Automated Writing Support for Swedish Learners

Yaroslav Getman
Department of Signal Processing and Acoustics
Aalto University
yaroslav.getman@aalto.fi

Abstract

This paper describes a tool developed for lexical and grammatical analysis of Swedish text and providing automated feedback for language learners. The system looks for words and word sequences that are likely to contain errors and suggests how to correct them using different non-neural models. The feedback consists of alternative word and word sequence suggestions and morphological features[1] which need to be corrected. Although the system is able to provide reasonable feedback which is believed to be useful for language learners, it still needs further improvements to address the drawbacks such as low precision.

1 Introduction

The majority of automatic error detection and correction systems focus on searching for mistakes and providing right solutions directly without any feedback. Instead, providing the feedback would be useful especially for non-native writers and help them to understand the mistakes and correct the errors on their own.

In the *DigiTala* project (2019–2023), financed by Academy of Finland, we are developing tools for automatic Finnish and Swedish spoken language proficiency evaluation of non-native speakers. This paper addresses a system built for lexical and grammatical analysis of Swedish and giving automatic supportive feedback for language learners.

For the analysis, the current version of the system involves non-neural models only, supposing that they are able to provide enough accuracy while requiring less training data than deep neural networks. However, the models can be replaced later by neural ones for future experiments.

The rest of the paper is organized as follows. Section 2 provides a brief overview of related research. Section 3 describes the system components and the error analysis. Section 4 presents an example analysis performed by the system. Section 5 concludes the paper with ideas for future work.

2 Related work

There are some systems which act as pedagogical tools and provide constructive feedback, such as one developed by Morgado da Costa et al. (2020) for assisting students in their scientific English writing. The system described in the paper uses computational parsers and general NLP techniques, e.g. checking for repeated words, sentence length, word capitalization, etc.

Several grammar checkers for second language writers of Swedish have been developed in the research project *CrossCheck* (Bigert et al., 2004). One of them, called *Granska* (Domeij et al., 2000), consists of a POS tagger, a spelling checker and manually constructed rules for error detection and correction. The second one, *ProbGranska* (Bigert and Knutsson, 2002), is a statistical method which searches for unlikely grammatical constructions using POS tag trigram frequencies. The third one, *SnålGranska* (Sjöbergh, 2005), is a weakly supervised machine learning based system trained on a text corpus with artificially created errors.

A system called *Revita* (Katinskaia et al., 2017, 2018) is designed to support language learning and focuses primarily on Finno-Ugric languages. The system automatically generates ”cloze” exercises from texts, where a language learner needs to fill in the missing words to the sentences.

For German language learners, there is a feedback mechanism developed by Rudzewitz et al. (2018) as a part of language learners' tutoring system *FeedBook* (Rudzewitz et al., 2017). *FeedBook*

[1] https://universaldependencies.org/treebanks/sv_talbanken/index.html#features

consists of short answer and fill-in-the-blanks exercises. The system compares student answers to the target answers stored for each task and generates feedback based on predefined error templates for five grammar error types: tenses, comparatives, gerunds, relative clauses, reflexive pronouns.

3 System description

The tool developed so far relies on a language model (LM), when looking for errors in the input sentences. While ngrams (contiguous sequences of *n* words) which are present in the LM are supposed to be correct, unknown ngrams and out-of-vocabulary (OOV) words can possibly contain errors. If OOV words or unknown bigrams (two-word sequences) are found from the sentences, they are examined by the system in more detail and the feedback is provided. For the OOV words found in the sentences, the tool proposes similar words. Also, it suggests most likely part of speech and morphological features, or, grammatical categories (grammatical case, person, number, etc.), to use when asking to replace the OOV word with another word. If unknown bigram is detected, the system searches for similar bigrams and asks to change the part of speech and/or correct morphological features, if needed.

3.1 Corpora and Models

In total, 6 models are used in this work for different purposes: a part-of-speech (POS) tagger, a morphological features tagging module, a word-level LM, a subword-level LM, a LM trained on POS tags and a model for word segmentation.

A pretrained morphological features tagging module from the Stanza library (Qi et al., 2020) was also used in this work. The module is based on the Swedish-Talbanken treebank[2]. The treebank has 6,026 sentences and 96,819 tokens. It was used also for training a Conditional Random Fields (CRF) POS tagger.

