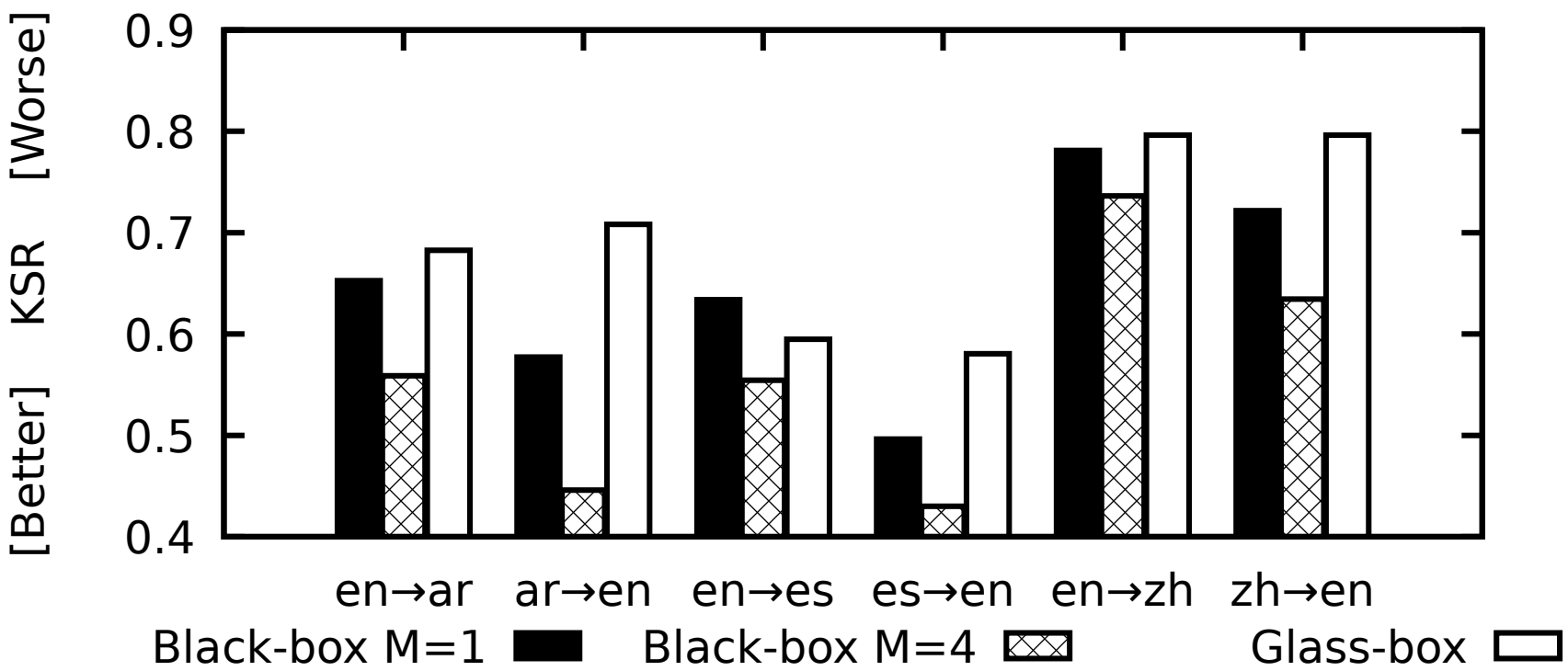
# Automatic evaluation

## Metrics

- We measure the keystroke ratio (KSR), the ratio between the number of keys typed and the length of the final translation
  - KSR$<1$ means we saved some of the keystrokes by using suggestions
  - If we type the translation without mistakes and use no suggestions, we get KSR$=1$
  - KSR$>1$ means we used extra keystrokes, e.g. the user mistyped or rethought the translation halfway

# Automatic evaluation

## Results: KSR




$M$ = maximum number of suggestions.

All differences between the values are statistically significant ($p \leq 0.05$).

# Automatic evaluation

## Results: Average number of words shown to the user

| | Avg. shown words |
|---|---|
| Black-box $M = 1$ | 1.4 $(2.3)^2$ |
| Black-box $M = 4$ | 5 (7.5) |
| Glass-box | 20 |

$M$ = maximum number of suggestions.

[2] Figures in parenthesis exclude the steps where no suggestion is shown.

