Automatic Diacritics Restoration for Hungarian

Attila Novák[1,2] **and Borbála Siklósi**[2]
[1]MTA-PPKE Hungarian Language Technology Research Group,
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
50/a Práter street, 1083 Budapest, Hungary
{novak.attila, siklosi.borbala}@itk.ppke.hu

Abstract

In this paper, we describe a method based on statistical machine translation (SMT) that is able to restore accents in Hungarian texts with high accuracy. Due to the agglutination in Hungarian, there are always plenty of word forms unknown to a system trained on a fixed vocabulary. In order to be able to handle such words, we integrated a morphological analyzer into the system that can suggest accented word candidates for unknown words. We evaluated the system in different setups, achieving an accuracy above 99% at the highest.

1 Introduction

Due to clumsy mobile device interfaces and reluctance of users to spend too much time entering their message, a great amount of text is generated in a format that lacks the diacritic marks normally used in the orthography of the language the text is written in. Whatever the causes for the missing accents are, NLP applications should be able to restore or generate the accented version of such texts prior to any further syntactic or semantic processing to avoid upstream errors.

In this paper, we aim at solving the problem of restoring accents in Hungarian texts with the combined application of a statistical machine translation system and a morphological analyzer. Our method can be applied to any other languages that have an accurate morphological analyzer.

2 Related work

For Hungarian, there have been some attempts at creating accent restoration systems. Zainkó et al. (2000) and Mihalcea and Nastase (2002) are examples for ML approaches, where the correct places of diacritics are predicted from the immediate grapheme-level context of the unaccented letter with an accuracy of 95%. Thus, unseen words can also be accented, but incorrect forms may also be introduced into the text. Dictionary-based approaches rely on large text corpora and the distribution of the different accented forms. Zainkó et al. (2000) report to have achieved a performance of 98% of accuracy with their dictionary-based method. Nevertheless, their system cannot recognize unseen wordforms quite common in Hungarian. Németh et al. (2000) have implemented a complex text processing system for TTS applications, applying morphological and syntactic analysis. The authors report that the performance of accent restoration depends very much on the performance of the analyzers (achieving 95% accuracy at best). Neither the implementations nor the resources used in these systems have been made publicly available.

A language-independent tool, Charlifter (Scannell, 2011), is based on statistical methods relying on a lexicon, a bigram contextual model and character distributions built from a training corpus. Its performance on Hungarian with its pre-built models is compared to our results in Section 5.

For other languages, similar methods are used. Yarowsky (1994) presents a comprehensive report on corpus-based techniques used for French and Spanish texts. The role of the context is emphasized in this report, however, both word form and accent variations are relatively moderate in the investigated languages compared to Hungarian. The study of Zweigenbaum and Grabar (2002) is also aiming at French, but in the medical domain, which contains a higher ratio of unknown words than general language. In their work, a tagging method is applied in combination with transducers, resulting in a tag sequence corresponding to each letter. The method is successfully (92% precision) applied to single headwords of a medical thesaurus (without exploiting any context). The most similar method to ours is that of Pham et al. (2013), who also applied SMT in order to au-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2286–2291,
Lisbon, Portugal, 17-21 September 2015.

tomatically restore accents in Vietnamese texts. In their case, the best results produced an accuracy of 93%. However, their system is augmented with a dictionary, and the distribution of accents and grammatical behaviour are also quite different from Hungarian.

3 Hungarian

Hungarian is an agglutinating language with an orthography that represents compounds as single word forms. These may result in rather complex word forms and words are often composed of long sequences of morphemes. Thus, agglutination and compounding yield a huge number of different word forms.

