Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-213

Noisy-Channel Spelling Correction Models for Estonian Learner Language Corpus Lemmatisation

Kairit SIRTS [1]
Institute of Cybernetics at Tallinn University of Technology

Abstract. Morphological analysis is an important task in Estonian learner language studies that gives information about the words and forms used by the learners. Because of the spelling errors frequently occurring in language learner texts, these texts should undergo some error correction step before applying the conventional morphological analysis tools because the morphological analyser fails to find the correct analysis for the misspelled words. In this paper we compare several different spelling correction models with the aim of improving the lemmatisation accuracy of learner language texts. Experiments show that the simplest non-word noisy-channel spelling correction model with a disambiguation model applied on top of the morphological analyser output performs the best while some of the more complicated models even fail to beat the baseline that does not include any spelling correction.

Keywords. spelling correction, learner languages analysis, lemmatisation

Introduction

Estonian learner language corpus[2] is a collection of written texts produced by the students learning Estonian as a foreign language. The texts in the corpus have been mostly written by learners with the first language as Russian (93%). It consists mainly of short essays (44%), answers to questions (18%) and personal letters (10%).

Morphological analysis is a task that maps to each word token in the corpus its lemma, part-of-speech (POS) and a set of morphological labels. Morphological analysis of the learner language corpus would be useful for researchers conducting corpus linguistic studies about the process of learning Estonian as the second language because it would enable to analyse the learner's usage of words and forms and the most common errors arising from using the different word forms.

There are tools available for morphological analysis and disambiguation for Estonian [1] and the conventional approach would just be to use these tools to obtain the analysis. Language learners, however, make mistakes that prevent using morphological analysis tools with the same accuracy as on the texts written by the native speakers.

[1]Corresponding Author: Kairit Sirts, Institute of Cybernetics at TUT, Akadeemia tee 21, 12618 Tallinn, Estonia; E-mail: kairit.sirts@phon.ioc.ee.

[2]`http://evkk.tlu.ee/?language=en`

Estonian language is highly inflective and derivational and thus the language learner might not know the correct inflected word form or derivation in a context and instead produce a word with a spelling error. Unusual word ordering in the sentence might cause confusions for the morphological disambiguator as well that, as a result, fails to find the correct lemma and part-of-speech (POS). Also, as Estonian is a heavily compounding language, language learners tend to do lots of mistakes in compounds either by compounding words that should be written separately or vice versa.

From the above it is clear that for morphological analysis of the Estonian learner language corpus some intermediate system addressing the above mentioned problems should be involved. In this work we will focus on the problems caused by the spelling errors and experiment with different noisy-channel spelling correction models to find out what works best for Estonian learner language corpus. In the next section we will explain in more detail the noisy-channel model for spelling correction and describe the different models we use in this work. Section 2 describes the experimental setting and gives the results. Section 3 follows with the discussion and section 4 concludes the paper.

1. Noisy-Channel Models for Spelling Correction

Noisy channel is a term originating from the information theory where it is used to model the process of sending a signal through a channel containing noise that possibly corrupts the signal. The idea of using noisy channel approach in spelling correction is already more than two decades old and was first proposed by [2] and [3]. In spelling correction, the channel is the writer who is intending to write a word but due to the noise in the channel produces some misspelling instead.

In this work, we will study different noisy-channel models for spelling correction with the aim of producing the best lemmatisation for the Estonian learner language corpus. We will compare three different models that we will call **pipeline**, **marginalised** and **HMM** models respectively and that are described in detail in the following subsections. We apply all models in two different settings: by using **non-word** and **real-word** spelling correction. Non-word spelling correction deals only with non-words, these are words that do not appear in the reference dictionary, while real-word spelling correction aims also to correct the words that are in the reference dictionary but are actually erroneous spellings of some other words.

1.1. Pipeline Model

In pipeline model we will first apply the spelling correction to the learner language corpus and then process the corrected texts with the morphological analyser. The spelling correction model used here is the standard noisy-channel model of the form:

$$\arg\max_W P(W|S) = \arg\max_W P(S|W)P(W), \tag{1}$$

where $P(S|W)$ is the error model giving the probability of the spelling S given the correct word W and $P(W)$ is the language model expressing the prior probabilities of the words. In this and also other models we use the trigram language model.

