Adapting Term Recognition to an Under-Resourced Language: the Case of Irish

John P. McCrae
Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
john@mccr.ae

Adrian Doyle
National University of Ireland Galway
A.DOYLE35@nuigalway.ie

Abstract

Automatic Term Recognition (ATR) is an important method for the summarization and analysis of large corpora, and normally requires a significant amount of linguistic input, in particular the use of part-of-speech taggers. For an under-resourced language such as Irish, the resources necessary for this may be scarce or entirely absent. We evaluate two methods for the automatic extraction of terms, based on the small part-of-speech-tagged corpora that are available for Irish and on a large terminology list, and show that both methods can produce viable term extractors. We evaluate this with a newly constructed corpus that is the first available corpus for term extraction in Irish. Our results shine some light on the challenge of adapting natural language processing systems to under-resourced scenarios.

1 Introduction

Automatic term recognition (ATR) is the task of identifying relevant and interesting terms from a text corpus. This can be useful for a wide range of text understanding tasks, however most of the work on this task has to date focused on term extraction for English. In contrast, there are up to 7,000 languages spoken in the world, most of which are severely under-resourced, and the task of adapting Natural Language Processing (NLP) tools to such languages is still not well explored. The principle issue for these language is the lack of resources available and as such they are called *under-resourced languages*. In this paper, we will focus on the development of automatic term recognition for the Irish language, an under-resourced Celtic language spoken primarily on the island of Ireland. In particular, we will base our work on the previously developed Saffron system (Bordea et al., 2014; Pereira et al., 2019). The main requirements for this are the development of a part-of-speech tagger, a lemmatizer and a large background corpus and we will detail in this paper how we constructed these models for Irish.

In particular, the largest challenge was the construction of a part-of-speech tagger and we base our work on two main systems that have been developed based on annotated corpora. Firstly, we look at the system of Uí Dhonnchadha and van Genabith (2006), which was developed on a general language domain and secondly we refer to the system of Lynn et al. (2015), which was developed specifically for tweets. We then looked at an alternative approach using the terminology database, Tearma[1], to provide an annotation over the Irish Wikipedia, 'An Vicipéid'[2]. For both the systems trained on part-of-speech corpora and those on the terminology database, we compare them for the challenge of recognizing terms. We show how we incorporate into our term recognition system morphology information extracted from Pota Focal (Měchura, 2018). To analyse this we developed a small gold standard dataset of Wikipedia articles and compared the two methods on this dataset[3]. We then describe the construction of the automatic

[1] https://www.tearma.ie/eolas/tionscadal.en

[2] https://ga.wikipedia.org/wiki/Pr\%C3\%ADomhleathanach

[3] Datasets and code developed in this work are available at https://github.com/jmccrae/irish_saffron

term recognition system and compare the results of these two methods on a small corpus of discussion related to the future of the National University of Ireland Galway. Our results show that both methods provide a viable method of constructing a term extraction system, however there is still a need for significant language specific knowledge in the development of such a system and that new generic methods would be necessary to scale this to more under-resourced languages.

2 Related Work

Automatic term recognition is an area that has seen interest for a long time (Kageura and Umino, 1996), and a number of supervised and unsupervised methods have been proposed. More recently this has led to a couple of mature toolkits for this including Jate (Zhang et al., 2016), ATR4S (Astrakhantsev, 2018) and Saffron (Pereira et al., 2019), the latter of which we use as the basis for this work. This work has been characterized in terms of filters that extract terms, either in terms of 'closed' filters that focus only on nouns (Arora et al., 2014) and open filters that include adjectives (such as in this work). Open filters capture more general terms consisting of adjective and nouns such as 'natural language processing', which cannot be captured by closed filters, which would only accept noun terms such as 'language processing'. The result of choosing an open filter is a trade-off that increases the recall of the system at the cost of precision. Thus, in order to ensure high-quality results, there are a number of methods of ranking that are performed in order to rank the terms and thus to improve the precision of the top ranked candidates. The initial methods in this area focused on the use of term frequency statistics such as TF-IDF (Evans and Lefferts, 1995), or the relative frequency of the term compared to a background corpus (Ahmad et al., 1999; Peñas et al., 2001; Church and Gale, 1999). A further approach has been based on the analysis of the term, and in particular the presence of subterms in the same domain, which can be indicative of termhood (Buitelaar et al., 2013). It has been shown that the best performance is generally obtained through a combination of these methods (Astrakhantsev, 2018).