In Hungarian, umlauts and acute accents are used as diacritics for vowels. Acute accents mark long vowels, while umlauts are used to indicate the frontness of rounded vowels *o*→*ö* [o→ø] and *u*→*ü* [u→y], like in German. A combination of acutes and umlauts is the double acute diacritic to mark long front rounded vowels *ő* [øː] and *ű* [yː]. Long vowels generally have essentially the same quality as their short counterpart (*i-í, ü-ű, u-ú, ö-ő, o-ó*). The long pairs of the low vowels *a* [ɔ] and *e* [ɛ], on the other hand, also differ in quality: *á* [aː] and *é* [eː]. There are a few lexicalized cases where there is a free variation of vowel length without distinguishing meaning, e.g. *hova*∼*hová* 'where to'. In most cases, however, the meaning of differently accented variants of a word is quite different. Table 1 shows all the possible unaccented-accented pairs of vowels in Hungarian together with their distribution in a corpus of 1 804 252 tokens.

a	a: 70.33%; á: 29.66%
e	e: 73.40% é: 26.59%
i	i: 86.04% í: 13.95%
o	o: 55.41% ó: 14.65% ö: 15.82% ő: 14.10%
u	u: 46.96% ú: 12.72% ü: 29.98% ű: 10.32%

Table 1: Possible accent variations in Hungarian

4 Method

In this research, we considered the problem of accent restoration as a translation task, where the source language is the unaccented version, and the target language is accented Hungarian. Since it is easy to come up with a parallel training corpus for this task, methods of SMT can be applied.

In our experiments, we used Moses (Koehn et al., 2007), a widely used SMT toolkit for building the translation models and performing decoding, and SRILM (Stolcke et al., 2011) to build the necessary language models. Moses was used with its default configuration settings and monotone decoding (i.e. reordering was not allowed), and without the alignment step, which was not needed in our case.

4.1 The baseline setup

In the baseline setup, only the translation and language models built from the training corpus were used. The input for the decoder was Hungarian raw texts with all the accents removed. The translation model contained only unigram phrases (larger n-grams were also tried, but did not change the results) and the language model contained phrases up to 5 grams. Thus, the translation model was responsible for predicting the distribution of accented forms and the language model exploited contextual information.

Another baseline was also created in order to monitor the effect of the SMT system. In this second baseline, each unaccented word form was replaced by its most frequent accented form in the training set.

4.2 Incorporating a morphological analyzer

In order to be able to restore accents in unseen words as well, a Hungarian morphological analyzer (Prószéky and Kis, 1999; Novák, 2003) was integrated. A special version of the analyzer was created that directly maps unaccented word forms to their possible accented variants while also marking morpheme boundaries and adding morphosyntactic category tags. The segmentation marks (e.g. compound and derivational suffix boundaries) and the tags are used when we assign a score to the accented candidates. We also reanalyze accented forms to retrieve lemmas not directly returned by the accenting analyzer. In our test set of 1 804 252 tokens, about 1% of the words were not found in the translation model even in the case of the largest, 440 million words, training set. Table 2 shows the ratio of unknown words (OOV) as a function of the size of the training set used for building the phrase table.

For these unknown words, all possible correct accented candidates were generated by the morphological analyzer. These candidates were then fed to the Moses decoder using its `-xml-input`

train	sentences	M words	OOV in test
100K	100 000	1.738	9.63%
1000K	1 000 000	18.078	3.44%
5000K	5 000 000	89.907	1.23%
10M	10 000 000	180.644	1.68%
ALL	24 048 302	437.559	0.81%

Table 2: Ratio of OOV after building a translation model from a training set of a certain size

parameter. In order to be able to use this feature of the decoder, a probability for each candidate form had to be estimated. First, we assumed uniform distribution among the candidates. However, this approach assigned the same probability to the most common and the most nonsensical (although grammatical) candidates as well. Thus, in some cases these forms showed up in the results. In order to avoid the system to make such errors, a more sophisticated distribution was estimated for the candidate set. For this, we applied a linear regression model based on corpus frequency data determined for the lemma and other features of the candidate word (since the actual wordform was not present in the corpus). Thus, for each candidate, its lemma frequency (LEM), the number of productively applied compounding (CMP), the number of productively applied derivational affixes (DER), and the frequency of the inflectional suffix sequence returned by the analysis were determined. Compounding and derivation were penalized (i.e. they were given a negative sign), because the morphological analyzer could suggest some nonsensical, though grammatical compound or derived forms. Sometimes such forms could be the correct ones, but the more productive compounding and derivation there is in a word, the lower score it should get. On the other hand, the frequencies of the lemma and the inflectional pattern should increase the score of a candidate, thus these components were given positive weights. Based on these components, a score was assigned to each candidate based on Formula (1).