# Outline

# Human evaluation

## Profiles

- The human evaluation was performed for translation from English to Spanish
- Profile of the 8 subjects
  - Native Spanish speakers
  - Computer science researchers
    - Limited working proficiency with English
    - Experienced typists (except one)
  - No experience as translators
- A more extensive evaluation with professional translators using a similar setup will be carried out soon.

# Human evaluation

## Method

- They had to translate 20 sentences arranged in 4 blocks $SB_1$ to $SB_4$
- The glass-box approach (Thot) offers 1 whole-sentence suggestion
- The black-box approach (Forecat) offers 4 multi-word suggestions

| Users | Induction | Unassisted | Black-box | Glass-box |
|---|---|---|---|---|
| $U_1, U_5$ | $SB_1$ | $SB_2$ | $SB_3$ | $SB_4$ |
| $U_2, U_6$ | $SB_4$ | $SB_1$ | $SB_2$ | $SB_3$ |
| $U_3, U_7$ | $SB_3$ | $SB_4$ | $SB_1$ | $SB_2$ |
| $U_4, U_8$ | $SB_2$ | $SB_3$ | $SB_4$ | $SB_1$ |

# Human evaluation

## Metrics I

- ΔKSR
  - The offset between the assisted KSR and the unassisted KSR: lower is better

