Towards Semantic Tagging For Irish

Tim Czerniak & Elaine Uí Dhonnchadha

Trinity College, Dublin

Contents

Background

Irish (Gaeilge)

- Goidelic branch of Celtic family
- First official language of Republic of Ireland
- Minority language
- Speakership:
  - 1.7M total
  - 77k daily
  - 40k-80k native
- Linguistic features:
  - VSO word order
  - Fusional morphology
  - Word-initial mutations
  - Nominative-accusative case marking

Goals

- Goals of the project:
  - Understand how semantic tag-sets can be applied to Irish
  - Begin building a corpus of semantically tagged Irish language data
  - Build an automatic semantic tagger that achieves high accuracy
- Part of a government-funded project to update the national corpus of Irish
  - Beta is at http://www.corpas.ie

Status of Irish NLP

- Standard POS-tagging systems
  - Corpora available with PAROLE tags
- Almost zero available semantically tagged data
- No standard semantic typing or semantic role labelling
- Some (incomplete) semantic networks
  - Líonra Séimeantach na Gaeilge (LSG) - https://cadhan.com/lsg/index-en.html
- Some (incomplete) FrameNet/VerbNet equivalents
  - Foclóir Briathara na Gaeilge (FBG) - http://www.potafocal.com/fbg
- Unsupervised word embeddings
  - gaBERT (Barry et al., LREC 2022)

Status of Irish NLP

- We can not (yet) apply ML methods to perform supervised Semantic Annotation
- Must start by:
  - Deciding on a semantic tag-set and tagging norms
  - Building a corpus of manually-checked semantically-tagged data
  - Deciding how to measure accuracy of automatic tagging
  - Building an automatic tagger to grow the corpus, using knowledge-based methods
- Once enough data has been tagged, we can then turn to ML methods

Methodology & Data

Plan

1. Select a semantic tag-set to use
2. Apply that tag-set to a corpus of Irish texts
3. Decide how to measure the accuracy of automatic tagging
4. Develop a tagging pipeline
5. Score the accuracy of the pipeline
6. Accumulate a gold standard of manually-checked semantically-tagged texts

Semantic Tag-Set Selection

- Various tag-set options:
  - USAS (Archer et. al., 2002)
  - Bick (2000)
  - Corpus Pattern Analysis Ontology (Hanks, 2004)
- USAS/PyMUSAS chosen because:
  - Hierarchical
  - Facilitates accuracy scoring method
  - Well integrated with SpaCy
  - Some lexicon resources already available for Irish

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

G GOVT. & THE PUBLIC DOMAIN	
G1	Government, Politics & elections
G1.1	Government etc.
G1.2	Politics
G2	Crime, law and order
G2.1	Crime, law and order: Law & order
G2.2	General ethics
G3	Warfare, defence and the army; Weapons

Applying USAS to Irish

- Selected some texts and manually tagged them with USAS tag-set
  - News/Wikipedia articles
- This produced:
  - a set of tagging ‘norms’ specific to Irish
  - test data for scoring the automatic tagger

Automatic Tagging & Disambiguation

'aire'

Sense	USAS tag
'care' (i.e. caring for someone)	S8+ "Helping/hindering"
'attention' (as in 'pay attention')	A1.3 "General: caution"
'minister' (government)	G1.1 "Government etc."

PyMUSAS Component

- Lemma/POS matcher
- Uses a lexicon:

```
lemma          pos    semantic_tags
...
aird           Nc     X5.1+ S8+ N5+ I2.2% O1.2 Q2.2 W3
airdeallach    Aq     X5.1+ X2.2+
aire           Nc     S8+ A1.3 G1.1
aireach        Aq     A1.3+ E6-
aireacht       Nc     G1.1c S9
aireachtrón    Nc     Y2 B1
...
```

Scoring Accuracy

- Article about political controversy
  - ‘aire’ used for governmental minister (USAS tag G1.1)
- Automatic tagger

Tag assignment	Accuracy
G1.1	Entirely accurate
S8+ A1.3 G1.1	???
K3	Entirely inaccurate

Scoring Accuracy

‘Correct’ assignment	Tag assignment	Score
G1.1	**G1.1**	1
G1.1	**G1.1** S8+	0.85
G1.1	**G1.1** S8+ A1.3	0.79
G1.1	S8+ **G1.1** A1.3	0.45
G1.1	S8+ A1.3 **G1.1**	0.33
G1.1	K3	0

