An exploration of the encoding of grammatical gender in word embeddings

Harterg Veeman  
Department of Linguistics and Philology  
Upssala University  
harterg.veeman.7544@student.uu.se

Ali Basirat  
Department of Linguistics and Philology  
Upssala University  
ali.basirat@lingfil.uu.se

Abstract—The vector representation of words, known as word embeddings, has opened a new research approach in the study of languages. These representations can capture different types of information about words. The grammatical gender of nouns is a typical classification of nouns based on their formal and semantic properties. The study of grammatical gender based on word embeddings can give insight into discussions on how grammatical genders are determined. In this research, we compare different sets of word embeddings according to the accuracy of a neural classifier determining the grammatical gender of nouns. It is found that the information about grammatical gender is encoded differently in Swedish, Danish, and Dutch embeddings. Our experimental results on the contextualized embeddings point out that adding more contextual (semantic) information to embeddings is detrimental to the classifier’s performance. We also observed that removing morphosyntactic features such as articles from the training corpora of embeddings decreases the classification performance dramatically, indicating a large portion of the information is encoded in the relationship between nouns and articles.

I. INTRODUCTION

The creation of embedded word representations through deep learning has been a key breakthrough in natural language processing [1], [2], [3], [4], [5]. Words embeddings have proven to capture information relevant to many tasks within the field. Previous research [6] has shown a neural classifier can learn to classify word embeddings on grammatical gender. The ability of a classifier to classify embeddings on grammatical gender can be interpreted as an indication of the presence of linguistically motivated information in word embeddings. The way this information is encoded in word embeddings is of interest because it can contribute to the discussion of how nouns of a language are classified into different gender categories.

This paper explores the presence of information about grammatical gender in word embeddings from three perspectives. The first is to examine how such information is encoded across languages. This examination is based on the model transfer between languages. The second is to determine the classifier’s reliance on semantic information by using contextualized word embeddings. The third is to test the effect of gender agreements between nouns and other words categories on the information encoded into word embeddings.

II. GRAMMATICAL GENDER

Grammatical gender is a nominal classification system found in many languages. In languages with grammatical gender, the noun forms an agreement with another aspect of the language based on the noun class. The most common grammatical gender divisions are masculine/feminine, masculine/feminine/neuter, uter/nuter an animate/inanimate, but many other divisions exist.

Gender is assigned based on a noun’s meaning and it’s form. Gender assignment systems vary between languages and can be based solely on meaning [7]. The experiments in this paper concern grammatical gender in Swedish, Danish and Dutch.

Nouns in Swedish and Danish can have one of two grammatical genders, uter and neuter, that are indicated through the article for definite form and through the suffix for the definite and plural form. Dutch nouns can technically be classified as masculine, feminine or neuter, however in practice the masculine and feminine genders are only used for nouns referring to people, with the distinction between masculine and feminine being the subject and object pronouns, while in other cases all non-neuter nouns can be categorized as uter. Grammatical gender in Dutch is indicated in the definite article, adjective agreement, pronouns and demonstratives.

III. MULTILINGUAL EMBEDDINGS AND MODEL TRANSFERABILITY

The first experiments explores if grammatical gender is encoded in a similar way between different languages. For this experiment we leverage multilingual word embeddings. Multilingual word embeddings are word embeddings that are mapped to the same latent space in such a way that the vector of words that have a similar meaning have a high similarity regardless of the source language. The aligned word embeddings allow for model transfer of the neural classifier. This is done by applying the neural classifier model to a different language than it is trained on. If the model is effective for classifying embeddings from a different language, it is likely the model and embedding language encode grammatical gender in the embeddings in a similar way.

A. Data & Experiment

The nouns and their genders were extracted from Universal Dependencies treebanks [8]. This resulted in about 4.5k, 5.5k and 7.2k nouns in Danish (da), Swedish
(sv) and Dutch (nl) respectively. Ten percent of each data set was sampled randomly to be used as test data. For each noun the word embedding was extracted from pre-trained aligned word embeddings published in the MUSE project [9]. The uter/neuter class distributions are 74/26 for Swedish, 68/32 for Danish, and 75/25 for Dutch.

