Unsupervised Multi–Word Term Recognition in Welsh

Irena Spasić
School of Computer Science & Informatics
Cardiff University
SpasicI@cardiff.ac.uk

David Owen
School of Computer Science & Informatics
Cardiff University
OwenDW1@cardiff.ac.uk

Dawn Knight
School of English, Communication & Philosophy
Cardiff University
KnightD5@cardiff.ac.uk

Andreas Artemiou
School of Mathematics
Cardiff University
ArtemiouA@cardiff.ac.uk

Abstract

This paper investigates an adaptation of an existing system for multi-word term recognition, originally developed for English, for Welsh. We overview the modifications required with a special focus on an important difference between the two representatives of two language families, Germanic and Celtic, which is concerned with the directionality of noun phrases. We successfully modelled these differences by means of lexico–syntactic patterns, which represent parameters of the system and, therefore, required no re–implementation of the core algorithm. The performance of the Welsh version was compared against that of the English version. For this purpose, we assembled three parallel domain–specific corpora. The results were compared in terms of precision and recall. Comparable performance was achieved across the three domains in terms of the two measures (P = 68.9%, R = 55.7%), but also in the ranking of automatically extracted terms measured by weighted kappa coefficient (κ = 0.7758). These early results indicate that our approach to term recognition can provide a basis for machine translation of multi-word terms.

1 Introduction

Terms are noun phrases (Daille, 1996; Kageura, 1996) that are frequently used in specialised texts to refer to concepts specific to a given domain (Arppe, 1995). In other words, terms are linguistic representations of domain-specific concepts (Frantzi, 1997). As such, terms are key means of communicating effectively in a scientific or technical discourse (Jacquemin, 2001). To ensure that terms conform to specific standards, they often undergo a process of standardisation. Such standards are commonly based on the following principles. First and foremost, a term should be linguistically correct and reflect the key characteristics of the concept it represents in concise manner. There should only be one term per concept and all other variations (e.g. acronyms and inflected forms) should be derivatives of that term. TermCymru, a terminology used by the Welsh Government translators, assigns a status to each term depending on the degree to which it has been standardised: fully standardised, partially standardised and linguistically verified.

Terms will still naturally vary in length and their level of fixedness, i.e. the strength of association between specific lexical items (Nattinger and DeCarrico, 1992), which can be measured using mutual information, z-score or t-score. Such variation of terms within a language may pose problems when attempting to translate term variants consistently into another language. Verbatim translations also often deviate from the established terminology in the target language, e.g. TermCymru in Welsh. Therefore, high-quality translations, performed by either humans or machines, require management of terminologies. Specialised text requires consistent use of terminology, where the same term is used consistently throughout a discourse to refer to the same concept. Very often, terms cannot be translated word for word. Therefore, most machine translation systems maintain a term base in order to support translations that use established terminology in the target language.

Given a potentially unlimited number of domains as well as a dynamic nature of many domains (e.g. computer science) where new terms get introduced regularly, manual maintenance of one-to-one term bases for each pair of languages may become unmanageable. Where parallel corpora exist, automatic term recognition approaches can be used to extract terms and their translations, which can then be embedded into the term base to support machine translation of other document from the same domain. To that end, we are focusing on

comparing the performance of an unsupervised approach to automatic term recognition in two languages, Welsh and English, as an important step towards machine translation of specialised texts in the given languages.

2 Methods

2.1 Method overview

FlexiTerm is a software tool that automatically identifies multi-word terms (MWTs) in text documents (Spasić et al., 2013). Given a domain-specific corpus of plain text documents, it will extract MWTs in a form of a lexicon, which links together different forms of the same term including acronyms (Spasić, 2018). The most recent version can arrange the lexicon hierarchically (Spasić et al., 2018). Table 1 provides examples of terms that were automatically extracted from patent applications from three different domains.

Domain	Term variants
Civil engineering	bottom hole assembly bottomhole assembly BHA
Computing	network functions virtualization NFV virtual network function VNF
Transport	lightning strike protection LSP protection against lightning strike

Table 1: Examples of domain-specific terms

FlexiTerm performs MWT recognition in two stages. First, MWT candidates are selected using lexico-syntactic patterns. This is based on an assumption that terms follow certain formation patterns (Justeson & Katz, 1995). Indeed, terms are associated with preferred phrase structures. They are typically noun phrases that consist of adjectives, nouns and prepositions. Terms rarely contain verbs, adverbs or conjunctions.