The Swedish YLE corpus[3] was used for training other models. The corpus is a collection of news articles published by Finland's national public broadcasting company in Swedish from the year 2012 to the year 2018. It consists of 6,810,509 sentences and 93,405,178 tokens. The vocabulary has 1,102,561 words. The data was converted to a lowercased plain text corpus with punctuation preserved. Keeping the punctuation in the text corpus is important, because otherwise, a lot of words would form grammatically incorrect bigrams, e.g. the last word of one sentence and the first word of the following sentence which are not necessarily related to each other. When it comes to evaluation of transcribed speech, post-processing techniques for restoring the punctuation in the transcripts can be considered in order for the system to provide more proper analysis results.

In addition to plain text, the source data of the Swedish YLE corpus contains positional attributes for each word such as number of the token within the sentence, lemma, POS tag, morphological analysis, dependency head number and dependency relation. The annotations were extracted separately to create a new corpus consisting of POS tag sequences which was then used for training a trigram POS LM.

Morfessor 2.0 (Smit et al., 2014) is a tool for unsupervised and semi-supervised statistical morphological segmentation. In this work, a Morfessor model was trained in unsupervised manner. The whole YLE corpus was then passed through the model to divide it into subwords and train a subword-level trigram LM.

3.2 OOV words analysis

An OOV word found from the text is first divided into segments using the Morfessor model. Depending on the word length, different number of possible segmentations is used for further analysis: 5 most likely segmentations are preserved for words consisting of 5 or more characters and N = *length of word* segmentations for words shorter than 5 characters. It should be taken into account that different configurations were tested and these numbers found to be optimal for Swedish and might need to be adjusted for other languages.

For each of these segmentations, a new word is formed by reducing the last segment from the OOV word. If the new word is not found from the vocabulary of the LM, it is tested for possible continuations by the subword-level trigram LM: the system searches for most likely next segment(s) based on the previous segment or two previous segments. If such segments are found, new words are formed. The tool then checks if these new words are found from the LM vocabulary.

[2] https://github.com/UniversalDependencies/UD_Swedish-Talbanken

[3] http://urn.fi/urn:nbn:fi:lb-2020021103

A trigram POS LM is used to find the most likely part of speech given two preceding POS tags. If no POS tag can follow the previous POS tags according to the POS LM model, the most likely POS tag given one preceding POS tag is used instead. Morphological features are suggested using the bigram LM and a morphological features tagging module: 20 words which are most likely to follow the word before the OOV word are collected using the LM. Then, only the words belonging to the most likely part of speech are preserved. For the these words, morphological features are extracted and the most frequent value for each feature is selected.

3.3 Unknown bigrams analysis

To find bigrams similar to an unknown bigram, 5 most likely word segmentations are collected using the Morfessor model for each word of the bigram. In each of these segmentations, only the longest segment consisting of at least 3 characters is preserved. If a word consists of less than 3 characters, the whole word is preserved. After that, similar words are collected by searching for words containing any of these segments from the LM vocabulary. Then, combinations of these words are formed, including the combinations of the first word of the bigram with words similar to the second word of the bigram. The LM analyzes each of these new bigrams: the ones that are possible according to the LM are then preserved and proposed by the system as bigrams similar to the initial unknown bigram. If no similar bigrams are found, the process runs from the beginning with the number of word segmentations to collect increased by one each time until at least one similar bigram is found or the manually set threshold for the maximum number of word segmentations to collect is reached.

The tool also suggests part of speech and morphological features to use when replacing the second word of the unknown bigram. It selects the POS tag and the values for the morphological features in a similar way as for OOV words. However, the part of speech of the second word of the bigram is also taken into account. The system compares the probability for the POS tag of the word to follow two preceding POS tags to the mean of the probabilities of all possible POS tags that can follow the corresponding POS tag sequence.

If the probability is above the average, the system supposes that this part of speech is likely enough and suggests to use another word belonging to the same part of speech. In this case, morphological features of the second word of the bigram are extracted and compared to the most likely morphological features. The most likely features are collected in a similar way as for OOV words. However, here all possible values are preserved instead of selecting the most common value for each feature. The most common value is selected and suggested to use only in case if the value of the feature for the word is not in the list of possible values.

If the probability is below the average, the tool suggests to use the POS tag which is most likely to follow two preceding POS tags. The most likely morphological features are then collected in the same way as for OOV words. If no POS trigrams are found from the POS LM, the system searches for POS bigrams instead.

