Temporal Domain Adaptation 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

The digitisation of historical texts has provided new horizons for NLP research, but such data also presents a set of challenges, including scarcity and inconsistency. The lack of editorial standard during digitisation exacerbates these difficulties.

This study explores the potential for temporal domain adaptation in Early Modern Irish and pre-reform Modern Irish data. We describe two experiments carried out on the book subcorpus of the Historical Irish Corpus, which includes Early Modern Irish and pre-reform Modern Irish texts from 1581 to 1926. We also propose a simple orthographic normalisation method for historical Irish that reduces the type-token ratio by 21.43% on average in our data.

The results demonstrate that the use of out-of-domain data significantly improves a language model's performance. Providing a model with additional input from another historical stage of the language improves its quality by 12.49% on average on non-normalised texts and by 27.02% on average on normalised (demutated) texts. Most notably, using only out-of-domain data for both pre-training and training stages allowed for up to 86.81% of the baseline model quality on non-normalised texts and up to 95.68% on normalised texts without any target domain data.

Additionally, we investigate the effect of temporal distance between the training and test data. The hypothesis that there is a positive correlation between performance and temporal proximity of training and test data has been validated, which manifests best in normalised data. Expanding this approach even further back, to Middle and Old Irish, and testing it on other languages is a further research direction.

1 Introduction

With the increasing digitisation of historical texts, more data becomes available for analysis alongside contemporary documents. However, such data poses a set of challenges for any NLP task as it tends to be both scarce and inconsistent. Apart from natural artefacts of language evolution, such as spelling variation and grammatical changes, working with historical languages is complicated by the lack of a linguistic / editorial standard when this data is being digitised (Piotrowski, 2012; Jenset and McGillivray, 2017; Bollmann, 2019). It is especially true for Early Irish, as Doyle et al. (2018, 2019) and Dereza et al. (2023) have pointed out.

In this work, we explore the possibility of temporal domain adaptation[1] on Early Modern Irish and pre-reform Modern Irish data. Although these are not the oldest stages of the Irish language, they are less resourced and more versatile than Modern Irish, which is itself a minority language. We conduct a set of experiments on the use of out-of-domain data, both later and earlier than the target time period, for pre-training embedding models to improve the quality of a language model at the said period. We also investigate the effect that temporal distance between embedding training data and test data has in such a setting. Finally, we propose a simple and efficient normalisation method for historical Irish.

2 Related Work

The surge of interest in distributional semantics has lately reached historical linguistics. A recently emerged concept of diachronic, or dynamic (Bamler and Mandt, 2017; Rudolph and Blei, 2018; Yao et al., 2018; Hofmann et al., 2020), embeddings transforms the task of language modelling into the task of modelling language change, which most papers in this field focus on (Kulkarni et al., 2015; Frermann and Lapata, 2016; Hamilton et al., 2016;

[1] We use the term 'temporal domain adaptation' to describe transfer learning between two different stages of the same language. We believe that this is an instance of domain adaptation, where the main difference between source and target domains is associated with the time when the texts were produced, hence 'temporal'.

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 55–66

Dubossarsky et al., 2017; Rosenfeld and Erk, 2018; Tahmasebi, 2018; Boukhaled et al., 2019; Rodina et al., 2019; Brandl and Lassner, 2019; Hu et al., 2022). In 2018, three comprehensive surveys of detecting and measuring semantic shifts with word embeddings came out (Kutuzov et al., 2018; Tahmasebi et al., 2018; Tang, 2018). In 2020, one of the SemEval shared tasks was dedicated to unsupervised lexical semantic change detection (Schlechtweg et al., 2020). At least two PhD theses on the topic, “Distributional word embeddings in modelling diachronic semantic change” (Kutuzov, 2020) and “Models of diachronic semantic change using word embeddings” (Montariol, 2021), have been defended in the last few years.

Less attention has been paid to addressing the challenges historical languages pose for training a robust embedding model, such as high spelling variation or substantial grammatical change over time. A good example of such a work is a paper by Montariol and Allauzen (2019), who discuss the effectiveness of different algorithms for embedding training in diachronic low-resource scenarios and propose improvements to initialisation schemes and loss regularisation to deal with data scarcity. Di Carlo et al. (2019) are suggesting to use atemporal compass vectors as heuristics while training diachronic word embeddings on scarce data.

On the other hand, the use of closely related languages or language varieties to improve word embeddings and language models in a low-resource setting has been a subject of active discussion. For example, Currey et al. (2016) model a low-resource scenario on Spanish data, using Italian and Portuguese as donor languages for training a statistical machine translation model. Abulimiti and Schultz (2020) work in real low-resource conditions, successfully using Turkish data to improve a language model for Uyghur. Kuriyozov et al. (2020) make another successful attempt at leveraging better-resource Turkic languages to improve the quality of the embeddings for related low-resource languages. Ma et al. (2020) achieve a better performance on the low-resource Tibetan language by training cross-lingual Chinese-Tibetan embeddings. Generally, transfer learning is a popular approach in neural machine translation when it comes to the lack of data, as described in Zoph et al. (2016); Nguyen and Chiang (2017); Kocmi and Bojar (2018); Maimaiti et al. (2019); Chen and Abdul-Mageed (2022). However, the cross-lingual transfer aimed at overcoming data scarcity is not limited to related languages (Adams et al., 2017; Agić et al., 2016). The problem of low-resource scenarios is also discussed in an extensive survey of the cross-lingual embedding models (Ruder et al., 2018).

