Icelandic Language Technology

Eiríkur Rögnvaldsson

June 4, 2004

I am:

- Eiríkur Rögnvaldsson
  - eirikur@hi.is
  - http://www.hi.is/~eirikur
- Professor of Icelandic Language
  - Department of Icelandic
  - Faculty of Humanities
  - University of Iceland
- Head of the MA Program in LT

Teaching and Research Fields

- Teaching
  - NLP
  - Corpus Linguistics
  - Syntax
  - Morphology
  - Phonology
  - Phonetics
  - Applied Linguistics
- Research
  - Corpus Linguistics
    - esp. frequency studies
  - NLP
    - esp. tagging, treebanks
  - Syntax of Old Icelandic
  - Modern Icelandic Syntax
    - esp. word order
  - Icelandic Word Formation

LT at the University of Iceland

- MA in Language Technology
  - started in 2002
- Interdisciplinary program
  - we accept students with either
  - a BS degree in computer science, or
  - a BA degree in the humanities
    - (Icelandic, foreign languages, general linguistics)
- 120 ECTS credits (two years)

Preparatory courses

- Students from the humanities take at least
  - 30 credits in computer science
    - courses in Programming (Java and C++), Mathematical Structures for Computer Science, and Database Theory
- Students from computer science take at least
  - 30 credits in Icelandic language and linguistics
    - courses in Phonetics, Phonology, Morphology, and Syntax

Compulsory and optional courses

- All students take at least
  - 30 credits of compulsory courses in LT
  - these courses can in part be taken at (N)GSLT
    - Natural Language Processing, Speech Technology, Computational Morphology, Statistics in LT, etc.
  - up to 30 credits of optional graduate courses in
    - the Department of Icelandic
    - the Department of Mathematics
    - the Department of Computer Science
    - the Department of Electrical and Computer Engineering

MA Thesis

- All students must write an MA Thesis
  - 30 ECTS credits
  - preferably in cooperation with a research institute or a private company working on language technology
- The students have had the opportunity to participate in some R & D projects
  - sponsored by the Icelandic LT Fund

The Icelandic LT Program

- In 2001, the Icelandic Government launched a special Language Technology Program
  - with the aim of supporting instutions and companies to build basic resources for Icelandic LT work
- This initiative has resulted in several projects
  - which are either finished or well underway
  - while some new projects are being planned

A morphological database

- The Institute of Lexicography at the University of Iceland has now finished building a full-form morphological database for Icelandic
  - The database contains around 180,000 lexemes and all their inflectional forms
  - It will be accessible for on-line lookup in the fall

A PoS tagger for Icelandic

- The Institute of Lexicography and the University have also been evaluating and adapting PoS taggers for Icelandic
  - Several taggers were evaluated
  - TnT, fnTBL, MXPOST selected for training
  - The highest precision/recall rate is 93.65%
    - which is quite acceptible, given the size of the tagset
    - more than 600 different tags

A syntactic parser

- A private software company, Frisk Software, is working on an HPSG-based syntactic parser
  - which they plan to use in developing a grammar checker for Icelandic
- This company has previously developed a spell checker, *Púki*
  - which has been on the market for several years

The *Hjal* project

- In the end of 2002, the University of Iceland and four leading companies in the telecommunication and software industry decided to join their efforts to build the first Icelandic speech recognizer
- The goal of the project , which was called *Hjal* (‘babble’), was to collect sufficient material to train a speaker-independent isolated word recognition system

Linguistic preparations

- The LT Program was responsible for the linguistic preparations
- We cooperated with ScanSoft, Inc.
  - Their role was to train the speech recognizer on the basis of the material that we prepared
- ScanSoft uses the SAMPA phonetic alphabet for phonemic transcription
  - so our first task was to develop a SAMPA transcription standard for Icelandic

Listing the phonemic inventory

- We had to make a detailed description of the phoneme inventory of Icelandic
  - including an exhaustive list of all possible di-phones and a list of the most common triphones
- No such list was available, so this took considerable effort
  - There turned out to be almost 800 different diphones in Icelandic

Designing caller sheets

- The main task in the preparatory phase was to design caller sheets
  - containing words, phrases and sentences for the participants to read
  - words and phrases that are likely to be used in ASR applications
  - “phonetically rich” sentences and strings of isolated letters

Constructing sentences

- The most difficult part was to construct the "phonetically rich" sentences
  - They were to be composed in such a way as to get enough samples of all occurring diphones and common triphones in Icelandic
- We started with a list of 90,000 sentences
  - and ended up with 1433 sentences containing a sufficient number of all occurring diphones and common triphones

Making a word frequency list

- We also had to make a word frequency list
  - This list was compiled from various sources; the newspaper *Morgunblaðið*, recent novels, and the Icelandic spoken language corpus *Ístal*
- The minimum size was 30,000 word forms
  - but due to the inflectional character of Icelandic, we concluded that a larger list would be feasible
  - thus, we ended up with a list of almost 50,000 word forms

Collecting recordings

- ScanSoft needed speech data from at least 2,000 native speakers
  - in order to be able to train a speech recognizer
- Almost 3,000 people volunteered to call in
  - which amounts to 1% of the whole population
- We collected 2,000 valid recordings
  - sufficiently well distributed with respect to gender, age groups, regional dialects, and type of telephone (mobile vs. fixed line)

Transcribing the data

- The recordings were distributed over 90,000 sound files
  - They were transcribed using normal Icelandic orthographic conventions
  - The wordlist was also transcribed using the SAMPA phonetic alphabet
- The transcriptions were done by students in Language Technology at the University of Iceland

The result

- ScanSoft delivered the speech recognizer in the beginning of November last year
  - comprehensive testing has now been carried out
  - the results are quite satisfying
- The recognition rate appears to be ≥ 97%
  - The system can tell apart very similar words
    - for instance, different inflectional forms of the same lexeme where the only difference lies in a vowel in an unstressed syllable
      - *hest<u>u</u>r* ‘horse’ vs. *hest<u>a</u>r* ‘horses’

New LT projects

- Two new LT projects are starting this summer
  - both sponsored by the Language Technology Fund
- The Institute of Lexicography is building a balanced tagged corpus of Icelandic
  - containing ca. 25 million running words
- The *Hjal* group is starting work on a new text to speech system for Icelandic
  - two LT students will write their theses in connection with that project