Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3991–3995
Marseille, 11–16 May 2020

Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic

Jón Friðrik Daðason[1], David Erik Mollberg[1], Hrafn Loftsson[1], Kristín Bjarnadóttir[2]
[1]Department of Computer Science, [2]The Árni Magnússon Institute for Icelandic Studies
[1]Reykjavik University, [2]University of Iceland
[1]{jond19, de14, hrafn}@ru.is, [2]kristinb@hi.is

Abstract

In this paper, we present a character-based BiLSTM model for splitting Icelandic compound words, and show how varying amounts of training data affects the performance of the model. Compounding is highly productive in Icelandic, and new compounds are constantly being created. This results in a large number of out-of-vocabulary (OOV) words, negatively impacting the performance of many NLP tools. Our model is trained on a dataset of 2.9 million unique word forms and their constituent structures from the Database of Icelandic Morphology. The model learns how to split compound words into two parts and can be used to derive the constituent structure of any word form. Knowing the constituent structure of a word form makes it possible to generate the optimal split for a given task, e.g., a full split for subword tokenization, or, in the case of part-of-speech tagging, splitting an OOV word until the largest known morphological head is found. The model outperforms other previously published methods when evaluated on a corpus of manually split word forms. This method has been integrated into Kvistur, an Icelandic compound word analyzer.

Keywords: compound splitting, morphology, BiLSTM

1. Introduction

Compounds are extremely common in Icelandic, accounting for over 88% of all words in the *Database of Icelandic Morphology* (DIM) (Bjarnadóttir, 2017; Bjarnadóttir et al., 2019). As compounding is so productive, new compounds frequently occur as out-of-vocabulary (OOV) words, which may adversely affect the performance of NLP tools. Furthermore, Icelandic is a morphologically rich language with a complex inflectional system. There are 16 inflectional categories (i.e., word forms with unique part-of-speech (PoS) tags) for nouns, for adjectives 120, and for verbs 122, excluding impersonal constructions. The average number of inflectional forms per headword in DIM is 21.7. Included in this average are all uninflected words as well inflectional variants, i.e., dual word forms with the same PoS tag.

Compounds are formed by combining two words, which may be compounds themselves. The former word is known as a modifier and the second as a head, assuming binary branching (Bjarnadóttir, 2005). Theoretically, there is no limit to how many constituents a compound can be composed of, although very long words such as *uppáhaldseldhúsinnréttingaverslunin* ‘the favorite kitchen furniture store’ (containing 7 constituent parts) are rare. The constituent structure of a compound word can be represented by a full binary tree, as shown in Figure 1.

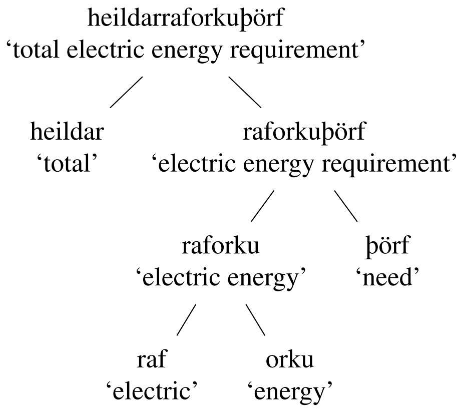

Figure 1: The constituent structure of an Icelandic compound word consisting of four constituent parts.

Compound splitting, or decompounding, is the process of breaking compound words into their constituent parts. This can significantly reduce the number of OOV words for languages where compounding is productive. Compound splitting has been shown to be effective for a variety of tasks, such as machine translation (Brown, 2002; Koehn and Knight, 2003), speech recognition (Adda-Decker and Adda, 2000) and information retrieval (Braschler et al., 2003).

In this paper, we present a character-based bidirectional long short-term memory (BiLSTM) model for splitting Icelandic compound words, and evaluate its performance for varying amounts of training data. Our model is trained on a dataset of 2.9 million unique word forms and their constituent structures from DIM. The model learns how to split compound words into two parts and can be used to derive the constituent structure of any word form. The model outperforms other previously published methods when evaluated on a corpus of manually split word forms. Our method has been integrated into Kvistur, an Icelandic compound word analyzer. Finally, preliminary experiments show that our model performs very well when evaluated on a closely related language, Faroese.