$$score = -\lambda_c \#CMP - \lambda_d \#DER + log_{10}LEM + \lambda_i log_{10}INF + MS \quad (1)$$

, where

$$MS = \begin{cases} |minscore| + 1 & if \quad minscore \leq 0 \\ 0 & otherwise \end{cases} \quad (2)$$

The MS component was used to scale up the scores by adding $|minscore| + 1$, i.e. the lowest score received for any candidate in the actual candidate set in order to evade negative scores. The λ weights were set by the `mert` tuning of the Moses system. We used a separate development set for this, on which we observed the distribution of compounds, derivational and inflectional suffixes in OOV words analyzed by the morphology and from which we sampled 1000 words approximating the observed distributions. The target of the optimization in the `mert` tuning was the accuracy of the system on these words, resulting in the optimal values for each λ. Even though, in linear regression, it is standard to use an additional bias weight, we did not find it necessary, because we did not need to bring our estimates in sync with estimates from other sources. And assuming one factor to have a fixed unit value was just another simplification that would not affect the overall ranking, just its scaling.

Even though, following an appropriate scaling of the scores, the ranked candidates could be used the same way as the entries in the translation table, the system would never select any accented form other than the most probable one, since the language model does not include any of these forms. Thus, only the candidate with the highest relative score was made available to the system.

5 Experiments and results

In our experiments, the Hungarian webcorpus (Halácsy et al., 2004) was used for training and testing purposes. A set of 100 000 sentences were separated from the corpus as the test set, and another 100 000 sentences were used as a development set. The rest were used for training in different settings. The size of each training set is shown in Table 2.

We evaluated the performance on all the 1 804 252 tokens of the test set (56.84% correct without accent) and on a subset of 1 472 200 words that included any vowels (47.09% correct without accent). The experiments were then performed for the baseline system using the most frequent form (BL-FREQ), for the baseline SMT system (BL-SMT) and for the one augmented by the morphology with the first-ranked candidate (RANK). Table 3 shows the detailed results for the smallest and largest training sets for all words (ALL) and for words that include vowels (VOWEL). It can be seen that the precision of the system is only

	100K			ALL		
system	prec	rec	acc	prec	rec	acc
BL-FREQ-ALL	98.25	82.82	92.34	98.37	96.26	98.13
BL-FREQ-VOW	98.25	82.82	90.62	98.37	96.26	97.71
BL-SMT-ALL	99.03	83.88	92.91	99.09	97.36	98.72
BL-SMT-VOW	99.03	83.88	**91.31**	99.09	97.36	**98.44**
RANK-ALL	98.81	98.08	98.99	99.01	98.56	99.23
RANK-VOW	98.82	98.08	**98.77**	99.02	98.56	**99.06**

Table 3: Performance results for each experimental settings and training size

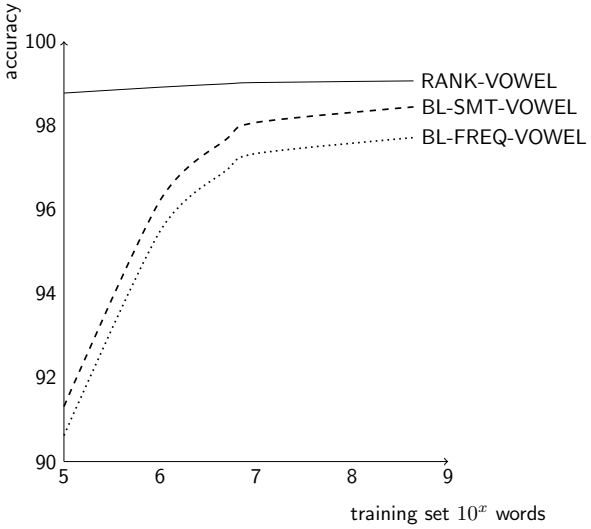

Figure 1: Accuracy as a function of the size of the training set for each system, measured on words containing vowels.