Once potential MWT are identified, they are ranked using a measure that combines their length and frequency with an aim of identifying the longest repetitive patterns of word usage. This is based on an assumption that MWTs are expected to demonstrate collocational stability (Smadja, 1993).

The original FlexiTerm method was implemented to support the English language. In the following sections, we describe the modifications that were required to support the same functionality in the Welsh language.

2.2 Linguistic pre-processing

FlexiTerm takes advantage of lexico–syntactic information to identify term candidates. Therefore, the input documents need to undergo linguistic pre–processing in order to annotate them with relevant lexico–syntactic information. This process includes part-of-speech (POS) tagging, sentence splitting and tokenisation. The original implementation of FlexiTerm uses Stanford CoreNLP library (Toutanova et al., 2003) to support such processing in English. In the Welsh version, text is processed using the Canolfan Bedwyr Welsh POS Tagger (Jones, Robertson, and Prys, 2015) to tokenise the text and tag each token with an appropriate lexical category including end-of-sentence annotations. A subset of relevant tags from Canolfan Bedwyr Welsh language tag set (Robertson, 2015) were mapped to tags compatible with the original version of FlexiTerm to minimise re-implementation (e.g. specific noun tags NM and NF were mapped to generic noun tag NN). This mapping was restricted to nouns, adjectives and prepositions only as these lexical classes are used to extract term candidates as explained in the following section.

2.3 Term candidate extraction and normalisation

Term candidates are extracted from pre-processed documents using pattern matching. The patterns describe the syntactic structure of targeted noun phrases (NPs). These patterns are treated as parameters of the method and as such can be modified as required. In general, NPs in Welsh and English follow different formation patterns. The main difference is concerned with headedness or directionality of NPs. Nearly all adjectives follow the noun in Welsh (Willis, 2006). For example, *gorsaf ganolog*, where the word *ganolog* means *central*, corresponds to the *central station* in English. Two lexico-syntactic patterns defined using regular expressions were used in our experiments, one to model simple (linear) NPs:

$$\text{NN (NN | JJ)}^{+}$$

and the other one to model complex (hierarchical) NPs:

$$\text{NN (NN | JJ)* IN NN (NN | JJ)*}$$

Here, NN, JJ and IN correspond to noun, adjective and preposition respectively.

Identification of term candidates is further refined by trimming the leading and trailing stop words. Stop word list has been created by automatically translating the English stop word list distributed with FlexiTerm (Spasić et al., 2013; Spasić, 2018), e.g. *unrhyw* (Engl. *any*), *bron* (Engl. *nearly*), etc. The translation was performed using the Canolfan Bedwyr Machine Translation Online API (Jones, 2015).

To neutralise morphological and orthographic variation, all term candidates undergo normalisation, which involves lemmatisation of each token and removal of punctuation, numbers, stop words and any lowercase tokens with less than 3 characters. To address syntactic variation, the order is ignored by representing each candidate as a bag of words (BOW). For example, term candidates *niwed i iechyd* (Engl. *damage to health*) and *iechyd niwed* (Engl. *health damage*) are both represented as {*niwed, iechyd*}.

Unlike English, Welsh syntax often requires words to inflect at the beginning depending on the preceding word or its role in the sentence (Harlow, 1989). These morphological changes are known as mutations. For example, *mwg tybaco* (Engl. *tobacco smoke*) can appear as *fwg tybaco* in some contexts where soft mutation occurs. Lemmatisation will neutralise various word mutations. In the previous example, both *mwg* and *fwg* would be lemmatised to *mwg*.

2.4 Lexical similarity

As mentioned before, many types of morphological variation can be neutralised by lemmatisation. For instance, *cerbyd* (Engl. *vehicle*) and *cerbydau* (Engl. *vehicles*) will be conflated to the same lemma *cerbyd*. However, previously normalised term candidates may still contain typographical errors or spelling mistakes. Lexical similarity can be used to conflate these types of variation. For example, two normalised candidates {*llywodraeth, cymru*} and {*llywrydraeth, cymru*} (where *llywrydraeth* is a misspelling of the correct word that means government) can be conflated into the same normalised form {llywodraeth, llywrydraeth, cymru}. In FlexiTerm, similar tokens are matched using the Cysill Ar-Lein (Spelling and Grammar Checker) API (Robertson, 2015).