1.2. Marginalised Model

The marginalised model integrates both the spelling correction and lemmatisation into one step. The idea is to find the most likely lemma and POS by marginalising over the possible corrections. For each candidate lemma-POS pair the model sums over the set of possible words that could have been given rise to this lemma:

$$\begin{aligned}\arg\max_{L,T} P(L,T|S) &= \arg\max_{L,T} \sum_{W} P(L,T|W)P(W|S) \\ &= \arg\max_{L,T} \sum_{W} P(L,T|W)P(S|W)P(W), \end{aligned} \tag{2}$$

where L and T are the lemma and POS respectively and $P(L,T|W)$ is the probability of the word having the lemma L and POS T. Most words will have only one possible lemma and tag in which case this term is equal to one but in case of ambiguous word forms this term helps to select the most likely lemma. $P(S|W)$ and $P(W)$ are the error model and language model terms as before.

Whereas in pipeline model each non-word was found its most likely correction and the trigram language model context could be taken over the sequence of corrected words, here we don't have a single most likely correction for each misspelling which causes difficulties in using the language model component if such a word happens to occur in the trigram context of another non-word. The exact solution would require summing over the set of candidate correction words of the non-word in context. However, we assume that this situation does not occur too often and so most of the times we don't have to worry about it. In those few cases where one non-word happens to occur in the trigram context of another non-word we resort to the approximation where the language model uses the sequence of original spellings as the context.

1.3. HMM Model

In HMM model we have the assumption that the text is generated according to a HMM where the hidden state values correspond to lemma-POS pairs. Each hidden state emits a word that is also latent and each word goes through the noisy channel generating the final spelling of the word that is observed. Similar to marginalised model, the HMM model sums over the set of possible corrections:

$$\arg\max_{L,T} P(L,T)P(S|L,T) = \arg\max_{L,T} P(L,T) \sum_{W} P(S|W)P(W|L,T), \tag{3}$$

where $P(L,T)$ is the n-gram transition model over lemma-POS pairs (we use again trigram model), $P(S|W)$ is the error model as before, and $P(W|L,T)$ is the probability distribution over different word forms of the same lemma.

1.4. Non-Word Spelling Correction Models

In non-word spelling correction scenarios all the described models assume the presence of some sort of oracle that is capable of telling words from non-words. We use the Esto-

nian morphological analyser in the role of this oracle. The words that are recognised by the analyser are treated as correct and their lemmas for marginalised and HMM models can be taken straight from the analyser output.

1.5. Real-Word Spelling Correction Models

Real-word spelling errors are such errors where the result of the misspelling is a valid word in the language. According to the literature the real-word spelling errors constitute 25-40% [4] of all spelling errors. We might assume that these numbers vary across different languages and it's hard to tell without further study how these numbers change when talking about language learners but these numbers indicate roughly of what we can expect.

Estonian is a highly compounding language meaning that basically any two nouns can be used to form a compound word and in such a manner one can easily create new words that have never been seen before. This legal creativeness causes problems in non-word spelling correction, because our morphological analyser oracle is incapable of detecting words that are perfectly valid but strange compounds that are actually misspellings of some other words. To bring an example, our learner language corpus test set contains a word *praaktikana* which is a misspelling of the word *praktikana* but the morphological analyser treats it as a compound word *praak_tikana* and analyses it accordingly. Ideally, real-word spelling correction system would catch such an error. Unfortunately, the real-word spelling correction models implemented in this paper failed to correct this specific error because the correct word form *praktikana* was missing from the language model vocabulary and thus it scored as low as the misspelled original form.

Turning a non-word spelling correction system into the real-word spelling correction system means essentially treating all words as potential misspellings. The main difference is that if previously we could calculate the score and make the decision for each misspelled word in isolation then with real-word spelling correction we must consider and score the whole sentence. The necessary computations grow exponentially because for each word token in the sentence we must consider all its possible corrections and find the sequence of words with the highest score. In practice, often a simplified assumption is made that in each sentence there is a maximum of one real-word spelling error [5] and the possible candidate sequences are generated by correcting only one word at a time. Although this assumption may be too simplistic for the learner language texts we still adopt it here, because the preliminary experiments on some more sophisticated real-word correction models did not yield better results.

We turn all the described models into real-word spelling correction models by first correcting the non-word spelling errors and then generating candidate sentences by treating each of the remaining word as a possible real-word spelling error.

1.6. Post-Processing Disambiguation

Even when the Estonian morphological analyser is run in the disambiguation mode there are words whose analysis remains ambiguous. The simplest approach in this situation would be to just choose the analysis listed first. There is no guarantee that this analysis is the best one according to the disambiguator and thus we decided to implement a small additional disambiguation model that would choose among the analyses proposed by the

morphological analyser. The disambiguation model chooses the most likely lemma-POS pair basically in the same manner as the HMM model but omits the marginalisation part:

$$\arg\max_{L,T} P(L,T)P(W|L,T) \tag{4}$$

where $P(L,T)$ is the trigram transition probability over sequences of lemma-POS pairs and $P(W|L,T)$ is the probability distribution over different word forms of the given lemma.