4 Examples of System Output

In this section, we will use the following example to explore the output of the tool:

(1) *Hej Peter! Jag försökte ringa du, men din
Hi Peter! I tried to call you, but your
mobilen var avstängd.
phone was switched off.
'Hi Peter! I tried to call you, but your phone was switched off.'

When the example is fed to the system, the output is:

"***Text to evaluate:*** *Hej Peter! Jag försökte ringa du, men din mobilen var avstängd.*

(For human evaluator) Text tagged with part-of-speech tags:

[[('Hej', 'IN'), ('Peter', 'PM'), ('!', 'MAD')], [('Jag', 'PN'), ('försökte', 'VB'), ('ringa', 'VB'), ('du', 'PN'), (',', 'MID'), ('men', 'KN'), ('din', 'PS'), ('mobilen', 'NN'), ('var', 'VB'), ('avstängd', 'PC'), ('.', 'MAD')]]

Analyzed text with annotations:

*hej peter ! jag försökte ringa du****1*** *, men din mobilen****2*** *var avstängd .*

Unknown words:

Uncommon ngrams:

1. *ringa du. Similar ngrams: placeringar du, placeringar fördubblades, beväringarna duschar, tillbringat dubbelt, tillbringade du. You used the pronoun du (Case: Nom, Definite: Def, Gender: Com, Number: Sing, PronType: Prs). You can also try to use some other pronoun instead of du. It is also recommended to correct the following morphological featutes: Case: Acc.*

2. *din mobilen. Similar ngrams: din mobiltelefon, din mobil, din mobila, din mobilbutik, din mobils. You used the noun mobilen (Case: Nom, Definite: Def, Gender: Com, Number: Sing). You can also try to use some other noun instead of mobilen. It is also recommended to correct the following morphological featutes: Definite: Ind.*"

In this example, there are two grammatical errors: ***"*ringa du"*** (*"call you"*), where the pronoun ***du*** should be used in the accusative case (***dig***), and ***"*din mobilen"*** (*"your phone"*), where the noun should be in the indefinite form (***mobil***). The system found both errors and marked the bigrams containing the errors as uncommon.

As can be seen from the list of similar ngrams, the system did not manage to provide the correct bigram (***"ringa dig"***). This happens because the morphological segmentation model defines the word ***du*** itself as its only morpheme which is then used as a search query in the LM vocabulary. However, the tool suggests correctly to change the case of the pronoun from *nominative* to *accusative*. For the second incorrect bigram, the system managed to provide both correct bigram (***"din mobil"***) and suggestion to change the form of the noun from *definite* to *indefinite*.

Here is another example for the analysis:

(2) *Ursäkta mig, jag kan inte komma eftersom
Excuse me, I can not come because
jag blev sjuk. Kan vi träffas
I got sick. Can we meet
i övermorgon eller kanske i
the day after tomorrow or maybe in
fredag?
Friday?
'*Excuse me, I can not come because I got sick. Can we meet the day after tomorrow or maybe in Friday?*'

and the corresponding output of the system:

"***Text to evaluate:*** *Ursäkta mig, jag kan inte komma eftersom jag blev sjuk. Kan vi träffas i övermorgon eller kanske i fredag?*

(For human evaluator) Text tagged with part-of-speech tags:

[[('Ursäkta', 'VB'), ('mig', 'PN'), (',', 'MID'), ('jag', 'PN'), ('kan', 'VB'), ('inte', 'AB'), ('komma', 'VB'), ('eftersom', 'SN'), ('jag', 'PN'), ('blev', 'VB'), ('sjuk', 'JJ'), ('.', 'MAD')], [('Kan', 'VB'), ('vi', 'PN'), ('träffas', 'VB'), ('i', 'PP'), ('övermorgon', 'NN'), ('eller', 'KN'), ('kanske', 'AB'), ('i', 'PP'), ('fredag', 'NN'), ('?', 'MAD')]]

Analyzed text with annotations:

*ursäkta mig , jag kan inte komma eftersom jag blev sjuk . kan vi träffas i övermorgon eller*1 kanske i fredag ?*

Unknown words:

Uncommon ngrams:

1. *övermorgon eller. Similar ngrams: imorgon eller, imorgon håller, imorgon ställer, imorgon heller. You used the conjunction eller. You can also try to use some other conjunction instead of eller.*"

The second sentence contains an error the system did not recognise at all. An incorrect preposition **i** is used before ***fretag***, while the correct answer is ***"på fretag"*** (*"on Friday"*).