A few works consider the transfer between different historical stages of the same language as a case of domain adaptation (Yang and Eisenstein, 2015; Huang and Paul, 2019; Manjavacas and Fonteyn, 2022), and we adopt this terminology. Manjavacas and Fonteyn (2022) compare adapting and pre-training large language models for historical English, concluding that pre-training on domain-specific (i.e. historical) data is preferable despite being costly and dependent on the amount of training data.

However, the effect on a language model’s performance produced by initialising it with temporarily distant pre-trained embeddings and by using the out-of-domain temporal data at the training stage has not been evaluated yet, to the best of our knowledge. Moreover, the Irish data has never been used in the research on diachronic word embeddings and temporal domain adaptation before.

3 Data

The data for the experiment is a collection of Early Modern Irish and Modern Irish texts spanning over 350 years, from the late 16^{th} to early 20^{th} century.

Irish belongs to the Celtic branch of the Indo-European language family. Like other Celtic languages, it is notable for initial mutations: sound changes at the beginning of a word happening in certain grammatical environments, which are reflected in spelling. These are combined with a rich nominal and verbal inflection at the end of a word. The four types of initial mutations in modern Irish and their effect on spelling is shown in Table 1.

Before becoming a grammatical feature of the language, mutations happened as historical phonetic processes.[2] For instance, a mutation called *lenition* in the intervocalic position turned Old Irish *cri**d**e* [ˈkʲrʲiðʲe] ‘heart’ into Middle Irish *croi**d(h)**e* / *cri**dh**e* / *crai**d(h)**e* [ˈk⁽ʲ⁾r⁽ʲ⁾iɣʲə] / [ˈk⁽ʲ⁾r⁽ʲ⁾ijə], which later became Modern Irish *croí* [krˠiː].[3]

[2]We apologise for this necessary simplification of historical Irish phonology to our Celticist readers.

[3]Our IPA transcriptions of Middle Irish forms are purely hypothetical. Not enough is known about spoken Middle Irish to say with any authority how things were pronounced, as

Letter	Lenition	Eclipsis	t-prothesis	h-prothesis
b	bh	mb	-	-
c	ch	gc	-	-
d	dh	nd	-	-
f	fh	bhf	-	-
g	gh	ng	-	-
p	ph	bp	-	-
t	th	dt	-	-
m	mh	-	-	-
s	sh	-	ts	-
vowels	-	n-V	t-V	hV

Table 1: Initial mutations in modern Irish.

3.1 Early and Pre-Reform Modern Irish

Early Modern Irish is a term used to describe a vast period in the history of the Irish language between Middle and pre-reform Modern Irish. It spans from the 13^{th} to the 18^{th} century (McManus, 1994) and is marked by multiple religious works (both original and translated), epic tales (both native and adapted from continental material), bardic poetry and historical writing, such as genealogical tracts.

Modern declension and conjugation systems were formed during this period, which makes Early Modern Irish relatively close to what Irish is today, and even closer to what it was before the spelling reform in 1947 and the introduction of the official standard, *An Caighdeán Oifigiúil*, in 1958 (Rannóg an Aistriúcháin, 1958), which is being regularly revised and updated (Tithe an Oireachtais, 2017).

However, both Early Modern Irish and pre-reform Modern Irish texts show considerable spelling variation and unstable grammatical changes, which makes them challenging for NLP tasks (Scannell, 2022).

3.2 Historical Irish Corpus

The data used in the experiments originates in a book subcorpus of the Historical Irish Corpus, or *Corpas Stairiúil na Gaeilge* (hereafter CSnaG), created by the Royal Irish Academy (Acadamh Ríoga na hÉireann; Uí Dhonnchadha et al., 2014). It includes texts from 1581 to 1926 and amounts to 13, 599, 882 tokens. It covers a wide variety of genres, such as bardic poetry, native Irish stories, translations and adaptations of continental epic and romance, annals, genealogies, grammatical and medical tracts, diaries, and religious writing. Each text is dated (both creation and publication dates are provided), and the majority of the texts are author-attributed. The data is available in different formats (plain text, TEI, ePub) along with the metadata on the CSnaG website.[4]

For our purposes, the data was continuously split into 10 parts, 99 texts each, except for the last one, which only includes 97 texts. The motivation for splitting the corpus by the number of texts as opposed to the number of tokens comes from the necessity to keep whole texts within a particular corpus subset to avoid the time, author, and genre interference. Cutting a text into several chunks would have created an overlap between the corpus parts and affected the results of the experiments. Table 2 shows the time frame of each corpus subset along with its size.

3.3 Preprocessing

The texts were split into sentences by the end-of-sentence punctuation marks; then, all sentence-level punctuation was removed and the texts were lowercased. No stemming, lemmatisation or part-of-speech tagging was applied.

