Name

Tino Didriksen

Contact information

- E-mail: mail@tinodidriksen.com
- Skype: jezral
- Everything else: http://tinodidriksen.com/convoke/

Why are you interested in machine translation?

I've always been interested in languages and computers independently, plus always liked deterministic behavior. So combining languages and rule-driven tech felt natural to me. At first this was mainly applied to language analysis, but in recent years I've been part of projects using it for machine translation.

Why are you interested in the Apertium project?

I believe the Apertium project represents the best current and future implementation of widely usable free open source quality machine translation software.

Why Google and Apertium should sponsor it?

This is a long overdue change to how the Apertium project operates. The current model of every pair being stand-alone works, but simply doesn't scale, and it is starting to show. So better to fix it now, before it gets even worse.

How and who it will benefit in society?

Indirectly, anyone who wants to understand or at least get a gist of what a text in a foreign unknown language says.
Directly, people developing the translation engines to realize the above.

Which of the published tasks are you interested in? What do you plan to do?

I am interested in the Data Decoupling (http://wiki.apertium.org/wiki/Ideas_for_Google_Summer_of_Code/Monolingual_and_bilingual_data_decoupling) idea.

Work plan

Coding challenge

I modified the testvoc script to use gzip'ed temporary files: https://ideone.com/dtzjYx

Total: zcat /tmp/tmp_testvoc.out.gz | wc
@: zgrep '@' /tmp/tmp_testvoc.out.gz | wc
#: zgrep '#' /tmp/tmp_testvoc.out.gz | grep -v '@' | wc

Testvocs w/ unmodified dixs

- pair - total - @ - #
- es-ca - ? - 57 - 3478
- ca-es - ? - 15520 - 206987
- pt-es - 23980325 - 37123 - 509093
- es-pt - 2671 - 0 - 1
- es-pt_BR - 244 - 0 - 0
- pt-ca - 16634364 - 15 - 665255
- ca-pt - 1573605 - 268 - 31547

Testvocs w/ modified dixs

- pt-es w/ pt-ca - 16634364 - 517634 - 264451

Using the pt dix from pt-ca in the pt-es testvoc chain caused the number of @ to jump from 0.15% to 3.11% (up by factor 20), and number of # from 2.12% to 1.59%.

Community Bonding Period

I consider myself quite well bonded with the Apertium project, having been hanging around its mailing list and IRC channel for years due to Apertium's use of CG-3 (http://beta.visl.sdu.dk/cg3.html) that I maintain.

Week 1 (Jun 17-23) & Week 2 (Jun 24-30)

Getting the basic decoupling and dependency resolution set up, with automatic trimming.

Week 3 (Jul 01-07)

- In Poznan
- In Riga

Week 4 (Jul 08-14)

- In Riga

Deliverable 1: Scripts to: decouple a pair; resolve dependency and trim during building

Week 5 (Jul 15-21) & Week 6 (Jul 22-28)

Finalize first stage decoupled repository structure and decouple all pairs. After this point, all monodixes should be in a separate folder, but otherwise no changes to pairs.

Week 7 (Jul 29-Aug 04) & Week 8 (Aug 05-11)

Work on second stage of decoupling, where the original surface form has to survive.

Deliverable 2: Progress report and proof-of-concept.

Week 9 (Aug 12-18) & Week 10 (Aug 19-25)

Get lt-proc to process data with surviving surface forms.

Week 11 (Aug 26-Sep 01) & Week 12 (Sep 02-08)

Debugging and polishing. At this stage, everything up to and including source language tagger and disambiguation should be in a separate folder, leaving only bidix and target language generation for each pair.

List your skills and give evidence of your qualifications

- http://tinodidriksen.com/curriculum-vitae/

My non-Summer-of-Code plans for the Summer

- Exam on June 20th or June 21st
- In Poznan from June 30th until July 3rd (*EU project meeting*)
- In Riga from July 5th until July 13th (*family vacation*)
- It's summer, so my daughter will not be in daycare during July, which will mean less time for work.
- During the whole period, I am still working both freelance and for University of Southern Denmark, but I get to choose my own hours so I can shrink or expand between Apertium and non-Apertium as needed.