Proceedings of the Language Technologies for All (LT4All) , pages 276–279
Paris, UNESCO Headquarters, 5-6 December, 2019.

Cardamom: Comparative Deep Models for Minority and Historical Languages

John P. McCrae and Theodorus Fransen
Data Science Institute / Insight Centre for Data Analytics
National University of Ireland, Galway
{john.mccrae, theodorus.fransen}@insight-centre.org

Abstract

This paper gives an overview of the Cardamom project, which aims to close the resource gap for minority and under-resourced languages by means of deep-learning-based natural language processing (NLP) and exploiting similarities of closely-related languages. The project further extends this idea to historical languages, which can be considered as closely related to their modern form, and as such aims to provide NLP through both space and time for languages that have been ignored by current approaches.

Keywords: natural language processing, under-resourced languages, deep learning

Achoimre

Tugtar léargas ginearálta sa pháipéar seo ar an tionscadal Cardamom; tionscadal taighde a bhfuil sé mar aidhm aige easnamh acmhainne a laghdú do theangacha mionlaigh agus do theangacha nach bhfuil mórán acmhainní ann ina leith trí phróiseáil teanga nádúrtha, atá bunaithe ar an domhainfhoghlaim, agus trí leas a bhaint as cosúlachtaí teangacha a bhfuil dlúthbhaint acu lena chéile. Áirítear mar chuid den tionscadal seo teangacha stairiúla de bharr go meastar go bhfuil dlúthbhaint acu lena bhfoirm nua-aimseartha. Dá réir sin, tá sé mar aidhm ag Cardamom próiseáil teanga nádúrtha a sholáthar ó thaobh spáis agus ama do theangacha a ndearnadh neamhaird orthu go dtí seo ó thaobh cur chuige an lae inniu de.

1. Introduction

There are estimated to be about 7,000 languages spoken in the world, but currently digital language tools support only a small fraction of these languages. Recent breakthroughs in natural language processing (NLP) have been based on the emergence of deep learning for processing texts and in particular in the use of vectors (*word embeddings*) to represent the meaning of words. Such representations have been shown to be truly interlingual and to allow translation between language pairs without any training data for that pair (Johnson et al., 2017). Deep learning has been enabled by the big data resources for NLP. However, it has been thought to be unsuitable for under-resourced languages as there is insufficient data to train these models. The comparative method, a keystone of modern linguistics (Schleicher, 1876), shows us that careful comparison of closely-related languages can give deep insight into the history, structure and semantics of a language. The key goal of the Cardamom project[1] is the creation of vector-based models of language that take into account the shared phonetic, etymological and semantic information of words in closely-related languages and their application to minority and historical languages. Speakers of minority languages are among the fastest growing communities on the Web and meeting their need is of major societal and commercial importance. Secondly, in order to meet the growing demand for text analysis in digital humanities, whereby access to large corpora text in languages such as Latin, Old English and Old Irish can enable new insights in the study of history and literature, we will develop technologies for historical languages.

The rest of this paper is organized as follows. Section 2. looks at the resource gap that underlies the motivation of the Cardamom project. Section 3. gives a short overview of the state-of-the-art in NLP, constituting the background for the novel methodology employed in our project, discussed in section 4. Advances and opportunities are the subject of section 5., followed by a conclusion in section 6.

2. The resource gap

Providing support for a new language to an existing language technology is by no means an easy goal. Few commercial or public institutes have a clear plan of how to scale beyond 100 languages, as most language technologies can only be developed with experts who speak the language. As such, new methods are required to develop NLP technologies that are viable for real-world applications, in that there is sufficient data and tools to enable enterprises to develop NLP applications for these languages and derive commercial benefits from addressing these minority populations. For example, it is generally considered that over 10 million words (MW) of parallel text are required to train a basic machine translation system, and in OPUS (Tiedemann, 2009)[2] there are only a few languages such as German (919.1MW parallel with English) or Hindi (13.2MW) where this is true. For most languages there is at least a moderate resource gap, such as for Irish (6.9MW), or more frequently a huge resource gap, such as for Scottish Gaelic (0.5MW). Modern algorithms for NLP claim to only require sufficient training data so they can adapt to any task or domain, yet all languages present unique challenges and most real-world systems have required language expertise in their development.