3 Methodology

The methodology for automatic term extraction as implemented by the Saffron system consists of the following steps

1. Part-of-speech tagging is applied to the text corpus.

2. The candidate terms are extracted using a simple regular expression over the part-of-speech tags. For English texts tagged with the Penn Treebank, this was `((NN|JJ|NNP|NNS)+ (IN|NN|JJ|NNP|NNS)*)? (NN|CD|NNS)`

3. A morphological engine is used to create a single normalized base form for the term, e.g., in English we turn plural nouns into singular nouns.

4. The frequency of the terms is recorded and from this a number of metrics are calculated (see Section 3.4).

5. The candidates are ranked according to the mean reciprocal rank of the metrics and top N candidates are returned.

From this it can be seen the key language-dependent elements are: part-of-speech tagging, term normalization and the inclusion of a background corpus for some of the metrics. We will explain how we adapted this procedure to Irish.

3.1 Morphology

Irish morphology is noticeably more complex than that of English and this presents a challenge for processing the language that should generally require more resources. For automatic term recognition it is not in general necessary to consider verbs as they do not generally occur in terms, which in the context of Irish is beneficial as verbal morphology is more complex than nominal morphology. On the other hand, verbal morphology is generally regular in Irish, whereas nominal morphology is mostly irregular with plural and genitive forms not generally being predictable from the lemma. As such, the only high accuracy approach to handling Irish nominal morphology is a dictionary approach and for this we used the Pota Focal dictionary (Měchura, 2018), as it provides an easy to parse XML version of the morphology for the basic vocabulary of the language. In total there are 4,245 lemmas (of which 3,488 are nouns) in Pota Focal, which we used in this work.

dia (7,747)	dé (14,450)	déithe (400)
dhia (2,671)	dhé (83)	dhéithe (59)
ndia (231)	ndé (33)	ndéithe (157)
ollscoil (4,189)	ollscoile (1,141)	ollscoileanna (265)
hollscoil (106)	hollscoile (1,438)	hollscoileanna (234)
n-ollscoil (7)	n-ollscoile (2)	n-ollscoileanna (41)
t-ollscoil (0)*	t-ollscoile (0)*	t-ollscoileanna (0)*

Table 1: Example of the forms that are lemmatized to 'dia' (god) and 'ollscoil' (university) and their frequency in the New Corpus for Ireland. *Ungrammatical forms.

However, a particular challenge with Irish (along with other Celtic languages) is initial mutation, that is the changing of initial consonant by lenition, eclipsis or prefixing of a consonant to a word starting with a vowel. We used hard-coded rules to generate the forms of each word with initial mutation as they were not included in Pota Focal directly, but could be easily and systematically derived. We over-generate forms including applying a t-prefix to feminine nouns such as 'ollscoil', on the principle that it is unlikely that we will generate any errors from recognizing too many forms of the noun. An example of all the forms is given in Table 1 and we give the frequency of each form in the New Corpus for Ireland (Kilgarriff et al., 2006), showing that all forms do occur in text, even those that may be considered ungrammatical. The morphology engine is then implemented by a simple lookup.

3.2 Part-of-speech Tagging

Corpus	Documents	Words	#POS
Uí Dhonnchadha	42	63,096	16
Lynn	3,032	52,279	22

Table 2: Analysis of part-of-speech Corpora used in this work. #POS refers to the number of distinct top-level part of speech categories.

