Do not Trust the Experts: How the Lack of Standard Complicates NLP for Historical Irish

Oksana Dereza and **Theodorus Fransen** and **John P. McCrae**
University of Galway
Insight Centre for Data Analytics
firstname.lastname@insight-centre.org

Abstract

In this paper, we describe how we unearthed some fundamental problems while building an analogy dataset to evaluate historical Irish embeddings on their ability to detect orthographic, morphological and semantic similarity. Low agreement among field experts and the absence of an editorial standard in available resources make it impossible to build reliable evaluation datasets for computational models and obtain interpretable results. We emphasise the need for a historical text editing standard, particularly for NLP applications, and prompt Celticists and historical linguists to engage in further discussion. We would also like to draw NLP scholars' attention to the role of data and its (extra)linguistic properties in testing new models and evaluation scenarios.

1 Introduction

Historical languages are known to present greater challenges to NLP due to high orthographic variation, diachronic morphological changes and lack of resources (Piotrowski, 2012; Jenset and McGillivray, 2017; Bollmann, 2019). Our initial goal was to compare different embedding architectures and hyperparameters for detecting morphological and spelling variation in historical Irish, but we unearthed some fundamental problems while we were building an evaluation dataset and testing our models on it.

2 Word Embedding Evaluation Scenarios

There are two main strategies for the evaluation of word embeddings: extrinsic and intrinsic (Schnabel et al., 2015; Bakarov, 2018; Torregrossa et al., 2021). Extrinsic evaluation involves using pretrained embeddings as input vectors in a model solving a downstream NLP task, such as part-of-speech tagging, named entity recognition, or sentiment analysis. The model's performance is believed to reflect the quality of the embeddings it was initialised with. Intrinsic evaluation is focused on assessing linguistic relations within the embedding model itself through solving specially designed mathematical problems: similarity and analogy. The similarity task entails comparing the similarity scores of two words yielded by an embedding model to those calculated based on experts' judgments. The analogy task is a vector proportion, where we ask an embedding model, "What is to b as a' is to a?", and expect b' as an answer.

Generally, task-driven extrinsic evaluation looks more feasible, because it allows the use of already existing evaluation datasets. However, the majority of downstream tasks have not been attempted yet for many minority and historical languages, which leaves us with no available datasets or baselines. As such, constructing a small dataset for intrinsic evaluation seems the best alternative. Both analogy and similarity datasets can be created automatically or semi-automatically by translating an existing dataset from another language, or with the help of a WordNet or a comprehensive dictionary of a language in question in a machine-readable format if there are any. Such a dataset would still require expert proofreading and evaluation, but the amount of manual work would not be as daunting as when a dataset is created from scratch.

3 Early Irish Analogy Dataset

Traditionally, analogy datasets are based on pairwise semantic proportion (Mikolov et al., 2013b), and therefore every question has a single correct answer. Given the high level of variation in historical languages, such a strict definition of a correct answer seems unjustified. Therefore, we follow the creators of the Bigger Analogy Test Set, or BATS (Gladkova et al., 2016). This dataset has highlighted the problems of popular embedding models, such as GloVe, and provided additional proof of the importance of subword information for capturing morphological relations. The origi-

The Fourth Workshop on Insights from Negative Results in NLP, pages 82–87

nal English BATS has successfully been adapted to Japanese (Karpinska et al., 2018) and Icelandic (Friðriksdóttir et al., 2022). Our Early Irish analogy dataset is not a full-scale adaptation of BATS but draws heavily upon the ideas behind it, providing several correct answers for each analogy question and evaluating the performance with set-based metrics proposed by BATS authors, such as an average of vector offset over multiple pairs (3CosAvg) and a logistic regression cosine similarity (LRCos):

$$b' = \text{argmax}_{b' \in V} \left(\cos\left(b', b + \text{avg\_offset}\right)\right), \quad \text{where avg\_offset} = \frac{\sum_{i=0}^{m} a_i}{m} - \frac{\sum_{i=0}^{n} b_i}{n} \quad (1)$$

$$b' = \text{argmax}_{b' \in V} \left(P_{(b' \in \text{ target\_class })} * \cos\left(b', b\right)\right) \quad (2)$$

The Early Irish analogy dataset consists of four parts: morphological variation, spelling variation, synonyms, and antonyms.