2. Compounding in Icelandic

In Icelandic, any of the open word classes can be combined to form a compound, although some combinations are more productive than others (noun-noun compounding, in particular). The DIM contains a total of 2.9 million unique inflected forms, of which approximately 2.5 million are compounds. Fully split, these compounds are composed of

169,000 distinct compound parts, with 146,000 functioning as heads and 40,000 as modifiers, with some overlap (Daðason and Bjarnadóttir, 2014). The figures are for word forms, not lemmas, and the discrepancy in numbers is due to the fact that the heads of compounds are inflected, whereas the modifiers rarely are (Bjarnadóttir, 2017).

Any word form can, in theory, appear as a head, and a compound always has the same grammatical features as its head. The form of modifiers is much more restricted, as reflected by the compound part frequencies in the DIM. Nominal modifiers are mostly limited to stems (e.g., *fótbolti* 'football', *fótur* 'foot' + *bolti* 'ball') and inflected forms in the genitive case, singular or plural (e.g., *umferðarskilti* 'traffic sign', *umferð* 'traffic' + *skilti* 'sign'). Less commonly, they may appear in the dative case (e.g., *snæviþakinn* 'snow covered', *snær* 'snow' + *þakinn* 'covered'), and they can also contain linking elements (e.g., *hæfnispróf* 'aptitude test', *hæfni* 'competency' + *próf* 'test'). Modifiers of other word classes are similarly restricted (Bjarnadóttir, 2002).

3. Related Work

Brown (2002) presents a method for splitting compound words using a bilingual dictionary between a compounding and a non-compounding language. This method identifies cognates, such as "Abdominalangiographie" in German and "abdominal angiography" in English, finding the split that maximizes the similarity between the constituent parts of the compound and the words in the non-compound.

Koehn and Knight (2003) find all ways that a potential compound word can be split into a sequence of known words, optionally joined by linking elements and inflectional suffixes. Each split is scored by the geometric mean of the word frequencies of its constituent parts. Additionally, PoS tags and a translation dictionary are used to filter out implausible splits. When evaluated on a collection of 3,498 German words that have been manually split, this method achieves a precision of 93.8% and a recall of 90.1%.

Schiller (2005) uses a morphological analyzer to find all possible splits for a given word. Each constituent part is weighted by its frequency in a morphologically annotated corpus. The probability of each potential split is calculated as the product of those weights. This method achieves a precision of 97.7% and a recall of 99.1%, when evaluated on a corpus of 30,891 compounds from German news articles.

Riedl and Biemann (2016) propose an unsupervised method for compound splitting based on distributional semantics. Their method finds the split which maximizes the semantic similarity between the compound parts and the compound itself. The method achieves a precision of 96.1% and a recall of 88.1%, when evaluated on a set of 158,653 German nouns from newspaper articles.

Tuggener (2018) evaluates various neural network architectures, both supervised and unsupervised, for splitting German compounds. The best performance is achieved by a model composed of an unsupervised BiLSTM network combined with a supervised multilayer perceptron (MLP), which obtains an accuracy of 95.1% when evaluated on a corpus of 75,000 manually split German compounds.

3.1. Kvistur 1.0

Kvistur 1.0 is a compound word analyzer, capable of determining the constituent structure of Icelandic word forms (Daðason and Bjarnadóttir, 2014). It is trained on a large corpus of manually split compound words from DIM. From this corpus, Kvistur learns the probability that any two compound parts can be combined to form a compound. The probability of a previously unseen combination of constituent parts is estimated from the probability of the former part appearing as a modifier in the corpus and the second part appearing as a head. Kvistur finds all possible ways to split a given word into a sequence of known constituent parts and uses these probabilities to find the likeliest constituent structure. This approach achieves a precision of 97.6% and a recall of 98.0%, when evaluated on a sample of 6,098 words from Icelandic Wikipedia articles.

One shortcoming of this method is that it cannot correctly split a compound word if any of its constituent parts are unknown. Additionally, it does not take any semantic information into account when estimating the probability of two parts being combined.

The aim of the work presented in this paper is to develop a neural network-based method for splitting compound words, to evaluate it, and to compare it to the original approach. This method will be integrated into Kvistur[1].

