Vocab : a dictionary plugin for web sites

Dewi Bryn Jones, Gruffudd Prys, Delyth Prys
Language Technologies Unit, Bangor University, Bangor, Wales, UK

{d.b.jones, g.prys, d.prys}@bangor.ac.uk

RÉSUMÉ

Vocab : un plugin dictionnaire pour les sites web
Ce document décrit un plugin dictionnaire, Vocab, qui peut être installé sur les sites Web plutôt que dans un navigateur. Le plugin permettre aux utilisateurs de passer le curseur de la souris au-dessus les mots, les entités à mots multiples et les phrases, et voir les entrées de dictionnaire pertinentes dans un pop-up sur la page web elle-même. Les filtres de Bloom et lemmatisation sont utilisés pour identifier les mots-vedette qui se trouvent dans une page. Vocab est disponible comme une ressource gratuite via un portail en ligne d'outils linguistiques. Les instructions sont faciles à suivre afin que les concepteurs de sites Web peuvent l'intégrer dans leurs propres sites Web. Vocab est utile comme aide à l'apprentissage pour les apprenants avancés et pour aider les utilisateurs couramment avec des mots techniques ou inconnus. Il a été principalement développé pour la traduction des mots et des phrases entre gallois et en anglais.

ABSTRACT

This paper describes a dictionary plugin tool, Vocab, that enhances websites by providing a rapid, integrated facility for users to hover the mouse cursor over or touch words, multi word entities and phrases and see relevant dictionary entries aggregated from a number of federated dictionaries as pop-up windows within the website itself. Bloom filters and lemmatization are used to identify dictionary entry headwords within a webpage's text. Vocab is made available as a free resource via an online portal of language tools, with easy to follow instructions on its deployment so that web designers can integrate and customize into their own websites. Vocab is useful both as a learning aid for advanced language learners and as an aid to vocabulary improvement. While primarily developed for word and phrase translation between Welsh and English it could be adapted for use with other language pairs through opportunities for collaboration

MOTS-CLÉS : gallois, dictionnaires en ligne, Les filtres de Bloom
KEYWORDS: Welsh, Online dictionaries, Bloom filters

1 Introduction

Vocab is an easy to install server-side tool that enables users to read the text in a website that they may not completely understand without having to resort to a translated version or to an external reference resource such as a dictionary

When activated, the plugin is able to highlight all words, multi word entities (such as technical terms) and phrases where they occur as entries from a number of dictionaries associated with the plugin. Users are able to simply hover over (with a mouse) or touch (using a touchscreen) any highlighted text in order to view the associated dictionary information in full. A user can also click through to search for related or similar words on the Welsh National Terminology Portal website[1].

It is available as a free resource via the Welsh National Language Technologies Portal[2] (Prys D., Jones., 2016). The Vocab plugin can currently be seen in use on popular Welsh language websites such as Golwg360[3] and the BBC CymruFyw[4] service. Vocab supports all modern desktop and mobile based web browsers.

Figure 1 shows a screenshot of the Vocab widget in action on the Golwg360 Welsh language news website. Recognized dictionary entities have been highlighted with subtle blue underlining. A popup with the dictionary definition for the multiple word phrase 'Cefn Gwlad' is displayed as a consequence of hovering over with the mouse.

Figure 1- Vocab in action the Welsh language newswebsite Golwg360

Other reading assisting plugins and products exist for a wide variety of languages (Shuttleworth, 2014) but only one or two support Welsh as well such as ReadLang[5], Geriaoeg[6]. Vocab is distinguishable from these offerings in that it is integrated into websites where users are more likely to use it and that it can also recognize multi word entities such as terms, placenames and phrases rather than only single words.

[1] http://termau.cymru
[2] http://techiaith.cymru/widgets/vocab/?lang=en
[3] http://www.golwg360.com
[4] http://www.bbc.co.uk/cymrufyw
[5] http ://readlang.com/cy/dashboard
[6] http://wiki.apertium.org/wiki/Geriaoueg

2 Vocab Architecture

Vocab's client is a Javascript library which operates within the containing web page. The Vocab server hosts dictionary data and provides RESTful APIs for dictionary search and lookup services. This means that Vocab exists in two decoupled parts, following the classic client-server model, so that processing is partitioned and distributed and communication can be completed rapidly resulting in a disruption-free user experience. The Vocab client is responsible for collecting all eligible texts and making multiple calls to the server. The server is responsible for recognizing the words, terms and phrases to be highlighted as well as providing the content of the pop-ups in the form of detailed dictionary and lexicographical information.

2.1 Vocab Client

This section provides a brief overview of the Vocab client's internal construction and operation.

