UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A turing test

To evaluate a complex summarization task

Alejandro Molina

alejandro.molina-villegas@alumni.univ-avignon.fr

http://molina.talne.eu/

Eric SanJuan – Ibekwe

eric.sanJuan@univ-avignon.fr

Juan Manuel Torres Moreno

juan-manuel.torres@univ-avignon.fr

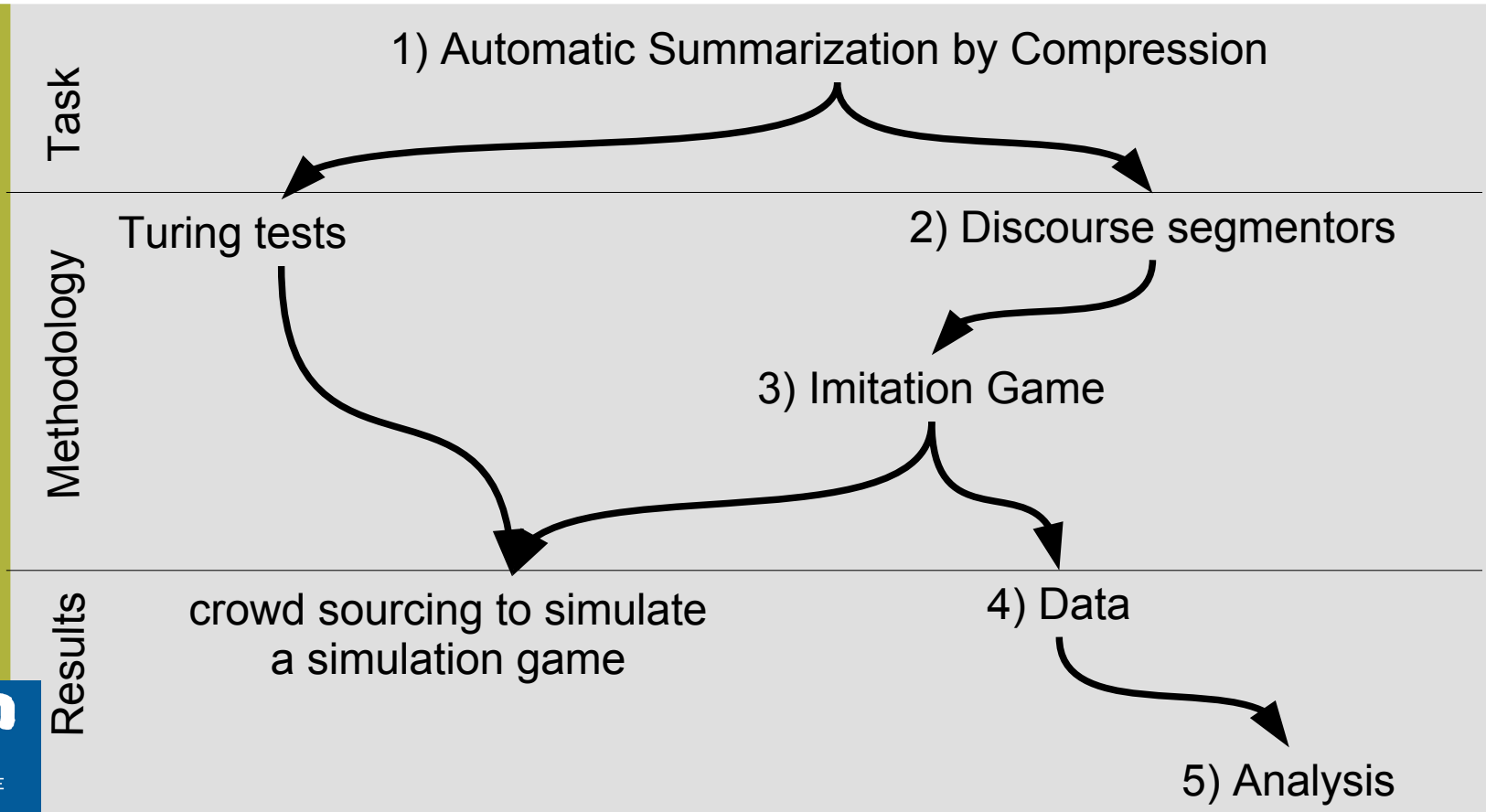
LABORATOIRE
INFORMATIQUE
D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A turing test

To evaluate a complex summarization task

Summary



LABORATOIRE
INFORMATIQUE
D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

1) Automatic Summarization by Compression (ASC)

- Automatic Summarization
  - by sentence extraction and scoring is easy unless breaking anaphora.
  - much more complex if computers are asked to cut and compress sentences like humans do.
- There are usually several correct ways to compress a sentence and human experts often disagree on which is the best one.