The most important step for the creation of the tool is the identification of terms from the text and this is achieved in English by means of a regular expression over the output of a part-of-speech tagger. For adapting this to Irish, there is the obvious challenge that there is much less available training data for a part-of-speech tagger and secondly that the part-of-speech tagset would naturally differ from that of English, as for example there is no tag for genitive noun in English. To our knowledge there are two part-of-speech corpora available for Irish of sufficient size to apply machine learning techniques. The first one is from Uí Dhonnchadha and van Genabith (2006) and this corpus consists of the annotation of a number of documents, while a more recent corpus is due to Lynn et al. (2015) and this was created on Twitter by annotating a number of tweets. The basic statistics of the two corpora are given in Table 2, and we can see that both corpora are similar in size (number of words) but there are differences in the number of documents due to the nature of the annotation as in the case of Lynn's corpus each tweet is considered a single document. Uí Dhonnchadha's corpus has more detailed part-of-speech types, however for the purpose of this work we consider only the top category part-of-speechs (e.g., 'noun', 'verb'). In order to adapt our ATR system to this task we further aligned the two corpora to use a single part-of-speech tagging using the following categories: **N**oun, **V**erb, **A**djective, Adve**r**b, Prepo**s**ition, **C**onjunction, **P**ronoun, Par**t**icle, **D**eterminer and demonstrative[4], Nu**m**eral and **O**ther[5]. Further, we considered verbal nouns as verbs as we do not wish them to be extracted as terms, however we note that this could cause issues as there are many cases where there would be ambiguity between nouns and verbal nouns, for example 'aistriú' means 'translation' as a noun, but 'moving' or 'translating' as a verbal noun. We expect that the original corpora have made this distinction consistently so as to enable ATR, but this is certainly an aspect that deserves further investigation. As such we can use the following regular expression to identify terms

[4]Actually determiners (e.g., 'an', 'na') and demonstratives (e.g., 'seo', 'sin', 'úd') are clearly distinct in Irish grammar also determiners, as determiners precede the noun and demonstratives follows the noun. In Uí Dhonnchadha's corpus they are distinct but Lynn confounds them, as such this was the only major failing in harmonizing the two tagsets.

[5]We merged many of Lynn's categories into this category as they were specific to Twitter, e.g., Lynn has two tags for hashtags.

in the text:

```
N((N|A|D)*(N|A)+)?
```

Note that this expression allows an article to occur in the middle of a term, which is quite common in Irish, for example in 'Banc na hÉireaan' (Bank of Ireland). In addition, we observe that it is common for terms in Irish to either start with an article, for example 'An Fhrainc' (France) or contain a preposition, such as 'aistriú focal ar fhocal' (translating word by word), however initial experiments suggested that including prepositions in the pattern lead to too many false positive terms.

3.3 Weak Supervision

While the part-of-speech tagging approach described above has been successful in English and our results show that it is an effective method also for Irish, there are some clear shortcomings of the approach. In particular, the corpora we train on are quite small and as such there is a necessity to make trade-offs for part-of-speech tags that rarely occur within a term. As an alternative, we considered the use of a large database on known terms which exists in the form of the Tearma database. As such we attempted to train a model that could work at identifying terms in context. To achieve this we collected a large corpus of Irish from the Irish Wikipedia, which was selected due to its size and availability but also due to its technical nature meaning that it is likely to contain the terms used in a similar manner to the Tearma database. We used the dump from April 2019 and in total we extracted 10,074 articles totalling 4,093,665 words and we identified all terms from the Tearma database that occur in this corpus of which we found 24,038 terms. We trained our tagging model based on a simple **IOB** tagging (Ramshaw and Marcus, 1999) where a word was tagged as **B** if it was first word from a term, **I** if it occurred in a non-initial position in term and **O** and if it was not in a term in the Tearma database. This naturally leads to a large number of false negatives as many terms that are used in An Vicipéid are not in Tearma, more concerningly we also found a large number of false positives as there were terms in the database that were similar to other common words. An example of this was 'IS', which is an abbreviation for 'Intleacht Shaorga' (Artificial Intelligence), but also matched a very common form of the copula. As such we also filtered the term database as follows:

- If the term occurred more than 3,000 times (this value was hand-tuned) in the corpus it was rejected,
- If the term occurred more than 100 times in the corpus it was accepted only if the first word was marked as a noun in Pota Focal,
- If the term occurred less than 100 times it was accepted as a term.

We also converted the corpora of Uí Dhonnchadha and Lynn to the IOB format so that we could compare the result.

3.4 Term Ranking

The goal of the previous task was to identify candidate terms from the text, and the next step is normally to provide a ranking of these terms so that those which are most relevant to the domain can be identified. A first step is then to provide some basic filters to remove some incorrect terms. In particular, we do the following:

- Filter by the length of the term (up to a maximum of 4 words)
- Remove all terms that consist solely of stopwords[6].
- Has a minimum number of occurrences in the corpus. However, given the size of the corpus we had, this number was set to 1, and so effectively this filter was ignored

We then carried out the scoring of each term according to multiple metrics, this has been shown in previous work (Astrakhantsev, 2018) to be very effective and allows the method to be adjusted to the task. To this extent, we consider a corpus, C, and consider $t \in C$ to a term extracted in the first step. Then, we develop a number of functions $f_i : T \to \mathbb{R}$ that produce a score for this.