The morphological and spelling variation data was automatically extracted from the eDIL (Toner et al., 2019), a digital historical dictionary of medieval Irish covering the period ca. 700 – 1700. Spelling variants were taken from the headwords, and the morphological variation subset was compiled from the 'Forms' field that covers both inflected forms of a headword and its derivatives. Unlike the original BATS, no division was made between different types of inflection, nor between inflection and derivation, within the morphological variation subset because the structure of eDIL does not allow for obtaining this division automatically. We would also like to point out that the eDIL sometimes lists spelling variants along with inflected forms and derivatives in the 'Forms' section, and we did not filter them out manually. The raw data amounted to 2,370 spelling variation and 9,690 morphological variation questions, from which 100 examples were randomly selected for each of the subsets to be comparable in size with the synonym and antonym subsets.

The synonym and antonym subsets are translations of the correspondent BATS parts proofread by four expert evaluators. The translations for each entry in the synonym subsets L07 (intensity, *cry : scream*) and L08 (exact, *sofa : couch*), and antonym subsets L09 (gradable, *clean : dirty*) and L10 (binary, *up : down*) were obtained by reverse-searching the eDIL. The translations were then organised in synsets, each labelled with an English keyword, which the expert evaluators were asked to review. The evaluators were allowed to consult eDIL but were advised not to rely on provided definitions, if in doubt, but instead to utilise their knowledge of how these words occur in texts. The task description also included the following guidelines:

- words in a synset must express the same concept and be of the same part of speech;
- words in a synset must be used in similar contexts and be of the same part of speech;
- a polysemous word can belong to several synsets;
- the annotators should not distinguish between language periods, i.e. an Old Irish and a Middle Irish word can belong to the same synset.

We obtained 98 entries in the synonym subset and 109 entries in the antonym subset, upon which three or more experts agreed. If a word had multiple spellings in the corresponding eDIL entry, we included all of them in these subsets.

4 Experiment, Evaluation and Epic Fail

Our initial goal was to compare different embedding architectures to measure the effect of leveraging subword information on detecting morphological and spelling variation along with semantic similarity in a diachronic scenario. We also aimed at finding the best embedding size for our low-resource and highly inconsistent data. For this purpose, we trained SkipGram (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2017) models with embedding sizes of 20, 50, 100 and 300 on Old and Middle Irish corpora, as well as on both of them combined. We refer to the combined Old and Middle Irish data as 'Early Irish' for convenience, although this term is usually used to describe a broader period, from Primitive Irish ($4^{th} - 6^{th}$ c. A. D.) to Middle Irish ($10^{th} - 12^{th}$ c. A. D.), according to Stifter and Griffith (2021). More information about the training data for embedding models is provided in Table 1. There was no orthographic normalisation (except lowercasing and sentence-level punctuation removal), lemmatisation, or POS-tagging applied. We then tested these embedding models on our analogy dataset using two different metrics, 3CosAvg (Equation 1) and LRCos (Equation 2), with the help of a Python library Vecto.[1]

[1] https://vecto.space/

Dataset	Source	Period	Tokens	Types	TTR
Old Irish	CELT + St. Gall	$8^{th} - 9^{th}$ c.	400,922	77,754	193.9
Middle Irish	CELT	$10^{th} - 12^{th}$ c.	1,071,640	170,851	159.4
Early Irish	CELT + St. Gall	$8^{th} - 12^{th}$ c.	1,171,439	202,172	172.6

Table 1: Embedding model data, periodisation according to Stifter and Griffith (2021). **CELT** = Corpus of Electronic Texts (Ó Corráin et al., 1997), **St. Gall** = Diplomatic St. Gall Glosses Treebank (Doyle, 2020). TTR scores are calculated as $TTR = \frac{types}{tokens} \times 1000$ according to Schlechtweg et al. (2020).

To our surprise, the scores that our embedding models achieved were low enough to be statistically insignificant regardless of the training corpus, hyperparameters and evaluation metrics: the highest accuracy score in the whole experiment was 0.08, achieved by a Middle Irish FastText model with an embedding size = 100 on the morphological variation subset. We carried out a qualitative evaluation to see if our embedding models really did not learn any linguistic patterns from the data, or if the problem lies somewhere else.