A text nodes selector component is responsible for discovering all valid text nodes underneath any given HTML element. Valid text nodes are those considered not to be included within 'iframe', 'script', 'noscript', 'style', 'object', 'input', 'textarea' and meta HTML elements. In the case of mobile based browsers, text nodes with 'a' HTML elements for hyperlinks are also considered invalid so that their touch still activates navigating to another webpage or site.

A server communication component receives all gathered texts. First all texts are split using a simple regular expression into segments and grouped into suitably sized payloads for requesting the services of the Vocab server's REST API. Payloads are packaged as HTTP GET requests with an optimal maximum size of 2048 characters. Larger sized requests are possible with current browser/server expectations, but a reasonable maximum payload was defined so as to limit latency in progressing through the server communication component's queue of requests.

An HTML injection component receives responses containing markup which is able to update original text node locations with new markup that provide highlighting and further Vocab client functionality. A pop-ups handler component attaches event handlers to each recognized dictionary entity's mouseover or touch triggers. When triggered, the Vocab client makes another call to the Vocab server API for the corresponding dictionary entries to be displayed with the pop-up. The amount of event handler attachments typical for a reasonably sized webpage can frustrate the user with its lack of responsiveness, whatever the qualities of the linguistic resource. Vocab client thus attaches delegated event handlers[7] in order to avoid such issues.

2.2 Vocab Server

The Vocab Server is a component of the wider dictionary and terminologies infrastructure developed by the LTU to support its activities in terminology standardization and lexicographical resource building and dissemination. Its Welsh National Terminology Portal allows users to search

[7] http://javascript.info/tutorial/event-delegation

over 20 terminology dictionaries and connect easily to search on other similar resources and services on the web.

Vocab server uses at present two of the largest dictionaries, namely *Y Termiadur Addysg*[8] - a technical dictionary of approximately 45700 standardized terms for the National Curriculum in Wales, and *Geiriadur Cyffredinol Cysgair*[9] – a general language dictionary containing approximately 30,000 entries.

The Vocab server provides two REST API endpoints to the Vocab client. The first provides a means by which recognized dictionary entities are noted as such in any given string of text. The second provides a simple and efficient dictionary entries lookup for a given word or term.

Welsh, in common with other Celtic languages, is a moderately inflected language, where the first letter of a word can change according to certain grammar rules. This together with internal vowel changes and conjugated verbs using different word endings cause complications for dictionary lookups where a root or lemma form is required. The only significant use of a natural language processing component by Vocab is therefore that of a lemmatizer. The lemmatizer used was originally developed for the Cysill spelling and grammar checker (Hicks, 2004) which can recognise over half a million mutated, verb and plural forms to return lemma forms of all words. For example, it has the ability to recognise 'ellir' as the mutated impersonal present tense of the verb 'gallu'.

Once all lemma forms have been derived, the next step is for fast and efficient identification of headwords and terms from the dictionaries associated with the Vocab service. This involves iterating through the given text and looking up sub-sequences of words in one or more dictionaries. Such an algorithm is feasible for such a service if querying the database, where dictionary data resides on disk or over a network, is avoided since it introduces latency and unnecessary iterations and lookups are eliminated. These requirements were addressed by deciding to use two caches implemented with Bloom filters for each dictionary as in-memory caches.

Bloom filters (Bloom, 1970) are highly efficient data structures that are ideal for determining membership queries of a given set. False positives are possible but their use is still beneficial if given sufficient size and tolerable error rate parameters. Bloom filters have traditionally been used in the implementation of spell checkers (Broder, Mitzenmacher, 2004) and more recently in the efficient utilization of massive language models (Talbot, Osborne, 2007) where memory resources are restrictive. Bloom filters were seen as a sensible approach given the LTU's limited server capacity along a need for future proofing for any possible expansion of the service that would include 20 or more dictionaries from its National Welsh Terminology Portal.

The first cache is a Bloom filter of dictionary headwords and multiple word entities split into their sequences of words. For example a standardized education term prescribed in the Termiadur Addysg such as 'gallu i ddatrys problemau' (translation: *'ability to solve problems'*) would be cached into 4 separate lemmatized entries: "gallu", "gallu i", "gallu i datrys", "gallu i datrys problem"

[8] http://www.termiaduraddysg.org
[9] http://geiriadur.bangor.ac.uk

The recognition algorithm iterates through the text a word at a time and is able to look ahead with the first Bloom filter as to whether subsequent sequences of lemmas constitute a possible a dictionary term or phrase. When a multiple word term or phrase has been identified, the algorithm is able to skip by the last value of the look ahead counter. A second Bloom filter contains all dictionary headwords and the lemmatized versions of terms and phrases in their entirety. For example, only one entry exists for 'gallu i ddatrys problemau' i.e. 'gallu i datrys problem'. This filter serves a double check against any false positives that may have arisen from the first Bloom filter.