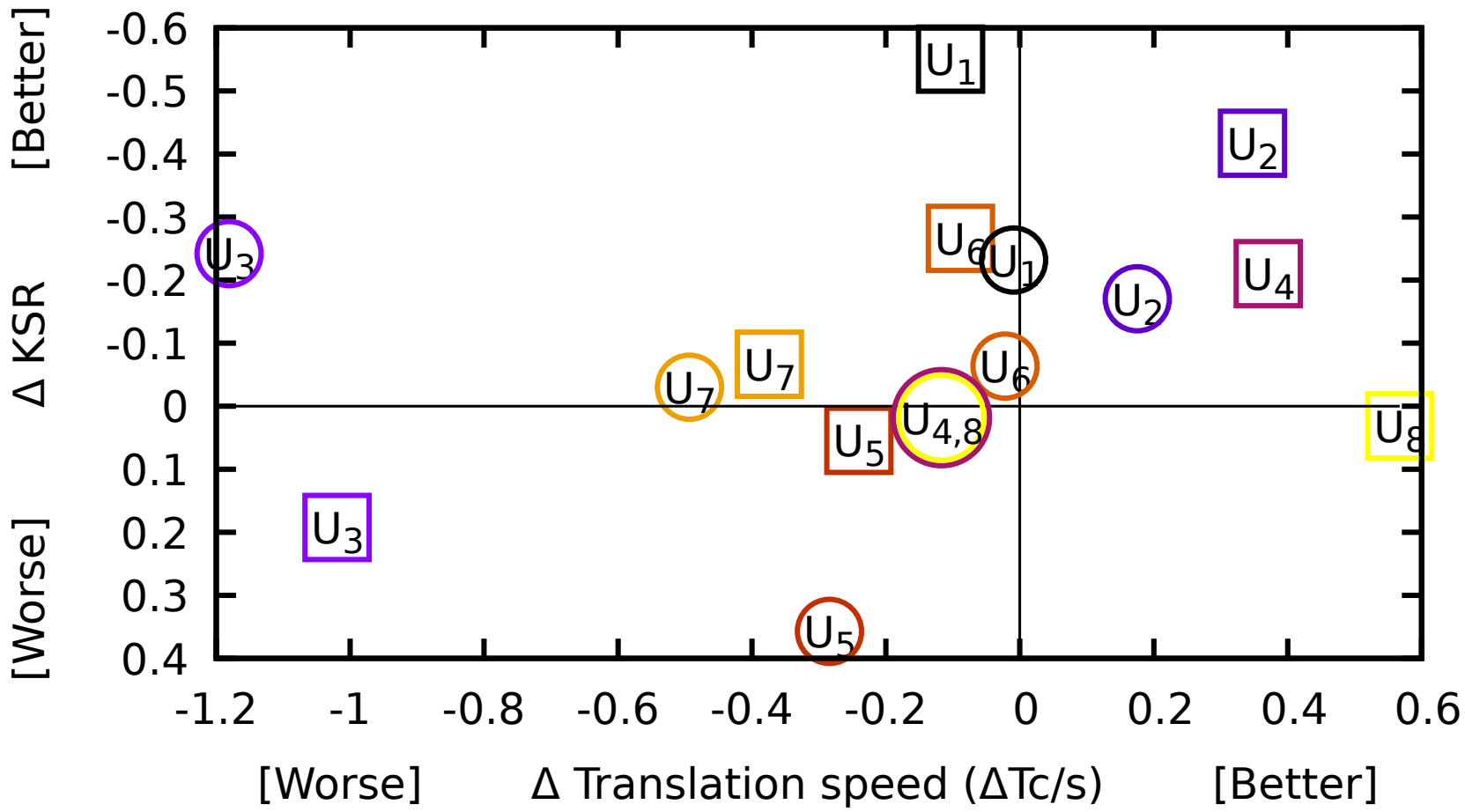
# Human evaluation

## Metrics II

- Translation speed
  - The ratio of the number of characters of the final translation to the time elapsed for typing it
    - ★ If creating a 100-character translation takes 100 seconds, including time spent reading the source sentence and thinking the translation, we have a translation speed of $1\,Tc/s$
  - Measured in target characters per second, $Tc/s$
- $\Delta\,Tc/s$
  - The offset between the assisted $Tc/s$ and the unassisted $Tc/s$: higher is better

# Human evaluation

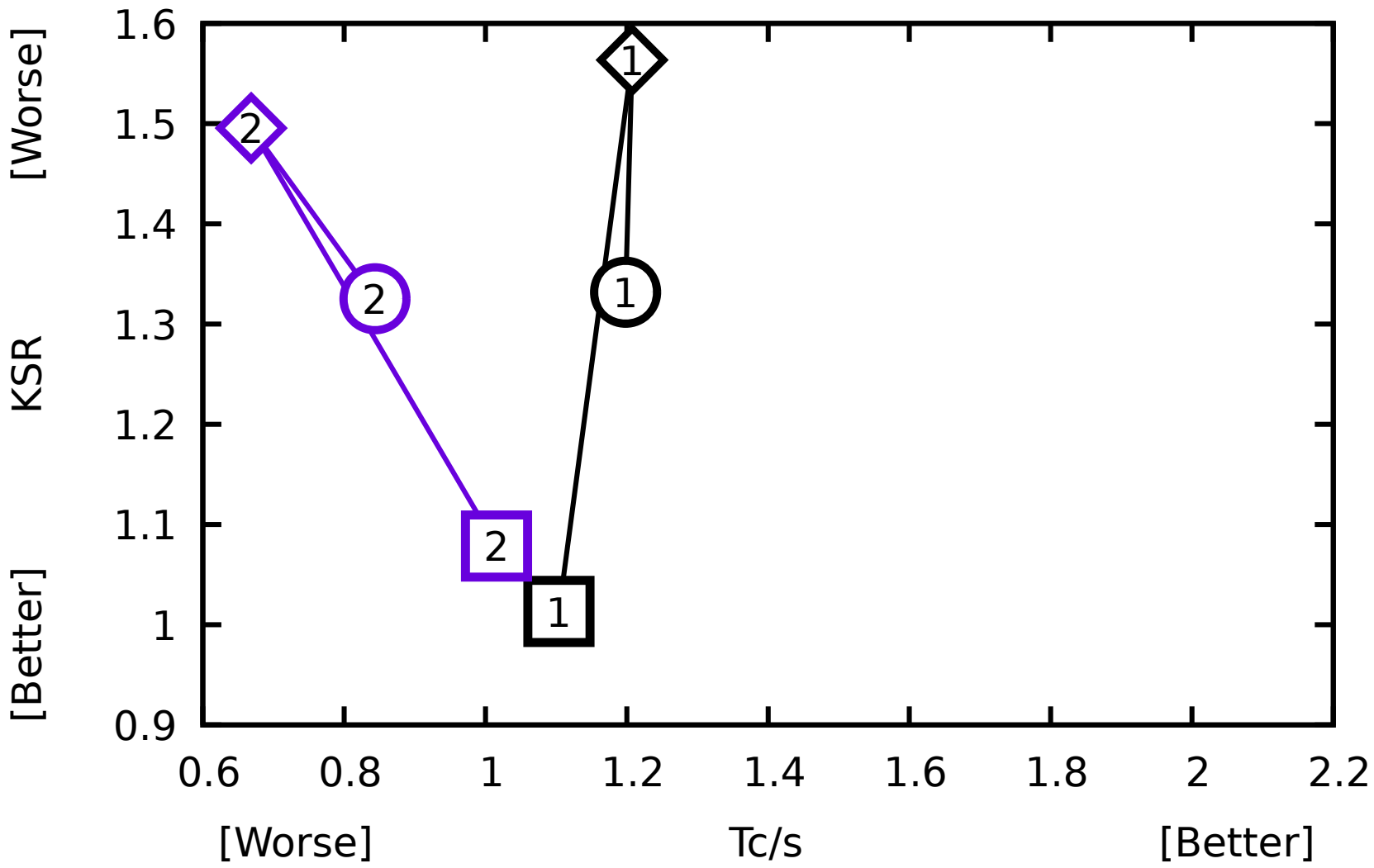
## Comparison with unassisted translation