First, we made a few queries to the biggest Early Irish FastText model[2] to see the word vectors nearest to these queries. For example, the closest words to *mainister* 'monastery' were its spelling variants (*mainistear, mainistir, mainisttir*), forms with suffixed demonstratives (*mainistir-si, mainistir-se, mainisttir-si, mainistir-sin*) and compounds (*cédmhainistir* 'early monastery, former monastery', *énmhainistir* 'individual monastery'). The name of a legendary Irish king, *Ailill*, yielded spelling variants (*Ailil, Oilill*), mutated and inflected forms (*hAilill, tAilill, Aililla*)[3] and another personal name, *Ailill Miltenga*. The Early Irish SkipGram model with the same parameters did not capture any morphological or spelling variation but detected semantic associates for personal names from the Early Irish literature.

Then, we used the TensorFlow projector (Smilkov et al., 2016) to see if there are any meaningful clusters in the 3D projection of the vector space of the aforementioned Early Irish FastText model. We found many interesting clusters of different sizes, such as nouns referring to peoples perceived as foreign in the Dat. pl. (*allmurachaib* 'to foreigners', *lochlannachaib* 'to Scandinavians', *saxanachaib* 'to Saxons', *paganachaib* 'to pagans') or verbal nouns ending in *-udh* (*etargnaghudh* 'interpreting, explaining', *cotludh* 'sleeping', *slonnudh* 'naming, mentioning' etc.). It is worth mentioning that the model learned subtle spelling differences: the first cluster mentioned above did not include the later spelling variants with the ending *-aibh*, and in the same way, the second cluster did not include earlier spelling variants ending in *-ud* rather than *-udh*. Moreover, nouns with a suffixed demonstrative *sin* formed two different clusters depending on whether the demonstrative was hyphenated (*fechta-sin, sliabh-sin, caislein-sin* etc.) or not (*ceilgsin, uairsin, curuchsin* etc.).

Thus, we witnessed that our models did learn a significant amount of spelling variation, as well as some inflectional and derivational morphology patterns and a limited quantity of semantic similarities. In this case, what factors may have contributed to the inadequate performance observed?

5 Discussion

5.1 Data Sparsity

The first reason, as one might have expected, is data sparsity combined with high variation. The type-token ratios in our Old, Middle and Early Irish datasets are 193.9, 159.4 and 172.6 respectively. A high TTR score means that a significant amount of words is only attested once or twice in the whole corpus. To put these numbers in context, Schlechtweg et al. (2020) report the TTR of 38.24 for Latin and 47.88 for $18 - 19^{th}$ c. Swedish.

The example of *ulchobchán* 'owl' from Table 2 suggests that there are simply not enough occurrences of this word and its variants in the corpus for the model to learn anything about it: the output we got for this query is completely unrelated to it, the most similar word being a special character for *ocus* 'and'. For the same reason, FastText models learned remarkably less semantic similarity than morphological and orthographic similarity, and SkipGram models could not capture much se-

[2]The hyperparameters of this model are the following: emb_size = 300, min_count = 2, window = 10.

[3]Like other Celtic languages, Irish is notable for initial mutations: sound changes at the beginning of a word happening in a certain grammatical environment. In historical Irish, mutations are marked in spelling in a few different ways and sometimes are not marked at all. The first two examples here demonstrate h-prothesis and t-prothesis.

Subset	Query	Translation	Expected Answer	Answer
Spelling	immairecc	conflict, battle	immairg	**immairec**, immaire, *h-immairecc*, immairi, immaircidi, immaircide
Spelling	ulchobchán	owl	ulchobcán, ulchubchán, ulchubcán, ulcachán	_&_, dhocum, puipli, goirti, disciplina, murruscaib
Morphology	asal	donkey	asaile, assail, asail, asala assail	róuasal, uasal, huasal, asalim, an-uasal, anuasal
Morphology	úasal	high, noble	ūassal, uasal, huasil, huasail, úaisliu, húaisliu, huaisliu, huaisle, huaisli, huaislimem, uasalathair, huasalsacart, huasalfichire, úasal-athraig, huasallieig, huasal-gabáltaid, huasalterchomrictid	**anúasal, ardúasal, úasal-nóeb, róuasal, n-úasal**, asal
Antonyms	dorchae	dark, gloomy	gel, gelbdae, gelmar, gleórach, soillsech, soillside, solus	*dorchatae, dorchai, dorcha, dorchato, dorchadu, dorchatu*
Antonyms	descert	south	túaiscert	*ndescert, descertaig, n-descert, descertach, descertaigi*, **túascert**
Synonyms	fliuch	wet	fliuchaide, uiscemail	imliuch, naliuch, fedliuch, nimliuch, **fliuchaidi**, coiuch
Synonyms	álaind	lovely, beautiful	cáem, cáemdae, cruthach, cruthamail, delbach, delbdae	*hálaind, roálaind, comálaind, n-álaind, com-álaind, frálaind*

Table 2: Answers of the biggest Early Irish FastText model compared to expected answers. The words in bold are correct answers that were not present in the evaluation dataset; the words in italic are related to the query, but would not have been correct answers to a particular question.

mantic similarity beyond personal names, as qualitative evaluation has shown.