4. Method

We propose a character-based BiLSTM model for splitting compound words. For every character in an input word, the model predicts whether it marks the beginning of the second part of a compound (i.e., the head). The model returns an output vector of equal length to the word form, as shown in Table 1.

Input	r	a	f	o	r	k	u	þ	ö	r	f
Output	0	0	0	0	0	0	0	1	0	0	0

Table 1: An example of an input word (*raforkuþörf*, 'electric energy requirement') and the expected output vector from the model, denoting where a binary split occurs. The output vector for a base word would consist entirely of zeros.

Each element of the vector is equal to 0, except for the element corresponding to the position where a binary split occurs, in which case it is equal to 1. The full constituent structure of a word form can be derived by repeatedly applying the model until no further splits can be made.

BiLSTM models have been shown to perform extremely well on morphologically complex languages, and are used in current state-of-the-art models for Icelandic PoS tagging (Steingrímsson et al., 2019) and named entity recognition (Ingólfsdóttir et al., 2019). Furthermore, character-level embeddings capture information on the internal structure of words and have proven to be very helpful when dealing with OOV words, improving performance for a variety of

[1] Kvistur is available at `https://github.com/jonfd/kvistur`

NLP tasks (Plank et al., 2016; Dos Santos and Zadrozny, 2014; Verwimp et al., 2017).

Our model uses character embeddings as input. These embeddings are input into a BiLSTM layer, whose output is fed into a dense output layer, which makes a binary prediction for each character. The model is shown in Figure 2. We also evaluate a version of this model with two BiLSTM layers.

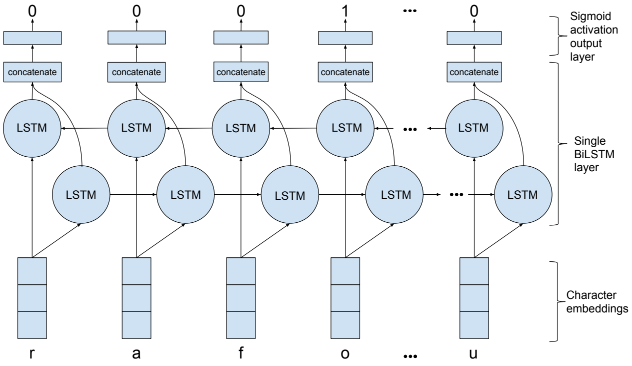

Figure 2: A representation of the model with one BiLSTM layer, showing where the compound word *raforku* 'electric energy' is split in two.

Tuggener (2018) evaluates various neural-network based models for splitting compound words, the best performing of which is an unsupervised BiLSTM model combined with a supervised MLP. The BiLSTM model is used to monitor the probability of an end-of-token symbol appearing after any character within a given token sequence. This probability distribution is passed to a supervised MLP, which is trained to identify the likeliest position of a binary split within a given token. Our model is based on the unsupervised BiLSTM model from this approach. However, our model is fully supervised, directly predicting the likeliest position of a binary split from the training data. With this in mind, we replace the MLP with a simpler sigmoid activation layer.

5. Experimental Setup

In this section, we describe the datasets used in our experiments and give an overview of various implementation details.

5.1. Datasets

We evaluate our model on Icelandic and German corpora that have been manually annotated with information on how compound words should be split.

5.1.1. Icelandic

For Icelandic, we evaluate our model using data intended for inclusion in *MorphIce* (Bjarnadóttir et al., 2019), a morphological database which will contain the constituent structure of every word form in DIM. In total, this preliminary (unpublished) version of MorphIce contains the constituent structures of approximately 2.9 million unique word forms, of which 2.5 million are compounds. For every unique word form in the data, we create a single training example consisting of the word form as input and a target vector as output, as shown in Table 1.

We observe that there are only approximately 500 word forms (out of 2.9 million) in the dataset with multiple constituent structures. One example is *heimsenda*, which can be split as *heim* 'home' + *senda* 'deliver' ('home deliver') or *heims* 'world's' + *enda* 'end' ('apocalypse'). Furthermore, we find that converting all words forms to lowercase does not introduce any additional ambiguity.