$\circ$ = Glass box. $\square$ = Black box.

# Human evaluation

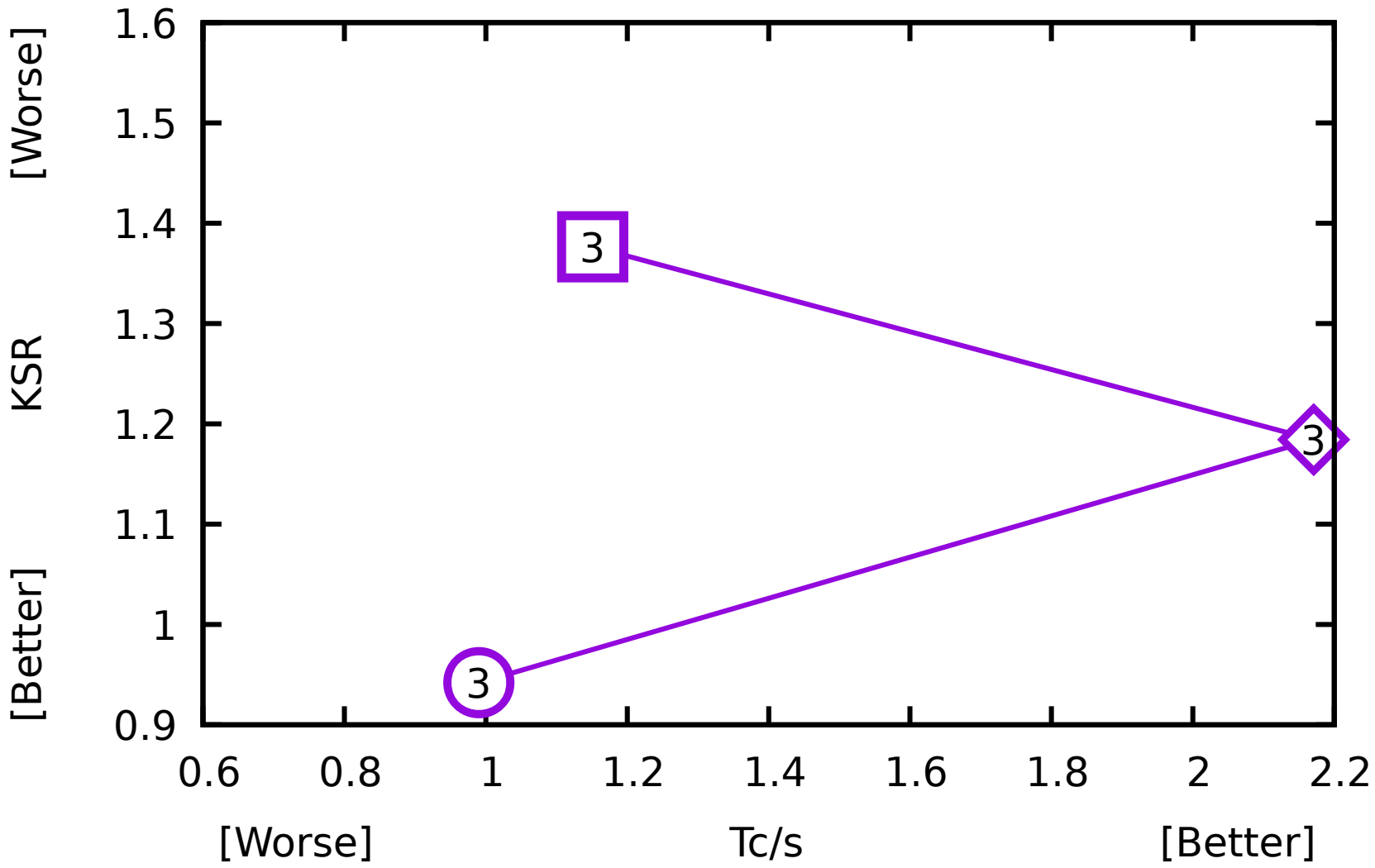
## Absolute values: most improvement (Users 1 and 2)