2. Experiments

2.1. Training Models

We trained the trigram language model on the corpus of newspaper texts[3] containing 87.9 mio word tokens with the lexicon containing 100000 most frequent word tokens. The out-of-vocabulary tokens are modelled in the standard way with the special word UNK. Lemma-tag pair model for HMM and disambiguation model uses also trigrams and is trained on the same newspaper corpus. We first disambiguated the whole corpus with morphological analyser and then extracted lemma-type pairs. We again use the lexicon of size 100000 containing the most frequent lemma-tag pairs. Both of these language models are trained with the SRILM toolkit[4] with modified Kneser-Ney smoothing.

The error model uses the Damerau-Levenshtein distance heuristic [6,7] and is trained on the list of misspelled-corrected word pairs that were collected from the newspaper texts[5]. In the error model, each edit operation on each character is assigned a probability based on the frequency of how many times this operation can be found in the spelling errors corpus. The probabilities of insertion and deletion operations are calculated in one-character context. For example, the error model will calculate the probabilities of inserting and deleting 'b' after 'a' etc. The probabilities learned from this corpus are smoothed with absolute discounting.

The candidate words to consider are chosen from the pool of assumably correct word forms that were extracted from the same newspaper corpus and contains ca 500000 tokens. For non-word spelling corrections we will choose as candidates the words that are up to edit distance 2 from the misspelled words. When there are no candidates within the edit distance 2, the word itself is proposed as candidate. The real-word spelling error correction candidates are generated within the edit distance one.

2.2. Results

We trained the six models described above and applied them on the Estonian learner language corpus. Development and evaluation will be based on a subset of learner language corpus consisting of 460 sentences containing ca 6500 running word tokens that have been manually annotated by a linguist[6]. Each word in this subset has been annotated with

[3] http://www.cl.ut.ee/korpused/segakorpus/epl/

[4] http://www.speech.sri.com/projects/srilm/

[5] This list can be downloaded from http://www.phon.ioc.ee/dokuwiki/lib/exe/fetch.php?media=people:sirts:est-spelling-errors.txt

[6] We thank Pille Eslon from Tallinn University for doing the annotations.

Table 1. Lemmatization results by using different spelling correction models.

	Without disambiguation			With disambiguation		
Model	**Lemmas**	**Types**	**Both**	**Lemmas**	**Types**	**Both**
Baseline	93.27	92.52	90.01	94.65	93.61	91.10
Non-word Pipeline	95.44	93.57	91.88	**95.70**	**93.83**	**92.15**
Non-word Marginalized	90.54	86.76	84.37	95.06	93.01	90.95
Non-word HMM	90.84	86.76	84.67	95.06	93.01	90.95
Real-word Pipeline	95.44	93.57	91.81	95.62	93.75	92.00
Real-word Marginalized	92.86	90.39	88.11	95.03	92.89	90.80
Real-word HMM	90.58	86.76	84.37	93.68	91.88	89.83

its correct lemma and POS. We divided this set into development and test set of equal sizes. Both language models are scaled by exponentiating them with the parameter λ and the development set is used to tune this scaling factor. Test set is used to report the final results.

We will compare the proposed models with each other and also with the baseline that just uses morphological analyser output on the original learner language texts. For each model we report the results with and without the post-processing disambiguation model. We report the accuracies of the lemmas and POS tags separately and also the accuracy of getting both lemma and POS right. The results are given in **Table 1**.

3. Discussion

The experimental results show that the simple non-word pipeline spelling correction model performs the best and that also real-word pipeline model performs better than the marginalising models. Marginalised and HMM model are restricted to choose lemmas from the list of most frequent 100000 items. If the true lemma happens to be missing from this lemma lexicon then the system has no way of inferring it and has to come up with a lemma proposal that is the word itself (which is most of the times wrong). The pipeline model will operate within the frames of possible corrections and does not need to know the lemmas of these words because, assuming that they all are proper words, the morphological analyser is able to find the correct lemma for each of these.

In all cases the models with post-processing disambiguation work significantly better that the same models without it. Using post-processing disambiguation on the morphological analyser output leads to a very high baseline that the marginalising models are even unable to beat.