In spite of the large potential impact of the development of language technologies for minority languages, there has

[1] The project website is at `http://cardamom.insight-centre.org/`.

[2] This is based on data collected in 2017.

been comparatively little related focused work on these languages in computer science or linguistics. There are some platforms for the collection of information about under-resourced languages, most notably PanLex (Westphal et al., 2015), Glottolog (Nordhoff, 2012), An Crúbadán (Scannell, 2007) and Ethnologue (Lewis et al., 2009), yet these platforms have not focused on the development of language technologies but instead on language preservation. In contrast, there have been a number of workshops organized in the field of NLP on the topic of 'under-resourced languages', however the focus of these workshops has often been on national languages that have direct support from national funding agencies. For example, a recent report on European languages classified all but 2 EU languages as 'severely' under-resourced (Rehm and Uszkoreit, 2013). Despite the potentially huge impact of work on minority languages, they have been surprisingly neglected and thus there would be significant benefit in providing language technology for these languages. Figure 1 shows a geographic distribution of speakers of digitally emerging languages.

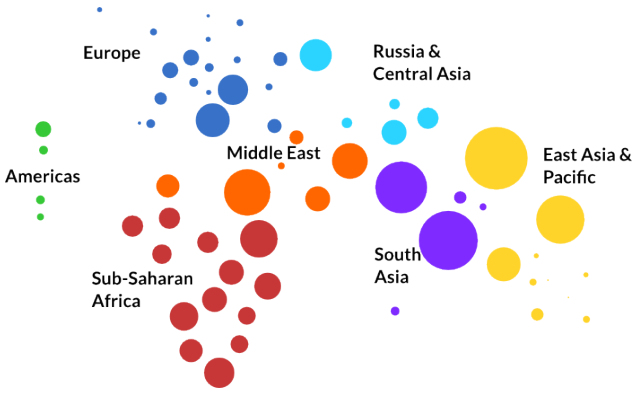

Figure 1: Geographic Distribution of Speakers of Digitally Emerging Languages.

Another important motivation for this project is the growth in digital humanities, where much research is focused on texts written in minority languages, dialects and, most importantly, in historical languages. Distant reading approaches, whereby literary analysis is done by comparing large volumes of text, have given humanities researchers new insights, for example into the linguistic style of a text (*stylometry*). Research in this area draws heavily from NLP techniques and, more recently, deep learning in particular (Brocardo et al., 2017), although a gap still exists between the state-of-the-art in NLP and the tools that are available to digital humanities researchers. Moreover, humanities researchers are often interested in pre-modern texts, where the language and grammar may not correspond to modern languages usage. As such, these texts can be considered to be written in an under-resourced language, which is closely related to an existing modern language. This underlines the interdisciplinary nature of this project and the input and collaboration with researchers in literature and history will be required.

The Cardamom project will address the resource gap with a big data approach to automatically produce resources and technologies for under-resourced languages by exploiting Web content and a deep learning approach grounded in linguistic theory that can leverage a wider range of input data. This will require developing a set of basic NLP approaches, which can build tools for all the world's languages.

3. State-of-the-art in NLP

Deep learning has revolutionized natural language processing technologies and the use of word embeddings such as *word2vec* (Mikolov et al., 2013) has become standard in the field. These methods, along with distributional methods that preceded them, associate a vector with each word, which is also called a word embedding. However, it has been shown that further breaking words up into smaller subword units (Sennrich et al., 2016) or acoustic units (Kamper et al., 2016) can improve quality in tasks such as machine translation. For minority languages, using many languages as a pivot can detect complex morphological phenomena (Asgari and Schütze, 2017). Furthermore, these vectors can very accurately induce semantic similarity (Tai et al., 2015) even in a cross-lingual setting (M^{c}Crae et al., 2013) and this has led to the development of bilingual lexicon induction (Haghighi et al., 2008), where translations are learnt without the need for any existing parallel translations. This work has been extended to full machine translation by a process known as linguistic decipherment (Ravi and Knight, 2011) and such approaches have been shown to enable machine translation systems to be trained on many languages simultaneously (Firat et al., 2017). It is our belief that these state-of-the-art deep models can be employed also for less-resourced languages, both modern and historical. Section 4. will outline the methodology of the Cardamom project in more detail.