◦ = Glass box. □ = Black box. ♢ = Unassisted

# Human evaluation

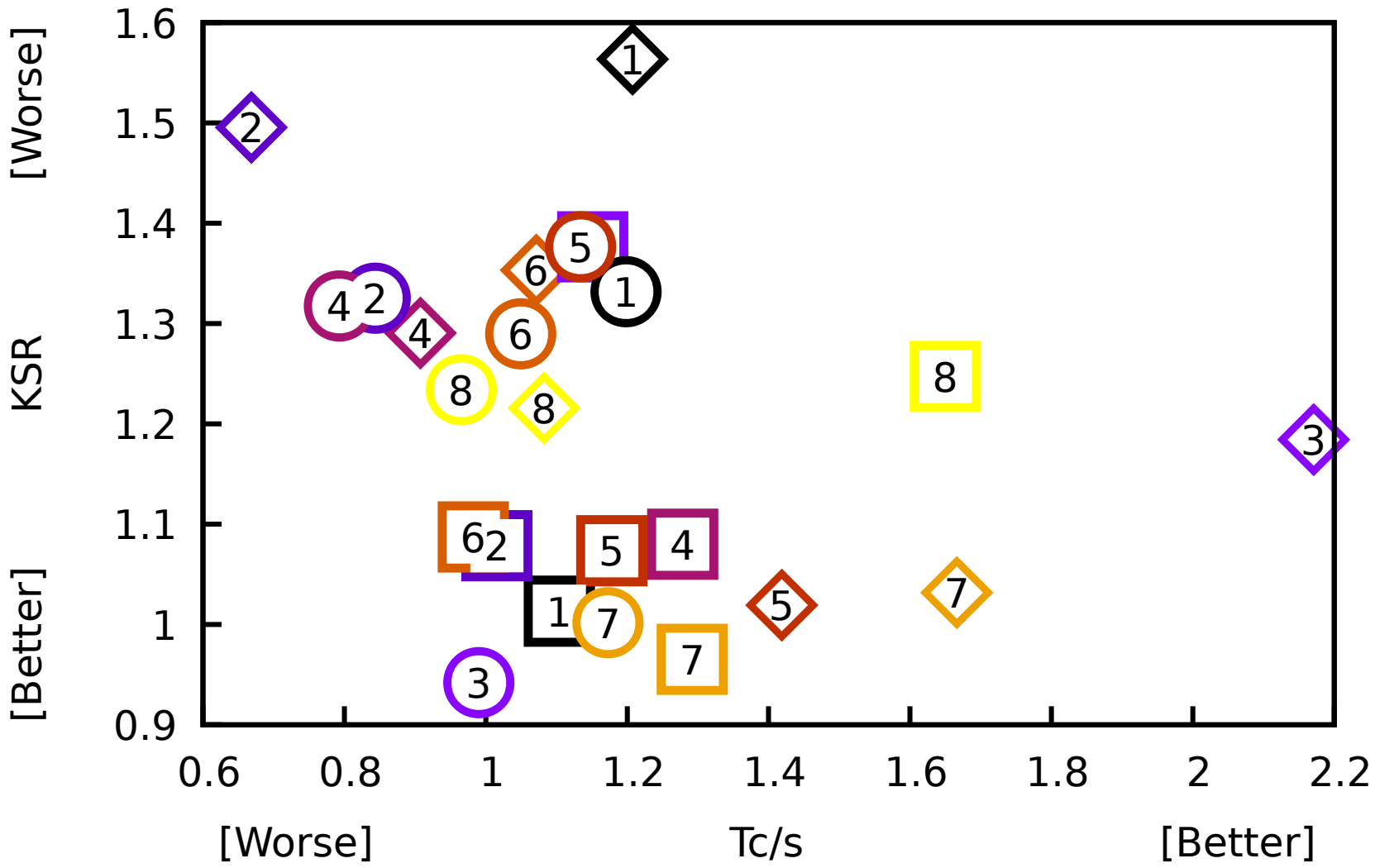
## Absolute values: most deterioration (User 3)