5.2 Lack of Standardisation of Resources

The second reason is a lack of a text editing standard between different resources for the same historical language, or even within the same resource, which is a case of CELT (Ó Corráin et al., 1997). The usual process of editing manuscript texts includes introducing word spacing, expanding contractions and abbreviations, adding punctuation and sometimes even combining different versions of a text from different manuscripts for linguistic clarity. However, the extent of these changes as well as the use of notation, such as brackets, may differ dramatically from editor to editor. For example, the digital corpora of historical Irish that came out in recent years, St. Gall Priscian Glosses Database (Bauer et al., 2017), Diplomatic St. Gall Glosses Treebank (Doyle, 2020) and CorPH (Stifter et al., 2021), all separate words by different linguistic standards.[4]

The digitised versions of old paper text editions usually include some updates and corrections but still reflect the original editor's ideas of what the text should look like. Moreover, this kind of variation is not reflected in the metadata, and you have to be familiar with each editor's practice to be able to take it into account. Therefore, it is usually almost impossible to use both text and metadata, such as manuscript datings or language periods (Old Irish, Middle Irish etc.), out-of-the-box for NLP applications. These issues have been discussed in Doyle et al. (2018, 2019) in more detail.

How did this lack of standard manifest in our data? About 65 % of morphological and spelling variation subsets, retrieved from eDIL, were not present in the entire Early Irish corpus retrieved from CELT, on which the biggest model was trained. As for synonym and antonym subsets, ca. 30 % are missing in the corpus (see Table 3 for more detail). In other words, a historical dictionary covering mostly Old and Middle Irish periods contains a very high percentage of forms that do not

[4]However, some steps are being made to initiate a standard as far as tokenisation is concerned: thus, the electronic edition of Würzburg glosses (Doyle, 2018) is deliberately tokenised to the same standard as the St. Gall Glosses Treebank.

Dataset	OIr	MIr	EIr	CELT
Morphology (full)	78.7	72.4	69.3	65.4
Morphology (100)	66.2	58.1	54.7	48.5
Spelling (full)	76.7	70.4	68.2	64.0
Spelling (100)	76.5	69.7	66.7	62.6
Synonyms	42.9	36.0	33.3	28.8
Antonyms	45.8	38.2	35.4	30.9

Table 3: The % of missing words from different parts of the analogy dataset (based on eDIL) in the texts from CELT that served as training data for embedding models. **OIr** = Old Irish, **MIr** = Middle Irish, **EIr** = Early Irish (Old + Middle Irish), **CELT** = all Irish texts from CELT, from Old Irish up to Early Modern Irish, including Classical Modern Irish.[5]

occur in real [edited] Old and Middle Irish texts. This also works in the opposite direction: many forms and spellings from the corpus are not listed in the dictionary and, therefore, did not make their way to the evaluation dataset. Such a discrepancy between the corpus on which they were trained and the historical dictionary, which became the source for the evaluation dataset, seriously affected the performance. Table 2 shows that the model often gives reasonable answers, but they are just not among the expected ones. For example, *anúasal, ardúasal, úasal-nóeb, róuasal* are derivatives of *úasal* 'high, noble', and *n-úasal* is its mutated form; thus, they should have been considered correct answers to a morphological similarity question.

5.3 Lack of Agreement between Experts

In addition to the inherent disagreement on fundamental linguistic questions, such as "What is a word?", and on editorial policies ("To what extent should we edit texts? What should the standard for normalisation be?"), scholars do not concur with each other on more specific tasks either.

All the experts who participated in the evaluation are actively working with Early Irish in their research and/or teaching. In addition to that, they were asked to evaluate their knowledge of Early Irish on a scale from 1 ("I did an introductory course") to 5 ("I am experienced in editing Early Irish texts and/or teaching Early Irish") before completing the task. Three of the participants answered with a 4, and one chose a 3, which suggests a profound level of expertise.

[5]Classical Modern Irish is a strict, highly formalised version of Irish used in bardic poetry, which has developed throughout the Middle Irish period and was fixed around the beginning of the 13^{th} century (McManus, 2005).