We can broadly group the ranking categories into four categories:

3.4.1 Frequency of Occurrences

These methods consider as primary evidence the frequency and distribution of the words, in particular focusing on words that are prevalent in only a few documents in the corpus. We define as usual

[6]This proved very useful as the system was lemmatizing 'bhfuil' (a form of the verb 'bí', to be) as 'fuil' (blood)

a set of documents, D, and for each word a frequency across all documents denoted, $tf(w)$. We can then define document frequency, $df(w)$, as the number of documents, $d \in D$, where the word occurs at least once. We can then define the following basic metrics:

Total TF-IDF is a well-established method for estimating the importance of a term based on how frequently occurs but penalizing terms that occur uniformly across the corpus.

$$\text{Total TF-IDF}(w) = tf(w) \log\left(\frac{|D|}{df(w)}\right)$$

Residual IDF (Church and Gale, 1995) compares the distribution of TF-IDF against an expectancy of it being randomly distributed.

$$\text{Residual IDF}(w) = tf(w) \times \left[\log_2\left(1 - \exp\left(\frac{tf(w)}{|D|}\right)\right) - \log_2\left(\frac{df(w)}{|D|}\right)\right]$$

3.4.2 Context of occurrences

These functions incorporate the distributional hypothesis (Harris, 1954), by including information about how terms occur within other terms. For this we define $T_{sub}(w)$ as the set of terms which are contained in w, that is all sub-sequences of the words of w and $T_{super}(w)$ as all terms that contain w occurring in the corpus. We can then defined the following metrics:

Combo Basic (Astrakhantsev, 2015) uses the count of both the super- and subterms as well as the length (in words) of the term, $|w|$:

$$\text{ComboBasic}(w) = |w| tf(w) + \alpha |T_{super}(w)| + \beta |T_{sub}(w)|$$

Similarly, cValue (Ananiadou, 1994) uses the subterm frequency as well:

$$\text{cValue}(w) = \log_2(|w| + 0.1) \times \left(tf(w) - \frac{\sum_{t' \in T_{sub}(w)} tf(t')}{|T_{sub}(w)|}\right)$$

The domain coherence measures the correlation, using probabilistic mutual information, of the term with other words in the corpus and then uses this to predict a score, in particular we use the PostRankDC method (Buitelaar et al., 2013).

3.4.3 Reference Corpora

Another important distinguishing factor about terms is that they are very frequent in their domain but not widely used outside that domain. We do measure this by taking a background corpus with term frequencies given as $tf_{ref}(w)$, let $T = \sum_t f(w)$ be the total size in words in the foreground corpus and T_ref be the total total size of the background corpus. We can define Weirdness (Ahmad et al., 1999) as:

$$\text{Weirdness}(w) = \frac{tf(w)}{tf_{ref}(w)}$$

And a second metric Relevance (Peñas et al., 2001) as:

$$\text{Relevance}(w) = 1 - \log\left(2 + \frac{tf(w) T_{ref} df(w)}{tf_{ref} w T |D|}\right)$$

3.4.4 Topic Modelling

Finally, the use of topic models has been suggested based on the success of Latent Dirichlet Allocation (Blei et al., 2003) in the form of the Novel Topic Model (NTM) (Li et al., 2013), although we did not in fact use this metric, as our previous experiments have shown it to perform poorly. NTM requires a probability distribution of a word being labelled to one of K topics, $p(w_i = w | z_i = k)$, the score is then calculated as

$$\text{NTM}(w) = tf(w) \sum_{v \in w} \max_k P(w_i = w | z_i = k)$$

3.4.5 Multi-metric scoring

Once all the scores for all candidate terms have been calculated, a ranking of the top terms is necessary. In general, these terms produce very different scores and as such, methodologies such as linear models (e.g., support vector machines) or simple classifiers (e.g., feed-forward neural networks) would not work well and would require significant training data. Instead, we have observed that the use of the unsupervised methods of *mean reciprocal rank* produces a very strong result without the need for training. For this we produce from each score a ranking function $R_i : T \to \mathbb{N}$ that produces the rank (from 1) of the score and then calculate the final score as:

$$score(t) = \sum_{i}^{n} \frac{1}{R_i(t)} \quad (1)$$

For our experiments we used a combination of metrics that has proven to work well across many settings that consist of the five scores: ComboBasic, Weirdness, TF-IDF, cValue and Residual IDF. Then we apply a filtering step to select the top n candidates; for our experiments we set $n = 100$.