In addition to that, a normalised (hereafter ‘demutated’) dataset was created where mutations were removed regardless of their type and position in the word. As a result of such normalisation, *ngrádhmhar* became *grádmar*, *t-ollmhughadh* became *ollmugad*, and so on. Mutations are one of the main sources of spelling variation, especially in the diachronic setting. Although we do lose some grammatical information and sometimes create lexical ambiguities by removing them at the beginning of a word, this change is not critically damaging and is comparable to lemmatisation. Scannell (2020) discusses demutation in modern Irish and the types of errors it can lead to in great detail.

Removing historical mutations that occur in the middle and at the end of a word may, in turn, lead to the conflation of dialectal and standard spellings (standard *d(h)éanfadh* vs. dialectal *d(h)éanfad*), as well as of unrelated words (*óige* ‘youth’ and *óighe*, ‘Gen. sg. Virgin [Mary]’). However, homonymy exists in non-normalised Irish texts too: for instance, *óige* not only means ‘youth’, but can also be a part of the analytical comparative and superlative forms of *óg* ‘young’. A slight increase in homonymy

the writing standard of the period was very archaic. Scribes were following the rules of Old Irish, leaving us with only occasional errors and innovations to conjecture the language they were speaking.

[4] http://corpas.ria.ie/index.php?fsg_function=1

Part	Years	Tokens	Mutated		Demutated		Improvement, %
			Types	TTR	Types	TTR	
0	1581 – 1640	1 669 581	54 748	32.79	42 411	25.40	22.53
1	1640 – 1690	1 524 344	49 658	32.58	39 434	25.87	20.59
2	1691 – 1728	775 412	28 967	37.36	23 425	30.21	19.13
3	1729 – 1771	875 635	33 038	37.73	26 367	30.11	20.19
4	1771 – 1817	688 900	28 708	41.67	22 995	33.38	19.90
5	1817 – 1836	1 094 053	36 048	32.95	28 361	25.92	21.32
6	1836 – 1875	634 692	21 981	34.63	17 468	27.52	20.53
7	1876 – 1908	1 562 576	33 833	21.65	26 185	16.76	22.61
8	1908 – 1919	2 294 943	38 548	16.80	29 132	12.69	24.43
9	1919 – 1926	2 479 746	46 117	18.60	35 501	14.32	23.02

Table 2: Reducing vocabulary size by removing mutations. TTR scores are calculated as $TTR = \frac{types}{tokens} \times 1000$ according to Schlechtweg et al. (2020).

Language	Period	TTR
English	1880 – 1860	13.38
German	1800 – 1899	14.25
Swedish	1790 – 1830	47.88
Latin	−200 – 0	38.24
CSnaG (original)	1581 – 1926	45.50
CSnaG (demutated)	1581 – 1926	33.15

Table 3: TTR scores of Early Modern Irish and pre-reform Modern Irish compared to other historical languages.

Part	Baseline	EX1.1	EX1.2	EX1.3	EX1.4	EX1.5
0	11.25	10.35	14.39	16.62	13.32	19.17
1	8.88	7.97	10.98	13.62	11.20	10.09
2	4.77	3.85	8.30	13.25	8.36	11.96
3	8.27	6.19	10.72	16.95	11.01	12.44
4	8.64	6.77	13.00	19.33	13.55	17.11
5	9.46	9.91	12.70	11.51	11.56	16.37
6	3.85	5.36	10.30	33.02	7.43	20.08
7	9.39	9.60	11.33	16.25	10.38	8.78
8	8.88	9.52	10.68	32.57	10.25	9.97
9	9.52	10.24	11.87	13.88	10.49	26.01
AVG	**8.29**	**7.98**	**11.43**	**18.70**	**10.76**	**15.20**

Table 4: The % of a language model's quality improvement (the decrease in perplexity) achieved by simple orthographic normalisation consisting in the removal of synchronic and historical mutations.

seems to be a justified tradeoff for a significant reduction of vocabulary size unless one is specifically interested in dialectal variation, pronunciation and spelling change, or rhyme patterns in bardic poetry.

Removing mutations from data reduces vocabulary size and type-token ratio (TTR) by 21.43% on average (see Table 2). Moreover, it helps to bridge the gap between Old Irish, where mutations were not marked in writing, and more modern stages of the language. To put these results into context, let us compare TTR scores calculated on the whole CSnaG, containing Early Modern Irish and pre-reform Modern Irish texts, with similar results for historical English, German, Swedish, and Latin provided by Schlechtweg et al. (2020), in Table 3.

Lower TTR has a positive effect on NLP models' performance: in our case, it leads to a notable drop in the perplexity of a language model. Table 4 shows the percent of improvement on demutated texts in comparison to the original ones in each of the experiments, described in more detail in Section 5.1.

4 Methodology

4.1 Embedding Model

We use a FastText (Bojanowski et al., 2017) embedding model that takes subword information into account, which is preferable due to the nature of historical language data. Due to a high degree of variation, which is explained both by the morphological complexity of historical languages and by the lack of standardisation, going down to the subword level is crucial for reducing the vocabulary and effectively dealing with out-of-vocabulary words at the same time. A similar approach is adopted in other works on low-resource data (Kuriyozov et al., 2020; Ma et al., 2020). During our initial set of experiments on non-normalised diachronic Early Irish data, embedding models learned mostly paradigmatic and derivational morphological relations, as well as spelling variation. Some semantic relations were also captured but to a lesser extent (Dereza et al., 2023).