We use 80% of the dataset for training, reserving 10% for validation and 10% for testing. Each set consists of a unique set of word forms (i.e., each word form only occurs in one of the three sets), with the exact same proportion of base words and compounds. This means that there is no overlap between any of the three sets, and all words in the validation and test sets are unknown to our model. Additionally, all inflected forms of the same word are placed into the same set.

5.1.2. German

For German, we use approximately 75,000 compound words from GermaNet (Henrich and Hinrichs, 2010). Each compound has been manually annotated to indicate the position where a binary split occurs. We split the dataset into a training, validation and test set with the same ratios as for Icelandic.

5.2. Implementation details

We built our model using Keras[2]. We train using the Adam optimizer, with a constant learning rate of 0.001. We train our model for 100 epochs, stopping early if the validation accuracy does not improve for 20 epochs, and select the model with the highest validation accuracy. The character embeddings have 128 dimensions and each BiLSTM layer has 128 hidden units in each direction. The model accepts 40 character long words as input, with shorter words being padded.

6. Results

In this section, we compare our proposed model against the statistical-based method used by Kvistur 1.0, which has been shown to achieve high accuracy for splitting Icelandic compounds (see Section 3.1.).

Table 2 summarizes our results on the validation and test data, showing the overall accuracy of the three models. The accuracy is computed as the number of compounds that were correctly split into two parts and base words that were not split, divided by the total number of words. We find that the BiLSTM models significantly outperform the statistical method implemented in Kvistur 1.0, with the two-layer model performing marginally better. Additionally, we find that the two-layer BiLSTM model obtains slightly higher accuracy on German than reported by Tuggener (2018), 96.2% compared to 95.1%. We note, however, that although we used the same corpus for training and evaluation, we use a different data split.

Table 3 shows the accuracy of the the three models on base words and compounds in the Icelandic test set. We observe

[2] `https://keras.io/`

Model	is	de
Kvistur 1.0	91.7% / 91.7%	-
BiLSTM (1 layer)	97.1% / 97.1%	96.4% / 95.4%
BiLSTM (2 layers)	97.4% / 97.5%	96.6% / 96.2%

Table 2: The validation and training accuracy (respectively) of the statistical method in Kvistur 1.0 and the BiLSTM models. Kvistur 1.0 was not evaluated on the German dataset as it requires the full constituent structure of compound words for training.

that in addition to a much higher overall accuracy, the BiLSTM models are significantly less likely to mistakenly split base words.

Model	Base words	Compounds
Kvistur 1.0	84.4%	92.9%
BiLSTM (1 layer)	96.2%	97.3%
BiLSTM (2 layers)	96.7%	97.6%

Table 3: The accuracy of the three models with regards to base words and compounds in the Icelandic test set.

We also evaluate the three models in terms of precision and recall, in a similar manner as done by Koehn and Knight (2003). The precision is measured as the number of compounds that were correctly split into two parts divided by the total number of words that were split by the model. The recall is measured as the number of compounds that were correctly split into two parts divided by the total number of compounds.

Model	Precision	Recall	F-score
Kvistur 1.0	93.73%	92.87%	93.30%
BiLSTM (1 layer)	98.30%	97.26%	97.78%
BiLSTM (2 layers)	98.05%	97.59%	97.82%

Table 4: The precision, recall and F-score of the three models, as measured on the Icelandic test set.

The results of the evaluation show that the BiLSTM models obtain a much higher precision and recall than the statistical method in Kvistur 1.0. However, adding more layers to the BiLSTM model does not appear to have a significant impact on precision and recall.

Finally, we evaluate the two-layer BiLSTM model using a varying amount of Icelandic and German training data. We start off by evaluating the accuracy of our model on a training set of 2,000 words, doubling the amount of training data thereafter. We add words to the training data from the complete training set in a decreasing order of frequency, using word frequencies from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018) and from the German Wikipedia[3]. The results of this evaluation can be seen in Figure 3.