◦ = Glass box. □ = Black box. ♢ = Unassisted

# Human evaluation

## Absolute values




◦ = Glass box. □ = Black box. ♢ = Unassisted

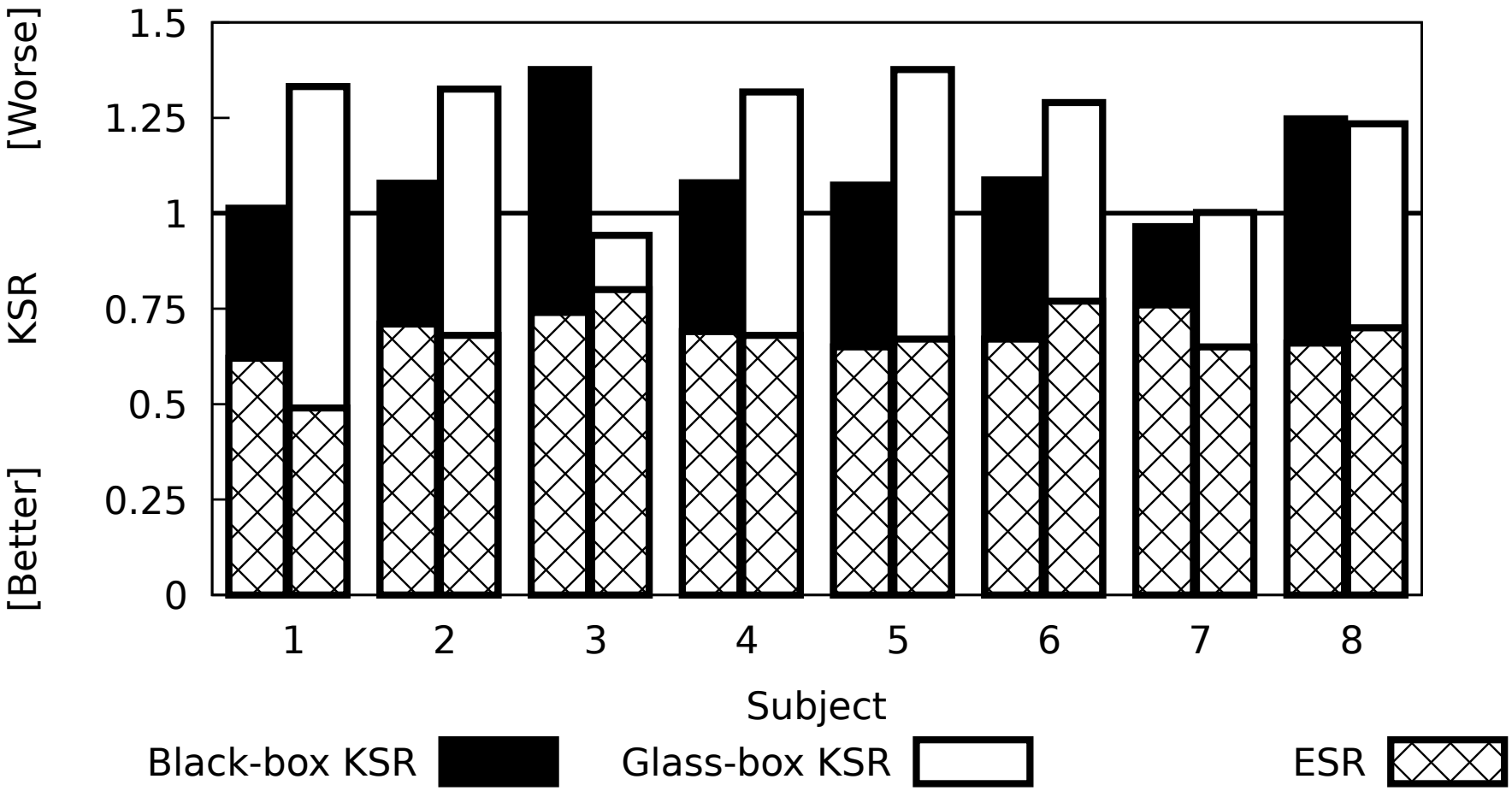
# Human evaluation

## Metrics III

- Emulated keystroke ratio (ESR)
  - ESR is the automatically measured KSR using the translations generated during the human test as references
  - It represents an upper bound for improvement if the user did accept the best suggestion each time some were offered