For both experiments described in this paper, all embedding models were trained with the following parameters: embedding size = 100, context window = 10, and minimal count = 2 regardless of vocabulary size. The embedding size is motivated by the experimental results demonstrating that a smaller embedding dimension reduces the model's sensitivity to noise when the data is scarce (Stewart et al., 2017). The low minimal word count is aimed at preserving as much information at each time step as possible.

4.2 Evaluation Scenario

Extrinsic evaluation of embeddings (Schnabel et al., 2015; Bakarov, 2018; Torregrossa et al., 2021) through language modelling seems preferable since it is language-independent and scalable. In addition to that, it does not require manual preparatory work such as dataset creation, unlike other popular downstream tasks, such as bilingual dictionary induction, part-of-speech tagging, or any kind of classification. Hypothetically, using pre-trained embeddings must lower the perplexity score of a language model, even if these were trained on a different period of the language in question.

Perplexity is a standard metric to evaluate language models, which can be defined as the inverse probability of the test set normalised by the number of words. The lower it is, the better.

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta\left(x_i \mid x_{<i}\right)\right\}$$

4.3 Language Model

The configuration of our language model is deliberately simple so that it would allow seeing the contribution that the pre-trained embeddings make to its performance more clearly. It is an LSTM (Hochreiter and Schmidhuber, 1997) with one hidden layer trained until convergence with the Adam optimiser using the early stopping technique, starting with the learning rate = 0.001. The minimum word count was set to 2 to match the pre-trained embedding models. The number of neurons on the hidden layer was calculated depending on corpus vocabulary size as $n_{hidden} = V \times 0.01$ regardless of whether pre-trained embedding models were used or not, and of their vocabulary size. The coefficient was devised empirically based on available computational resources. The pre-trained embeddings were not fixed during the language model training to allow for domain adaptation. More information on vocabulary sizes for each experiment can be found in Tables 9 and 8 in Appendix A.

5 Experimental Results

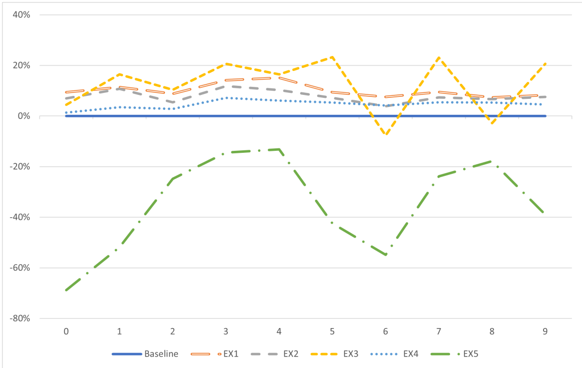

Figure 1: Experiment I: the % of a language model's quality improvement / deterioration in comparison to the baseline, original texts without orthographic normalisation.

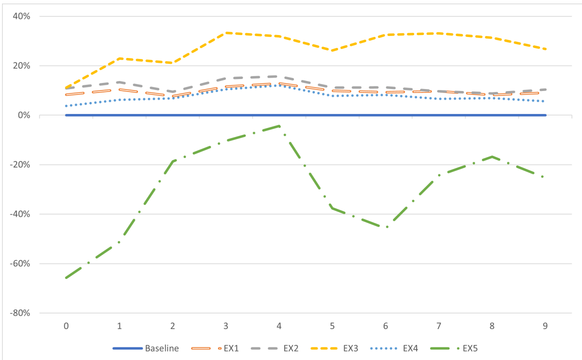

Figure 2: Experiment I: the % of a language model's quality improvement / deterioration in comparison to the baseline, orthographically normalised (demutated) texts.

5.1 Experiment I

Experiment I consisted of 5 tasks summarised in Table 5. Each of these tasks was aimed at answering a particular question about pre-training, such as *"Does the use of an embedding model pre-trained on related data without the target [temporal] domain help to lower the perplexity of a language model at timestamp t_i?"*. The perplexity of a language model trained on a target temporal domain data t_i (i.e. one of the corpus parts № 0-9) without pre-training was taken as a baseline.

№	LM train data	LM test / valid data	Pre-training	Research Question
1.0	t_i	t_i	—	Baseline
1.1	t_i	t_i	t_i	Does pre-training on the target temporal domain t_i help to lower the perplexity of a language model for the timestamp t_i?
1.2	t_i	t_i	T	Does using a bigger pre-trained embedding model, containing more than the target domain, help to lower the perplexity of an LM for the timestamp t_i?
1.3	T	t_i	T	Does the use of out-of-domain data along with in-domain data at both the pre-training and the LM training stages help to lower the perplexity of an LM for the timestamp t_i?
1.4	t_i	t_i	T_{-i}	Does the use of an embedding model pre-trained on related data without the target domain t_i help to lower the perplexity of an LM for the timestamp t_i?
1.5	T_{-i}	t_i	T_{-i}	If we do not have any in-domain data for training, does the use of related data at both the pre-training and the LM training stages help to lower the perplexity of an LM for the timestamp t_i?