4. Methodology

We propose the development of a new model for machine learning over natural languages, that will break the paradigm of learning models independently for each language, but instead learn models for closely-related languages simultaneously. The primary goal of this is to overcome the lack of data for minority and history languages, thus developing new tools and insights for researchers in computer science, linguistics and the humanities. This project will be primarily focused on three key areas: Firstly, as a data science project, we will attempt to find as much information on specific languages in as many forms as possible and combine this using linguistic linked data methods (M^{c}Crae et al., 2013), which have been acknowledged as a key technique for under-resourced languages (Westphal et al., 2015).

Secondly, for deep learning we will apply existing unsupervised methods, such as word embeddings, and develop them further into generic methods that can process minority and historical languages. Thus, we will develop comparative algorithms to exploit similarities between closely-related languages, to overcome the data gap for under-resourced languages. We will approach this first by identifying a small set of about 100 languages in four families of closely-related languages (Celtic, Germanic, Indic and

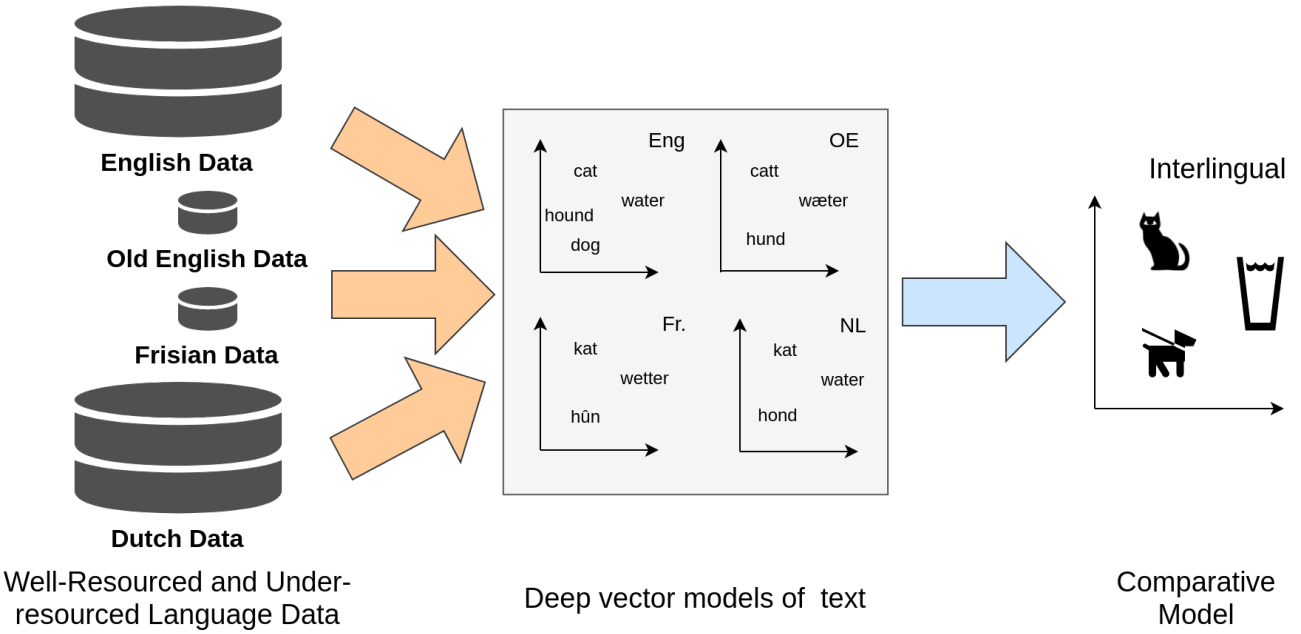

Figure 2: The architecture of Cardamom.

Dravidian) which include a mixture of minority languages (e.g., Irish, Frisian, Tulu) and historical languages (e.g., Old Irish, Old English, Sanskrit). This approach is illustrated in Figure 2, where we see how languages with less data (Old English and Frisian) can be learned simultaneously with well-resourced languages (English and Dutch) to produce a single representation. This model needs to consider orthographic differences ('cat' vs. 'kat'), phonetic changes ('water' vs. 'wetter') and semantic changes (the change of 'hound' in English to refer to only some types of dogs), which will be handled by unsupervised machine learning.