The Vocab server's second REST API endpoint is called upon the Vocab client when a user has hovered or touched over a highlighted range of text and replies with the result of a normal query on dictionary data residing in databases.

The Vocab Server keeps logs of all of its API usage. User privacy and anonymity is respected as described in the Vocab service's terms and conditions[10] so that no information can be used to individually identify the user. Vocab server logs consist of the webpage URL that a user has used with Vocab; each source text submitted for headword, term or phrase recognition along with the consequent result of recognized (or not) headwords, terms and phrases as well as each word or term the user has hovered or touched on for triggering popups that display further dictionary and lexicographical information.

3 Performance and Uptake

Section 2 described how Vocab's architecture was designed so as to ensure viable performance and usability despite its operation involving a substantial amount of communication and computation. The figures in Table 1 demonstrate that Vocab performs with sufficient performance that user's only experience a 'small perceptible delay' (Grigorik, 2013) when Vocab is initialized on a typical news webpage.

Webpage URL	Word Count	Sentences	Total time	No. Of Requests	Average Request Time
http://golwg360.cymru/newyddion/cymru/221283-siarad-cymraeg-gyda-chyfrifiaduron	2848	583	286 ms	9	31.7ms
http://www.bbc.co.uk/cymrufyw/36092710	6769	1443	323 ms	9	35.8ms
http://golwg360.cymru/blog	13932	1320	7.22 s	113	63.89ms

Table 1- Performance of Vocab with variously sized webpages

[10] http://techiaith.cymru/api/terms-and-conditions/?lang=en

In its first year of general availability, Vocab has been used on over 6300 distinct URL webpages. Also to date, users have hovered over or touched 209,000 recognised dictionary entries. This compares quite favourably with the usage statistics recorded for other websites and services provided by the LTU. (Prys D., Prys G., Jones D.B., 2015)

4 Future Work

A substantial amount of work has already been done on developing Vocab as a means of applying and disseminating terminological and lexicographical resources maintained by the LTU. Due to its success and uptake by significant and popular Welsh language websites a number of ideas have been suggested and opportunities identified for expanding its use to other languages and media that users consume. However all further work would be dependent on successfully obtaining further funding.

That said, the number of dictionaries that Vocab supports can be easily extended if the requirement ever arose and Vocab as such could be utilised on webpages to push technical terminologies only. The number of recognised dictionary entries could be made to be more focused by adding controls and expanding the Vocab server API to filter all but difficult or unusual words.

This idea has been recently considered for a version of Vocab that would operate on subtitles with browser-based catch up services or news video clips. In such a use case, where the viewer does not want translated subtitles and is not able to hover or touch a word, term or phrase he/she doesn't understand, a Vocab for Video would choose on behalf of the user and display in real-time any difficult word or term used in the source language subtitles.[11] Further research is required to identify the words that are perceived by users as being difficult or unfamiliar, and which are not. The content of the search logs may provide a useful indication of the words that are generally found challenging by users, and this may enable the Vocab server in future to suggest only those words that exceed a general threshold of unfamiliarity.

Acknowledgements

Whilst Vocab itself has received no external funding, we wish to acknowledge grant aid from the Welsh Government towards the establishment of the Welsh National Language Technologies Portal from which the Vocab is available for free to all website developers. We wish to thank Golwg360 and BBC Cymru Wales for their help and support in developing and testing Vocab.

[11] https://vimeo.com/160714756 (Vocab for Video)

References

BLOOM B. (1970) *"Space/Time Tradeoffs in Hash Coding with Allowable Errors."* Communications of the ACM 13:7 (1970), 422-426

BRODER A, MITZENMACHER M (2004) "*Network applications of Bloom filters: A survey"*, Internet Mathematics, vol 1 no. 4, pp. 485-509, 2004.

GREGORIK, I. (2013) *High Performance Browser Networking*. O'Reilly Media.

HICKS W.J. (2004) "*Welsh Proofing Tools: Making a Little NLP go a Long Way*." Proceeding of the1st Workshop on International Proofing Tools and Language Technologies. Greece: University of Patras

PRYS D., JONES D. B. (2016) "*National Language Technology Portals for LRLs: A Case Study"* Language Technologies in Support of Less-Resourced Languages, (LRL 2015) 28 November 2015, Poznan, Poland

PRYS D., PRYS G., JONES D. B. (2016) "*Quantifying the Use of Digital Welsh-language Language Resources"*. Language Technologies in Support of Less-Resourced Languages, (LRL 2015) 28 November 2015, Poznan, Poland

SHUTTLEWORTH M. (2014) "*Approaches to language learning: Blending tradition with innovation"* Presented at: Symposium on International Languages and Knowledge, Penang, Malaysia (2014)

TALBOT D., OSBORNE M. (2007) *"Smoothed Bloom filter language models: Tera-Scale LMs on the Cheap"* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 468-476, Prague, June 2007.