The non-word spelling correction models perform slightly better than real-word correction models. In order to explain it note that the real-word correction models first do the non-word error correction and then on top of that the real-word correction part. It means that the real-word error correction part must induce some more errors than it can correct. Comparison of outputs with non-word spelling correction revealed that most of the induced errors are related to proper names that the real-word correction part tries to correct to nouns. These results show that the adopted strategy for real-word spelling correction did not justify itself and that further studies are needed to find the proper models to reveal and correct the real-word spelling errors in language learner texts.

Some systematic errors are due to the inconsistencies between the hand-annotated gold standard data and the output of morphological analyser. For example, some words that are labelled as adverbs in gold standard are systematically tagged as prepositions or postpositions by the analyser. As these errors are mostly due to some finite set of words then the simplest way of reducing these errors would be to compile a list of error-causing words and their types according to the linguists studying the learner language and use it to override the decisions made by the system.

The error model is trained on the spelling errors collected from the newspaper texts that have been assumably written by the native speakers. We could have collected the errors also from the learner language corpus but we decided that collecting the list of spelling errors introduced by native speakers is more general and could be useful also for other purposes. One could try to collect some amount of spelling errors from the learner language corpus and use the obtained edit counts to adapt the error model. However, we leave this as a future work for now.

The proposed models were developed to be used for lemmatising Estonian learner language corpus. It is clear that none of these models can get everything right but the aim of applying these models is to reduce the necessary manual work as much as possible. One could imagine at least two scenarios: 1) the whole learner language corpus is processed with the model and the fact that some amount of words have wrong lemmas and/or POS is accepted; 2) the model is implemented in an interactive environment where the words with the least confidence and the most confusion are detected and passed over to the user for manual inspection and analysis. The first approach would enable the full automatic processing while the second one would need the assistance of a linguist but would result in the higher accuracy.

The linguists studying the Estonian language learning would also be interested in the full morphological analysis of the corpus that beside the lemma and POS would also contain the morphological labels for each word. This poses no problem to the well-formed sentences because the Estonian morphological analyser is in general capable of doing good analysis and disambiguation. Things get much more complicated when the sentences contain misspelled words paired with strange word order or other ungrammaticalities. In such cases the morphological analyser might err even on correctly spelled words because of the unusual context. Currently, the models we developed do not provide any systematic way of predicting the correct morphological analysis of the misspelled words because the focus is on getting the lemma right. Finding correct lemma does not necessarily require finding the exact true correction for the word while the task of identifying morphological labels does.

4. Conclusion

We experimented with several spelling correction schemes with the aim of improving the accuracy of Estonian learner language corpus lemmatisation. The experiments showed that the simplest non-word noisy-channel spelling correction model together with the simple post-processing disambiguation model lead to the best results. In general, the post-processing disambiguation model was improving the results in most cases. The competing models that summed over possible corrections did not achieve as good results with the marginalized model being in most cases better than the HMM model. Also, the

non-word spelling correction schemes tended to achieve slightly better results than the more complicated real-word spelling correction models.

One could think of several possibilities for further improving the lemmatisation results. One way would be to introduce a set of heuristic rules that could solve some more frequently occurring systematic errors with high accuracy. Another approach to consider would involve human interaction, so as to find a set of words that are most confusing to the system and ask for human help for lemmatising them. In such a manner, most of the lemmatisation would be done automatically, but the user can help to improve the results by manually analysing the words the model finds too hard.

There are further issues that have not been touched at all in this paper, such as problems with unusual word ordering and errors in writing compound words and addressing these questions would potentially further help to improve the lemmatisation results.

References

[1] H.-J. Kaalep and T. Vaino, "Complete morphological analysis in the linguist's toolbox," in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, (Tartu, Estonia), pp. 9–16, 2001.

[2] E. Mays, F. J. Damerau, and R. L. Mercer, "Context based spelling correction," *Information Processing and Management*, vol. 27, pp. 517–522, Sept. 1991.

[3] M. D. Kernighan, K. W. Church, and W. A. Gale, "A spelling correction program based on a noisy channel model," in *Proceedings of the 13th conference on Computational linguistics - Volume 2*, COLING '90, (Stroudsburg, PA, USA), pp. 205–210, Association for Computational Linguistics, 1990.

[4] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.

[5] L. A. Wilcox-O'Hearn, G. Hirst, and A. Budanitsky, "Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model.," in *CICLing* (A. F. Gelbukh, ed.), vol. 4919 of *Lecture Notes in Computer Science*, pp. 605–616, Springer, 2008.

[6] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, pp. 171–176, Mar. 1964.

[7] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.