slightly improved when increasing the size of the training corpus, but the values of recall and accuracy do dramatically improve in the case of the baseline system. However, the integration of the suggestions of the morphology can make up for the lack of information due to the small training set improving recall a great deal while only slightly affecting precision. Even for the biggest 437.6M-word training corpus, incorporating the morphological analyzer with ranking yielded a relative error rate reduction of 39.74%, reducing the word error rate from 1.56% to 0.94%. For the smallest 1.74M-word training corpus tested, the relative error rate reduction was 85.85%. The system including the morphological analyzer performs better even with the smallest training corpus in terms of word accuracy than the baseline Moses system with the biggest corpus. Figure 1 shows the learning curves for each system with accuracy as a function of training set size.

Comparing our results to those we obtained using Charlifter (89.75% with most frequent accented form baseline, 90.00% with the lexicon-lookup+bigram contextual model and 93.31% with lookup+bigram context+character-n-gram-based model), the results reveal that both the contextual model in the SMT system improves accuracy better than the bigram context model of Charlifter, and the performance boost we get by incorporating morphology vastly exceeds the accuracy improvement yielded by the incorporation of the character-n-gram-based model used in Charlifter.

6 Error analysis

We performed a detailed error analysis on a 5000-sentence (87786-token) fragment of the test set. The results of the error analysis are presented in Table 4.

The detailed analysis showed that 14.7% of mismatches between the original and the system output is in fact not due to the latter being erroneous. 3.55% are equivalent forms , while the rest is correct in the output and erroneous in the reference, i.e. the system corrected errors in the original.

Another part of the reference (17.91%) is likewise erroneous, however, since the error in these cases was not in the accents, the system was not able to correct it. Missing or substituted letters are the most common mistakes (10.81%), and further 6.42% of the errors is due to punctuation errors in the original.

About $2/3$ of the mismatches are real errors. 5.57% of these could be attributed to the stem of the word missing from the database of the morphological analyzer. In 3.55% of the cases, the system transforms a name to a more frequent word: sometimes to another name, but more often to some common frequent word. A similar case is when some common noun is transformed to a more frequent name (another 1.35%). The number of these errors could be reduced to some extent by making the system rely on case information (in the case of some proper name-common noun ambiguities), however this could make the system perform worse elsewhere due to increased data sparseness. 2.20% of the errors is due to errors in the training corpus. Since rare word forms are quite frequent in Hungarian, the chances are high that a specific form is more often mistyped

Mismatch type	Ratio	Examples
Output correct	**14.70%**	
Equivalent forms	3.55%	lévő→levő fele→felé áhá→aha periférikus→periferikus
Corrected erroneous name	1.01%	USÁ-ban→USA-ban Szóladon→Szóládon
Other corrected erroneous	10.14%	un.→ún. kollegánk→kollégánk lejto→lejtő lathato→látható
Real errors	**67.40%**	
Missing from MA	5.57%	hemokromatózis-gén→hemokromatozis-gen
Correct name to erroneous output	3.55%	MIG→míg Bösz→Bősz Ladd→Ládd Márton→Marton
Other correct original to some erroneous form	2.20%	megőrzést→megorzést routeréhez→routerehez
Other correct original to contextually inadequate name	1.35%	logó→logo eperjeskein→eperjeskéin
Other correct original to some contextually inadequate form	51.01%	még→meg termek→termék gépét→gépet címét→címet vágyók→vagyok érméket→érmeket képé→képe
Original is a filename or a url containing accents	3.72%	latok→látok viz→víz szantok→szántók telepok→telepók felhasználó@profinter.hu→felhasznalo@profinter.hu www.valamicég.hu→www.valamiceg.hu
Uncorrected error in original	**17.91%**	
Punctuation error in original	6.42%	közalk.tan→kozalk.tan 1922.évi→1922.evi
Hyphenation error in original	0.68%	bemuta-tásra→bemuta-tasra
Other error in original	10.81%	véri→veri ra→rá gonolkozásában→gonolkozasaban imátkoztok→imatkoztok hirújsághoz→hirujsasghoz változaban→változabán környezetkíméli→kornyezetkimeli