4 Gold Standard Creation

	B	I	O
Uí Dhonnchadha	22%	17%	61%
Lynn	19%	10%	71%
Tearma	19%	2%	80%
Gold	16%	11%	73%

Table 3: The comparative tagging of each of the corpora using the IOB scheme.

In order to evaluate this approach we manually annotated a small section of the Wikipedia corpus. In total we annotated 11 documents consisting of 5,178 words and found among those 846 terms. This annotation was carried out by a single annotator and while this makes it difficult to estimate the quality of the annotation, this is unfortunately a typical issue with developing resources for under-resourced languages. In Table 3, we see the proportion of words marked with the IOB schema and see that the corpus of Lynn is most similar in terms of composition of the corpus. Moreover, we see that the distant supervision by Tearma while producing a similar ratio of terms, has far fewer words marked as `I`, suggesting that there are more one-word terms in this corpus than the part-of-speech tagging based corpora. An example of this annotation is given in Figure 1.

5 Results

In order to evaluate the effectiveness of our automatic term recognition approach we evaluated the accuracy of the extraction in various settings. For the part-of-speech-based extraction we considered the two corpora of Uí Dhonnchadha and Lynn separately as well as in a 'merged' mode, where we aligned the part-of-speech tags between the two corpora. We also considered each of these corpora where we converted the tagging from the part-of-speech tags to the IOB scheme and then trained

Is í **an tSomáilis** an **teanga** a labhraíonn formhor*[sic]* muintir **na Somáile** agus na **Somálaigh** sna tíortha in aice láimhe . Is **teanga Cúiseach** í agus í an dara **teanga Cúiseach** is mó a labhraítear ar domhan í (i ndiaidh **na hOraimise**).

Term	Translation
an tSomáilis	Somali (language)
teanga	language
an tSomáil	Somalia
Somálach	Somali (person)
teanga Cúiseach	Cushitic Language
an Oraimis	Oromo

Figure 1: An example from the gold standard annotated corpus with terms in bold and the extracted terms with translations

the model on the IOB tags. In addition, we considered the weakly supervised training scheme by using the Tearma-based model and finally we concatenated all corpora with IOB tags to produce a corpus called 'All'. We trained all models with the OpenNLP toolkit using the standard maximum entropy model[7]. In the case of using the part-of-speech tagged corpora the data was trained using the default parameters of the models and the top-level part-of-speech tags as described in Section 3.2, which for the Tearma database and the models using IOB we again used the default parameters with each word being tagged as either 'I', 'O' or 'B'. We note that the maximum entropy model implemented by OpenNLP is probably not state-of-the-art and does not take advantage of word embeddings or other techniques. This implementation is used by Saffron due to it being a reasonable trade-off between accuracy and computational cost, as well as being openly licensed without any copy-left restrictions, however this will likely be revised in the future. In Table 4 we show the results of the extraction presented in terms of precision, recall and F-Measure on each of the classes. We see that no training corpus performs best on all classes, for the B and I class the part-of-speech based system is best when both corpora are combined with only a minor difference between the part-of-speech tags and the IOB tag scheme. For the O class, however the Tearma corpus performs best, and the effect of adding the part-of-speech tagged corpora seems to be very marginal.

[7] As implemented by `POSTaggerME` in OpenNLP

	B			I			O		
	P	R	F	P	R	F	P	R	F
Random (baseline)	0.163	0.163	0.163	0.111	0.111	0.111	0.726	0.726	0.726
Uí Dhonnchadha	0.676	0.458	0.546	**0.777**	0.327	0.460	0.648	0.951	0.771
Lynn	0.707	0.490	0.579	0.739	0.446	0.556	0.759	0.948	0.843
Merged	**0.722**	0.498	0.589	0.747	0.468	**0.576**	0.770	**0.953**	0.851
Uí Dhonnchadha (IOB)	0.656	0.432	0.521	0.725	0.302	0.427	0.625	0.933	0.748
Lynn (IOB)	0.670	0.467	0.551	0.537	0.485	0.510	0.801	0.905	0.850
Merged (IOB)	0.707	0.506	**0.590**	0.681	0.480	0.563	0.790	0.933	0.855
Tearma	0.612	0.506	0.554	0.101	0.806	0.180	0.906	0.834	**0.869**
All	0.618	**0.507**	0.557	0.098	**0.824**	0.174	**0.907**	0.835	**0.869**

Table 4: Per-class performance of term extraction for various training inputs evaluated on the gold standard.