The classification is performed using a feed forward neural network. The network has an input size of 300 and a hidden layer size of 600. The loss function used is binary cross entropy loss. The training is run until no further improvement is observed.

For every language, a network was trained using the data described above. Every model was then applied to its own test data and the test data of the other languages, to test transferability of the model.

All data was also combined in a multi-lingual data set to create a multilingual model and test set. The same classifier as the previous experiment was trained on this data, without explicitly differentiating between the source language of the data. The purpose of this experiment was to find any language independent information on grammatical gender.

B. Results

The performance of the Swedish model for classifying Dutch grammatical gender and vice versa performed worse than the majority baseline. This indicates that little of how the grammatical gender is encoded in the word embeddings, is common between these languages.

The Danish model applied to the Swedish test data produces the best result of all model transfers, achieving an accuracy of 81.18%. Danish and Swedish are very closely related languages, so this model transfer represents a near best-case scenario. Considering this, the models performance on Swedish is surprisingly underwhelming.

The multilingual model was able to generalise well. The accuracy of this model was 89.37%. This somewhat contradicts the conclusion from the language specific model transfer experiment that grammatical gender is encoded very differently in word embeddings in different languages. It could be the case that the model learns to differentiate between the language sources or there could be common ways of encoding information that the monolingual models fail to identify. The full results can be found in table 1.

IV. CONTEXTUAL WORD EMBEDDINGS

Adding contextual information to a word embedding model has proven to be effective for semantic tasks like named entity recognition or semantic role labeling [10]. This addition of semantic information can be leveraged to measure it’s influence. In an attempt to quantify the role of semantic information in the noun classification a contextualized word embedding model was used in to control the semantic information in the embeddings.

A. Embeddings from Language Models (ELMo)

The ELMo model [4] is a multi-layer biRNN that leverages pre-trained deep bidirectional language models to create contextual word embeddings. The ELMo model has 3 layers, where layer 0 is the non-contextualized token representation, which is a concatenation of the word embeddings and a character based embeddings created with an CNN or RNN. This token representation is fed into a biRNN. Afterwards the resulting hidden state is concatenated with the states of both directions of the language model and fed into another biRNN layer.

The word representation layer of this model has been shown to capture morphology faithfully while encoding little semantics [10]. The semantic information is instead represented in the contextual layers. Comparing the results for embeddings extracted from different layers therefore allows us to compare less semantically rich embeddings (non-contextualized word representations) with more semantically rich embeddings (contextualized embeddings). A comparison of output of these different layers was made to discover the influence of this difference in semantic information.

B. Data & Experiment

The comparison was performed using a Swedish pretrained ELMo model [11]. The nouns and their gender labels were extracted from UD treebanks. The noun embeddings were generated using their treebank sentence as context. The output is collected at the word representation layer, the first contextualized layer and the second contextualized (output) layer. This resulted in a set of a little over 10k embeddings for every layer, from which 10% was randomly sampled and split as test data. The embeddings have a size of 1024, and the hidden layer size of the classifier has a size double the input size, 2048. The results for this comparison are shown in table 2.

C. Results

A clear decrease in accuracy and increase in loss is observed when classifying gender of the contextualized word embeddings. The added semantic and contextual information is not only unhelpful, but even detrimental to the classifiers performance. Based on these results it

---

TABLE I

|       | SV  | DA  | NL  | Multi |
|-------|-----|-----|-----|-------|
| SV    | 93.55 | 73.89 | 73.37 | 80.01 |
| DA    | 81.18 | 91.81 | 78.50 | 82.38 |
| NL    | 71.32 | 78.54 | 93.34 | 82.84 |
| Multi | 89.60 | 87.34 | 90.42 | 89.37 |

---

TABLE II

| Model Layer                              | Loss  | Accuracy |
|-----------------------------------------|-------|----------|
| Word representation layer                | 0.168 | 93.82    |
| First layer                             | 0.260 | 92.13    |
| Second layer                            | 0.274 | 91.24    |
can be argued that the classifier uses very little semantic information for classifying grammatical gender, and that the semantic information added in this experiment acted like noise.

