Celtic Language Technology Workshop (CLTW2014)

Co-located with Coling 2014, August 23rd-29th Dublin, Ireland

CNGL
CENTRE FOR GLOBAL INTELLIGENT CONTENT
FIONTAR
Ollscoil Chathair Bhaile Átha Cliath
DCU
Coling
Dublin | 2014

An Introduction to Celtic Language Technology

A tall order!

Elaine Uí Dhonnchadha

23 August 2014

COLING, DCU

Introduction

- Thank you to the organising committee John Judge, Teresa Lynn, Brian O' Raghallaigh, Monica Ward
- for the CLT workshop

- Exciting time for Language Technology

Outline

- Introduction
- Benefits of a workshop like this
- Inspiration from some unlikely sources …
    - Elephants
    - Giants
    - Seagulls
- Language Processing
- Sharing and Cooperation

Benefits of this workshop

- 1. we can get to know each other
- 2. learn from each other
- 3. share ideas, resources and methodologies

Inspiration

Elephants

What does this means for CLT is …

What this means for Celtic Language Technology is …

…that by combining our knowledge we will each get a better understanding of the biger picture

LT – the big picture

- Language technology – multi facetted
- Diversity of work and viewpoints
- Facilitated by
  - CLTW - this workshop,
  - CIGILT mailing list (Core Interest Group for Irish Language Technology),
  - Have/Want page https://docs.google.com/spreadsheets/d/1pvdPBg12bq1Fvuemj-et9zDmE_sMO-W7ygdmPinFhl4/edit?pli=1#gid=0

CLT: An Overview

- A **Fragmentary** Overview
  - Breton
  - Gaelic
  - Welsh
  - Cornish
  - (Manx)
  - Irish

- There are papers/posters covering almost every aspect of LT and almost every Celtic language.

Irish Language Technology

- Terminology, Lexicography – Fiontar, DCU
- Speech and Language Technology – TCD , abair.ie,
- Treebank/parsing – Lynn, DCU
- CALL – Ward, DCU,
- MT – Scannell, USA; DCU; TCD;

Irish Language Technology

- My part ..
  - Corpus development – written/spoken
  - Morphological analyser/POS tagger/chunker
  - Rule-based

On the shoulders of giants

- Using existing language resources

- Irish authors and speakers
- Irish language resources: An Gúm
- Xerox FST: Beesley, Karttunen
- VISL Constraint Grammar: Bick
- Foma FST: Hulden
- Apertium RBMT: Tyers, Forcada
- Geillatekno Lang Tech: Trosterud

On Language Processing

- Psycholinguistically
  - The nature of language processing
- Computationally
  - Rules
  - Statistics/Probability

On Language Processing

- Idiom Principle
- Take ‘as a matter of fact’ for example
  - Sinclair, J. (1991) Corpus, Concordance, Collocation.

- Over 50% of language is formulaic
  - Erman, B. & Warren, B. (2000): "The idiom principle and the open choice principle". *Text* 20, 1: 29–62.

- Chunks are useful for parsing and translation and speech recognition/generation etc.

On Language Processing

- Hierarchical/Sequential Processing
  - Frank, S. L., et al. (2012). "**How hierarchical is language use?**" Proceedings of the Royal Society of Biological Sciences 279:1747
  - Jelinek, IBM

- Sequential processing
  - local grammar rules,
  - bigrams and trigrams,
  - chunking, shallow parsing
  - Parallel corpora

What does this mean for CLT?

- Rules, statistics – we need both
- Learning from one another and about one another
- Sharing of resources and skills
- Teamwork – Meitheal
- Co-operation especially important for lesser used and lesser resourced languages

On the importance of sharing

- Sharing Resources
- Sharing Skills
- On not reinventing the wheel
- Need for discussion and cooperation

First Celtic LT Workshop

- The papers/posters to be presented today
  - Breton – FST morphology
  - Scottish Gaelic - POS tagger; Grammar
  - Welsh – Speech recognition; Corpora
  - Irish – CALL, Terminology DB, Parser, Web tools and resources
  - Celtic Languages – Dictionaries, MT, Language detection

- Enjoy the First Celtic Languages Technology Workshop!