We then ran the full pipeline embedded in the Saffron system and described in Section 3.4, using An Vicipéid as a background corpus. This was applied to a set of chat dialogues that concerned plans for the future of National University of Ireland Galway. Considering each comment as a single document we used a corpus of 239 documents totalling 9,313 words. We considered two of the best scoring settings for this from the previous experiment and the top 20 extracted terms for each settings are shown in Table 5.

6 Discussion

The results presented show that both the extraction using a part-of-speech tagged corpus and using the weak supervision by using a term database can be effective at developing a term extraction system. The principle difference can be seen from the corpus, in that the Tearma based approach extracted many more one word terms than the part-of-speech-based approach, and this is probably due to the inclusion of many short words as terms, that may have a specific meaning as domain terminology but are also frequently used in general. This can be seen from the higher prevalence of the 'B' tag in Table 3 and by the comparatively better performance on the 'O' class on the gold standard in Table 4. This is further clearer in the top 20 extracted terms in Table 5, where we can see that the Tearma based system extracted many more one-word terms but only extracted one multiword term (excluding those terms that erroneously contain the definite article 'an'). However, the corpus developed by the Tearma approach was much larger than that which has part-of-speech tags, so performance of this methodology may be impaired.

As such, it seems clear that both methods are viable approaches and in the context of an under-resourced language both options could be used as the basis for creating a term extractor. As the list of terms is a resource that in general requires less specialist expertise to be created and may be more available for languages with even fewer resources than Irish, for example by using the page titles of Wikipedia articles, it is good to see that for the task of automatic term recognition it may not be necessary to engage in the expensive process of annotating a corpus with part-of-speech tags. That said, given the relatively small size of the part-of-speech tagged corpus, it may follow that effort spent here more directly translates into improvement in the quality of automatic term recognition.

We were not able to provide a good quantitative evaluation of the quality of the extracted terms as this would require a significant and costly analysis of the corpus as well as creating a ranked list of highly relevant terms that is difficult to achieve. However we have provided the top 20 terms in Table 5, and will provide a qualitative evaluation of them here. Both lists contain a similar number of non-terms (four each). This is also based on the assumption that 'déan' is an error, which while a very relevant term in this context, referring back to the corpus suggests that this was actually a from of the verb, e.g., the verbal noun 'déanamh', and so should not have been extracted, a similar case may apply to 'úsáid' which can be both a noun and a verbal noun. In much the same way, it seems that 'cónaí' was entirely used in the phrase 'i gcónaí' (always) rather than as an independent term. Moreover, there are a number of errors in the lemmatization in both lists in particular with relation to the rather specialized term 'ollscolaíocht', which does not occur in Pota Focal. Also, in a few

Part-of-speech		Tearma	
Irish	Translation	Irish	Translation
gaeilge	Irish	ollscoil	university
mac léinn	student	foireann	staff
ollscoil	university	ceart	right
ionad ghaeltachta	Irish centre	an phobail†	the public
teanga	language	dátheangach	bilingual
duine	person	obair	work
mac	son	cúrsa	course
scéim teanga	language plan	seirbhís	service
gaeltacht	Irish-speaking area	ceist	question
foireann na hollscoile	university staff	bliain	year
pobal	public	deis	opportunity
áras na gaeilge	Irish Building at NUIG	easpa ceannaireachta	lack of leadership
pobal na hollscoile	people of the university	leanúnach	successor
léann	learning	*déan*	dean/'to do'
foireann	staff	iarraidh	request
cuid na hollscoile	part of the university	oifigeach	officer
níos mó	more	dualgas	duty
meán	media	*cónaí*	residence/always
cúrsa	course	pleán†	plan
leath na gaeilge	for Irish	scéim	plan
hOllscolaíochta gaeilge†	Irish Language University Education	comhrá	conversation
seirbhís	service	úsáid	usage
cónaí	residence/always	*inbhuanaithe*	sustainable
deis	opportunity	cultúr	culture
ball foirne	member of stafff	an gclár†	the programme
oifigeach na gaeilge	Irish language officer	plean	plan
hOllscolaíochta †	university education	*an rud*	the thing
oifigeach	officer	ról	role
ceist	question	oideachas	education
acadamh na hóige	youth academy	an domhan	the world