Automatic Summarization by Compression (ASC) requires to handle a high level of incertitude in the decision process since there is not a best way to compress a sentence, only observations that sometimes humans prefer one way rather than another one

LABORATOIRE
INFORMATIQUE
D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

2) Discourse segmentors

- Discourse structure among other implicit semantic relations play a key role in ASC
  - humans tend to remove complete discourse units from sentences when they try to compress them:

Molina, A., Torres-Moreno, J.M., SanJuan, E., da Cunha, I., Martinez, G.E.S. *Discursive sentence compression* *(CICLing 2013)*

- We propose ASC systems based on a regression analysis of the way that assessors agree or not to remove a discourse unit.
  - Each discourse segmentor induces a different system.
  - How to compare them ?

LABORATOIRE INFORMATIQUE D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

3) Imitation game

- two Discourse segmentors DiSeg and CoSeg used to generate compressed sentences.
  - Available questionnaire data for regression analysis.
- 12 texts selected from the RST Spanish Tree Bank at random.
  - Summaries of these texts have been written down by post graduate students in linguistics from the UNAM.
  - Three summaries of different length (short, medium and long) were generated using DiSeg, and three other ones also of different length were generated using CoSeg.
- Assessors to guess if the system is human were 54 other post graduate students.

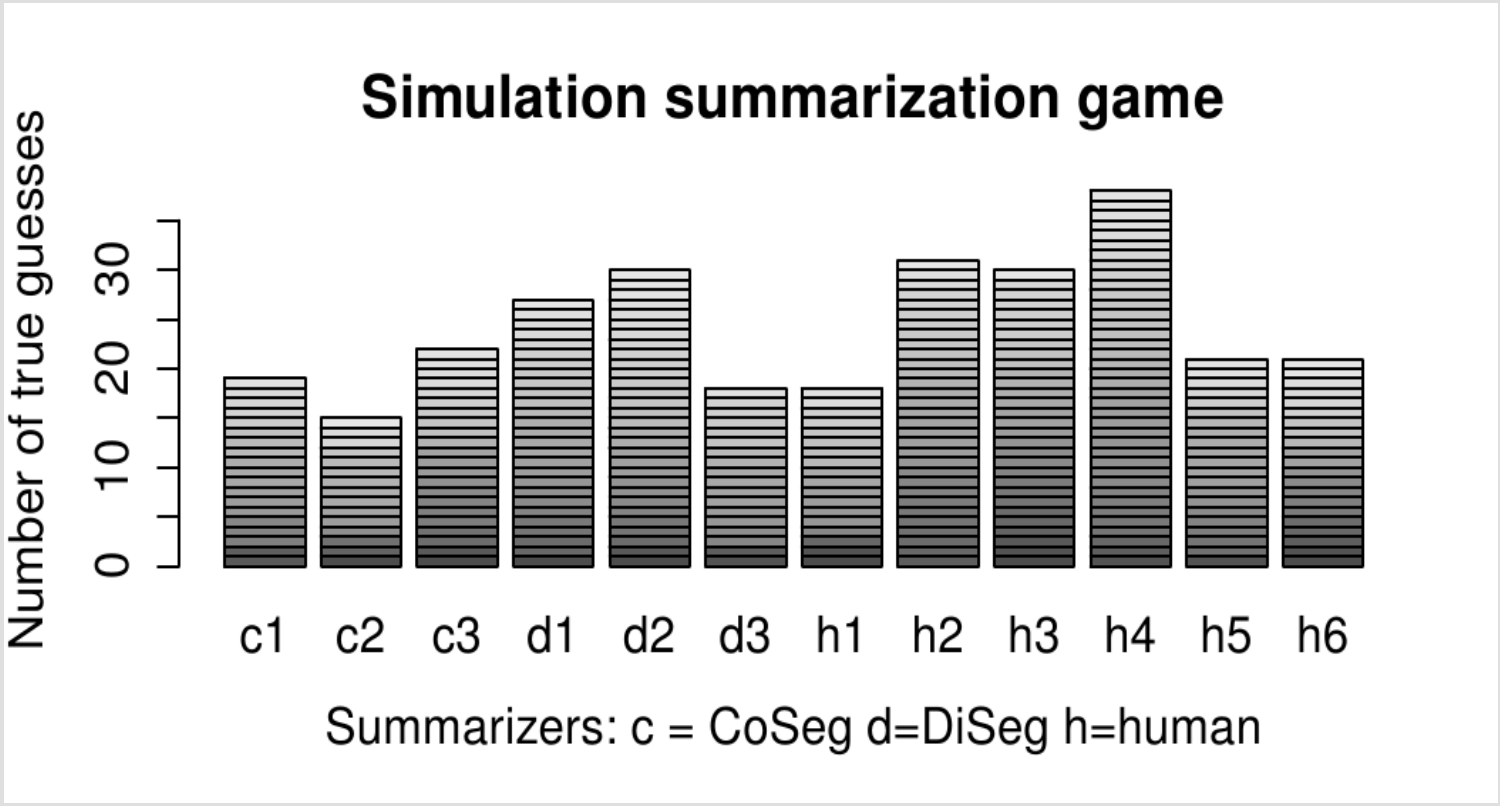
LABORATOIRE INFORMATIQUE D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