There is also a bigram marked as incorrect in the second sentence of the example: ***"övermorgon eller"*** (*"the day after tomorrow or"*). The system proposed to try another conjunction instead of ***eller***, as well as to replace the whole bigram with another one. It can be noticed that there is a bigram in the list of similar ngrams very close to the original one: ***"imorgon eller"*** (*"tomorrow or"*). Although both of them are grammatically equally correct, the original one was not found from the training set of the language model and erroneously marked as incorrect.

5 Discussion

The current implementation of the tool is able to analyze words and sentences at grammatical and lexical level and provide reasonable feedback. In addition, the system can be applied for other languages by replacing the models. The models can be changed also to neural ones, if needed in future. However, the work is still in progress and further improvements are needed to overcome the existing drawbacks.

Many correct words and bigrams are not recognized by the system due to morphological richness of Swedish language. However, the same word in other word form(s) can be proposed as word(s) similar to the OOV word. Many compound words are also unknown to the system. On the other hand, some common types of grammatical errors might be skipped by the system. For example, while in the previous section ***"*din mobilen"*** was recognized as incorrect bigram, the system would not recognize the error in the phrase ***"*din nya mobilen"*** ("your new phone"), since both ***"din nya"*** and ***"nya mobilen"*** are correct bigrams according to the LM. Switching to trigrams and fourgrams has helped to increase the recall of the system. However, it resulted in increased number of false positives. In other words, a lot of trigrams were recognized as incorrect due to the limited size of the training text data of the LM. Larger text corpora can help to reduce the amount of unknown words and ngrams.

The text corpus which is used for training the LM might contain lexical or grammatical errors. However, it is quite unlikely that the same error would occur many times in the corpus. Therefore, one possible solution could be to set different threshold for the log probability of a bigram. The current threshold is set to minus infinity, which means that only unknown bigrams are recognized by the system as incorrect. Setting it to a very low number close to minus infinity would possibly help to filter out some erroneous bigrams.

The tool is able to detect bigrams containing lexical errors like wrong word choice, but it cannot provide the most suitable word based on the context. Instead, the system tries to find bigrams which are similar to the original one. One possible solution to address this drawback is to use more advanced LMs, for example Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). BERT can be used in the masked language modeling task, where an inappropriate word is masked and needs to be predicted by the LM. BERT uses both left and right context of a word and therefore is believed to make more accurate word predictions compared to the ngram LMs which look only at the preceding context. There are several BERT models available which are pretrained on Swedish text corpora, for example KB-BERT (Malmsten et al., 2020).

Because the system focuses on providing feedback, it is difficult to evaluate how well it works. In addition, there is lack of labeled data for grammatical error correction task in general, and no such a dataset was found for Swedish. A corpus of Swedish learner essays where errors are manually annotated has been presented as part of the SweLL (Volodina et al., 2016) project. Unfortunately, at the time of writing, the corpus was not yet available for public access. However, one way to evaluate efficiency of the tool would be to compare its feedback to the one provided by human annotators. Another option would be organizing a survey for Swedish learners and asking how useful they find the feedback.

References

Johnny Bigert, Viggo Kann, Ola Knutsson, and Jonas Sjöbergh. 2004. *Grammar Checking for Swedish Second Language Learners*, pages 33–47.

Johnny Bigert and Ola Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *In Proceedings RO-MAND-02*, pages 10–19.

Luís Morgado da Costa, Roger V P Winder, Shu Yun Li, Benedict Christopher Lin Tzer Liang, Joseph Mackinnon, and Francis Bond. 2020. Automated writing support using deep linguistic parsers. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 369–377, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rickard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska - an efficient hybrid system for swedish grammar checking. In *I Proceedings of the 12 th Nordic Conference in Computational Linguistics, Nodalida-99*.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. *Linköping electronic conference proceedings*, 134:27–35. Volume: Proceeding volume: ; Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition ; Conference date: 22-05-2017 Through 22-05-2017.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of its and call. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4084–4093, France. European Language Resources Association (ELRA). International Conference on Language Resources and Evaluation, LREC 2018 ; Conference date: 07-05-2018 Through 12-05-2018.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 36–46, Gothenburg, Sweden. LiU Electronic Press.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–136, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Sjöbergh. 2005. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. pages 506–512.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).