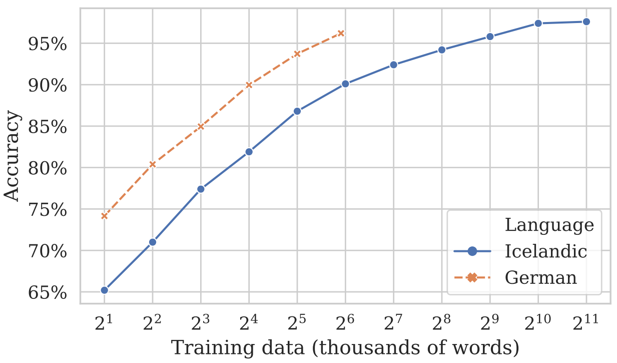

Figure 3: Accuracy of the two-layer BiLSTM model for a varying amount of training data.

Our evaluation shows that increasing the amount of training data has a positive impact on the accuracy of our model up to a certain point. For Icelandic, there are negligible gains from adding more training data beyond two million words. We also observe that our model obtains approximately 10% higher accuracy for German than for Icelandic, given the same amount of training data. This can perhaps be explained by the more complex morphology of Icelandic.

7. Conclusion

In this work, we presented a BiLSTM model for splitting compound words. We evaluated it on manually annotated corpora of Icelandic and German words and found that it outperformed previously evaluated methods.

For future work, we intend to experiment with using constituent part embeddings, derived from MorphIce, in addition to character embeddings. We also want to examine the possibility of applying our model on other languages using transfer learning. We note that without any additional adjustments or training, our model performs very well on Faroese, a low-resource language that is closely related to Icelandic. In a trial run on a sample of 166 Faroese compounds derived from the translations of Icelandic headwords starting with “dr-” in ISLEX, an Icelandic-Faroese dictionary, Kvistur 2.0 returned 12 errors: 6 incorrect splits and 6 compounds that were not split. The remaining 154 compounds were correctly split. The sample was hand-checked, but it is too small for evaluation. It should be noted that Faroese spelling differs quite a bit from Icelandic spelling, and so do the rules of compound formation to a degree.

8. Bibliographical References

Adda-Decker, M. and Adda, G. (2000). Morphological decomposition for ASR in German. In *Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Germany.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computional Linguistics*, NODALIDA 2019, Turku, Finland.

Bjarnadóttir, K. (2002). A short description of Icelandic Compounds. Retrieved from http://www.lexis.hi.is/comp-short.pdf.

[3] https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

Bjarnadóttir, K. (2005). *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum*. Orðabók Háskólans, Reykjavík, Iceland.

Bjarnadóttir, K. (2017). Phrasal compounds in Modern Icelandic with reference to Icelandic word formation in general. In Carola Trips et al., editors, *Further investigations into the nature of phrasal compounding*. Language Science Press, Berlin, Germany.

Braschler, M., Göhring, A., and Schäuble, P. (2003). Eurospider at CLEF 2002. In Carol Peters, et al., editors, *Advances in Cross-Language Information Retrieval*, pages 164–174, Berlin, Heidelberg. Springer Berlin Heidelberg.

Brown, R. D. (2002). Corpus-Driven Splitting of Compound Words. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, Keihanna, Japan.

Daðason, J. F. and Bjarnadóttir, K. (2014). Utilizing constituent structure for compound analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavík, Iceland.

Dos Santos, C. N. and Zadrozny, B. (2014). Learning Character-level Representations for Part-of-speech Tagging. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, Beijing, China.

Henrich, V. and Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC 2010, Valletta, Malta.

Ingólfsdóttir, S. L., Þorsteinsson, S., and Loftsson, H. (2019). Towards High Accuracy Named Entity Recognition for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computional Linguistics*, NODALIDA 2019, Turku, Finland.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany.

Riedl, M. and Biemann, C. (2016). Unsupervised Compound Splitting With Distributional Semantics Rivals Supervised Methods. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA.

Schiller, A. (2005). German Compound Analysis with wfsc. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 239–246. Springer.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Steingrímsson, S., Örvar Kárason, and Loftsson, H. (2019). Augmenting a BiLSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2019, Varna, Bulgaria.

Tuggener, D. (2018). Evaluating Neural Sequence Models for Splitting (Swiss) German Compounds. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018)*, Winterthur, Switzerland.

Verwimp, L., Pelemans, J., hamme, H. V., and Wambacq, P. (2017). Character-Word LSTM Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2017, Valencia, Spain.