2.5 Termhood calculation

Calculation of termhood is based on the C-value formula (Frantzi et al., 2000), which is based on the idea of a cost criteria-based measure originally introduced for automatic collocation extraction (Kita et al., 1994):

$$C-value(t) = \begin{cases} \ln|t| \cdot f(t) & , \text{ if } S(t) = \varnothing \\ \ln|t| \cdot (f(t) - \frac{1}{|S(t)|}\sum_{s \in S(t)} f(s)) & , \text{ if } S(t) \neq \varnothing \end{cases}$$

In this formula, $|t|$ represents the number of content words in term candidate t, $f(t)$ is the overall frequency of occurrence of term t which aggregates occurrences of the corresponding term variants. $S(t)$ is a set of all other term candidates that are proper supersets of t. The termhood calculation module is language independent and as such required no modification for Welsh.

2.6 Output

Given a corpus of text documents, FlexiTerm outputs a ranked list of MWTs together with their termhood scores. Within this list, all term variants that share the same normalised form represented as a BOW are grouped together and ordered by their frequency of occurrence. Table 2 provides a sample output. We added English translation manually for the benefit of non-Welsh readers.

Rank	Term variants	Translation	Score
1	mwg ail-law fwg ail-law	second-hand smoking	3.4657
2	fwg tybaco amgylcheddol mwg tybaco amgylcheddol	environmental tobacco smoke	3.2958
3	cerbyd preifat cerbydau preifat	private vehicle	2.7726
4	niwed difrifol i iechyd niwed i iechyd iechyd niwed	damage to health	2.0794
5	Llywodraeth Cymru Lywodraeth Cymru	Welsh Government	1.3863

Table 2: Sample output

3 Results

3.1 Data

We assembled three parallel corpora from three domains: education, politics and health. For each domain, a total of 100 publicly available documents were downloaded from the Welsh Government web site (Welsh Government, 2019). The Welsh Language Act 1993 obliges all public sector bodies to give equal importance to both Welsh and English when delivering services to the public in Wales. This means that all documents we collected from the Welsh Government web site were available in both languages. We collected a total of 100 documents in both languages for each of the three domains considered (600 in total). All documents were pre-processed to remove HTML formatting and stored in a plain text format for further processing by FlexiTerm. Table 3 describes the properties of each corpus whose name consists of two letters – first indicating the language and the second indicating the domain (e.g. WH stands for Welsh+Health).

Data set	Size (KB)	Sentences	Tokens	Distinct lemmas
EE	138	869	24,580	2,517
WE	141	913	27,847	2,204
EP	116	831	21,406	2,444
WP	120	877	23,884	2,352
EH	92	596	16,614	2,063
WH	96	615	18,975	1,960

Table 3: Three parallel domain-specific corpora

3.2 Silver standard

FlexiTerm had previously been thoroughly evaluated for English using the standard measures of precision and recall (Spasić et al., 2013). Their values were calculated against term occurrences that were annotated manually in five corpora used for evaluation. In this particular study, we are focusing on the actual terms extracted as a ranked list and not their specific occurrences in text. This simplifies the evaluation task as it does not require manual annotation of term occurrences in the three corpora (WE, WP and WH). Instead, only the ranked term lists need to be inspected. Moreover, the goal of this study is not to evaluate how well the Welsh version of FlexiTerm performs in general, but rather examine how it compares relative to the English version. In other words, by already knowing the performance of the English version of FlexiTerm from the previous study (Spasić et al., 2013), we can use its output on English versions of the three corpora (EE, EP and EH) as the "silver standard". The results obtained from the Welsh versions of the three corpora (WE, WP and WH) can then be matched against the silver standard. The only manual effort this approach requires is to map each automatically extracted term in Welsh to its equivalent in English (if an equivalent term has been recognised by FlexiTerm) and vice versa. Such mapping was performed by a Welsh-English proficient bilingual speaker.

3.3 Evaluation

We ran two versions of FlexiTerm against the three parallel corpora. Table 4 specifies the number of automatically recognised terms in each language. The Welsh output was evaluated against the corresponding English output (used here as the silver standard) in terms of precision and recall (also specified in Table 4). In other words, to calculate precision, for every Welsh term candidate, we checked whether its equivalent (i.e. translation) appeared in the English output. Vice versa, to calculate recall, for every English term candidate, we checked whether its equivalent appeared in the Welsh output.