This project aims to revolutionize NLP, which has so far been overwhelmingly applied to major languages such as English, which is a language with comparatively low morphological complexity and standardized spelling and grammar. In contrast, minority languages frequently have complex morphology and significant variation among dialects. As such, the study of such languages raises questions for processing that have not been considered so far. Moreover, the comparative nature of this work requires us not to consider words as independent units, as has typically been done in existing word embedding work, but instead to look into the phonemes that compose the words to establish relationships between dialects (Sennrich et al., 2016).

5. Advances and opportunities

The Cardamom project will open up new research areas in language processing that have not yet been treated. Moreover, the uniquely broad nature of this study covering languages from different areas of the globe as well as different period of times leads to new research opportunities. Given that there are over 2,000 languages being used on the Web and the increasing economic importance of speakers of these languages, it is likely that the unique challenges of these languages are going to become increasingly important for research, societal and commercial applications. This represents a shift in viewpoint that requires the development of new algorithms that tackle problems with novel methods, for example the development of unsupervised morpho-syntactic systems. It is expected that the development of these tasks will provide new viewpoints on existing tasks in NLP. In particular, it is expected that this work will create a shift in approaches in machine translation by providing data in a wide range of languages, spurring development of novel approaches to handle these languages.

The algorithms developed in this project for under-resourced languages will lead to novel developments in the wider context of artificial intelligence in a number of ways: Firstly, the challenges of developing robust algorithms that can work on limited data with a potential highly complex output space (e.g., identifying a wide range of languages) will require the development of novel applications of machine learning. Secondly, the large number of languages studied will make interesting new developments in cognitive sciences by allowing for comparison in, for example, the meaning of words, to be examined in a new and wider situation. Finally, the development of computer-aided language learning software and the primary role of social media will be of interest to researchers in fields such as e-Government, in particular as this work aims to develop interaction with speakers of languages in the G77. Finally, the development of computer-aided language learning technology will have an impact on the field by widening the area of study and providing real case studies on teaching languages that are severely under-resourced. It will also have a wider societal impact in encouraging young learners to adopt languages that are currently mostly only used by older members of their communities.

6. Conclusion

This paper has described the Cardamom project, the aim of which is to use NLP and deep learning applied to a set of minority and historical languages primarily in four language families: Celtic, Germanic, Indic and Dravidian. The methodology involves a big data approach with largely unsupervised models that are simultaneously applied to closely-related languages, in order to overcome the data gap for under-resourced languages. The results are expected to advance the current state-of-the-art computational models and translate into societal and commercial applications.

7. Bibliographical References

Asgari, E. and Schütze, H. (2017). Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Brocardo, M. L., Traore, I., Woungang, I., and Obaidat, M. S. (2017). Authorship verification using deep belief network systems. *International Journal of Communication Systems*, 30(12):e3259.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., and Bengio, Y. (2017). Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kamper, H., Jansen, A., and Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):669–679.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2009). Ethnologue: languages of the world, Dallas: SIL International. *Online version: http://www.ethnologue.com*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*.

M^{c}Crae, J. P., Cimiano, P., and Klinger, R. (2013). Orthonormal explicit topic analysis for cross-lingual document matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1732–1740.

Nordhoff, S. (2012). Linked data for linguistic diversity research: Glottolog/Langdoc and ASJP online. In C. Chiarcos, et al., editors, *Linked Data in Linguistics*, pages 191–200. Springer.

Ravi, S. and Knight, K. (2011). Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21.

Rehm, G. and Uszkoreit, H. (2013). *Strategic research agenda for multilingual Europe 2020, presented by the META Technology Council*. Springer.

Scannell, K. P. (2007). The Crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Schleicher, A. (1876). *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Hermann Böhlau, Weimar.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Tiedemann, J. (2009). News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent advances in natural language processing*, volume 5, pages 237–248. John Benjamins.

Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with PanLex and the Web of Data. *Semantic Web*, 6(4):347–353.