Table 5: A overview of Experiment I: t_i refers to a single corpus part from 0 to 9, T stands for the whole corpus, and T_{-i} is the whole corpus excluding a single corpus part from 0 to 9.

Part	EX1.1	EX1.2	EX1.3	EX1.4	EX1.5
0	+9.35	+6.98	+4.43	+1.34	–68.82
1	+11.45	+10.70	+16.49	+3.50	–51.84
2	+8.77	+5.44	+10.40	+2.82	–24.85
3	+14.13	+11.82	+20.67	+7.15	–14.43
4	+15.14	+10.23	+16.49	+6.08	–13.19
5	+9.37	+7.20	+23.27	+5.32	–42.44
6	+7.57	+3.84	–7.69	+4.18	–54.89
7	+9.44	+7.35	+23.03	+5.40	–23.81
8	+7.39	+6.66	–2.80	+5.28	–17.82
9	+8.18	+7.51	+20.64	+4.51	–38.96
AVG	**+10.08**	**+7.77**	**+12.49**	**+4.56**	**–35.10**

Table 6: Experiment I: the % of a language model's quality improvement / deterioration in comparison to the baseline; original texts without orthographic normalisation.

Part	EX1.1	EX1.2	EX1.3	EX1.4	EX1.5
0	+8.25	+10.90	+11.16	+3.76	–65.76
1	+10.36	+13.32	+22.90	+6.21	–51.19
2	+7.72	+9.49	+21.19	+6.85	–18.72
3	+1.60	+14.89	+33.27	+10.46	–10.35
4	+12.83	+15.75	+31.92	+12.10	–4.32
5	+9.92	+11.19	+26.13	+7.82	–37.68
6	+9.29	+11.30	+32.50	+8.19	–45.73
7	+9.69	+9.69	+33.10	+6.57	–24.32
8	+8.15	+8.81	+31.34	+6.89	–16.83
9	+9.04	+10.38	+26.74	+5.64	–25.35
AVG	**+9.68**	**+11.57**	**+27.02**	**+7.45**	**–30.02**

Table 7: Experiment I: the % of a language model's quality improvement / deterioration in comparison to the baseline; orthographically normalised (demutated) texts.

Every corpus part covering a particular period in the history of the Irish language, as shown in Table 2, was split into training (80%), validation (10%), and test (10%) subsets. Validation and test subsets have not been seen by the language model at any stage, including pretraining (i.e. word embeddings were trained only on the training subset of each corpus part).

The results of this experiment are reported in Tables 6 and 7, where each number shows an improvement (marked with a $+$) or a drop (marked with a $-$) in the performance of a language model compared to the baseline. For example, in *Experiment 1.3*, the use of additional out-of-domain data both at the pre-training and training stages results in a 11.16% improvement (i.e. the language model's perplexity drops by 11.16%) in comparison to the baseline on the corpus part № 0 with orthographic normalisation. In other words, adding the texts from $1640 - 1926$ to those from $1581 - 1640$ at both the pre-training and training stages improves the results of the model on the $1581 - 1640$ test data by 11.16%. Generally, *Experiment 1.3* demonstrates that providing a model with additional input improves its quality by 12.49% on average on non-normalised texts and by 27.02% on average on normalised texts.

Similarly, in *Experiment 1.5*, pre-training and training a language model on the whole normalised corpus excluding part № 0 and testing its performance on part № 0 makes the resulting score 65.76% worse (i.e. the language model's perplexity rises by 65.76%). Still, it is not as discouraging as it may seem: it means that we are still able to obtain 34.24% of the baseline model quality even if we do not have the target data from $1581 - 1640$ in our training corpus at all. This number is even higher for later stages of the language, where using related data for training allows to achieve up to 86.81% of the baseline model quality on non-normalised texts and up to 95.68% on normalised texts.

As expected, both pre-training on the same data and using additional out-of-domain data only at the pre-training stage leads to the improvement of a language model's performance despite the shallow architecture of a language model. Naturally, language models trained on earlier texts or on texts with genre-specific language are more sensitive to the absence of in-domain data. For example, parts 5 and 6 include a substantial amount of poetry, which often exhibits a richer, more archaic vocabulary compared to prose.

Figures 1 and 2 provide a graphical overview of the effect that the pre-training data makes on the performance of a language model in comparison to the baseline. Raw sentencewise perplexity scores for the experiment are given in Tables 10 and 11 in Appendix B.

5.2 Experiment II

The second experiment was aimed at observing the effect of the temporal distance between the pre-training and the training/test data. It consisted in the training of language models on each of the 10 corpus subsets initializing them with embeddings pre-trained on each of these corpus parts in all possible combinations. We hypothesised that smaller temporal distances would result in better performance than bigger ones. Our hypothesis has proven correct, as shown in Figures 3 and 4. This correlation is most pronounced when evaluating orthographically normalised (demutated) texts. Naturally, language models fed with embeddings pre-trained on the same data yield the best results. Table 12 in the Appendix C provides the results of this experiment run on non-normalised texts, where all mutations are preserved, and Table 13 presents similar results for demutated texts. Columns correspond to embedding models, and rows are corpus parts they were tested on. For the reader's convenience, we cite *normalised inverse perplexity* instead of the original sentence-wise perplexity scores. It shows how well a model performed in comparison to the best result, where 100% is the best result.