Despite that, the highest pairwise inter-annotator agreement score between experts, measured using Cohen's kappa, was 0.35, which constitutes only "fair agreement" according to Viera et al. (2005). The Fleiss' kappa score between all four annotators was as low as 0.17, which corresponds to "slight agreement" in Viera et al.'s classification.

6 Conclusion

We discussed an attempt at building an analogy dataset to evaluate historical Irish embeddings on their ability to learn orthographic, morphological and semantic similarity. However, the performance of our models was extremely poor regardless of the architecture, hyperparameters and evaluation metrics, while the qualitative evaluation revealed positive tendencies. Several factors have contributed to it, including a low agreement between experts on fundamental lexical and orthographic issues, and the lack of a unified editorial standard for the language.

These problems are by no means caused by poor scholarly practice. Each of the electronic resource creators pursues a particular, perfectly justifiable editorial approach that dictates their choices. However, the necessity of a text editing standard, especially for NLP applications, has not been properly debated and investigated by the historical Irish academic community. We suspect that this may be the problem of historical languages in general. Through this paper, we would like to highlight this issue and invite Celticists and historical linguists to engage in further discussion.

7 Acknowledgements

This publication has emanated from research in part supported by the Irish Research Council under grant number IRCLA/2017/129 (CARDAMOM – Comparative Deep Models of Language for Minority and Historical Languages). It is co-funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 P2 (Insight 2). We would also like to thank Dr. Elisa Roma, Dr. Eystein Thanisch and Adrian Doyle who took part in the evaluation task together with Dr. Theodorus Fransen.

References

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv:1801.09536*.

Bernhard Bauer, Rijcklof Hofman, and Pádraic Moran. 2017. St. Gall Priscian Glosses, version 2.0. Accessed: 19-02-2023.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Marcel Bollmann. 2019. A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3885–3898.

Adrian Doyle. 2018. Würzburg Irish Glosses. Accessed: 19-02-2023.

Adrian Doyle. 2020. Diplomatic St. Gall Glosses Treebank. Accessed: 19-02-2023.

Adrian Doyle, John P McCrae, and Clodagh Downey. 2018. Preservation of original orthography in the construction of an Old Irish corpus. *Sustaining Knowledge Diversity in the Digital Age*, pages 67–70.

Adrian Doyle, John Philip McCrae, and Clodagh Downey. 2019. A character-level LSTM network model for tokenizing the Old Irish text of the Würzburg glosses on the Pauline Epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79.

Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, Steinþór Steingrímsson, and Einar Sigurdsson. 2022. IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4227–4234.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Gard B Jenset and Barbara McGillivray. 2017. *Quantitative historical linguistics: A corpus framework*, volume 26. Oxford University Press.

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37.

Damian McManus. 2005. Irish Literature [3]. Classical Poetry. In John Thomas Koch, editor, *Celtic Culture: A Historical Encyclopedia*, pages 1003–1005. abc-Clio.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Michael Piotrowski. 2012. *Natural language processing for historical texts*, volume 5 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online).

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon.

Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. In *Proceedings of the Workshop on Interpretable Machine Learning in Complex Systems @ NIPS 2016*.

David Stifter, Bernhard Bauer, Fangzhe Qiu, Elliott Lash, Nora White, Siobhán Barret, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021. Corpus PalaeoHibernicum (CorPH). Accessed: 19-02-2023.

David Stifter and Aaron Griffith. 2021. Lecture notes in Old Irish. Accessed: 19-02-2023.

Gregory Toner, Sharon Arbuthnot, Máire Ní Mhaonaigh, Marie-Luise Theuerkauf, and Dagmar Wodtko. 2019. eDIL 2019: An Electronic Dictionary of the Irish Language, based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913-1976). Accessed: 19-02-2023.

François Torregrossa, Robin Allesiardo, Vincent Claveau, Nihel Kooli, and Guillaume Gravier. 2021. A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11(2):85–103.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Donnchadh Ó Corráin, Hiram Morgan, Beatrix Färber, Gregory Toner, Benjamin Hazard, Emer Purcell, Caoimhín Ó Dónaill, Hilary Lavelle, Seán Ua Súilleabháin, Julianne Nyhan, and Emma McCarthy. 1997. CELT: Corpus of Electronic Texts. Accessed: 19-02-2023. Data downloaded: 15-03-2021.