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

© 2014 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-442-8-57

Comparison of Rule-based and Statistical Methods for Grapheme to Phoneme Modelling

Ilze AUZIŅA[a,1], Mārcis PINNIS[b], Roberts DARĢIS[a]
[a]*Institute of Mathematics and Computer Science, University of Latvia, Latvia*
[b]*Tilde, Latvia*

Abstract. Grapheme to phoneme modelling is one of the key features in automated speech recognition and speech synthesis. In this paper, the authors compare two different approaches: a statistical machine translation based method using the phonetically transcribed Latvian Speech Recognition Corpus and a rule-based method for phonetic transcription of words from grammatically correct forms. The paper provides 10-fold cross-validation results and error analysis for both methods.

Keywords. Grapheme to phoneme modelling, Latvian language, method comparison

Introduction

Grapheme to phoneme modelling is one of the key features in automated speech recognition and speech synthesis. In informal conversation, people often do not pronounce all phonemes in order to talk faster, not all graphemes are realised in phonetic transcription, and the same graphemes may correspond to different phonetic symbols, depending on the surrounding context. Often, because of contexts that are difficult to pronounce, multiple phonemes may also merge into one phoneme. In automated speech recognition, it is important to generate pronunciation as close as possible to the most widely used pronunciation, despite what grammar rules require. In this paper, we focus on methods for grapheme to phoneme modelling for speech recognition purposes. There are various methods that have been proposed in order to obtain pronunciation variants from the grapheme representation of words. All methods can be grouped into data-based and knowledge-based methods [1]. We compare two different approaches: a statistical machine translation (SMT) based method using a phonetically transcribed Latvian Speech Recognition Corpus (LSRC) [2] and a rule-based method for phonetic transcription of words from grammatically correct forms. The latter method is based on the phonetic transcription rules used in the Latvian speech synthesis platform [3]. For automatic evaluation, we use the phonetically transcribed corpus in a 10-fold cross-validation scenario.

[1] Corresponding Author: Ilze Auziņa, Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia; E-mail: ilze.auzina@lumii.lv.

1. Data Sets

In our experiments, we use the recently developed Latvian Speech Recognition Corpus [2]. The corpus consists of 100 hours of orthographically annotated speech and 4 hours of phonetically annotated speech that represent the major speaker base of Latvian. That is, it contains balanced proportions of speech of people from both genders, different age categories, with different accentual and dialectal characteristics, and different speech styles (prepared and spontaneous speech). The corpus is both phonetically balanced and phonetically rich, which means that it is well suited for evaluation of different grapheme to phoneme models. The work reported in this paper is based on the phonetically annotated part of the LSRC.

The phonetically annotated corpus is provided in the broad transcription (or phonemic transcription), however, additional information about phonetic variations of some specific allophones in utterances is also marked: 1) lengthened consonants (e.g., *kase* “booking-office” [kɑsːĕ], *mežs* “forest” [meʃˑ], *apburt* “to charm” [ɑbːurt]), 2) extra short vowels (e.g., *māsa* “sister” [mɑːsă], *māsas* “sisters” [mɑːsăs]), 3) non-syllabic vowels (e.g., *tēvs* “father” [tæːus], *klājs* “deck” [klɑːi̯s]). The set of symbols used in the phonetically annotated corpus contains 48 phoneme and more than 20 allophone models. The phonetically annotated corpus also allows to identify cases where the main stress does not occur on the first syllable of a word.

The rule-based grapheme to phoneme modelling method produces less rich transcriptions, and therefore, we use two different phoneme sub-sets of the broad transcription that is used in LSRC:

1) In the first set, we omitted markings about exceptions in stressed syllables.
2) In the second set, we also omitted allophones, e.g., long and prolonged consonants (kk [kː]→ k, KK [cː]→ K, zz [z] → z, SS [ʃ] → S etc.)), non-syllabic vowels (i^ [i̯] → i, u^ [u̯],]→ v), and extra short vowels at the end of words (ax [ă]→ a, ix [ĭ]→ i, ux [ŭ] → u, ex [ĕ] → e) are not included.

2. Evaluation Methodology

To evaluate the two different grapheme to phoneme modelling methods, we use automated evaluation methods that are used to evaluate the quality of automatic speech recognition, i.e., Phoneme Error Rate (PER) and Word Error Rate (WER). For evaluation, we split the phonetic corpus in 10 sub-sets and performed a 10-fold cross-validation. For the SMT-based method, 8 sub-sets were used for SMT model training, 1 sub-set was used for tuning, and 1 sub-set was used for evaluation in each fold.

3. Rule-based Grapheme to Phoneme Modelling

For rule-based transcription, we use a custom built system which transcribes text by first tokenising it into words and then transcribing each word independently of the surrounding context. The rule-based transcription is performed as follows:

1) At first, each word is looked up in an exception dictionary which consists of approximately 1,300 exceptions that have pre-defined phonetic transcriptions.

2) If the word is not found among common exceptions, we apply transcription rules in order to transcribe the word. There are approximately 250 expert defined rules, which are sorted by priority. Each rule is context dependent with respect to nearby letters and the position in the word.