$$NIP = \frac{best\_score}{score} \times 100$$

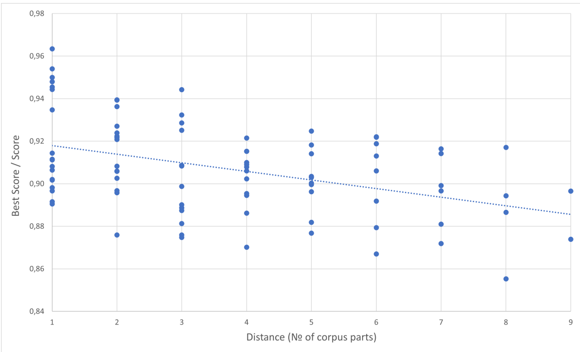

Figure 3: The effect of temporal distance between the pre-training (embedding) data and the language model training and test data; original texts without orthographic normalisation.

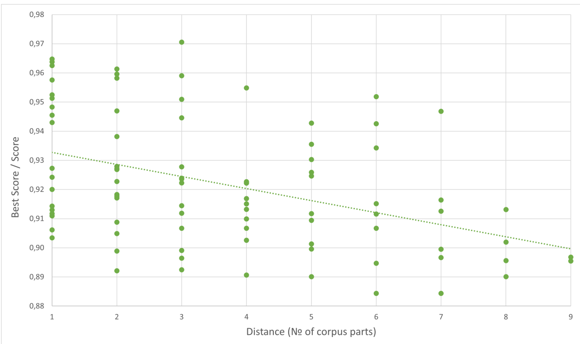

Figure 4: The effect of temporal distance between the pre-training (embedding) data and the language model training and test data; orthographically normalised (demutated) texts.

6 Conclusion

The results cited above testify that using out-of-domain temporal data in the pre-training and training of a language model for a historical language can significantly improve its performance. This is extremely valuable in low-resource scenarios, where we may only have a few texts dating back to a particular period, which would not be enough to train a robust language model. Providing a model with additional input improves its quality by 12.49% on average on non-normalised texts and by 27.02% on average on normalised texts even if this information is retrieved from data covering a different — no matter later or earlier — period in the history of a language. Most importantly, using only out-of-domain data at both pre-training and training stages allows for achieving up to 86.81% of the baseline model quality on non-normalised texts and up to 95.68% on normalised texts without any target domain data.

Our hypothesis that there is a positive correlation between the performance of language models and the temporal proximity of training and test data has been validated. This effect manifests best in orthographically normalised texts. Expanding this approach even further back, to Middle and Old Irish, and testing it on other languages is a further research direction.

Finally, we proposed a simple yet very effective orthographic normalisation method for historical Irish that reduced the type-token ratio by 21.43% on average in our data and allowed for up to 33.02% drop in a language model's perplexity.

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 also wish to acknowledge the Irish Centre for High-End Computing (ICHEC) and our colleague Adrian Doyle for the provision of computational facilities and support.

References

Ayimunishagu Abulimiti and Tanja Schultz. 2020. Building language models for morphological rich low-resource languages using data from related donor languages: the case of Uyghur. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 271–276, Marseille, France. European Language Resources association.

Acadamh Ríoga na hÉireann. Corpas Stairiúil na Gaeilge 1600-1926. Accessed: February 19, 2023. Data downloaded: June 10, 2022.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

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

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas 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.

Mohamed Boukhaled, Benjamin Fagard, and Thierry Poibeau. 2019. Modelling the semantic change dynamics using diachronic word embedding. In *11th International Conference on Agents and Artificial Intelligence (NLPinAI Special Session)*.

Stephanie Brandl and David Lassner. 2019. Times are changing: Investigating the pace of language change in diachronic word embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 146–150.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2022. Improving neural machine translation of indigenous languages with multilingual transfer learning. *arXiv preprint arXiv:2205.06993*.

Anna Currey, Alina Karakanta, and Jon Dehdari. 2016. Using related languages to enhance statistical language models. In *Proceedings of the NAACL Student Research Workshop*, pages 116–123.

Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023. Do not trust the experts: How the lack of standard complicates NLP for historical Irish. In *Proceedings of the 3d Workshop on Insights from Negative Results in NLP, EACL 2023*. Upcoming.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6326–6334.

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.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2020. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*.

Hai Hu, Patrícia Amaral, and Sandra Kübler. 2022. Word embeddings and semantic shifts in historical Spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461.

Xiaolei Huang and Michael Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123.

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

Tom Kocmi and Ondrej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *WMT 2018*, page 244.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW'15*, pages 625–635, Republic and Canton of Geneva, Switzerland.

Elmurod Kuriyozov, Yerai Doval, and Carlos Gómez-Rodríguez. 2020. Cross-lingual word embeddings for Turkic languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Andrey Kutuzov. 2020. Distributional word embeddings in modeling diachronic semantic change. [PhD thesis].

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*.

Wei Ma, Hongzhi Yu, Kun Zhao, Deshun Zhao, and Jun Yang. 2020. Tibetan-Chinese cross-lingual word embeddings based on MUSE. In *Journal of Physics: Conference Series*, volume 1453, page 012043. IOP Publishing.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource NMT using multiple high-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, pages 1–19.