	Welsh terms	**English terms**	**P**	**R**	**F**	**κ**
Health	90	120	75.0	55.1	63.5	0.6300
Education	107	136	63.8	46.3	53.7	0.8425
Politics	124	127	68.0	65.6	66.8	0.8550
Average	**107**	**128**	**68.9**	**55.7**	**61.3**	**0.7758**

Table 4: Evaluation results

Across the three domains, the Welsh version of FlexiTerm performed more consistently in terms of precision, which was relatively high (i.e. >60%). However, the recall varied significantly across the three corpora ranging from as low as 46.3% to as high as 65.6%.

3.4 Discussion

We investigated the plausible causes affecting the sensitivity of the method in Welsh, which are associated with different steps of the FlexiTerm algorithm: (1) term candidate selection, (2) term candidate normalisation, (3) termhood calculation.

First, term candidate selection depends on a set of lexico-syntactic patterns. If their coverage does not cover certain term formation patterns, then the corresponding terms will fail to be recognised. For example, the structure NN DT NN of the term *rheoliad y cyngor* (Engl.*council regulation*) does not match any of the patterns specified in Section 2.2, so further investigation is needed into the Welsh term formation patterns.

Furthermore, term candidate selection depends on linguistic pre-processing (see Section 2.1). For example, even if a term's internal structure does comply with the given patterns, for the term to be selected that structure needs to be correctly recognised. In practice, a term's constituents may consistently be tagged incorrectly or ambiguously with POS information. For example, the term *data biometrig* (Engl. *biometric data*) was tagged as NN ? (where ? denotes an unknown tag) instead of NN JJ. Such cases may fail to be matched with any of the given patterns, and, therefore, will also fail to be recognised.

Once term candidates have been selected, their formal recognition as terms will depend on their frequency of occurrence. The overall frequency may be underestimated when different term variants fail to be conflated into a single term representative used to aggregate their individual frequencies. Term conflation depends on term normalisation, which involves (1) lemmatisation of individual words and (2) lexical similarity of their lemmas. The performance of the Welsh lemmatiser was found to be poorer than that of its English counterpart. Further, term normalisation depends on matching lexically similar tokens (see Section 2.4). Welsh orthography uses 29 letters out of which eight are digraphs. Morphology of the words is also more likely to vary than English depending on the dialect (e.g. northern vs. southern dialects). For example, *hogyn* is the northern variant of *bachgen* (Engl. *boy*). While the same approach to term normalisation is still valid for Welsh, it requires further investigation into adjusting the lexical similarity threshold.

Finally, other than frequency, the calculation of the termhood also depends on the length of the term candidate (see Section 2.5). The equivalent terms in

the corresponding languages may not necessarily have the same number of content words due to compounding. For example, *ansawdd gofal iechyd* has got three content words whereas its English translation *quality of healthcare* has got two content words. This means that their termhood calculated using the C-value formula may have significantly different values. If this value does not meet the termhood threshold, the candidate will fail to be recognised as a term. In the worst case scenario, a MWT in one language (e.g. *gofal iechyd*) may be a singleton in the other language (e.g. *healthcare*), and as a single-word term it will fail to be identified as a term candidate.

To check how well the respective terminologies are aligned, we compared whether the ranking of terms was similar. The C-value scores are replaced by their rank when they are sorted in the descending order. Note that such ranking represents a weak order because different terms may have the same C-value and, therefore, the same rank. We can view the ranking of terms as an ordinal classification problem. This allows us to compare the differences in the ranking using weighted kappa coefficient (Cohen, 1968), which is traditionally used to calculate inter-annotator agreement. Unlike the original kappa coefficient, the weighted version accounts for the degree of disagreement by assigning different weights w_i to cases where annotations differ by i categories.

We reported the values of this statistics in Table 4 for the terms recognised in both languages. In other words, the missing values, i.e. terms not recognised in one of the languages, were ignored. These values have already been accounted for by means of precision and recall. For the common terms in the domains of education and politics, at $\kappa > 0.8$ the agreement of ranking is almost perfect. In the health domain, the agreement is still substantial at $\kappa > 0.6$.