Table 4: Analysis of mismatches between the system output and the input on a 5000-sentence test sample

than not (this is especially true for word forms that occur only once in the training data). 3.72% of the errors in the analyzed test data was due to either transforming arbitrary unaccented letter sequences used as file names in the text being transformed to some meaningful words or to accented words being used in an url in the original text.

The most common error (51.01% of all mismatches) is the case where the system is simply unable to correctly disambiguate the word in context, and this is not due to some other error or information loss. Interestingly, more than half (51%) of these errors belong to a single type where the system is unable to distinguish a possessive and a non-possessive form of the same nominal lemma: *gyereket∼gyerekét* 'the child (accusative)' vs.'his/her child (accusative)', *gyereken∼gyerekén* 'on/about the child' vs. 'on/about his/her child', and *gyereke∼gyereké* 'his/her child' vs.'(belongs to a) child' (anaphoric possessive).

Another 26% of the mismatches is due to a similar problem concerning verbs. In Hungarian, transitive verbs agree with their object in definiteness. Certain past, present and conditional verb forms differing in definiteness are only distinguished by an accent: *hajtottak∼hajtották* 'they drove' vs. 'they drove it'; *hajtanak∼hajtanák* 'they drive' vs. 'they would drive it'; *hajtana∼hajtaná* 'he/she would drive' vs. 'he/she would drive it'.

A factored model could in theory improve the recognition of these structures. It is questionable however, whether the improvement would justify the costs.

7 Conclusion

We have described a method to restore accents in Hungarian texts. The baseline method using only a fixed training corpus to build translation and language models for a statistical machine translation system, which is limited to handling word forms present in the training corpus achieved an accuracy of 98.44% at best. In order to process unknown words, a morphological analyzer was integrated to produce accented candidates for these unknown words as well, resulting in an improved accuracy of 99.06%. This performance could only be achieved by a system that is able to produce correct word forms and takes context into account. Our method can be applied to any other languages for which a training corpus and a morphological analyzer are available.

References

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for hungarian. In *LREC*. European Language Resources Association.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague. Association for Computational Linguistics.

Rada Mihalcea and Vivi Nastase. 2002. Letter level learning for language independent diacritics restoration. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Géza Németh, Csaba Zainkó, László Fekete, Gábor Olaszy, Gábor Endrédi, Péter Olaszi, Géza Kiss, and Péter Kis. 2000. The design, implementation, and operation of a Hungarian e-mail reader. *International Journal of Speech Technology*, 3(3-4):217–236.

Attila Novák. 2003. What is good Humor like? [Milyen a jó Humor?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Luan-Nghia Pham, Viet-Hong Tran, and Vinh-Van Nguyen. 2013. Vietnamese text accent restoration with statistical machine translation. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 423–429. Department of English, National Chengchi University.

Gábor Prószéky and Balázs Kis. 1999. A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin P. Scannell. 2011. Statistical unicodification of african languages. *Language Resources and Evaluation*, 45(3):375–386.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii, December.

D. Yarowsky. 1994. A comparison of corpus-based techniques for restoring accents in spanish and french text. In *Proceedings of the 2nd Annual Workshop on Very Large Text Corpora*, pages 19—-32, Las Cruces.

Cs. Zainkó, G. Németh, G. Olaszy, and G. Gordos. 2000. Eljárás adott nyelven ékezetes betűk használata nélkül készített szövegek ékezetes betűinek visszaállítására.

Pierre Zweigenbaum and Natalia Grabar. 2002. Accenting unknown words in a specialized language. In Stephen Johnson, editor, *ACL Workshop on Natural Language Processing in the Biomedical Domain*, pages 21–28. ACL.