Damian McManus. 1994. An Nua-Ghaeilge Chlasaiceach. In K. McCone; D. McManus; C. Ó Háinle; N. Williams; L. Breatnach, editor, *Stair na Gaeilge: in ómós do Pádraig Ó Fiannachta*, pages 335–44. Maynooth: Department of Old Irish, St. Patrick's College.

Syrielle Montariol. 2021. *Models of diachronic semantic change using word embeddings*. Ph.D. thesis, Université Paris-Saclay.

Syrielle Montariol and Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 795–803, Varna, Bulgaria. INCOMA Ltd.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301.

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

Rannóg an Aistriúcháin. 1958. Gramadach na Gaeilge agus litriú na Gaeilge: An caighdeán oifigiúil. *Baile Átha Cliath/Dublin: Oifig an tSoláthair*.

Julia Rodina, Daria Bakshandaeva, Vadim Fomin, Andrei Kutuzov, Samia Touileb, and Erik Velldal. 2019. Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian. Association for Computational Linguistics.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*.

Maja Rudolph and David Blei. 2018. Dynamic Bernoulli embeddings for language evolution. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM.

Kevin Scannell. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Kevin Scannell. 2020. Neural models for predicting Celtic mutations. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 1–8.

Kevin Scannell. 2022. Diachronic parsing of prestandard Irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 7–13.

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). International Committee for Computational Linguistics.

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, Portugal. Association for Computational Linguistics.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. 11(1):672–675.

Nina Tahmasebi. 2018. A study on word2vec on a historical Swedish newspaper corpus. In *DHN*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. In *Computational Linguistics*, volume 1.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Tithe an Oireachtais. 2017. Gramadach na Gaeilge. An Caighdeán Oifigiúil. Accessed: February 19, 2023. Data downloaded: July 6, 2022.

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.

Elaine Uí Dhonnchadha, Kevin Scannell, Ruairí Ó hUiginn, E. Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair, and Niall O'Leary. 2014. Corpas na Gaeilge 1882–1926: Integrating Historical and Modern Irish Texts. In *LREC 2014 Workshop LRT4HDA: Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage*, pages 12–18, Reykjavik, Iceland.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 672–682.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

A Vocabulary Sizes

Part	EX1.1		EX1.2		EX1.3		EX1.4		EX1.5	
	Original	**Normalised**	**Original**	**Normalised**	**Original**	**Normalised**	**Original**	**Normalised**	**Original**	**Normalised**
0	60,042	47,688	60,042	47,688	210,537	161,958	60,042	47,688	183,439	141,804
1	53,202	43,103	53,202	43,103	209,507	161,323	53,202	43,103	187,557	144,540
2	30,847	25,358	30,847	25,358	206,508	159,109	30,847	25,358	197,197	151,883
3	36,141	29,205	36,141	29,205	207,025	159,467	36,141	29,205	195,729	150,838
4	31,829	25,796	31,829	25,796	206,679	159,233	31,829	25,796	196,818	151,708
5	39,330	31,268	39,330	31,268	207,517	159,726	39,330	31,268	194,385	150,004
6	24,738	19,962	24,738	19,962	205,936	158,630	24,738	19,962	198,647	153,164
7	39,286	30,811	39,286	30,811	207,110	159,570	39,286	30,811	194,832	150,355
8	44,870	34,039	44,870	34,039	207,301	159,558	44,870	34,039	190,256	150,169
9	53,400	41,417	53,400	41,417	208,567	160,595	53,400	41,417	189,740	146,577

Table 8: Corpus vocabulary sizes. The data used in the Experiment II is the same as in the Experiment 1.1

<table>
<tr><th rowspan="2">Part</th><th colspan="2">EX1.1</th><th colspan="2">EX1.2</th><th colspan="2">EX1.3</th><th colspan="2">EX1.4</th><th colspan="2">EX1.5</th></tr>
<tr><th>Original</th><th>Normalised</th><th>Original</th><th>Normalised</th><th>Original</th><th>Normalised</th><th>Original</th><th>Normalised</th><th>Original</th><th>Normalised</th></tr>
<tr><td>0</td><td>51,302</td><td>41,268</td><td rowspan="10">204,290</td><td rowspan="10">157,402</td><td rowspan="10">204,290</td><td rowspan="10">157,402</td><td>175,325</td><td>135,366</td><td>175,325</td><td>135,366</td></tr>
<tr><td>1</td><td>45,554</td><td>37,176</td><td>181,140</td><td>139,403</td><td>181,140</td><td>139,403</td></tr>
<tr><td>2</td><td>26,497</td><td>21,909</td><td>194,791</td><td>150,019</td><td>194,791</td><td>150,019</td></tr>
<tr><td>3</td><td>30,872</td><td>25,175</td><td>192,775</td><td>148,533</td><td>192,775</td><td>148,533</td></tr>
<tr><td>4</td><td>27,073</td><td>22,064</td><td>194,493</td><td>149,911</td><td>194,493</td><td>149,911</td></tr>
<tr><td>5</td><td>33,609</td><td>26,931</td><td>190,620</td><td>147,236</td><td>190,620</td><td>147,236</td></tr>
<tr><td>6</td><td>21,274</td><td>17,298</td><td>196,621</td><td>151,648</td><td>196,621</td><td>151,648</td></tr>
<tr><td>7</td><td>34,108</td><td>26,901</td><td>191,514</td><td>147,827</td><td>191,514</td><td>147,827</td></tr>
<tr><td>8</td><td>39,220</td><td>30,063</td><td>190,256</td><td>147,416</td><td>147,416</td><td>147,416</td></tr>
<tr><td>9</td><td>46,447</td><td>36,260</td><td>183,939</td><td>142,114</td><td>142,114</td><td>142,114</td></tr>
</table>

