Lemmatization for Under-Resourced Languages with Sequence-to-Sequence Learning: A Case of Early Irish

Part of a PhD project "Deep Learning for Morphological Analysis of Low Resource Languages"

Oksana Dereza, oksana.dereza@gmail.com, github.com/ancatmara

Lemmatization is a crucial text preprocessing task, which is considered solved for most modern resource-rich languages. But what about morphologically complex languages that lack data, and especially annotated data?

- Dump your medieval texts and minority languages, who even needs them?!
- Let's manually annotate our texts! That's not that much after all...
- Hold on, what if there is a dictionary of language X that lists all the lemmata and some forms for each of them? Why not use it as training data?

Yay! Technology!

Early Irish: Source of Data

The Dictionary of the Irish Language [Toner et al. 2007] covers Old and Middle Irish periods. Each of 43,345 entries consists of a headword (lemma), a list of forms including different spellings and compounds, and examples of use with a reference to source text. **The DIL does not cover everything and sometimes is inconsistent!**

Early Irish: Challenges

- Spelling variation
- Initial mutations *(N.sg. céile 'servant' > N.pl. ind chéili 'the servants')*
- Infixed pronouns *(caraid 'he / she / it loves' > rob-car-si 'she has loved you')*
- Complex verbal morphology *(do-beir 'gives' – ní tab(a)ir 'does not give')*

Table 1: Contracted, restored and missing forms and spellings from the DIL

DIL	Restored	Missing
carpat, cairpthiu, -thib, -tiu, -tib	carpat, cairpthiu, caipthib, cairptiu, cairptib	carbad, carbat, carbait, carpait, carput, carpti...
carat(r)as	caratas, caratras	caratrad, caradras, caradrus, caradruis, caratrais...
cruimther, -ir	cruimther, cruimthir	cruimter, crumther, cruimthear, crumper, crumpir, cromthar, crumthirech
anmoth-aig[thig]e	anmothaige, anmothige	anmothaigthech, anmotuighe...
aball, a.	aball	abhull, aboll, ubull, abaill, abla, abhla, ubla, ubhaill...

Table 2: Some forms of the verb 'do-beir'

Form	Deutero-tonic	Prototonic (after preverb)	Translation
INDIC PRES 3SG	do-beir	(ní) thabair	'does (not) give / bring'
SUBJ PRES 3SG	do-bera	(ní) thaibrea	'if does (not) give / bring'
PRET 3SG	do-bert	(ní) thubart	'did (not) give / bring'
FUT 3SG	do-béra	(ní) thibéra	'will (not) give / bring'
PERF 3SG	do-rat	(ní) tharat	'did (not) give'
PERF2 3SG	do-uic	(ní) thuicc	'did (not) bring'

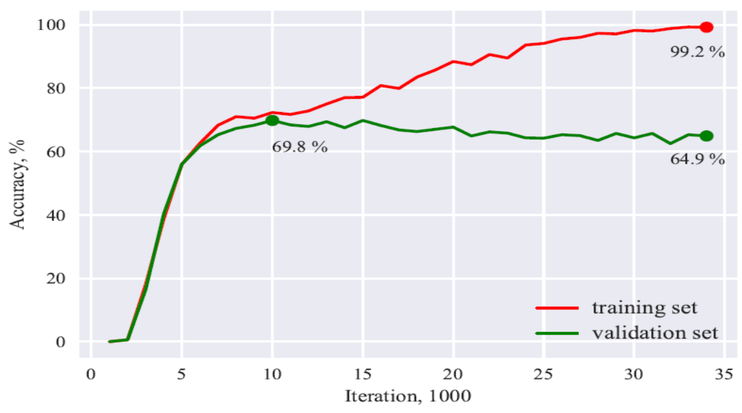
What do we do with that???

If we reformulate lemmatization task as taking a sequence of characters (form) as input and generating another sequence of characters (lemma), we can forget about tens of verbal and nominal inflection classes, let alone spelling variation. Going down to character level might also help to overcome data scarcity.

Experiment

- Character-level sequence-to-sequence model
- 83,155 unique form-lemma pairs from the DIL, split into train, validation and test sets
- Baseline: demutated form

Model	Accuracy (unknown)	Accuracy (known)
baseline	57.5 %	57.5 %
rule-based	45.2 %	71.6 %
char2char	64.9 %	99.2 %



Sequence-to-Sequence Modelling

A sequence-to-sequence model is an ensemble of recurrent neural networks (RNNs) that takes a sequence of a dynamic length as input and produces another sequence of a dynamic length. A basic sequence-to-sequence model consists of two modules, an encoder and a decoder [Cho et al. 2014, Sutskever et al 2014].

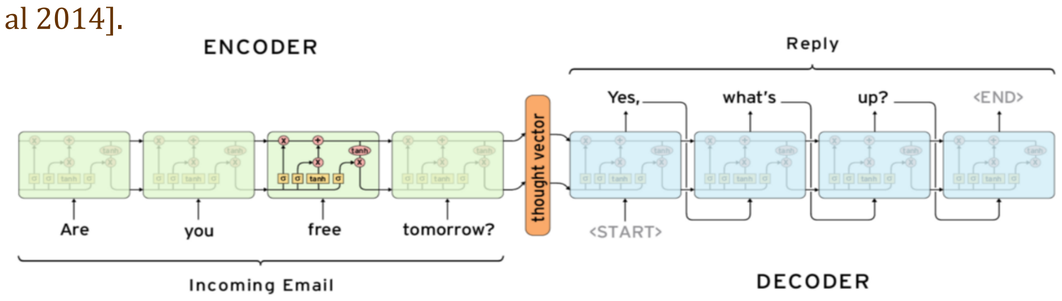


Image source: http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/

Related Tasks

- OCR post-correction and spelling correction: 62.75% to 74.67% accuracy
- Grapheme to phoneme translation: 44.74% to 72.23% accuracy

[Schnober et al., 2016]

Mistakes

form	real lemma	predicted lemma
ar-com-icc	ar-cóemsat	ar-coimcin
dáirfiniu	dáirine	dáirfinu
folortadh	folortad	folortaid
fris-tasgat	fris-tasgat	fris-taig
ithear	ithir	íthra
n-etarcnaigedar	etargnaigidir	etarncaigedar
t-iarrath	íarrath	dírarth

Background image: the Gospels of Máel Brigte (Harley MS 1802), f. 10r