4. SMT-based Grapheme to Phoneme Modelling

Using the different training data folds of the phonetically annotated corpus and the two different phoneme sets, we trained 20 character-based SMT systems using the Moses SMT toolkit [4]. The task of the SMT systems is to translate words from their grapheme representations into their phoneme representations by treating separate characters as single words. Language models for the SMT systems were built using IRSTLM [5] on the phonetic transcriptions. In order to simulate more data, we split the utterances in n-grams from one to four words and added them to the corpus. This was done so that the language models could better generalise phonetic variations at word beginnings and endings. The SMT models consist of 7-gram translation models and 5-gram language models. When training the SMT translation model, we disallowed phoneme reordering and performed extracted phrase pair filtering in order to ensure that word separators "*(w)*" and silence and non-verbal filler symbols ("*(.)*", "*(h.)*", "*(.h)*", etc.) are equally present in both source and target phrases. The filtering allows restricting insertion errors in rarely occurring contexts.

5. Evaluation Results

The 10-fold cross-validation results for grapheme to phoneme modelling are given in Table 1. The results show that the SMT-based method achieves lower error rates in both phoneme and word levels, which are statistically significant results with a confidence of 99%. In total, there are 188,638 phonemes annotated in the phonetically annotated corpus.

Table 1. Phoneme level and word level 10-fold cross-validation evaluation results.

Method	**Phoneme set**	**Error rate**	
		Phoneme	**Word**
Rule-based	Larger	9.831±0.292	33.372±0.693
	Smaller	9.143±0.272	31.236±0.651
SMT-based	Larger	8.893±0.265	31.594±0.811
	Smaller	8.167±0.288	29.092±0.857

Table 2. Error analysis.

	Rule-based - larger		**SMT-based larger**		**Rule-based - smaller**		**SMT-based smaller**	
Errors:	18,547		16,774		17,246		15,409	
Top errors	**Error type**	**%**	**Error type**	**%**	**Error type**	**%**	**Error type**	**%**
1	Ins: *ax*	13.4%	Del: *ax*	6.4%	Ins: *a*	17.1%	Del: *a*	8.8%
2	Ins: *ux*	9.4%	Ins: *ax*	4.8%	Ins: *u*	12.1%	Ins: *a*	8.2%
3	Sub: *O←uo*	8.8%	Ins: *ux*	4.8%	Ins: *i*	11.5%	Ins: *u*	6.9%
4	Ins: *ix*	7.4%	Del: *ux*	4.2%	Sub: *O←uo*	9.4%	Del: *i*	6.4%
5	Sub: *E←e*	4.8%	Del: *ix*	4.0%	Sub: *E←e*	5.1%	Ins: *i*	6.3%
Top 5 %	43.8%		24.1%		55.3%		36.7%	

The error analysis in Table 2 shows the top five errors for each cross-validation scenario. It is evident that the rule-based method does not capture phonemes missing in pronunciation. On the other hand, the SMT-based method allows learning generalisations. This results in up to 40% less insertion errors than for the rule-based method. The error analysis has also shown that a major proportion of errors occur in the prediction of word ending phonemes. Deletion errors in the rule-based method account for less than 1% compared to approximately 30% for the SMT-based method. The rule-based method also produces up to 20% more substitution errors than the SMT-based method.

6. Conclusion

In this paper, we have compared two grapheme to phoneme modelling methods: a rule-based method and an SMT-based method. The evaluation on a phonetically annotated corpus from the LSRC has shown that the SMT-based method allows achieving a lower phoneme error rate than the rule-based method. The performance of grapheme to phoneme modelling methods has also been evaluated in an ASR use case [6]. In further work, our aim is to merge both methods into one, adding context dependant transcription for the rule-based method and fixing incorrectly generalised rules for the SMT method.

7. Acknowledgements

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Centre” of EU Structural funds, contract nr. L-KC-11-0003 signed between the ICT Competence Centre (www.itkc.lv) and Investment and the Development Agency of Latvia, Research No. 2.9 “Speech corpus creation, principles, methods, realisation”.

8. References

[1] Karanasou, P., & Lamel, L. Comparing SMT methods for automatic generation of pronunciation variants. In Advances in Natural Language Processing, (2010), 167-178, Springer Berlin Heidelberg.

[2] Pinnis, M., Auziņa, I., & Goba, K. Designing the Latvian Speech Recognition Corpus. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC’14), (2014), Reykjavik, Iceland, European Language Resources Association (ELRA).

[3] Pinnis, M., & Auziņa, I. Latvian Text-to-Speech Synthesizer. In I. Skadiņa & A. Vasiļjevs (Eds.), Proceedings of the 2010 conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010, (2010), 69–72, Riga, Latvia, IOS Press.

[4] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., B., Cowan, W., Shen, C., Moran, R., Zens, C., Dyer, O., Bojar, A., Constantin, & Herbst, E. Moses: open source toolkit for statistical machine translation. In Proceedings of ACL 2007 on Interactive Poster and Demonstration Sessions, (2007), 177–180, Stroudsburg, PA, USA, Association for Computational Linguistics.

[5] Federico, M., Bertoldi, N., & Cettolo, M. IRSTLM: an open source toolkit for handling large scale language models. In Interspeech, (2008), 1618-1621.

[6] Salimbajevs, A., & Pinnis, M. Towards Large Vocabulary Automatic Speech Recognition for Latvian. In Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014, (2014), Kaunas, Lithuania, IOS Press.