4 Conclusions

In this paper we presented an adaptation of a MWT recognition algorithm, originally implemented for English, for Welsh. We compared the performance of the Welsh version relative to the original English version. The results demonstrate that the brute-force adaptation, which is concerned only with the modules that support linguistic pre-processing (e.g. POS tagging), will successfully recognise the majority of MWTs proposed by the English version (P = 68.9%, R = 55.7%). It is expected that fine tuning the internal parameters of the method (e.g. lexico-syntactic patterns and lexical similarity threshold) as well as improving the performance of external parameters (e.g. POS tagging) would further improve the performance in Welsh. Successfully mapping MWTs between Welsh and English would improve the performance of machine translation of specialised texts, whose quality of translation depends largely on using established terminology instead of verbatim translations.

5 Availability

The software is shared under the BSD-3-clause license on GitHub:
https://github.com/ispasic/FlexiTermCymraeg

References

Arppe, A. (1995) Term Extraction from Unrestricted Text. 10th Nordic Conference of Computational Linguistics, Helsinki, Finland

Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70(4):213-20.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In P. Resnik & J. Klavans (Eds.), The Balancing Act - Combining Symbolic and Statistical Approaches to Language, MIT Press, 49-66.

Frantzi K., Ananiadou S. (1997) Automatic term recognition using contextual cues. 3rd DELOS Workshop on Cross-Language Information Retrieval, Zurich, Switzerland.

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 2000, 3:115–130.

Harlow, S. (1989) The syntax of Welsh soft mutation. Natural Language & Linguistic Theory 7(3):289–316.

Hersh, W., Campbell, E., Malveau, S., (1997). Assessing the feasibility of largescale natural language processing in a corpus of ordinary medical records: a lexical analysis. Annual AMIA Fall Symposium, 580–584.

Jacquemin, C. (2001). Spotting and Discovering Terms through Natural Language Processing. Cambridge, Massachusetts, USA: MIT Press.

Jones, D.B. (2015). Machine Translation Online API [https://github.com/PorthTechnolegauIaith/moses-smt/blob/master/docs/APIArlein.md#moses-smt-machine-translation-online-api].

Jones, D. B., Robertson, P., Prys, G. (2015). Welsh language lemmatizer API service [http://techiaith.cymru/api/lemmatizer/?lang=en].

Jones, D. B., Robertson, P., Prys, G. (2015). Welsh language Parts-of-Speech Tagger API Service [http://techiaith.cymru/api/parts-of-speech-tagger-api/?lang=en].

Justeson, J. S., Katz, S. M. (1995) Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1): 9-27.

Kageura, K., Umino, B. (1996). Methods of automatic term recognition - A review. Terminology 3(2): 259-289.

Kita K, Y. Kato, T. Omoto and Y. Yano (1994) A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. Journal of Natural Language Processing 1:21-33.

Nattinger, J., DeCarrico, J. (2011) Lexical phrases and language teaching. Oxford University Press.

Robertson, P. (2015). Cysill Ar-lein [https://github.com/PorthTechnolegauIaith/cysill/blob/master/doc/README.md#cysill-online-api].

Robertson, P. (2015). POS Tagger API [https://github.com/PorthTechnolegauIaith/postagger/blob/master/doc/README.md#results].

Smadja, F. (1993) Retrieving collocations from text: Xtract. Computational Linguistics 19(1):143-177.

Spasić, I., Greenwood, M., Preece, A., Francis, N., Elwyn, G. (2013) FlexiTerm: a flexible term recognition method. Journal of Biomedical Semantics 4: 27.

Spasić, I. (2018) Acronyms as an integral part of multi-word term recognition - A token of appreciation. IEEE Access 6: 8351-8363.

Spasić, I., Corcoran, P., Gagarin, A., Buerki, A. (2018) Head to head: Semantic similarity of multi-word terms. IEEE Access 6: 20545-20557.

Toutanova, K., Klein, D., Manning, C., Singer, Y. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. North American Chapter of the Association for Computational Linguistics on Human Language Technology, 173-180.

Welsh Government (2019). Catalog Cyhoeddiadau Llywodraeth Cymru / Welsh Government Publications Catalogue

[http://welshgovernmentpublications.soutron.net/publications/]

Willis, D. (2006) Against N-raising and NP-raising analyses of Welsh noun phrases. Lingua 116(11): 1807-1839.