V. WORD EMBEDDINGS FROM STRIPPED CORPUS

In the previous experiment we have observed that the classifier does not seem to strongly rely on semantic data to classify grammatical gender in word embeddings. This would lead to the hypothesis that it relies on information on form and agreement instead.

Form and agreement could be encoded in word embeddings through the noun’s relation with agreed words. A neuter noun in Swedish would have a high co-occurrence with ‘ett’, thus leading to a strong relationship between the vectors for the noun and the word ‘ett’.

To test this hypothesis embeddings have been created from a corpus that has all forms of agreement removed through the removal of articles and the stemming of all words.

FastText [3] was used to create Swedish word embeddings from a corpus consisting of all Swedish articles on Wikipedia. Another set of embeddings was created from the same corpus, but with all articles removed from the corpus. A third set of embeddings was created from the same corpus, after stemming it with the Snowball stemmer [12]. A classifier was trained on these embeddings in the same configuration as the previous experiments. The results can be found in table 3.

The classifier only manages an accuracy of 85.61% on the embeddings from the no articles corpus. This is almost a 6% drop caused by the missing information, which is very significant considering the 70% majority baseline. This indicates that for Swedish, the relationship between nouns and articles in word embeddings is a large part of what encoded information on grammatical gender in word embeddings.

When classifying the stemmed embeddings the accuracy falls to 84.66%. It could be argued this is in part due to the decrease in quality of the embeddings overall that comes with stemming the corpus. It is however still a clear indicator that information on form is a source of information for the classifier.

VI. CONCLUSION

In this exploration of information on grammatical gender in word embeddings it was shown through model transfer of a neural classifier that it is probable that grammatical gender is encoded differently between languages, even when languages are closely related. We do however observe that a classifier trained on multilingual data can be effective.

It has also shown through a comparison of embeddings from different layers of an ELMo model, that adding semantic information to embeddings is detrimental to a grammatical gender classifier’s performance.

Lastly it has shown through creating embeddings from a corpus that is stripped of information on form and agreement, that a noun’s form and relationship to gender specific articles is an important source of information for grammatical gender in word embeddings.

These results indicate that for Swedish word embeddings grammatical gender is encoded through information based on form rather than meaning.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[2] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.

[4] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jan. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[6] A. Basraut and M. Tang, “Linguistic information in word embeddings,” in Agents and Artificial Intelligence, J. van den Herik and A. P. Rocha, Eds. Cham: Springer International Publishing, 2019, pp. 492–513.

[7] G. G. Corbett, “Systems of gender assignment,” in The world atlas of language structures online, M. S. Dryer and M. Haspel, Eds. Max Planck Institute for Evolutionary Anthropology, 2013.

[8] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hacij, C. D. Manning, R. McDonald, S. Petrov, S. Pysyslo, N. Silveira, R. Tsofaty, and D. Zeman, “Universal dependencies v1: A multilingual treebank collection,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mato, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[9] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” arXiv preprint arXiv:1710.04087, 2017.

[10] M. E. Peters, M. Neumann, L. Zettlemoyer, and W. Yih, “Dissecting contextual word embeddings: Architecture and representation,” CoRR, vol. abs/1808.08949, 2018. [Online]. Available: http://arxiv.org/abs/1808.08949

| TABLE III | ACCURACY AND LOSS FOR EMBEDDINGS WITH DIFFERENT SOURCE CORPORA |
|-----------------|-----------------|-----------------|
|                 | Loss | Accuracy |
| Wikipedia corpus | 0.247 | 91.37 |
| Wikipedia corpus, no articles | 0.368 | 85.61 |
| Wikipedia corpus, stemmed | 0.397 | 84.66 |
[11] W. Che, Y. Liu, Y. Wang, B. Zheng, and T. Liu, “Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation,” in Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium: Association for Computational Linguistics, October 2018, pp. 55–64. [Online]. Available: [http://www.aclweb.org/anthology/K18-2005](http://www.aclweb.org/anthology/K18-2005)

[12] M. F. Porter, “Snowball: A language for stemming algorithms,” Published online, 2001. [Online]. Available: [http://snowball.tartarus.org/texts/introduction.html](http://snowball.tartarus.org/texts/introduction.html)