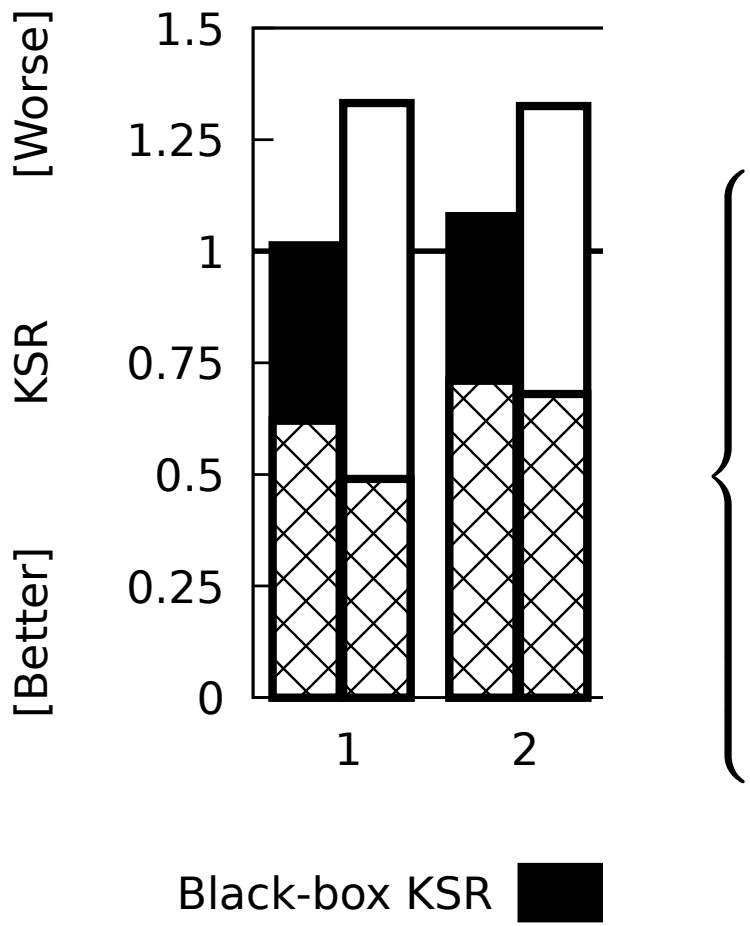
# Human evaluation

## Potential improvement



# Human evaluation

## Analysis of the gap between KSR and ESR




- Related to ITP
  - Bad interface
  - User was not paying attention
  - User distrusted the suggestions
- Unrelated to ITP
  - Typing mistakes
  - Rethought translations

# Human evaluation

## Questionnaire results

- Users mostly agree that the suggestions helped them to translate (Median of 4 using 5-level Likert scale)
- The first glass-box suggestion (a whole translation of the sentence) was praised as very useful
- Test subjects complained about suggestions being offered too often

# Human evaluation

## Questionnaire results II

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ | $U_7$ | $U_8$ |
|---|---|---|---|---|---|---|---|---|
| 1st | **B** | **G** | **G** | **B*** | **G** | **G** | **G** | **B** |
| 2nd | **G** | **B*** | **B** | **U** | **U*** | **B** | **B** | **G*** |
| 3rd | **U*** | **U** | **U*** | **G** | **B** | **U*** | **U*** | **U** |

B=black box. G=glass box. U=unassisted.

Systems ranked according to the perceived speed of translation. The task with the highest translation speed for each user is marked with *.

- Only 3 subjects were faster with assistance:
  - cognitive load may make users think they are translating faster when they are actually translating slower
  - slower translators get the most benefit

# Outline

# Summary

- 20–50% keystrokes can potentially be saved using either the black-box or the glass-box approach
  - Compared to faultlessly typed unassisted translation, KSR=1
  - Up to 60% for some sentences the test subjects translated
- Most test subjects mostly agree in that both methods are useful...
- but are divided when choosing which system is better

# Outlook

- Outlook
  - Explore how only offering the best suggestions affects the performance
  - More extensive evaluation with professional translators and different languages

- Software developed for this paper
  - Forecat: github.com/transducens/forecat
  - Forecat for OmegaT: github.com/transducens/forecat-omegat
  - Thot for OmegaT: github.com/transducens/thot-omegat
- Software used for this paper
  - OmegaT: www.omegat.org
  - Thot: daormar.github.io/thot
  - Omegat SessionLog: github.com/mespla/OmegaT-SessionLog

Slides: tinyurl.com/eamt2017dtr

dtorregrosa@dlsi.ua.es