Acquiring NLP Data by means of Games

RASLAN 2010

Marek Grác, Zuzana Nevěřilová

NLP Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

December 3–5, 2010

Why do we collect data?

- the data is hard to obtain
- the collected data can verify the data obtained by other means

Why do we collect data?

- the data is hard to obtain
- the collected data can verify the data obtained by other means

Why do we collect data by means of games?

- experts vs. volunteers
- public-made collections: it works!
  (Wikipedia, OpenMind, GWAP, Amazon Mechanical Turk)
  under certain conditions
- games are popular
- we can make the best of existing resources (databases, corpora, . . .)

Why do we collect data by means of games?

- experts vs. volunteers
- public-made collections: it works!
  (Wikipedia, OpenMind, GWAP, Amazon Mechanical Turk)
  under certain conditions
- games are popular
- we can make the best of existing resources (databases, corpora, . . .)

Why do we collect data by means of games?

- experts vs. volunteers
- public-made collections: it works!
  (Wikipedia, OpenMind, GWAP, Amazon Mechanical Turk)
  under certain conditions
- games are popular
- we can make the best of existing resources (databases, corpora, . . .)

Why do we collect data by means of games?

- experts vs. volunteers
- public-made collections: it works!
  (Wikipedia, OpenMind, GWAP, Amazon Mechanical Turk)
  under certain conditions
- games are popular
- we can make the best of existing resources (databases, corpora, . . .)

Why do we collect data by means of games?

- experts vs. volunteers
- public-made collections: it works!
  (Wikipedia, OpenMind, GWAP, Amazon Mechanical Turk)
  under certain conditions
- games are popular
- we can make the best of existing resources (databases, corpora, . . .)

Games background – players' point of view

- is it fun?
  (good design, high-scores, advance to new levels. . .)
- single player vs. multiplayer
- turn based vs. time-limited
- doing a useful thing

Games background – players' point of view

- is it fun?
  (good design, high-scores, advance to new levels. . .)
- single player vs. multiplayer
- turn based vs. time-limited
- doing a useful thing

Games background – players' point of view

- is it fun?
  (good design, high-scores, advance to new levels. . .)
- single player vs. multiplayer
- turn based vs. time-limited
- doing a useful thing

Games background – players' point of view

- is it fun?
  (good design, high-scores, advance to new levels. . .)
- single player vs. multiplayer
- turn based vs. time-limited
- doing a useful thing

Games background – developers' point of view

- web-based vs. desktop applications
- use of large databases, corpora etc.
- language specific resources (lemmatization)
- latency time response measure
- voluntary and involuntary errors

Games background – developers' point of view

- web-based vs. desktop applications
- use of large databases, corpora etc.
- language specific resources (lemmatization)
- latency time response measure
- voluntary and involuntary errors

Games background – developers' point of view

- web-based vs. desktop applications
- use of large databases, corpora etc.
- language specific resources (lemmatization)
- latency time response measure
- voluntary and involuntary errors

Games background – developers' point of view

- web-based vs. desktop applications
- use of large databases, corpora etc.
- language specific resources (lemmatization)
- latency time response measure
- voluntary and involuntary errors

Games background – developers' point of view

- web-based vs. desktop applications
- use of large databases, corpora etc.
- language specific resources (lemmatization)
- latency time response measure
- voluntary and involuntary errors

Reliability

- involuntary errors (rules misunderstanding, typing errors)
- vandalism
- language specific errors (writing without diacritics)
- e-communication specific features (abbreviations, smileys)

Reliability

- involuntary errors (rules misunderstanding, typing errors)
- vandalism
- language specific errors (writing without diacritics)
- e-communication specific features (abbreviations, smileys)

Reliability

- involuntary errors (rules misunderstanding, typing errors)
- vandalism
- language specific errors (writing without diacritics)
- e-communication specific features (abbreviations, smileys)

Reliability

- involuntary errors (rules misunderstanding, typing errors)
- vandalism
- language specific errors (writing without diacritics)
- e-communication specific features (abbreviations, smileys)

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

Case studies: Scrabble, X-plain

- games based on existing desktop games
- difficult to play
- extremely difficult for non-native speakers
- competitive vs. cooperative games
- human–computer or human–human interaction
- suitable for occasional vs. regular players

X-plain

- human plays with computer
- one player has to explain a word the other by means of templates
- time limit 3 mins.
- collection: triples
  `<something> <isrelated> <something else>`

X-plain

- human plays with computer
- one player has to explain a word the other by means of templates
- time limit 3 mins.
- collection: triples
  `<something> <isrelated> <something else>`

X-plain

- human plays with computer
- one player has to explain a word the other by means of templates
- time limit 3 mins.
- collection: triples
  `<something> <isrelated> <something else>`

X-plain

- human plays with computer
- one player has to explain a word the other by means of templates
- time limit 3 mins.
- collection: triples
  `<something> <isrelated> <something else>`

Scrabble

- players attempt to create valid words
- unlike X-plain we can't focus players on what we want to receive due to random nature of drawing
- turns have to be validated by the other player $\rightarrow$ higher quality
- verification of existing morphological database as players tend to use rare wordforms
- collection: rare but valid word forms, new words

Scrabble

- players attempt to create valid words
- unlike X-plain we can't focus players on what we want to receive due to random nature of drawing
- turns have to be validated by the other player $\rightarrow$ higher quality
- verification of existing morphological database as players tend to use rare wordforms
- collection: rare but valid word forms, new words

Scrabble

- players attempt to create valid words
- unlike X-plain we can't focus players on what we want to receive due to random nature of drawing
- turns have to be validated by the other player $\rightarrow$ higher quality
- verification of existing morphological database as players tend to use rare wordforms
- collection: rare but valid word forms, new words

Scrabble

- players attempt to create valid words
- unlike X-plain we can't focus players on what we want to receive due to random nature of drawing
- turns have to be validated by the other player $\rightarrow$ higher quality
- verification of existing morphological database as players tend to use rare wordforms
- collection: rare but valid word forms, new words

Scrabble

- players attempt to create valid words
- unlike X-plain we can't focus players on what we want to receive due to random nature of drawing
- turns have to be validated by the other player $\rightarrow$ higher quality
- verification of existing morphological database as players tend to use rare wordforms
- collection: rare but valid word forms, new words

Future Work

- evaluation of the collections
- compare X-plain associative network with other associative networks (CZWN)
- implementing other games to collect different types of linguistic data

Future Work

- evaluation of the collections
- compare X-plain associative network with other associative networks (CZWN)
- implementing other games to collect different types of linguistic data

Future Work

- evaluation of the collections
- compare X-plain associative network with other associative networks (CZWN)
- implementing other games to collect different types of linguistic data