Table 9: Vocabulary sizes of the pre-trained embedding models. The models used in the Experiment II are the same as in the Experiment 1.1

B Experiment I

Part	Baseline	EX1.1	EX1.2	EX1.3	EX1.4	EX1.5
0	336.35	307.58	314.40	322.07	331.90	1078.61
1	337.98	303.26	305.32	290.13	326.54	701.80
2	361.98	332.79	343.32	327.89	352.05	481.70
3	412.06	361.04	368.50	341.49	384.55	481.53
4	542.83	471.44	492.45	465.98	511.74	625.31
5	351.83	321.69	328.19	285.42	334.07	611.22
6	266.43	247.67	256.58	288.62	255.75	590.64
7	230.54	210.66	214.76	187.38	218.73	302.57
8	180.49	168.07	169.22	185.69	171.44	219.63
9	222.64	205.81	207.08	184.55	213.03	364.72
AVG	**324.31**	**293.00**	**299.98**	**287.92**	**309.98**	**545.77**

Table 10: Experiment I: sentencewise perplexity scores; original texts without orthographic normalisation.

Part	Baseline	EX1.1	EX1.2	EX1.3	EX1.4	EX1.5
0	298.50	275.75	269.15	268.53	287.69	871.87
1	307.98	279.08	271.79	250.60	289.96	630.99
2	344.70	319.99	314.81	284.44	322.61	424.11
3	377.99	338.7	329.01	283.62	342.20	421.61
4	495.91	439.51	428.44	375.91	442.40	518.32
5	318.56	289.82	286.51	252.56	295.45	511.15
6	256.16	234.39	230.16	193.33	236.76	472.01
7	208.89	190.44	190.43	156.94	196.02	276.00
8	164.46	152.07	151.14	125.22	153.86	197.73
9	201.44	184.74	182.50	158.94	190.68	269.85
AVG	**297.46**	**270.45**	**265.39**	**235.01**	**275.76**	**459.36**

Table 11: Experiment I: sentencewise perplexity scores; orthographically normalised (demutated) texts.

C Experiment II

Part	0	1	2	3	4	5	6	7	8	9
0	100.00	90.65	87.59	87.59	87.02	88.18	86.70	87.19	85.53	87.39
1	91.44	100.00	89.05	89.67	87.48	88.62	87.68	87.94	88.11	88.66
2	93.94	95.00	100.00	93.48	92.71	93.23	90.60	91.41	92.19	91.64
3	88.87	90.26	91.12	100.00	89.81	90.83	88.74	89.54	90.35	91.88
4	90.23	88.13	90.59	89.67	100.00	90.81	90.58	89.01	90.79	91.83
5	90.03	89.45	90.87	92.39	90.18	100.00	90.20	89.57	90.86	90.90
6	92.20	90.30	92.49	94.43	92.24	94.43	100.00	91.15	92.18	92.51
7	89.91	89.19	89.96	91.53	90.83	92.08	89.16	100.00	94.55	93.94
8	91.71	91.43	91.30	92.47	92.15	92.86	92.18	95.40	100.00	96.34
9	89.65	89.43	89.67	90.61	89.63	91.00	89.87	93.62	94.80	100.00

Table 12: Experiment II. Original texts, normalised inverse perplexity scores in %, where 100% is the best score. Columns correspond to embedding models, and rows are corpus parts they were tested on.

Part	0	1	2	3	4	5	6	7	8	9
0	100.00	91.30	89.21	89.24	89.07	89.96	88.44	88.43	89.01	89.55
1	92.42	100.00	91.17	90.88	89.64	90.26	89.01	89.47	89.95	90.20
2	96.14	95.76	100.00	94.55	93.82	95.09	92.22	93.55	95.18	94.68
3	91.19	91.77	92.01	100.00	91.43	92.72	91.44	90.99	92.59	93.42
4	91.69	92.22	91.83	94.29	100.00	92.73	89.88	94.46	92.27	94.28
5	90.94	91.51	90.67	92.79	90.34	100.00	90.62	91.71	92.78	92.22
6	94.25	93.02	95.48	97.06	94.70	96.25	100.00	95.13	95.96	95.90
7	91.64	91.52	91.17	92.25	92.36	92.27	91.10	100.00	96.48	95.82
8	91.31	91.25	91.16	92.46	91.31	92.38	90.49	95.25	100.00	96.38
9	89.68	89.56	89.67	90.67	90.13	90.67	89.91	92.69	94.83	100.00

Table 13: Experiment II. Demutated texts, normalised inverse perplexity scores in %, where 100% is the best score. Columns correspond to embedding models, and rows are corpus parts they were tested on.