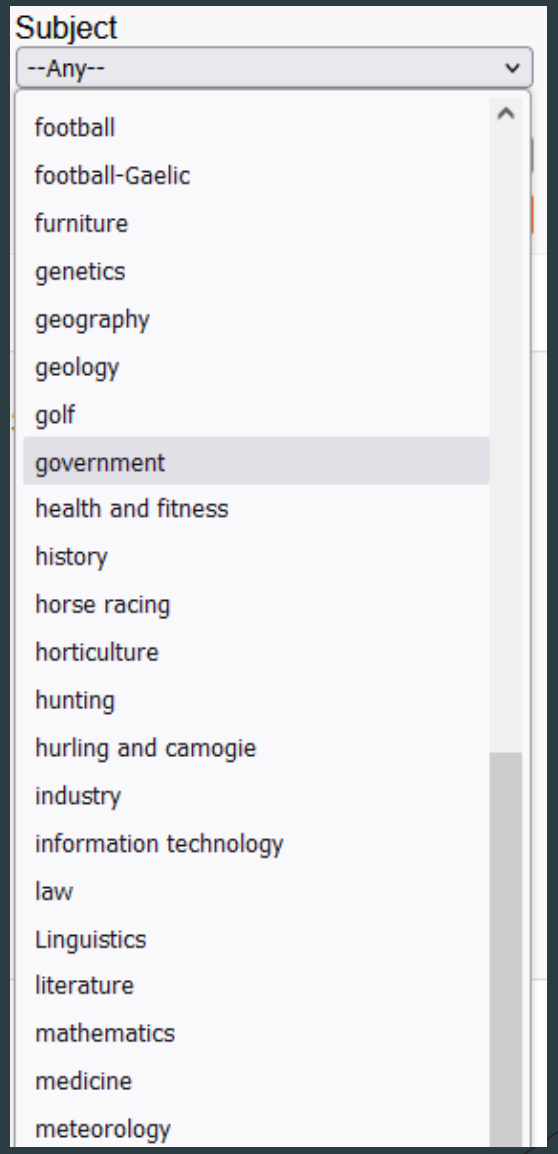
Lexicons

- Manually-built

Token	Lemma	PAROLE	MWE index	USAS Tags
Níor	níor	Q	(9, 10)	Z6
chuir	cuir	Vm	(10, 11)	A1.1.1
sé	sé	Pp	(11, 12)	Z8
cur	cur	Nc	(12, 14)	Q1
síos	síos	R	(12, 14)	Q1
ar	ar	Sp	(14, 15)	Z5
chabhair	cabhair	Nc	(15, 16)	S8
a	a	Q	(16, 17)	Z5
thug	tabhair	Vm	(17, 18)	A9
...				

Lexicons

- Manually-built
- From the *New English-Irish Dictionary*
  - https://www.focloir.ie/en/search/adv



Lexicons

- Manually-built
- From the *New English-Irish Dictionary*
  - https://www.focloir.ie/en/search/adv
- English USAS lexicon
  - Foclóir Gaeilge-Bearla (Ó Dónaill)
  - English-Irish Dictionary (De Bháltraithe)
  - New English-Irish Dictionary (Ó Mianáin)
  - “Core” WordNet
    - 5000 most frequently used senses, cross-linguistically
    - https://wordnet.princeton.edu/download/standoff-files

Disambiguation

- Current pipeline:
  1. PyMUSAS
     - Lexicon(s)
  2. Disambiguation: Document category frequency
     - E.g. for ‘aire’
     - “S8+ A1.3 G1.1” reordered to “G1.1 S8+ A1.3”
  3. Disambiguation: Year detector
     - T1.3 “Time: Period”

Results

Results

- Lexical coverage: 70-85 %
  - Best result: Combined lexicon
- Accuracy: 55-70 %
  - Best result: USAS-en lexicon with all components
  - Observation: NEID lexicon had a negative impact on accuracy

Future Work

Future Work

- Expanding the manually-tagged & checked corpus
- Adding more disambiguation methods
- Named Entities
- Phrasal/Prepositional Verbs, Modals
- ML

Questions