Table 5: The Top 20 ranked terms extracted using the part-of-speech tagged corpus and the distant supervision via Tearma. Italics indicate terms that are likely incorrect terms, † indicates terms with a lemmatization issues.

cases, we see terms that were extracted were possibly also used as adjectives, and hence would not be terms, in particular 'dátheangach' and 'leanúnach', which are very rarely used as a noun. Finally, we note that the Tearma-based system extracted the spelling error '*pleán' (which should likely be 'plean'), which while incorrect is interesting given that this misspelled form did not occur in training suggesting that the system has been able to generalize effectively.

7 Conclusion

We have analyzed two methods for the construction of an automated term recognition system for an under-resourced language. We have found that both methods make effective methods for training a system that is significantly better than a random baseline, however our analysis shows that there are still weaknesses with each system, suggesting that performance is being limited by the availability of resources. Further, it seems that basic linguistic facts such as the length of the term are being affected by the the resources and methods we are using to create the system and this could be a focus of further study.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Unions Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

Ahmad, Khurshid, Lee Gillam, Lena Tostevin, et al. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *TREC*, pages 1–8.

Ananiadou, Sophia. 1994. A methodology for automatic term recognition. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*, volume 2.

Arora, Chetan, Mehrdad Sabetzadeh, Lionel Briand, and Frank Zimmer. 2014. Improving requirements glossary construction via clustering: approach and industrial case studies. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, page 18. ACM.

Astrakhantsev, Nikita. 2015. *Methods and software for terminology extraction from domain-specific text collection*. Ph.D. thesis, Ph. D. thesis, Institute for System Programming of Russian Academy of Sciences.

Astrakhantsev, Nikita. 2018. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. *Language Resources and Evaluation*, 52(3):853–872.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bordea, Georgeta, Paul Buitelaar, and Barry Coughlan. 2014. Hot Topics and schisms in NLP: Community and Trend Analysis with Saffron on ACL and LREC Proceedings. In *Proceedings of the Ninth LREC, Reykjavik, Iceland, ACL Anthology: L14-1697*.

Buitelaar, Paul, Georgeta Bordea, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *The 10th international conference on terminology and artificial intelligence (TIA 2013), Paris, France*. 10th International Conference on Terminology and Artificial Intelligence.

Church, Kenneth W and William A Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

Church, Kenneth and William Gale. 1999. Inverse document frequency (IDF): A measure of deviations from Poisson. In *Natural language processing using very large corpora*, pages 283–295. Springer.

Evans, David A and Robert G Lefferts. 1995. CLARIT-TREC experiments. *Information processing & management*, 31(3):385–395.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Kageura, Kyo and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Kilgarriff, Adam, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language resources and evaluation*, 40(2):127–152.

Li, Sujian, Jiwei Li, Tao Song, Wenjie Li, and Baobao Chang. 2013. A novel topic model for automatic term extraction. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 885–888. ACM.

Lynn, Teresa, Kevin Scannell, and Eimear Maguire. 2015. Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8.

Měchura, Michal Boleslav. 2018. Maidir le Pota Focal. URL: `http://www.potafocal.com/_info/`.

Peñas, Anselmo, Felisa Verdejo, Julio Gonzalo, et al. 2001. Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics*, volume 2001, page 458. Citeseer.

Pereira, Bianca, Cecile Robin, Tobias Daudert, John P. McCrae, and Paul Buitelaar. 2019. Taxonomy Extraction for Customer Service Knowledge Base Construction. In *Submitted to SEMANTICS 2019*.

Ramshaw, Lance A and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Uí Dhonnchadha, Elaine and Josef van Genabith. 2006. A Part-of-Speech tagger for Irish using finite state morphology and constraint grammar disambiguation.

Zhang, Ziqi, Jie Gao, and Fabio Ciravegna. 2016. JATE 2.0: Java Automatic Term Extraction with Apache Solr. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*.