4) Data (http://molina.talne.eu/)



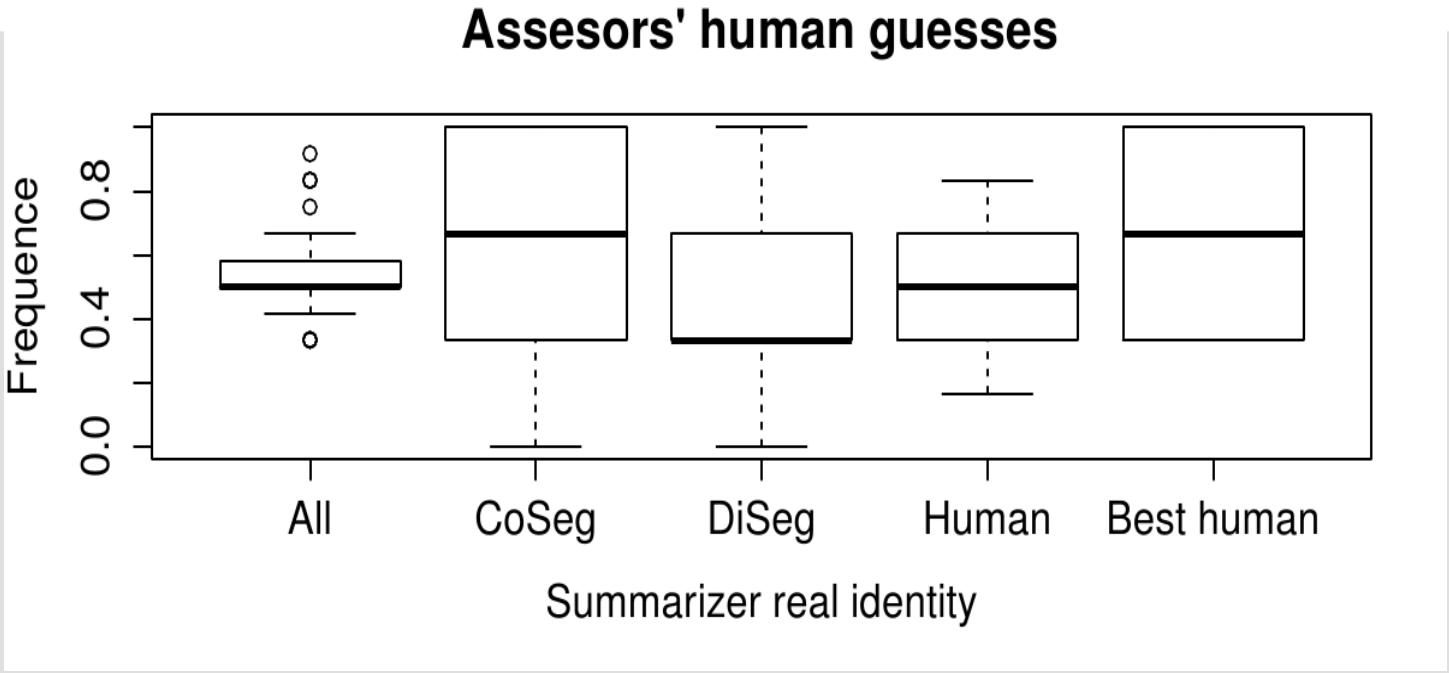
LABORATOIRE
INFORMATIQUE
D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

4) Analysis



Median number of times that an assessor thought it was a summary. Shows that CoSeg based summaries outperform DiSeg ones (p-value < 0.05)

LABORATOIRE INFORMATIQUE D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

A Turing test

To evaluate a complex summarization task

Conclusions & perspectives

- *Back to Turing's idea of simulation game,we used crowd sourcing to simulate a simulation game to evaluate two state of the art automatic summarizers.*
  - *Usual evaluation protocols failed to differentiate between quality levels among the two system outputs.*
  - *The experiment set up here with 60 human players gives statistical evidence that one outperforms the other.*
- *Human ability to differentiate between a summary automatically generated and summary written by an author is less than expected on such complex task.*
  - *needs to be checked out by setting up a larger crowd sourcing task.*
- *Mixing human and machine outputs in the evaluation process seems to be a promising way to improve discriminative power of evaluation protocols.*

LABORATOIRE
INFORMATIQUE
D'AVIGNON