Projects at GF Summer School 2015

Kaarel Kaljurand

4th GF Summer School 2015, Gozo

2015-07-24

Projects

- Hands-free "always on" offline HTML5-based multilingual kitchen timer

  - proposed during GF Summer School 2013

  - implemented late 2013, no changes since then

  - just mentioning it briefly as a fun application project

- Wide-coverage grammar for Estonian

  - Estonian resource grammar implemented in late 2013

  - now looked into extending it to the wide-coverage grammar

  - managed to get it to compile (with many functions commented out)

Multilingual kitchen timer (proposal 2013)

Hands-free "always on" offline HTML5-based multilingual kitchen timer

- speech recognition: <u>pocketsphinx.js</u>
    - CMU Pocketsphinx automatically converted from C to Javascript
    - comes with English acoustic models, also works with Estonian models
    - uses finite-state language models
    - modern web browser
- GF grammar
    - time expressions, e.g. <u>Alarm</u> (currently in Estonian and English)
    - dialogue/commands: `start`, `stop`, `pause`, etc.
    - converted to finite state automaton
- GF translator to map the raw recognition result to timer commands
    - ideally Javascript-only
    - possibly/temporarily Javascript to GF server on localhost
- some jQuery timer plugin to execute the timer commands

Multilingual kitchen timer (now)

- speech input (command) and speech output (repeats the command)

- grammar-based recognition and translation to commands

- multilingual: to add a new language: (1) implement concrete syntax for the Alarm-grammar; (2) provide acoustic models; (3) provide pronunciation lexicon

- runs completely offline in a web browser (also on smart phone)

- problems

  - poor speech recognition: does not work from a distance, misrecognizes noise for commands

  - no support for dialogue (i.e. not completely hands-free)

  - see more on the TODO-list on the next slide

- links

  - Estonian app: https://touchless-timer.appspot.com/timer/et.html

  - English app: https://touchless-timer.appspot.com/timer/en.html

  - documentation: https://touchless-timer.appspot.com/timer/about/en/index.html

  - source: https://github.com/Kaljurand/touchless-timer

Multilingual kitchen timer (TODO-list)

- make it work with the latest pocketsphinx.js

- speech recognition: better acoustic models, better noise handling

- more languages

- multiple (named) alarm clocks and timers running at the same time

- dialogue: canceling the alarm clock, pausing/restarting the timer, asking for the remaining time

- listening to multiple languages at the same time (currently can be done by launching multiple instances of the app into separate browser windows, which is not possible on the smart phone)

Estonian wide-coverage grammar

- goals

  - Estonian as part of the GF wide-coverage translation demos (e.g. GF Offline Translator app)

  - test the Estonian RG

  - compare with other approaches to Estonian MT

- progress during the summer school

  - followed the instructions in `lib/src/translator/steps.t2t`

  - copied some files from German and Finnish + commented many functions out

  - focused on translation English->Estonian (i.e. the wide-coverage Estonian grammar is used to linearize)

  - lexicon based on `lib/src/translator/DictionaryEst.gf`

- https://github.com/GF-Estonian/GF-Estonian

Estonian grammar (status)

- regression testing on 100 of ACE parser's regression test sentences (~3000 in total)

  - simple/short sentences

  - easy to tokenize (avoids many parse errors)

  - 33 seconds to parse+linearize (Intel i7)

- is there an official testsuite for the GF wide-coverage translation?

- parse errors: 22; at least 1 function not linearized: 60; fully translated: 18

- a new customer inserts 2 valid cards.

  - uus meesklient sisestab 2 kehtivat kaarti.

- there is a laundry. the laundry is clean.

  - on olemas must pesu. must pesu on laitmatu. (dirty laundy exists. dirty laundry is impeccable.)

Estonian grammar (TODO-list)

- improve the lexicon (both word senses and morph. paradigms)

- implement missing functions

- compare with existing MT systems for Estonian